Министерство образования и науки Российской Федерации
Московский государственный институт электроники и математики

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Рекомендовано УМО вузов
по университетскому политехническому образованию
в качестве учебного пособия для студентов высших учебных заведений,
обучающихся по направлению 231300 — «Прикладная математика»

Москва, 2011

УДК 681.4
ББК 32.813
Б 79

Рецензенты: д.т.н. В.А. Галактионов (зав. отделом Института прикладной математики им. М.В. Келдыша РАН),
к.филол.н., доцент Е.Б. Козеренко (зав. лабораторией «Компьютерной лингвистики и когнитивных технологий обработки текстов» ИПИ РАН)

Б 79 Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

ISBN 978–5–94506–294–8

В учебном пособии рассматриваются базовые вопросы компьютерной лингвистики: от теории лингвистического и математического моделирования до вариантов технологических решений. Дается лингвистическая интерпретация основных лингвистических объектов и единиц анализа. Приведены сведения, необходимые для создания отдельных подсистем, отвечающих за анализ текстов на естественном языке. Рассматриваются вопросы построения систем классификации и кластеризации текстовых данных, основы фрактальной теории текстовой информации.

Предназначено для студентов и аспирантов высших учебных заведений, работающих в области обработки текстов на естественном языке.

УДК 681.4
ББК 32.813

Оглавление

ЧАСТЬ I. ОСНОВЫ ТЕОРЕТИЧЕСКОЙ, ВЫЧИСЛИТЕЛЬНОЙ И ЭКСПЕРИМЕНТАЛЬНОЙ ЛИНГВИСТИКИ, ИЛИ РАЗМЫШЛЕНИЯ О МЕСТЕ ЛИНГВИСТА В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ *(ЯГУНОВА Е.В.)*

Предисловие (несколько слов от себя)

В лингвистических главах представлена минимальная терминология и предложены цели, задачи, методы и термины компьютерной лингвистики. Главы ориентированы на экспериментально-теоретическую парадигму сочетающую, по возможности, методы вычислительных экспериментов и экспериментов с информантами. В текст вошли материалы докторского исследования и результаты разноплановых экспериментальных работ последних лет, большинство из них соавторские. Изложение ориентировано на специалистов, работающих с языковым и текстовым материалами, вне зависимости от исходного образования читателей. Сверхзадачей является привлечение специалистов к лингвистическому и экспериментально-теоретическому осмыслению тех объектов и процедур, которые они моделируют. Хочется надеяться, что в результате уровень оценки работающих систем только повысится, а главное – повысится качественный уровень лингвистического знания.

Хочу поблагодарить моего научного консультанта В.Б.Касевича, которого постоянно цитирую в своем тексте, моего главного верного соавтора Лидию Пивоварову и многих моих дорогих друзей-коллег-соавторов последних лет, прежде всего, Дмитрия Ландэ, Александра Антонова, Эдуарда Клышинского.

Глава 1. Язык. Текст. Основы лингвистики и теории речевой коммуникации

Первая глава неизбежно вводная, она посвящена основным целям, задачам, гипотезам, методам и терминам. Работа с терминологией – особо тонкое место в междисциплинарной области, т.к. представители каждой из сторон имеют свою терминологию и свое представление об «общей терминологии», которая должна использоваться в этой области.

§ 1.1. Язык. Введение

Первый из заявленных терминов – язык. В своем тексте я буду в максимальной степени опираться на идеи В.Б.Касевича, для начала приведу краткий реферат из цитат его произведений. Такого рода цитатник – своего рода доказательная база, построенная по принципу «доказательство, основанное на авторитетности мнения». «Обобщая различные определения, можно сказать, что **язык — это знаковая система, предназначенная для порождения, передачи и хранения информации** /здесь и далее п/ж шрифт маркирует то, что выделено Е.Я./. **Информация**, передаваемая языковыми средствами, **всегда воплощается в некотором тексте**, поэтому передача информации — создание, или порождение текста, с одной стороны, и восприятие, «прием» текста — с другой. Система речевых действий и операций,

выполняемых в процессах порождения и восприятия текста, — это речевая деятельность. Первым и естественным условием ее реализации является наличие языковой системы.

Говоря о том, что язык — знаковая система, имеют в виду, что основной элемент такой системы — знак. Знак служит средством отражения того или иного элемента действительности. Благодаря наличию в языке данного знака этот элемент не только получает представительство в системе знаний о мире, присущей носителю языка[1], — возникает возможность передать эти знания другому. Знания становятся коммуницируемыми. Знак <…> обладает экспонентом, или означающим, т. е. материальной оболочкой, и сигнификатом, или означаемым, т. е. мыслительным содержанием, значением. Иными словами, языковой коллектив, вычленяя данный элемент действительности и осмысляя его определенным образом, закрепляет за таким осмыслением ту или иную материальную форму, материальный способ выражения; в результате и возникает знак» [108: 660-661].

Продолжим: «язык представляет собой знаковую с и с т е м у. Это сложная функциональная система. В данной части определения языка («части» — потому что язык здесь не отграничен от других сложных функциональных систем) существенно все: и то, что язык — система, и то, что система функциональная и, наконец, сложная. Система как таковая — это любое целостное образование, части (элементы) которого объединены отношениями, теряющими силу за пределами данного целого» [108: 661].

«Каждая система имеет, таким образом, относительно замкнутый характер. Системы соотносятся друг с другом именно и только как целостные образования. <…> Ни одна система не существует как нечто абсолютно изолированное. Принято говорить о системе и среде, в которой существует данная система. Но среда, в свою очередь, тоже системна, и реально мы имеем дело с вхождением одной системы в другую, нередко — в другие, т. е. некоторая система является подсистемой по отношению к другой или другим; в последнем случае происходит пересечение, «переплетение» систем. <…>

Для функциональной системы (напомним, что это понятие введено П. К. Анохиным [85; 86]) сказанное выше действительно в полной мере, однако здесь добавляется новый системообразующий фактор, гораздо более «мощный», чем фактор замкнутости. Это результат (или функция), для достижения которого (которой) существует данная совокупность элементов. Именно необходимость обеспечения некоторого результата, который не может быть достигнут «разрозненными усилиями» отдельных элементов, и служит причиной объединения последних в единое целое, — такое, какому «под силу» соответствующая задача. Это и имеется в виду, когда говорится, что функция выступает системообразующим фактором для системы, а последняя, соответственно, функциональна.

По существу, любая «работающая» система – живая или неживая – функциональна, поскольку «работать» и означает, в конечном счете, «получать результат» [108: 662].

Под **сложными** системами обычно понимаются такие, которые удовлетворяют двум условиям:

- налицо достаточно большое число подсистем,
- часть подсистем носит дублирующий характер.

[1] Знания о мире не всегда «означены», т. е. представлены соответствующими знаками и их структурами, но знаковое представительство знаний — несомненно высшая, наиболее развитая форма знания.

Дублирование может проявиться двояким образом. Один тип представлен тогда, когда подсистемы имеют более или менее одинаковую функцию. Параллельное сосуществование объясняется особой важностью этой функции: дублирование (неэкономность, избыточность) в системе обеспечивает выполнение требуемого результата в любых условиях, даже при выходе из строя каких-то подсистем. Другой тип дублирования (относительного) — это уровневое, иерархическое строение системы. «Здесь также можно говорить — с определенной долей условности — о дублировании, так как в выполняющей сложные виды деятельности иерархической системе на каждом следующем уровне происходит возвращение к той же задаче, только взятой в другой степени конкретности (подробнее см. [108; 109])» [107 : 663].

«Наиболее важные черты системы и любого образования в ее составе определяются функцией. Для чего, для выполнения каких задач существует сама система, тот или иной ее компонент (подсистема), отдельный элемент — ответ на этот вопрос является решающим для определения качественной специфики интересующих нас объектов. **Функция языковой системы** как таковой, как уже отмечалось выше, заключается в том, чтобы **служить средством порождения, хранения и передачи информации**. Порядок перечисления «подфункций», заметим сразу же, отражает реальную последовательность процессов: информация сначала должна быть порождена, а затем передана — с промежуточным хранением, если это необходимо. Что же касается иерархии «подфункций», то главенствующей и определяющей выступает как раз последняя из перечисленных — передачи информации, т. е. коммуникативная.

Нелишне подчеркнуть, что язык является именно средством передачи информации: информация заключена в тексте, а не в языке, а уже текст «построен» с использованием языка, языковой системы[2]. Поэтому характеристики языка в принципе определяются следующим вопросом: чем должен обладать язык, чтобы эффективно обеспечивать продуцирование несущего информацию текста (и извлечение информации из последнего)?» [108: 664].

«Разнообразие способов отражения действительности, присущих конкретным индивидуумам, потенциально бесконечно ввиду уникальности каждого индивидуума, бесконечно разнообразны и конкретные условия, в которых имеет место процесс отражения и, на его основе, формирования информации. Отсюда следует, что для передачи именно той информации, с которой имеет дело каждый индивидуум, в данный момент времени в данной точке пространства требуется бесконечное число некоторых информационных единиц, бесконечный алфавит, бесконечный код (и, вероятно, бесконечный канал связи). Информация, следовательно, должна быть как-то модифицирована, ограничена, подвержена своего рода компрессии, чтобы она могла быть передана (и воспринята).

Процедуры компрессии как преобразования информации в принципе могут быть выполнены по-разному: за счет разных фрагментов подлежащей передаче информации и присвоению разных весов информационной значимости. Первичная переработка информации с целью сделать ее «пригодной» для коммуникации должна ориентироваться именно на **общезначимость** передаваемого, на его адекватность

[2] Никак нельзя признать корректными обычные утверждения о том, что система языка «реализуется» в тексте (речи) как абстрактное в конкретном. Так можно было бы сказать, например, о некотором языке или диалекте по отношению к идиолекту, которые и соотносятся как система с системой по принципу большей/меньшей абстрактности (скажем, русский язык соответствующего периода и язык Пушкина или Горького). Язык и речь (текст) соотносятся, скорее, как «механизм» и «продукт» работы последнего.

задачам, решаемым данным обществом. Язык возникает и функционирует только в обществе, обслуживает наиболее важные ситуации (с точки зрения общества, в т.ч. некоторой социальной группы). Для языка естественна функция кодирования: преобразовывания информации, чтобы она была коммуницируема. При этом информация усредняется, обедняется, огрубляется. Компрессия информации (ее огрубление, обеднение) в каждом языке (подъязыке, см. следующий параграф) происходит к тому же по-своему. Язык участвует в порождении информации, является средством не только передачи, но и порождения информации: ведь «окончательный вид», который приобретает передаваемая информация, в известной — и немалой — степени определяется именно языком» [108].

§ 1.2. Язык или языки. Текст или тексты. Основы речевой коммуникации

Как уже было сказано, язык – средство передачи информации, информация заключена в тексте (не в языке), текст «построен» с использованием языка, языковой системы. Характеристики языка определяются задачей эффективно обеспечивать порождение и анализ текста (извлечение информации из текста), т.е. речевую коммуникацию[3]. Изменяются ли эти характеристики в зависимости от особенностей коммуникативной ситуации? Коммуникация может быть устной или письменной. Язык, обеспечивающий эффективную устную коммуникацию, не может не отличаться от языка, обеспечивающего письменную коммуникацию. Каждый из носителей письменного языка (успешно овладевший письменным языком) может по праву называться билингвом: человеком, владеющим двумя – устным и письменным – языками и умеющим переключаться с одного языка на другой (с одного кода на другой) в зависимости от требований коммуникации.

Следующий тезис: информация заключена в тексте (не в языке), но текст строится и анализируется с использованием языка. Значит, легко допустить, что тексты существенно разного типа накладывают свои требования на используемый язык. Речь идет, прежде всего, о текстах, различающихся по степени и типу информационной нагруженности: о текстах разных функциональных стилей.

Сначала приведем несколько цитат, как принято при опоре на авторитеты. «Функциональная стилистика рассматривает функциональный стиль (функциональную разновидность языка, функциональный тип речи) как исторически сложившуюся, общественно осознанную речевую разновидность, … которая складывается в результате отбора и сочетания языковых средств» [105: 43]. Среди стилеобразующих факторов выделяются, в целом, те же факторы, что и для формирования коммуникативной ситуации: цели коммуникации, сфера коммуникации (и шире – деятельности), функции языка и пр. (см., например, [117; 148: 581] и др.). Существенно, что характеристики функциональных стилей «создаются не столько за счет … стилистически маркированных средств, сколько **за счет различной частоты употребления тех или иных языковых единиц**…» [148: 581] и за счет различий в предпочтительной сочетаемости этих языковых единиц.

[3] В рамках этих лекций мы не рассматриваем терминологические вопросы, интересующие многих традиционных лингвистов: где граница между языком и формой языка, где граница между языком, вариантом языка и диалектом и т.д. Вместо разнообразия терминов мы используем термин «язык», подчеркивая тем самым тот факт, что разным языкам будут приписаны разные характеристики, позволяющие эффективно обеспечивать коммуникацию на данном языке в тех или иных коммуникативных ситуациях.

Обычно выделяют следующие функциональные стили (одна из самых грубых классификаций): разговорный (бытовой диалог), литературно-художественный, газетно-публицистический (новостной), научный, деловой (официально-деловой). Нас интересует, прежде всего, (1) степень и тип информационной насыщенности, (2) основной тип контекста и (3) жесткость композиционной структуры (два последних фактора рассматривается в следующей главе).

Вне контекста коммуникативной ситуации текст первого функционального стиля – разговорного, или бытового диалога, – воспринимается как искаженный (своего рода восприятие в условиях помех). Основным контекстом для текстов данного типа будет именно ***контекст коммуникативной ситуации***, а контекст собственно текста занимает до некоторой степени подчиненное положение. Это и есть основное отличие текстов этого функционального стиля. Соотношение информационной насыщенности и реализованности в тексте других функций языка (напр., воздействия на адресата, контакто-устанавливающей и контакто-поддерживающей функций) зависит от конкретного типа коммуникативной ситуации и текста. В этом смысле этот функциональный стиль «перпендикулярен» основной шкале функциональных стилей.

В качестве основной шкалы мы рассматриваем шкалу степени (и типа) информационной насыщенности. Два полюса этой шкалы занимают ***литературно-художественный*** vs. ***официально-деловой*** стили.

Литературно-художественный (художественный) стиль неоднороден с точки зрения своей функциональности, в нем реализуется практически ***вся палитра функций языка***. Даже исключив из рассмотрения поэтические тексты, сложно единообразно структурировать множество художественных текстов. Для текстов художественного стиля невозможно выделить приоритет именно информационной насыщенности (в ущерб, например, воздействию на адресата или эстетической функции). Для текстов ***делового*** стиля, напротив, безусловен приоритет именно информационной составляющей. В качестве примеров текстов официально-делового стиля приведем тексты законов, договоров (тексты, имеющие юридическую силу и требующие однозначного понимания, см. об этом в главе 2).

Деловой и научный стили имеют значительное число общих характеристик. В обоих стилях доминирует информативная функция языка. Однако для текстов ***делового*** стиля в целом характерна более жесткая смысловая и коммуникативная структурированность текста (композиция, структура фрейма). Язык официально-делового стиля должен позволить однозначно закодировать и декодировать коммуницируемый смысл текста.

Множество научных текстов неоднородно. С одной стороны – эта неоднородность определяется тем, что при общем доминировании информативной функции языка в текстах смешанного научного стиля по-разному реализуется взаимодействие информативной функции и функции воздействия на адресата: например, в научной публицистике или учебной литературе. С другой стороны, это связано с неоднородностью самих предметных областей. Во множестве научных языков сосуществует множество языков, различающихся именно в соответствии с предметной областью: языки математики, физики, техники, лингвистики, философии и т.д.

Специфика языков публицистического стиля складывается из взаимодействия информативной функции и функции воздействия. Соответственно, множество текстов

публицистического стиля неоднородно, и эта неоднородность в большей степени связана с их функциональностью, а не тематикой (предметной областью). Например, язык новостных лент и язык политической публицистики (напр., интервью и даже аналитики) существенно различаются в отношении частоты встречаемости языковых единиц и их сочетаемостных предпочтений. На основной шкале – шкале степени информационной насыщенности – тексты публицистического стиля будут занимать промежуточное положение между художественными текстами и научными текстами. Более того, в дальнейшем вместо традиционного наименования «публицистические» тексты мы будем использовать наименование «новостные» тексты, подчеркивая тем самым общую для этих текстов задачу коммуникативных ситуаций: задачу сообщения новостей. При этом тексты новостных лент будут находиться на предлагаемой шкале ближе к научным (информационно более насыщенным) текстам, а собственно публицистические (активно реализующие функцию воздействия) – к художественным. В этом же ключе, но гораздо подробнее этот вопрос рассматривается в главе 2 (и в некоторых наших работах [159 – 161] и т.д.).

Продолжая идею функциональности языка, мы сможем выделить большое количество языков (в рамках одного национального языка): устный или письменный; язык художественной прозы, язык новостной ленты, научный язык, деловой язык и т.д. и т.п. «Носитель языка», обладающий высокой коммуникативной компетенцией, таким образом оказывается полилингвом (даже полиполилингвом). Невозможно представить себе человека, владеющего всеми языками (коммуникативными знаниями и умениями). Например, знание научного языка не сможет распространиться на все многообразие языков науки.

Для человека полилингвальность является несомненным достижением, т.к. один и тот же человек – особенно публичный человек – оказывается включенным в разные коммуникативные ситуации, требующие своего языка. Но в каждой конкретной ситуации носитель языка «выбирает» – из всего многообразия – по возможности единственный язык. Для автомата (системы автоматической обработки текста), напротив, специализация является законным правом и преимуществом автоматической обработки текста. Ведь в случае с автоматом мы можем (и должны) «заточить» его под определенные задачи: коммуникативные ситуации и языки.

§ 1.3. Лингвистика и лингвистики. Принцип моделирования. Цели, методы, задачи

Приведем цитату лингвиста-теоретика, высоко ценящего принцип моделирования. «Всякая теоретическая дисциплина имеет перед собой задачу построения *модели* того объекта, который она изучает. Модель — это тоже объект, но специально построенный исследователем[4] с целью познания того или иного фрагмента действительности, т. е. объекта-оригинала. Правильная, адекватная модель создается тогда, когда два эти объекта — модель и оригинал — обладают структурным и/или функциональным подобием. Наличие структурного подобия означает, что объект-оригинал и модель имеют одинаковую структуру, в таких случаях говорят, что они изоморфны друг другу. Наличие функционального подобия

[4] Если не учитывать естественных моделей, которые «стихийно» создаются мозгом, психикой человека при отображении внешней действительности.

говорит о том, что модель способна выполнять те же функции, что и объект-оригинал» [109: 26]. Сложно не согласиться с этим утверждением. Более того, на наш взгляд, построение модели объекта свойственно и большинству прикладных задач. В данном параграфе зададимся двумя вопросами:

- в чем принципиальное отличие между теоретическим и практическим подходом;
- в чем сходство и различие отношения к принципу моделирования у лингвистов и представителей технических областей.

И попутно решим проблемы, связанные с ролью носителя языка как «субъекта» такого моделирования

«При узколингвистическом характере моделирования основным критерием адекватности модели является адекватность текста <…>, т. е. результата функционирования этой модели. Однако один и тот же текст может быть порожден (иногда и интерпретирован) разными способами. Поэтому собственно лингвистическая модель может оказаться не изоморфной внутренней системе носителей языка, т. е. языку в психолингвистическом смысле, точно так же структура функционирования модели может не воспроизводить структуру деятельности человека.

При психолингвистическом характере моделирования следует добиваться именно такой изоморфности, используя для этого все доступные наблюдению факты речевого поведения, ставя специальные психолингвистические эксперименты» [109: 28].

В современном понимании часто говорят скорее об антропоморфной – а не психолингвистической – адекватности модели, хотя по сути имеется в виду ровно то, о чем говорит В.Б. Касевич: носитель языка выступает в роли идеального субъекта, воплощающего в ходе своей коммуникативной деятельности функциональные возможности языковой системы.

Действительно ли носитель языка выступает в роли идеального субъекта коммуникации? На этот вопрос нельзя ответить точно раз и навсегда, к этому вопросу мы будем возвращаться во второй, третьей и четвертой главах. Для меня ответ на этот вопрос будет «скорее положительным» или даже «безусловно положительным» (в зависимости от решаемой задачи) в силу того экспериментального направления, которое я представляю. Более или менее очевидный плюс такого подхода заключается в ориентации на точность моделирования процесса (точность результата является следствием точности модели, а не главной задачей). Хотя, конечно, такая модель может не дать быстрый вариант получения требуемого (например, в техническом задании) результата.

По [89] важнейшим свойством методов прикладной – в отличие от теоретической – лингвистики является оптимизация. Термин «прикладная лингвистика» выступает здесь в российском или даже скорее советском значении: как синоним компьютерной, вычислительной, об этом будет чуть ниже. Под оптимизацией понимается такая модель языковой системы (или подсистемы), при которой этот объект сохраняет в результирующем представлении только те существенные свойства, которые необходимы для данной практической задачи. Иными словами, если для теоретического исследования предполагается полная модель, например, полное описание этого объекта со всеми его характеристиками, сложностями и т. п., то прикладное оптимизированное описание должно быть удовлетворительным только для данной конкретной задачи.

Анатолий Николаевич Баранов приводит в качестве примера категорию времени (пример приводится полностью по [89]). Теоретический подход, в зависимости от выбранной концепции, будет требовать:

- описание грамматической категории времени (выделение граммем, морфологических способов выражения граммем, сочетаемость граммем категории времени с граммемами других грамматических категорий), классификация лексики со значением временных отношений, классификация синтаксических конструкций;
- в рамках уровневой модели языка — семантика временных отношений → способы выражения на синтаксическом уровне; → способы выражения на лексическом уровне; → способы выражения на морфологическом уровне.

Прикладное описание будет выглядеть совершенно по-другому:

- составление технического задания (определяется заказчиком);
- анализ проблемной области (сколько типов временных отношений представлено в проблемной области и каковы формальные способы выражения темпоральных отношений в данном подъязыке);
- формирование метаязыка, способов описания проблемной области, совместимых с другими привлекаемыми метаязыками;
- применение метаязыка → результирующее представление (модель) проблемной области;
- проверка результирующего представления (объяснительная и предсказательная сила модели; компьютерная реализация или эксперимент).

Прикладные модели ориентированы на конкретные коммуникативные ситуации, конкретные языки (подъязыки), в существенно большей степени огрубляют моделируемый объект и допускают широкие возможности выбора инструмента моделирования.

Обещанные несколько слов про терминологию. Как-то укоренилось, что термин «прикладная лингвистика» в зарубежной и отечественной науке имеют существенно различное значение. В зарубежной науке лингвистика «прикладывается», прежде всего, к такой безусловно прикладной задаче, как обучение языку. Наше понимание термина ближе всего к компьютерной или вычислительной / машинной / инженерной лингвистике[5] (наша специальность «Прикладная и математическая лингвистика» (в номенклатуре ВАК) за рубежом скорее всего найдет себе аналоги на факультетах Computer Science).

Впрочем, неоднородность толкования этих терминов в отечественной науке налицо. Приведу в качестве примера два толкования «компьютерная лингвистика»[6]

[5] Недаром при обсуждении названия школы так долго шли обсуждение и выбор наименования.

Про термин «инженерная лингвистика» стоит сказать отдельно. Только что мы говорили об одном его понимании (в широком смысле). Однако иногда можно встретить этот термин и в узком смысле: для того, чтобы подчеркнуть заведомо суженную задачу обработки текстовой информации. Например, однократное решение такой задачи (даже на материале одной коллекции), не претендующее на построение долговременной модели, но требующее получения быстрого результата (иногда с минимальными требованиями к точности). В случае такой сильно зауженной постановки задачи рассуждения о принципе моделирования могут казаться избыточными, не нужным теоретическим довеском.

[6] Термина «вычислительная лингвистика» в этом тезаурусе нет, что понятно в силу выбора источников для работы.

(автор Е.Г.Соколова) из Русско-английского тезауруса по компьютерной лингвистике (Доступен на http://uniserv.iis.nsk.su/thes/) [147].

Дескриптор	
название	компьютерная лингвистика
язык	русский
релятор	
определение 1	направление в прикладной лингвистике, ориентированное на использование компьютерных инструментов – программ, компьютерных технологий организации и обработки данных – для моделирования функционирования языка в тех или иных условиях, ситуациях, проблемных сферах и т.д., а также вся сфера применения компьютерных моделей языка в лингвистике и смежных дисциплинах.
определение 2	область Искусственного Интеллекта, занимающаяся компьютерным моделированием владения языком с целью передачи информации, а также решением прикладных задач автоматической обработки текстов и звучащей речи
автор словарной статьи	Соколова Е.Г.

И далее толкование оригинального английского термина.

Дескриптор	
название	computational linguistics
язык	английский
релятор	
определение 1	Computational linguistics is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition. Computational linguistics has applied and theoretical components.
автор словарной статьи	Соколова Е.Г.

Нужны ли комментарии? Как видите, то, что в англоязычной традиции связывает и объединяет «компьютерную лингвистику» в единую междисциплинарную область, в отечественной науке часто оказывается представленным на двух противоположных полюсах. Поэтому нам всегда приходится «во первых строках» определить, что мы (лично, в данной работе и т.д.) понимаем под «компьютерной лингвистикой». Конечно, в наших лекциях мы будем склоняться к «определению 1». Особо обращаю внимание читателя на то, что «Computational linguistics has applied and theoretical components», именно «вычислительная» теория (модель) является для меня ведущим аспектом, а прикладной аспект в идеальном случае является естественным результатом этой модели.

Когда и зачем нужны лингвисты? Лингвисты несколько лучше представляют себе «физическую» природу объекта моделирования. Языковая система уникальна в том смысле, что она полностью не починяется законам ни естественнонаучного, ни гуманитарного познания. Язык многие рассматривают как творение человека, но это в существенной степени заблуждение. Пожалуй, так никогда не скажет лингвист. Язык

– объект принципиально особого свойства. Он сосуществует в природе совместно с человеком (ср. разнообразные варианты рассмотрения гипотезы лингвистической относительности, т.е. степени взаимообусловленности человека, языка и социума (цивилизации). Для моделирования языковой системы используются инструменты моделирования, пришедшие из физики, из экономики (и/или социологии), из физиологии, из философии и семиотики (теории о знаках). Лингвистика – хорошая лингвистика – должна уметь оценить рассматриваемый объект во всех этих плоскостях (быть междисциплинарной), конечно, если лингвистика – действительно наука о языке. Вернее сказать, это наука о языке (языковой системе) и тексте, формах и способах функционирования этой системы. Может ли на начальном этапе – этапе постановки задачи – компьютерная лингвистика обойтись без лингвиста? Вряд ли. Может ли хотя бы на начальном этапе лингвист обойтись без инструментария смежных дисциплин? Безусловно, нет.

Возможны ли чисто вычислительные методики оценки эффективности работы модели (в конкретной ее реализации для реализации конкретных задач в соответствии с требованиями технического задания)? На наш взгляд, скорее «нет», чем «да». Эти методики могут дать результаты экспресс-анализа. Однако окончательное слово, на наш взгляд, остается за лингвистом: лингвистическим анализом результатов и лингвистическим экспериментом. Забегая вперед, зададимся вопросом: лишает ли лингвиста работы все большее применение методов статистического анализа? И сразу же ответим: нет. В современном информационном мире лингвистика расширяет сферу своих интересов. Статистические закономерности функционирования языка – и текста – и раньше были (должны были быть) предметом лингвистики. Сейчас они становятся все более и более значимыми, также как и анализ тех единиц, которые выделяются на основании этих статистических закономерностей. Наряду с единичными текстами, которыми и раньше занимались лингвисты, объектом лингвистики становятся и коллекции текстов, и информационные потоки как объекты нового информационного пространства (см. подробнее главу 2 и 4).

Попробую сформулировать свое собственное ощущение от изменения парадигмы лингвистики, во всяком случае – компьютерной лингвистики[7]:

- изменился главный объект исследования, перестроилась перспектива – компьютерная лингвистика могла (должна) была поставить во главу угла исследование информационных объектов, как минимум, текстов;
- компьютерная лингвистика оказалась максимально включенной в экспериментальную парадигму;
- компьютерная лингвистика стала максимально междисциплинарной;
- компьютерная лингвистика стала предъявлять повышенные требования к знаниям в области математического моделирования, теории сложных систем и психофизиологии (обработке информации у человека);
- у компьютерной лингвистики появились новые объекты изучения (коллекции, кластеры и т.д.) и новые экспериментальные возможности (возможности современных информационных технологий).

[7] Это произошло по следам Круглого стола по проблемам автоматического извлечения лингвистической информации («Лингвистика без лингвистов?»). Вед. Наталья Лукашевич на конференции «Диалог-2011» http://www.dialog-21.ru/dialog2011/materials.asp?id=159065, во время которого я вдруг почувствовала себя не прототипическим лингвистом и захотела дополнительно сформулировать свои представления о «современном лингвисте».

Глава 2. Слово — коллокация – синтаксические конструкции – текст. Единица анализа и контекст.

Во второй и третьей и четвертой главах мы рассмотрим не только общие подходы, но приведем конкретные примеры и те данные, которые были получены в ходе наших экспериментов с информантами и/или вычислительных экспериментов. Ключевым для этих глав является представление о вариативности и неединственности. Каждый текст (и шире – информационный лингвистический объект) обладает неединственной структурой. В зависимости от задачи анализа (человеком и/или автоматом) должна выбираться (и далее – извлекаться) требуемая структура. Вариативность (и сам по себе набор вариантов) в существенной степени зависят от тех параметров, которые мы уже начали обсуждать в первой главе: функционального стиля, жанра (подстиля), предметной области и т.д.

§ 2.1. Инвентарные и конструктивные единицы. Понятие «текущего словаря»

Основными вопросами, рассматриваемыми в этом параграфе, является два вопроса

- об единицах анализа текста;
- о понятии «текущего словаря», учитывающего максимальную подстройку под особенности конкретного текста (в дальнейшем – информационного лингвистического объекта).

В качестве единицы анализа (письменного) текста в работах используются, прежде всего, такие стандартные единицы, как лексема и словоформа. Когда и какая из этих единиц важнее – решать исследователю, и выбор задается целью и задачами работы. Впрочем, отметим, что роль словоформы как основной единицы восприятия (анализа) текста подтверждается психолингвистическими экспериментами (особенно для звучащей речи)[8]. Для звучащего текста в качестве основной единицы первичного анализа используются фонетические слова. Однако приведем немного теории.

«Положение о слове как единице словаря означает, что именно словам принадлежит роль тех базовых элементов, которые образуют язык как систему. В самом деле: язык есть система, система — это элементы, связанные определенными отношениями (словарь) и функционирующие в соответствии с определенными правилами (грамматикой) для выполнения некоторой задачи, и элементами оказываются, прежде всего, именно слова. Все остальные виды единиц языка существуют либо в отвлечении от слов, которое осуществляется непосредственно или опосредованно (на нескольких уровнях), либо в результате соединения слов по правилам. И лишь слова непосредственно образуют тот инвентарь, который служит источником всего в языке и речевой деятельности. Именно поэтому, несмотря на многочисленные и постоянно повторяющиеся попытки «упразднить» слово, оно

[8] Экспериментальная проверка гипотезы о том, что основной единицей перцептивного словаря является словоформа, осуществлялась с помощью нескольких серий свободного устно-устного ассоциативного эксперимента (эксперименты осуществлялись мной [158] и в рамках диссертационного исследования (Бочкарева 2006)). Стимулами для такого эксперимента служили словоформы (в словарной и несловарных формах) и предложно-падежные конструкции. Результаты эксперимента дают основания утверждать, что в условиях дефицита времени испытуемые ***непосредственно*** переходили от словоформы как стимула к словоформе как реакции, минуя дополнительную процедуру лемматизации.

сохраняет свои позиции в языкознании до сегодняшнего дня» [108: 819-820]. Введем вслед за В.Б.Касевичем понятие инвентарных и конструктивных единиц языка [108]. Круг проблем возникает для языков наподобие русского с развитой и морфологией и неоднозначностью парадигм. Слово как единица словаря и как единица морфологии не всегда совпадают. Что является инвентарной единицей: словоформа или лексема? Уменьшительные существительные вроде *домик*, *кошечка* несомненно являются словами по любым морфологическим критериям, но являются ли они инвентарными или создаются по мере надобности в процессе порождения текста с помощью простейших правил и единиц, принадлежащих грамматике? В общем случае их следует отнести к конструктивным, но существуют подъязыки (ребенка или обращенный к ребенку), для которых это правило, возможно, не выполняется. И, конечно, то, что эти единицы являются конструктивными при порождении текста, не значит, что они выступают в этом же качестве при анализе текста. При построении динамической языковой системы для анализа текста нам может быть гораздо разумнее (и выгоднее) отнести эти единицы к инвентарным.

К инвентарным единицам относят также единицы, размерностью больше, чем слово. Инвентарными единицами являются безусловные фразеологизмы (например, *бить баклуши*). Однако степень фразеологизации и идиоматизации в языке может быть разной. Поэтому правильнее было бы сказать, что фразеологизмы и идиомы расположены на шкале от инвентарных к конструктивным единицам. Кроме того большую проблему представляют составные слова: «*в отличие от*» в современном языке является инвентарной единицей, но состоит из трех пробельных слов (текстоформ). Каждый прикладник на своей шкуре испытал всю сложность и неоднозначность решения задачи разделения на слова (графематического анализа и парсинга). К этой же проблеме относится, например, задача выделения (объединения?) компонентов сложных номинаций. Обо всем этом пойдет речь в данной главе.

«С морфологической точки зрения слова — конечные составляющие высказывания, т. е. такие структурные единицы, взаимодействие которых и создает высказывание **безотносительно** к его устройству. Это значит, что, во-первых, по отношению к высказыванию слова́ являются мельчайшими интегрантами и, во-вторых, статус слова предполагает лишь (относительную) цельность и автономность в составе высказывания» [113: 821]. Именно по этой причине сложно, а подчас и невозможно анализировать информационную (или коммуникативную) структуру текста на уровне этих мельчайших интегрантов.

В [98] было выдвинуто понятие «текущего словаря»: подобно тому, как в самом начале восприятия осуществляется фаза ориентировки (знакомство с коммуникативной ситуацией, подстройка под нее, подстройка под диктора), имеет место и своего рода подстройка под лексико-семантические особенности воспринимаемого текста, что позволяет сузить рабочую область словаря: перейти от общего словаря к текущему. Соответственно облегчаются и становятся более эффективными процедуры идентификации (поиска в текущем словаре), ведь словарь слушающего переструктурировался.

Остановимся подробнее на следующих вопросах:

- как происходит подстройка слушающего под структурные особенности текста;
- как формируется «текущий словарь» в процессе восприятия речи;

- как соотносится формирование «текущего словаря» с извлечением смысловой структуры текста и ключевых слов как наиболее ярких представителей этой структуры.

Согласно [99: 136] «…«общий» словарь разбит на потенциальные «текущие» по тематическому принципу, примерно так же, как это имеет место для словарей идеографического или тезаурусного типа, создаваемых лексикографами». Соотношение общего словаря и потенциальных текущих, вероятно, соответствует соотношению словаря, полученного на репрезентативном корпусе, и подсловарей, полученных на соответствующих подкорпусах[9].

В процессе восприятия речи «один из тематических словарей может активироваться, в результате чего и появляется возможность обращения к «текущему» словарю» (там же). Активация рассматривается в цитируемой работе, по-видимому, в традиционном контексте сетевых моделей как активация по некоторому заданному семантическому стимулу-признаку (в частности на материале работ по семантическому прайммингу (ср. [101])). «…Уровень активации не используемых в данный момент подсловарей, будучи существенно ниже, как бы временно выводит их из игры, тем самым поле поиска словарных единиц существенно сужается» [99: 136]. Процедуры поиска в таком переструктурированном словаре по всей видимости должны быть наиболее легкими и быстрыми. Однако даже столь прямолинейно решаемая задача разбиения всего общего словаря на потенциально текущие может быть достаточно сложно реализуемой:

- как правило, возникают сложности при отнесении к какому-либо тематическому подсловарю сравнительно частотной лексики;
- возможны сложности при определении степени дробности такого рода тематических словарей;
- вероятно, возможность осуществления такого рода структурирования словаря (построения системы вложенных словарей) зависит, во-первых, от функционального стиля рассматриваемых текстов (жанра, типа и т.д.) и, во-вторых, от анализируемых предметных областей.

Например, можно представить себе тезаурусного типа систему вложенных словарей научного (ср., например, библиотечные рубрикаторы и классификаторы) или делового функциональных стилей. В какой-то степени подобную схему можно представить и для новостных текстов (новостных сообщений, новостных лент).

Формирование «текущего» словаря осуществляется на этапе восприятия первых композиционных фрагментов текста. В дальнейшем «текущий» словарь, будучи уже сформированным, претерпевает изменения по мере узнавания структуры текста, таким образом, активированная сеть отвечает на каждый новый квант информации. Функционирование этой сети тоже в значительной степени зависит от стиля текста.

Использование представления о роли «текущего» словаря в процедурах анализа текста неизбежно ставит вопрос о том, насколько при этом оказываются взаимосвязанными «текущий» словарь и ключевые слова. «Можно сказать, вероятно, что набор ключевых слов для заданного текста представляет собой подмножество словарных единиц, которые принадлежат «текущему» словарю…» (там же: 137). Предложенное А.В. Венцовым и В.Б. Касевичем решение вопроса заключается в том,

[9] Построение тематических словарей разных уровней вложенности является обязательным компонентом многих моделей автоматического понимания текста. В качестве единиц такого рода словарей выступают не только словоформы и лексемы, но и сложные номинации.

что «…«текущий» словарь задает широкую тематику, всю предметную область…, а набор ключевых слов очерчивает в ней определенную подобласть» (там же: 137). По-видимому, это решение соответствует основным особенностям текстов научного, делового, отчасти новостного функциональных стилей. Вариативность возможных пересечений подмножеств «текущего» словаря и набора ключевых слов является прямым следствием вариативности стратегий анализа и типа текста. Кроме уже названных параметров, таких как функциональный стиль и предметная область, на мой взгляд, стоит указать еще два:

- степень статичности-динамичности текста,
- степень информационной насыщенности, которой противопоставляется функция воздействия на адресата (и другие возможные функции).

Под такой характеристикой как «динамичность» понимается наличие в тексте нескольких ситуаций, сменяющих друг друга. Под статичностью, соответственно, – минимальное количество ситуаций (одна-две). Все три перечисленных функциональных стиля имеют, казалось бы, явно выраженную статическую природу. Они занимают на шкале «статичность» vs. «динамичность» положение близкое к статичности, однако могут отстоять от этого полюса. Аналогично обстоит дело с информационной насыщенностью.

Существенное пересечение «текущего словаря» и набора ключевых слов характеризует, прежде всего, статичные и информационно насыщенные тексты. Максимальное число ключевых слов такого текста вводится в начальном композиционном фрагменте, таким образом, и область, и подобласть задаются в самом начале анализа[10].

Тексты, относящиеся к научной публицистике, учебной литературе, новостной аналитике, интервью и т.д. оказываются в более уязвимом положении. Более того, часть ключевых слов таких текстов может вообще никогда не оказаться на пересечении с «текущими словарями», ориентированными на разные предметные области.

С другой стороны – при решении вопросов о соотношении «текущего словаря» и ключевых слов при анализе (письменных) текстов, принадлежащих некоторым коллекциям и подколлекциям, мы выходим на более высокий уровень анализа и сопоставляем:

- «текущие словари», принадлежащие тексту и коллекциям разной степени однородности;
- ключевые слова, характеризующие текст и коллекции (подколлекции).

Переход на этот уровень анализа позволяет получить представление о наличии пересечений как характеристике степени тематической однородности коллекций и центральном/периферийном положении текста в информационном пространстве коллекций не только для информационно насыщенных текстов (напр., [138]), но и художественных текстов [156], см. чуть подробнее в параграфе 4 главы 3.

[10] Научный текст, представляющий описание работы программы, скорее всего, будет менее статичным, чем текст, в котором идет обсуждение некоторого положения дел.

§ 2.2. Избыточность. Контекстная предсказуемость

При исследовании процессов восприятия и понимания текста – устного или письменного – неизбежно обращение к вопросам, связанным с информационной избыточностью как неотъемлемому свойству любого текста. Употребляя термин «информационная избыточность» мы подчеркиваем, что для нас подход к исследованию избыточности связан с тем направлением в лингвистике, которое наследует идеи теории информации. Информационная избыточность является тем свойством любого текста, которое обеспечивает возможность успешного восприятия речи (особенно актуальным для звучащей речи). Подчеркиваем, что с этой точки зрения любой текст на естественном языке характеризуется информационной избыточностью, в противном случае он не может быть воспринят и понят адресатом[11]. Для того, чтобы исследовать информационную избыточность, необходимо опираться не только на ее качественные, но и на количественные признаки. Они могут быть определены в результате проведения вычислительных экспериментов (ср. многочисленные работы Р.Г Пиотровского, напр., [139-141]) и экспериментов с информантами (прежде всего, экспериментов по восприятию текста).

Для звучащего текста в современной теории восприятия речи стало естественным опираться на представление о том, что фонетические характеристики текста не могут содержать того количества информации, которое достаточно для полной фонемной интерпретации всего текста (всех слов текста). Положение о том, что в тексте сосуществуют сегменты полного и неполного типа произнесения, из которых только первые могут распознаваться за счет анализа фонетических характеристик, впервые было сформулировано в [97]. Прочие сегменты могут интерпретироваться только в результате контекстной предсказуемости, то есть предсказываться на основании знания контекста. Соотношение сегментов полного и неполного типа произнесения в рамках текста определяется самыми разными характеристиками, прежде всего – функциональным стилем текста. Очевидно, однако, что даже подготовленное дикторское чтение содержит большое количество сегментов неполного типа произнесения (слогов, слов, возможно, синтагм и даже фраз), восстанавливающихся на основании присущей тексту избыточности.

Возможно, наиболее иллюстративным примером функционирования избыточности при восприятии звучащего текста является роль морфонологических явлений (см. [107: 266–267, 280–282]). Так, например, большая часть морфологических характеристик слова может приходиться на безударные сегменты (слоги), тогда – в силу сегментной редукции – собственно морфологическая информация не может быть извлечена из соответствующего сегмента слова, но лишь на основании более широкого контекста. В этом случае, по-видимому, на первый план могут выступать интегральные характеристики фонетического слова (ФС), которое наряду со знаменательной словоформой может включать и служебные слова (напр., предлоги в предложно-падежных конструкциях). В особенной степени сказанное следует учитывать при исследовании восприятия на материале русского языка, т.к. для него характерны свободный порядок слов, морфологическая сложность, подвижное разноместное ударение и высокая степень сегментной редукции.

[11] Примером текста без информационной избыточности является текст программы, написанный на одном из языков программирования.

С другой стороны, человек, как известно, не в состоянии проводить пофонемное декодирование слов звучащего текста в силу ограничений своей психофизиологической организации (памяти и быстродействия). В результате этих ограничений и благодаря возможностям контекстной предсказуемости в процедурах восприятия текста человек оперирует сравнительно большими единицами: как минимум, словами, а чаще – коллокациями и конструкциями, т.е. последовательностью таких слов, совместная встречаемость которых существенно превышает случайный уровень. В условиях благоприятной коммуникативной ситуации и знания предметной области (и/или стиля) – когда уровень избыточности текста превышает некий средний, необходимый для восприятия – такого рода оперативные единицы могут приобретать еще больший формат: синтагм и целых фраз.

Увеличение формата подобных единиц может значительно увеличивать скорость восприятия и понимания (см., например, Грановская 1974). Поэтому увеличение формата характеризует восприятие и звучащего, и письменного текста. Мы – владеющие письменным языком – не читаем не только побуквенно, но даже пословно (кроме исключительных ситуаций). Однако значительное укрупнение единиц (и ускорение восприятия) возможно лишь в определенных коммуникативных ситуациях. Эти ситуации могут задаваться задачей коммуникации извлечь основной смысл и, как уже было сказано, благоприятной коммуникативной ситуацией, позволяющей максимально включать процедуры контекстной предсказуемости.

Избыточность – это свойство, неотъемлемое от естественного текста (и любого естественного языка), однако существенно зависящее от функционального стиля. Напомним основные функциональные стили, расположенные на шкале «степень информационной насыщенности» (в порядке возрастания): литературно-художественный, новостной, научный и официально-деловой. Какой текст будет более требователен к условиям «readability»: литературно-художественный или официально-деловой? Ответ очевиден. Прагматически задача успешности восстановления структуры и смысла текста закона значительно важнее задачи восстановления смысла художественного текста. На самом деле речь идет не столько о восстановлении, сколько об однозначном восстановлении структуры и смысла текста закона. В противном случае каждый из нас – носителей официально-делового языка – вправе понимать один и тот же текст закона по-своему.

Успешность восстановления зависит от типа и степени компрессии текста, что определяется условиями коммуникации. К сожалению, в русском языке нет эквивалента термину «readability», однако само явление несомненно присутствует. В данном случае речь идет о «readability» в зависимости от тех или иных параметров (см. главу 3).

Любой естественный текст характеризуется компрессией как результатом эллиптирования некоторого количества информации. Эллиптирование может происходить на самых разных уровнях – от фонетического до смыслового. Эллиптирование говорящим тех или иных смысловых фрагментов зависит от коммуникативной ситуации, прежде всего – функционального стиля текста и от соответствия «баз знаний» говорящего и слушающего (адресанта и адресата): если слушающий знает предметную область, владеет темой разговора, то говорящий (в силу закона экономии усилий), как правило, опускает ту информацию, которая может

быть восстановлена слушающим на основании этого знания. Таким образом, восстановление компрессированного текста адресатом в процессе восприятия оказывается обязательным компонентом, обеспечивающим успешность коммуникации. «Требуемая» адресату информация восстанавливается на основании контекста[12].

§ 2.3. Единица анализа и контекст. Коллокации и конструкции.

При восприятии и порождении (анализе и синтезе) текста неизбежно используются единицы разного масштаба, разной степени связанности и разных уровней иерархии. Эти единицы «задаются» характеристиками языка и контекста, предпочтение тех иных единиц имеет ярко выраженную вероятностную природу. В качестве такого рода оперативных единиц могут выступать как синтаксические, так и лексические единицы (под последними понимаются разнообразные обороты, единицы, эквивалентные слову и т.д. – см., напр., [143] и словарь оборотов www.ruscorpora.ru/obgrams.html).

Однако начнем с попытки разобраться в вопросах терминологии.

В современной лингвистике, ориентированной, с одной стороны, на функциональность и антропоцентричность описания, а с другой стороны – на возможности корпусной лингвистики, уже практически очевидна необходимость использования основных положений грамматики конструкций и близких к ней научных направлений. Подход «GxC» (грамматики конструкций) начал разрабатываться с 1970х годов и чрезвычайно популярен в разных направлениях современной лингвистики: [23; 24; 34; 37; 38; 65] и многие другие; подробную библиографию см. в http://constructiongrammar.org/.

Так что же такое «конструкция»? Кажется, стало уже традицией опираться на те свойства конструкций, которые были указаны Филмором [26]. Сформулируем основные (во всяком случае для наших исследований) признаки:

- конструкции состоят из «родительских» и «дочерних» элементов, отношения между которыми могут различаться по степени жесткости;
- конструкции могут определять не только синтаксические, но и лексические, семантические, прагматические параметры;
- в конструкцию могут быть включены лексические единицы;
- конструкции могут (и в некоторых случаях должны) быть идиоматичными, тогда семантика конструкции как целого будет шире семантики составляющих элементов.

Множество таким образом определяемых конструкций очень неоднородно: они будут различаться степенью и типом идиоматичности, жесткостью и закрепленностью определенных лексем (классов лексем).

При широком понимании такого подхода любая синтаксическая единица является конструкцией, статус такой единицы-конструкции зависит от классификации по названным параметрам.

[12] Мы понимаем контекст широко: от того контекста, который не выходит за пределы текста, до контекста коллекции (базы текстов) или коммуникативной ситуации. «Требуемая» информация заключена в кавычки, т.к. адресат может приписывать тексту тот смысл, который не был ему присущ (в силу сильного желания носителя языка или ошибки обработки у автомата).

Однако наиболее важным с точки зрения функциональности конструкции является ее положение в дихотомиях лексикон vs. синтаксис, инвентарные vs. конструктивные единицы (по В.Б.Касевичу [108]), номинации vs. предикативные единицы. Эти дихотомии (шкалы) функционально близки, но все же они не тождественны. Наиболее «типовые» (на наш взгляд) конструкции оказываются, прежде всего, синтаксическими и предикативными единицами, возможно, они являются конструктивными, но высокочастотными единицами. Степень жесткости отношений между компонентами конструкции может существенно различаться.

В предельном случае мы имеем дело с ориентацией на радикальный вариант грамматики конструкций У.Крофта (Radical Construction Grammar), отрицающий композициональность конструкций, т.е. не конструкции конструируются из элементов более низких уровней иерархии (напр., слов), а слова могут вычленяться в результате последующих процедур обработки из целостной конструкции [15; 16].

Другой вариант грамматики конструкций у Филлмора, реализующего проект «Конструктикон» как продолжение идей и принципов лексикографического проекта FrameNet на материале корпуса предложений с разметкой конструкций [25]. Филлмор вводит свою терминологию и – главное – схему описания конструкций: «Constructions are the rules that license ‘new’ linguistic signs based on other linguistic signs. The structures licensed by one or more constructions are called CONSTRUCTS, following the terminology of Sign-based Construction Grammar. A construction can be described formally, in Attribute-Value Matrix form, or informally in prose, but annotation must be of constructs: each annotation captures the properties of a particular construct with respect to a particular construction that licenses it»[13] [25: 9]. В его проекте делается попытка скорее сблизить синтаксис и лексикон: «There were numerous reasons for trying to articulate a lexicon with a constructicon: serious work in lexical description was unable to escape the need to appeal to features of grammar that go beyond the basic structures that define ordinary valence satisfaction…»[14] [там же: 47].

В рамках парадигмы корпусных и когнитивных исследований нас интересует изучение лексико-грамматических явлений (вернее было бы даже сказать: лексических и морфолого-синтаксических явлений) при восприятии и порождении (анализе и синтезе) текста. Поэтому для нас наиболее интересным является объединение идей, заложенных в моделях грамматики конструкций и различных контекстно-ориентированных моделях (от широко известной «Контекстуальной теории значения» (Contextual Theory of Meaning) Ферса (см., напр., [29; 30] до современных Usage based models (см. обзор в [5]).

Как известно, в процедурах обработки текста происходит максимальная опора на контекст. Причем понятие «**контекст**» также рассматриваться в разных смыслах. Для нас контекст предполагает широкое понимание:

•**минимальный контекст**, в котором реализуются лексические и морфолого-синтаксические явления;

---

[13] Конструкции – это правила, которые легализуют «новые» языковые знаки на основе других языковых знаков. Структуры, легализованные одной или более конструкцией именуются КОНСТРУКТОМ, следуя терминологии основанной на знаках грамматики конструкций. Конструкции могут описываться формально, в виде матрицы «*атрибут-значение*», или неформально с помощью текстового описания, но аннотироваться должны именно конструкты: каждая аннотация описывает свойства конкретного конструкта с отсылкой на конструкцию, которая его лицензирует.

[14] Существует множество причин, чтобы пытаться связать лексикон с конструкциями: серьезная работа по описанию лексикона не может избежать привлечения грамматических свойств, которые выходят за пределы базовых структур, описывающих простое заполнение валентностей.

•**текстовый контекст**, включающий в себя фрагменты текста вплоть до текста целиком;

•**контекст коллекции** (базы текстов), предполагающий учет текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т.д.) (подробнее см. [158].

Можно было бы добавить еще одно понимание контекста: как совокупности текстового опыта человека, а также тем самым – знание языка (на основании опыта по восприятию и порождению текстов). Такое понимание «широкого контекста» в известной степени моделируется в создании и последующем изучении Национальных корпусов.

Процедуры обработки текста носят вероятностный характер. Безусловно вероятностный характер носит обработка (восприятие, понимание) текста человеком (начиная со старых работ, напр., [100; 142] и т.д.). О вероятностном характере процедур обработки текста мы можем говорить в отношении многих систем автоматического понимания текста (ср., напр., системы кластеризации новостных текстов на новостных порталах или машинный перевод, основанный на статистическом анализе). Возможны, наконец, процедуры автоматического анализа текста, моделирующие стратегии обработки текста человеком.

Степень связанности конструкций, по всей видимости, зависит от вероятностной модели, описывающей появление этой конструкции в ходе процедур обработки текста. Вероятные оценки могут быть получены лишь на основании статистических данных. Причем статистические характеристики должны описывать данные в зависимости от перечисленных выше типов контекста.

Что же из себя представляет «**коллокация**»? Сравним несколько определений этого понятия. «Collocations of a given word are statements of the habitual or customary places of that word[15]» [29: 181]. «A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things[16]» [64: 141].

В отечественной литературе достаточно часто встречается понимание лингвистами коллокаций как несвободных сочетаний, не относящихся к идиомам, когда, с одной стороны, ключевое слово этих сочетаний может появляться в контексте разных языковых единиц, с другой стороны эти единицы (т.е. контекст ключевого слова) можно перечислить в виде закрытого («полузакрытого») списка (ср., напр., работы Л. Н. Иорданской, И. А. Мельчука и их последователей по исследованию лексических функций и моделей управления[17]).

Термины «открытый / закрытый список» в нашу работу пришли из традиций исследования восприятия речи и обработки информации человеком или автоматом. Закрытый список традиционно задается в форме перечисления всех потенциальных ответов системы, например, отраженный в словарях список неоднословных служебных слов. Более сложный вариант закрытого (или вернее «полузакрытого») списка задается с помощью закрытого списка значений признака (или набора признаков). Например, набор лексических вариантов, или лексических функций, или моделей управления и т.д. Открытый список не предполагает никаких заранее

[15]«Коллокации заданного слова – это установление обычных или привычных мест этого слова».

16«Коллокация – это выражение, состоящее из двух или более слов, которое соотносится с некоторым способом говорения».

17См. подробнее в [106; 47]; сейчас такие работы ведутся на основе Национального корпуса русского языка (НКРЯ), в частности, представленные на http://dict.ruslang.ru/ [121; 95].

заданных ограничений. Для нас исследование лексических вариантов, лексических функций, моделей управления или других параметров является этапом интерпретации данных, полученных в виде открытого списка (прежде всего, на основании статистических характеристик). Чаще всего принцип выделения коллокаций (в идеале список) отражает традицию определенной школы (и собственную интуицию исследователя) или узко заданную изучаемую тему. Даже в традициях русистики существует огромное количество терминологических и теоретических сложностей, что отражается в различии трактовок в словарях и грамматиках. В качестве примера позволим себе цитату из предисловия к электронному ресурсу «Словарь русской идиоматики» (это один из словарных ресурсов, создаваемых на основе Национального корпуса русского языка [121: 2],): «... в отечественной традиции принято различать собственно фразеологизмы (идиомы), в которых исходное значение полностью переосмысляется (*медведь на ухо наступил, ломиться в открытую дверь*), и коллокации, в которых одно слово выступает в своем обычном значении, а другое – во фразеологически связанном (*плакать* ***навзрыд***, ***в стельку*** *пьяный*)». Это предисловие как бы примиряет отечественные традиции и современные парадигмы корпусной лингвистики. Все чаще приходится признавать, что, несмотря на явную неоднородность выделяемых списков, границы между классами оказываются проницаемыми. В словаре представлены «наряду с настоящими идиомами (фразеологизмами, ср. *круглый сирота*) и коллокациями (ср. *плакать навзрыд, диаметрально противоположный*), менее идиоматичные (ср. *глубоко огорчен*), а также свободные (семантически мотивированные, ср. *чрезвычайно огорчен*) сочетания со значением высокой степени)» [121: 2]. Такое решение создателей ресурса отвечает основным задачам контекстно-ориентированных и корпусных исследований.

Попытки последовательно учитывать контекст (причем – как указывалось выше – разные типы контекстов) ставят перед исследователем дополнительные задачи. Обычно получаемые в работах списки коллокаций лишь в некоторой степени могут быть соотносимы с исследованием тех особенностей, которые не просто заложены в языке (всех текстах на этом языке), но в существенной степени зависят от типа контекста (напр., от функционального стиля текстов, конкретной коллекции или отдельного текста по отношению к этой коллекции).

Реализовать контекстно-ориентированный подход можно с использованием различных статистических мер, позволяющих автоматически выделить из текстов коллокации и ранжировать их по степени неслучайности в соответствии со значениями выбираемых мер [76]. При этом нечеткое и интуитивное понятие контекста принимает черты объективности – в узком смысле под контекстом понимается та коллекция, на которой проводится исследование. Возможность варьировать коллекции (например, выбирая коллекции текстов разных функциональных стилей или даже отдельные тексты из этих коллекций) позволяет получать списки коллокаций, различающие различные контексты. Именно текстовый материал, реализация лексико-грамматических и синтаксических проявлений, оказывается базой для исследования.

Понимание терминов «коллокация» и «конструкция», как уже было сказано, оказывается различным в зависимости от той или иной парадигмы. Во многих случаях одни и те же единицы могут быть названы и «коллокацией», и «конструкцией». Отдельно стоит прагматический признак: в прикладных

исследованиях автоматической обработки текста, как правило, можно встретить термин «коллокация». В настоящее время появляются первые попытки использовать «конструкции» в прикладных исследованиях: ср. [69; 55][18].

Если пытаться разделить эти термины «по совокупности пониманий», то получится некоторое градуальное противопоставление: т.е. «скорее конструкция» vs. «скорее коллокация».

Мы предлагаем некоторую схему классификации, задающей основные параметры такого разделения. В ходе наших исследований эта схема оказалась плодотворной. Однако на настоящем этапе положения данной классификации представляются набором гипотез, которые, несомненно, надо верифицировать, и верификация должна происходить именно с опорой на контекст как материал анализа.

Чаще всего, термин «коллокация» используется при решении задачи выделения и описания неоднословных ***номинаций*** (не только в прикладной области). Ср. примеры из [45: 150]: *strong vs. powerful tea* '*сильный* vs. **сильный чай*', т.е. сочетаемостные ограничения, диктующие выбор прилагательного *strong* для '*сигарет, чая и кофе*' (*cigarettes, tea and coffee*), но *powerful*, напр., для '*героина*' (*heroin*). Неоднословные номинации наподобие *белый медведь*, *белый гриб*, *белое вино* или *проливной дождь*, *заклятый враг* очевидным образом ложатся в таким образом понимаемую идею коллокаций. Более того, такие традиционные признаки как «устойчивость» и «идиоматичность» (ср. [128]) в известной степени переосмысляются. Колокации выходят за пределы исследования «чистой фразеологии», зачастую их целостность как единой номинации оказывается более значимым признаком, а под устойчивостью понимается скорее степень неслучайности совместной встречаемости слов. Такое понимание устойчивости ощущается носителем языка и может быть выявлено в ходе экспериментов с информантами. Так, например, для анализируемых нами новостных и научных текстов среди таких коллокаций выступают самые разные с лингвистической точки зрения неоднословные номинации: *непосредственная близость*, *стихийное бедствие*, *Нижний Новгород*, *Саудовская Аравия*, *Бритни Спирс*, *Невский экспресс* и *корпусная лингвистика*, *речевой акт*, *именительный падеж*, *речевой сигнал*, *концептуальный граф*, *внешний посессор* соответственно.

Таким образом, коллокации достаточно часто выступают в качестве важной и частотной единицы словаря. Ср. цитату «Lexical unit is a word or collocation[19]» в начале аннотации к статье [19]. Действительно, практические задачи автоматической обработки текста (напр., информационный и фактографический поиск) чаще всего связаны с поиском и идентификацией разнообразных сложных номинаций. Таким образом выделяются неоднословные термины, могут определяться предметные области и ключевые словосочетания, характеризующие заданную коллекцию текстов или ее подвыборку, и т. п. Именно коллокации, соответствующие неоднословным номинациям, по всей видимости могут претендовать на статус «ядерных коллокаций». В этом смысле можно было бы представить себе даже более представительную шкалу: от слова до коллокации, от колокации к конструкции. Тогда «коллокация» будет представляться как бы в виде промежуточного звена и перевалочного пункта при движении от слова к конструкции.

[18] Впрочем, показательно, что даже в этих и других работах «Workshop on extracting and using constructions in NLP» активно используется именно термин «коллокация».

[19] «Лексические единицы – это слова или коллокации».

Конструкции, напротив, чаще всего представляют собой единицы скорее синтаксического плана. Таким образом, типовые или ядерные коллокации и конструкции часто могут оказаться противопоставленными как парадигматические vs. синтагматические единицы; инвентарные vs. конструктивные единицы; единицы, принадлежащие лексикону vs. синтаксису; номинации vs. предикативные единицы. Предикативность анализируемых единиц понимается, прежде всего, как потенциальная возможность занять позицию предиката в предложении. Таким образом, наиболее явная предикативность будет у сочетаний с вершиной в виде глагола в личной форме (хотя, конечно, не исчерпывается этим типом сочетаний).

Впрочем, и здесь проявляется неоднозначность, т. к. предикативные образования, обладающие высокой степенью воспроизводимости и/или идиоматичности, будут, по всей видимости, распределены по шкале(-ам) движения от коллокации к конструкции ближе к конструкциям. Приводимые выше *медведь на ухо наступил*, *ломиться в открытую дверь*, *плакать навзрыд*, *в стельку пьяный* и т.д. окажутся в зоне конструкций именно благодаря ярко выраженной предикативности. Однако для того, чтобы о них зашла речь, необходимо, чтобы они оказались реализованными в текстах и – соответственно – выделимыми с помощью статистических мер. Те, кто работает с коллекциями и корпусами, знают, что многие фразеологизмы в текстах встречаются довольно редко.

Особое внимание обратим на одно из традиционных свойств конструкций по Филмору [26]: лексические единицы могут быть включены в конструкцию. Следовательно, существует противопоставление с точки зрения включенности фиксированных лексем (вернее словоформ) или лексем, принадлежащих фиксированной лексико-семантической группе: напр., *А еще **N** называется!* (*А еще **друг** называется!*) (один из многочисленных примеров «синтаксических фразем», собранных и проанализированных в диссертационном сочинении М. Копотева [118: 125]). К данному типу конструкций относятся многие клише: высокочастотные конструкции, характерные для определенного типа текстов (например, сообщений из новостных лент), которые носят скорее казенный характер и возможно, воспринимаются как излишне навязчивые. Однако группа клише выделяется, прежде всего, на основании стилевого (и стилистического) набора признаков: к клише относятся те сочетания, которые маркируют специфический стиль («казенный», подчеркнуто навязчивый). Поэтому среди клише мы можем найти не только типовые конструкции (клишированные конструкции) с ярко выраженной предикативностью. Среди клише могут оказываться также предложно-падежные сочетания (напр., *со ссылкой*, *по данным*, *в настоящее время*), дискурсивные слова, производные служебные слова, если эти единицы высокочастотны для рассматриваемой коллекции, и их отличают особые стилевые характеристики. Под устойчивыми сочетаниями понимаем, прежде всего, дискурсивные слова, производные служебные слова, наречные образования и предложно-падежные сочетания наподобие *со ссылкой*, *по данным* и т.д. Таким образом, клише пересекается и с конструкциями, и с устойчивыми сочетаниями. Использование термина клише в нашей статье целесообразно именно в силу того, что материал анализируется по многим факторам; клишированность сочетаний выступает как своеобразный дополнительный параметр анализа, с одной стороны, необходимый в силу того, что он очевидно связан с частотностью, а с другой – как бы «перпендикулярный» заявленной шкале (-ам) «от коллокации к конструкциям».

Забегая вперед, упомянем, что конструкции-клише – напр., «введения источника информации» – высокочастотны в текстах портала lenta.ru: *сообщает РИА* 17081, *сообщает агентство* 10590, *пишет газета* 7722, *передает агентство* 7683, *передает РИА* 4487 (эта часть нашего анализа осуществлялась на коллекции [116], около 300 миллионов словоупотреблений; приведенные числа обозначают частоту встречаемости). Для информационно насыщенных коллекций (наподобие портала lenta.ru, подробнее см. следующий пункт) конструкции, выделяемые на основании статистических мер, могут достигать длины более 5 словоупотреблений (напр., «*сообщает Интерфакс со ссылкой на* ***источник в правоохранительных органах»*** из «*сообщает Интерфакс со ссылкой на* ***N***»). Полагаем, что именно такой тип единиц занимает место «прототипической конструкции» на шкале(-ах) «от колокации к конструкциям»: она частотна, синтаксична, предикативна и синтагматична, в вершине («родитель») глагол в личной форме.

Отдельного внимания заслуживает производная служебная лексика (напр., предлоги *в течение*, *в качестве*) и дискурсивные слова (напр., *по крайней мере*, *может быть*). Они чаще всего выступают под маркой «сочетаний, эквивалентных слову», хотя степень устойчивости этих единиц может существенно различаться, что, в частности, находит отражение в словарях (напр., [96]). Где они должны быть сосредоточены на шкале(-ах) движения от коллокации к конструкции? Полагаем, что в качестве условного приближения можно допустить, что они расположены в некоторой серединной зоне, равноудаленной и от «ядерных коллокаций», и от «ядерных конструкций». Это зона распределения соответствующих «сочетаний, эквивалентных слову» (термин заимствован из «Толкового словарь сочетаний, эквивалентных слову» Р.П. Рогожниковой [143], но, конечно, принципы выделения и множество единиц существенно отличается от того, что представлено в словаре). Чем выше предикативность (особенно для дискурсивных слов и наречных образований), тем они оказываются ближе к конструкциям. Другим параметром является степень устойчивости: чем выше она, тем эти единицы оказываются ближе к полюсам сосредоточения коллокаций как целостных единиц словаря (мы сейчас абстрагируемся от лингвистического анализа процессов фразеологизации).

Напомним, что предикативность понимается нами как возможность занять позицию предиката в предложении, что сравнительно часто может относиться к дискурсивным словам и наречным образованиям.

В качестве условного приближения мы сочли, что производная служебная лексика, наречные образования, а также дискурсивные слова находятся в некоторой серединной зоне. Для данной статьи это разумное допущение, т.к. в ней анализируются коллекция и неоднословные единицы (от коллокаций до конструкций), характеризующие коллекцию в целом. На следующих этапах анализа и интерпретации, когда рассмотрению подлежат характеристики как коллекций, так и конкретных текстов, составляющих эти коллекции, шкалы конкретизируются. На следующих этапах анализа оценивается то, насколько степень удаленности от «ядерных коллокаций» и/или от «ядерных конструкций» зависит от конкретной шкалы. Так, например, *по крайней мере*, *может быть* оказываются ближе к коллокациям в шкалах словарь vs. грамматика и инвентарные vs. конструктивные единицы, но ближе к конструкциям в шкале номинация vs. предикативная единица, парадигматика vs. синтагматика.

Цели исследования и способы решения поставленных задач вынуждают нас двояко рассматривать анализируемые единицы с точки зрения того, включают ли они слоты или представлены в виде фиксированного лексического наполнения. Слоты или, другими словами, лексические элементы, которые могут варьироваться, нас интересуют в тех конструкциях (или «скорее конструкциях»), в которых наличие слотов – и варианты их заполнения – важны для решения определенных задач (прежде всего, задач анализа текстов). Сошлемся на приведенные выше примеры конструкций введения источника информации, где слот представляет собой тот самый источник информации: *сообщает X, сообщает Интерфакс со ссылкой на* ***N***. В случае исследования, например, производной служебной лексики мы останавливаемся на варианте представления в виде фиксированного лексического наполнения: *в зависимости от*, а не *в зависимости от X*. Причина выбора такого варианта рассмотрения в предполагаемой информационной незначимости возможных видов заполнения слота – для решения задач анализа текстов. Если при анализе какой-либо коллекции выявляется явное статистическое предпочтение одного или нескольких вариантов заполнения потенциального слота X, производный предлог «сдвинется» в сторону конструкции со слотом (напр., представим себе такую коллекцию, где в конструкции *в зависимости от X*, X предпочитает принимать значение *цели*, *задачи* или *гипотезы*).

§ 2.4. Типы коллокаций и конструкций. Принцип шкалирования

Описание материала

Главное требование к материалу и методике в экспериментальном исследовании – в данном случае это вычислительный эксперимент – адекватность целям и задачам. Применительно к лекциям это требование дополняется еще важностью доказательной силы и **наглядности**. В качестве основного материала в наших иллюстративных примерах использовались три коллекции новостных и научных текстов:

- портала www.lenta.ru 2009; общий объем проанализированных текстов: более 66000000 «токенов» (словоупотреблением и знаков препинания);
- материалов конференции «Корпусная лингвистика» 2004-2008 года (монотематическая коллекция); объем коллекции составляет около 220000 «токенов»;
- материалов международной конференции «Диалог» «Компьютерная лингвистика и интеллектуальные технологии» за 2003-2009 годы; объем коллекции составляет около 2500000 «токенов».

Привлекался также дополнительный материал (новостные источники, отличающиеся от Ленты.ру по жанру, предметной области, стилевым и прочим характеристикам, связанным со степенью информационной насыщенности): «РИА Новости», «РосБизнесКонсалтинг», «Компьюлента», «Независимая газета»[20]. Дополнительный материал анализируется только тогда, когда описанные на материале Лента.ру особенности характеризуют новостные тексты только одного жанра (напр., текстов сообщений новостной ленты), и отличаются при смене жанра (или других стилевых параметров).

[20] Эта часть работы подробно описывается в [162].

Морфологическая разметка коллекций осуществлялась В.В. Бочаровым при помощи свободно распространяемого программного обеспечения АОТ (www.aot.ru). Для разметки использовался, в первую очередь, модуль морфологической анализа; модуль синтаксического анализа использовался для частичного снятия морфологической омонимии. В тех случаях, когда полностью снять омонимию не удавалось, для анализа использовалась первая из предложенных анализатором лемм, т.е. неоднозначность разбора просто игнорировалась. При выделении коллокаций учитывались знаки препинания: рассматривались любые последовательности слов в тексте, не разделенных знаками препинания.

Главной задачей методики было намерение разделения биграмм – уже на этапе применения статистических мер – на указанной шкале от коллокаций к конструкциям[21]. Нами использовались две меры: *MI* [10] и t-score [11].

$$MI = log_2 \frac{f(c_1, c_2) \times N}{f(c_1) \times f(c_2)}, \quad (1)$$

$$t - score = \frac{f(c_1, c_2) - \frac{f(c_1) \times f(c_2)}{N}}{\sqrt{f(c_1, c_2)}} \quad (2)$$

где

c_i – коллокаты;

$f(c_1,c_2)$ – абсолютная частота встречаемости коллокации c_1 c_2, с учетом порядка коллокатов внутри биграммы;

$f(c_1)$, $f(c_2)$ – абсолютные частоты c_1 и c_2 в корпусе;

N – общее число словоупотреблений в корпусе.

С точки зрения теории вероятности, мера MI (mutual information, коэффициент взаимной информации) является способом проверить степень независимости появления двух слов в тексте — если слова полностью независимы, то вероятность их совместного появления равна произведению вероятностей появления каждого из них, т. е. произведению частот, а значение меры MI равно нулю.

Недостатком меры MI является ее свойство завышать значимость редких словосочетаний. Чем более редки слова, образующие коллокацию, тем выше будет для них значение MI, что делает данную меру совершенно «беззащитной» перед опечатками, окказионализмами, иностранными словами и другим информационным шумом, который неизбежен в большой коллекции. Поэтому для данной меры используется порог отсечения по частоте. К сожалению, правильный подбор порога отсечения оказывается чрезвычайно сложной задачей. Верно и обратное: мера MI оказывается беззащитной в том случае, если хотя бы один из коллокатов имеет (сверх)высокую частоту встречаемости, напр., она не сможет выделить такие предлоги как *в качестве*, *в зависимости*, *в отличие (от)* в силу того, что предлог «в» всегда имеет сверхвысокую частоту.

Другой мерой, которая использовалась в данном исследовании, стала мера *t-score*, которая учитывает частоту совместной встречаемости ключевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами.

[21] Подробнее о методике для рассматриваемого типа исследования см. (Ягунова, Пивоварова 2011; Пивоварова 2010).

Данная мера используется гораздо реже, чем мера MI, в частности, потому что она является лишь несколько модифицированным ранжированием коллокаций по частоте. Очевидно, что значение данной меры тем выше, чем выше частота коллокации в коллекции. Хотя данная мера содержит коррекционный компонент — вычитание деленного на размер коллекции произведения частот коллокатов, однако эта поправка отражается лишь на самых частотных словах. Stubbs [Stubbs 1995] показывает (на примере английского языка), что значение меры t-score для знаменательных слов примерно равно $\sqrt{f}(n, c)$ и лишь для служебных заметно меньше этого значения. В литературе эта особенность часто трактуется как малопригодность этой меры для поиска терминологических словосочетаний и номинаций; для этой цели она, как правило, не используется. Естественно, что мера t-score, в отличие от MI, не преувеличивает значимость редких коллокаций и не требует использования порогов отсечения.

В нашем исследовании мы учитывали порядок коллокатов внутри биграммы.

Меру MI можно обобщить для любого числа коллокатов, в данном случае мы рассматриваем результаты, полученные с помощью [72]:

$$MI = \log_2 \frac{f(c_1, c_2, ..., c_i) * (N^{(i-1)})}{f(c_1) * f(c_2) * ... * f(c_i)}, \quad (1a)$$

где i – число коллокатов, остальные условные обозначения те же, что и для формул 1 и 2.

Обобщение меры t-score для коллокаций длиннее, чем биграммы, в литературе не встречается. Причиной этого может быть тот факт, что мера t-score является аппроксимацией частоты, которая за счет поправочного коэффициента «понижает» значимость словосочетаний, состоящих из двух очень частотных слов (например, двух союзов или союза и предлога). Поскольку сами коллокаты очень частотны, такие коллокации становится частотными просто в силу вероятностных причин. Однако чем больше число коллокатов входит в коллокацию, тем меньше сила этого эффекта (не говоря уже о сомнительности появления в тексте, например, трех союзов подряд). Поэтому для многословных коллокаций использование t-score не представляется осмысленным, а сама частота становится более надежным источником информации, чем для биграмм. В нашей работе для многословных сочетаний используется собственно частота коллокации (вместо расширенного варианта t-score).

Вопрос о выборе первичной лексической единицы анализа – лексемы и/или словоформы – для русского языка (как языка с развитой морфологией) всегда решается неоднозначно; эти единицы отражают разные аспекты и уровни лексико-грамматической информации об исследуемых единицах (см. ниже).

MI-коллокации

Как уже говорилось, под типичными коллокациями в нашей классификации мы понимаем прежде всего неоднословные номинации и сложные термины. Более того, такие колокации зачастую выходят за пределы «чистой фразеологии», их целостность как единой номинации оказывается более значимым признаком, а под устойчивостью понимается скорее степень неслучайности совместной встречаемости слов.

Коллокации достаточно часто выступают в качестве важной и частотной единицы словаря. В этом смысле «ядерные» колокации могут рассматриваться не только на шкале от «коллокации до конструкции», но и на дополнительной шкале «от слова до коллокации».

А что такое «слово»? Не углубляясь в неоднозначность определения – казалось бы – ведущей единицы языка и речи, вспомним о наличии противоречий даже на этом уровне. Что является единицей анализа текста: лексема или словоформа? Можно считать более чем обоснованным и экспериментально доказанным положение о том, что словоформа является ведущей единицей анализа русского текста (лексема выполняет роль дополнительной единицы анализа, востребуемой лишь в особых случаях) [112; 115]. Вероятно, такое противопоставление роли лексемы и словоформы, отчасти обусловлено типологическими характеристиками русского языка как флективного языка с богатой морфологией.

При работе с коллокациями выбор основной единицы анализа представляет собой дополнительный вопрос: лексема или словоформа? [22]

На материале новостных текстов был проведен предварительный сопоставительный анализ списка сочетаний, выделяемых для лексем (но не словоформ), списка сочетаний, выделяемых для словоформ (но не лексем) и списка сочетаний, выделяемых и для лексем, и для словоформ (подробнее см. статью [159])[23].

Биграммы, выделяющиеся и для лексем, и для словоформ, оказываются, как правило, наиболее информативными.

В список (только) лексемных биграмм попадают составные номинации, характеризуемые максимальной свободой (максимальным разнообразием, минимальной ограниченностью) набора выполняемых ими в предложении семантико-синтаксических ролей. Примеры этих биграмм, каждая единица сочетания приведена в нормализованном виде (прописными буквами – здесь и далее):

- для новостных текстов – *КУРМАНБЕК БАКИЕВ, АЛИШЕР УСМАНОВ, БЕНЕДИКТ XVI, УСЕЙН БОЛТ, СЕРДЕЧНЫЙ ПРИСТУП, ОСАМА БИН, СТИХИЙНЫЙ БЕДСТВИЕ, ЛАМПА НАКАЛИВАНИЕ, РАДОВАН КАРАДЖИЧ, ПОЛЕЗНЫЙ ИСКОПАЕМОЕ, ДЖОННИ ДЕПП, ФИДЕЛЬ КАСТРО, ДОЛИНА СВАТ, САДДАМ ХУСЕЙН, СИМФОНИЧЕСКИЙ ОРКЕСТР, КРОВНЫЙ МЕСТЬ, и т.д.;*
- для научных текстов – *ВИНИТЕЛЬНЫЙ ПАДЕЖ, ИМЕНИТЕЛЬНЫЙ ПАДЕЖ, АКТУАЛЬНЫЙ ЧЛЕНЕНИЕ, ИНСТРУМЕНТАЛЬНЫЙ СРЕДА*.

Показательна высокая доля, которую имеют в этом классе наименования лиц. Такие номинации, условно говоря, **можно сопоставить со словом**, которое характеризуется достаточно полной парадигмой формоизменения.

Словоформные биграммы, как правило, относятся к номинации в определенной синтаксической позиции. Примеры биграмм:

- для новостных текстов – *парниковых газов, Соединенных Штатов, Женской Теннисной, кредитном портфеле, Палестинской автономии, встречную полосу, Нижнем Новгороде, Федеральную трассу*;
- для научных текстов – *речевой акт, речевых актов, именная группа, именных групп, коммуникативного акта, коммуникативных актов, просодических характеристик, прошедшего времени, речевого сигнала*.

[22]Хочется отметить, что различные аудитории, обсуждавшие наши доклады на эту тему, высказывались весьма категорично: некоторые аудитории лишь лексемные коллокации считали достойными внимания, другие – напротив – только словоформные. Безусловно, основные особенности, рассмотренные на примере биграмм-коллокаций, действуют и при увеличении объема сочетания.

[23]Во всех трех случаях под «списком» имеется в виду первая сотня словосочетаний, выявленных тем или иным способом. Нас интересует, однако, словосочетания с наибольшим значением меры, т.е. верхние части списков, которые мы в дальнейшем для краткости именуем просто списками.

Кроме того, биграммы этого подкласса могут относиться к части целостной номинации, например, сочетание *речевых актов* часто является частью триграммы «*теории речевых актов*».

В этих списках в обоих случаях некоторая составная номинация или термин резко тяготеет к выполнению некоторой типичной (излюбленной) для неё семантико-синтаксической роли (то есть «излюбленная» роль для этой номинации оказывается гораздо употребительнее остальных возможных для неё ролей). Такое тяготение является частным проявлением более общего закона тяготения номинативных единиц некоторого грамматико-семантического разряда к выполнению некоторой типичной для них семантико-синтаксической функции. Такое тяготение оказывается важным и для однословных номинаций, и для неоднословных.

Если данная составная номинация входит в состав некоторого более крупного – трёхсловного или даже более протяжённого, напр., *(Женской теннисной) ассоциации*, *теории (речевых актов)* – сочетание является более устойчивым на синтагматической оси, чем в случае прочих словоформных биграмм (допускающих более свободные связи с соседями на синтагматической оси).

Таким способом мы выделяем наиболее информационно-нагруженные и точные сочетания, характеризующие данную коллекцию (см. напр., биграммы в Таблицах 1, 2 и 3). Для простоты восприятия в таблицах биграммы представлены в виде сочетаний словоформ (соответствующей словоформной биграмме). Ведущее место в ней отводится интересующим нас «ядерным коллокациям». Однако в таблице присутствуют и сочетания, рассматриваемые нами в следующем пункте **MI-конструкции** (особенно для научных коллекций).

Таблица 1. Пример пересечения между биграммами для лексем и для словоформ (для первой сотни, в порядке убывания значения меры). Материал портала lenta.ru 2009 года

ранг (для лексем)	ранг (для словоформ)	биграммы
1	1	Бритни Спирс
2	2	Эльвира Набиуллина
3	23	Ле Бурже
9	36	Лионель Месси
10	4	мысе Канаверал
11	43	бин Ладена
14	9	Норильского никеля
15	7	дельты Нигера
17	50	Ак Барс
18	28	тротиловом эквиваленте
19	20	тройскую унцию
20	70	Ролан Гаррос
26	49	дель Торо
27	87	дель Потро
29	33	Арбат Престиж
31	96	РАО ЕЭС
32	35	Салават Юлаев
34	51	Арсений Яценюк
36	42	голубых фишек

Таблица 2. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ (в порядке убывания значения меры). Материал конференции «Корпусная лингвистика»[24]

ранг	Биграммы	ранг	Биграммы
2	наш взгляд	36	одной стороны
3	(по) крайней мере	37	таким образом
4	речевой деятельности	40	разрешения неоднозначности
5	художественной литературы	41	английский язык
7	первую очередь	43	кроме того
9	общим объемом	47	Национальный корпус
11	корпусная лингвистика	48	грамматических категорий
13	имена собственные	52	устная речь
15	математической лингвистики	54	база данных
16	словарной статьи	58	во многих
17	свою очередь	61	лексических единиц
18	предметной области	62	дает возможность
19	машинного перевода	63	зависит от
20	точки зрения	64	отличие от
22	за счет	65	русский язык
24	речь идет	67	корпусные данные
25	прежде всего	68	отличается от
26	большое количество	71	зависимости от
28	настоящее время	72	работы над
31	представляет собой	79	частей речи
32	млн словоупотреблений	80	во всех
34	другой стороны	84	при помощи
35	семантических состояний	86	морфологической разметки

Таблица 3. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ (в порядке убывания значения меры). Материал конференции «Диалог».

ранг	Биграммы	ранг	Биграммы
1	ударном слоге	28	интеллектуальные технологии
2	концептуальных графов	30	корпусная лингвистика
4	внешним посессором	33	отглагольных существительных
5	оперативной памяти	37	знаки препинания
8	вокального жеста	38	педагогической коммуникации
14	крайней мере	42	основного тона
16	XIX века	46	машинного перевода
17	лингвистического процессора	61	устойчивых словосочетаний
21	положение дел	63	точки зрения
22	первую очередь	70	меньшей мере
25	картине мира	72	вряд ли
26	множественного числа	73	предметной области
		85	вплоть до

[24] Большую длину списка мы связываем с большей однородностью данной коллекции.

MI-конструкции

Большинство клише и конструкций выделяется с помощью меры t-score. Однако некоторые типы клише и конструкций хорошо извлекаются с помощью меры MI (т.е. основываясь на выраженных сочетаемостных ограничениях). Особенно эти разные типы противопоставлены для новостной коллекции. Прежде всего, эти MI-клише и MI-конструкции носят более казенный и (квази)терминологический характер: *злоупотребление должностными полномочиями, причинение тяжкого вреда* и т.д.

Если для новостных биграмм отмечены лишь штучные варианты: конструкция *НАЧИНИТЬ ВЗРЫВЧАТКА* для лексем и *обогащению урана* для словоформ, то в списках триграмм для новостной коллекции клише и конструкции составляют более 30%.

Примеры:

для лексем – *УМЫСЛИТЬ ПРИЧИНЕНИЕ ТЯЖКИЙ, КРАТКИЙ ИЗЛОЖЕНИЕ ПРИВОДИТЬСЯ, ПОДРЫВ НЕВСКИЙ ЭКСПРЕСС, ПРЕВЫШЕНИЕ ДОЛЖНОСТНОЙ ПОЛНОМОЧИЕ, ПСИХОЛОГИЧЕСКИ ВАЖНЫЙ ОТМЕТКА, ДА ПРИЙТИ СПАСИТЕЛЬ, ТЯЖКИЙ ВРЕД ЗДОРОВЬЕ, ВРЕМЕННО НЕДЕЙСТВУЮЩИЙ ЧЕМПИОН, ЗАСЛУГА ПЕРЕД ОТЕЧЕСТВО, ЭКОНОМИЧЕСКИ АКТИВНЫЙ НАСЕЛЕНИЕ* и т.д.;

для словоформ – *злоупотреблении должностными полномочиями, причинение тяжкого вреда, написания данной заметки, превышении должностных полномочий, краткое изложение приводится, совершил аварийную посадку, покончил жизнь самоубийством, превышение должностных полномочий* и т.д.

Приведенные примеры иллюстрируют то, что многие из конструкций имеют явно выраженную предикативность.

Граница между клише и конструкциями во многих случаях нечеткая. Так, напр., *должностные полномочия* могут сочетаться с *злоупотреблением* или *превышением*, с *злоупотреблять* или *превышать*. Общая логика заставляет предполагать чуть большую близость к конструкциям в случаях с глагольной вершиной. По-видимому, можно выделить два фактора, в какой-то степени разводящих клише и конструкции: глагольность и интуитивно ощущаемый казенно-канцелярский аромат сочетаний. Наиболее «правильными» среди выделяемых сочетаний полагаем конструкции типа *такому выводу пришли, фондовые индексы завершили, выглядит следующим образом.*

Граница между коллокациями и клише также нечеткая. Результаты анализа полученных списков позволяют предполагать, что признаками, которые можно считать условно разделяющими коллокации и клише, являются казенный колорит и референциальный статус. Под последним признаком мы понимаем то, что «коллокации» чаще всего включают в себя сложные номинации, обозначающие уникальный объект (или чрезвычайно информационно важный – для рассматриваемого контекста, напр., коллекции – класс объектов) внеязыковой действительности, коллокации-«клише», как правило, относятся к «традиционным» и сравнительно большим классам объектов внеязыковой действительности, напр., коллокациями-клише будут *ВЕТЕРАН ВЕЛИКИЙ ОТЕЧЕСТВЕННЫЙ, КОЛОНИЯ СТРОГИЙ РЕЖИМ, САМОДЕЛЬНЫЙ ВЗРЫВНОЙ УСТРОЙСТВО.*

В целом, можно рассматривать термин «клише» как «перпендикулярный» к шкале «коллокация-конструкция» – он отражает скорее стилистические характеристики, а с морфосинтаксической точки зрения, как ясно из вышеприведенного обсуждения, клише может являться как коллокацией, так и

конструкцией. Отметим также, что клише являются неотъемлемой частью газетного стиля, их обилие в новостных текстах, как нам кажется, можно объяснить следующими условиями:

- большое количество информации, полученной из официальных источников, и как следствие, сильное влияние официально-делового функционального стиля;
- требование оперативности, высокая скорость порождения текстов, которая приводит к многократному использованию одних и тех же шаблонов;
- высокие требования к скорости и качеству усвоения информации, которая для этого должна быть представлена в узнаваемой, всегда одной и той же форме.

Все эти условия приводят к известной шаблонности новостных текстов, существенно облегчающей их обработку в системах автоматического анализа, которые довольно плохо справляются с художественными и художественно-публицистическмими текстами.

t-score-конструкции

Биграммы, выделяемые с помощью меры t-score, кажутся сравнительно легко интерпретируемыми. Даже для новостной коллекции в 80% случаев мы наблюдаем пересечение списка словоформных и лексемных биграмм (ср. табл. 4).

Данная мера позволяет выделять высокочастотные коллокации (в частности, коллокации с высокочастотыми компонентами – прежде всего, предлогами). Она эффективна при поиске «общеязыковых устойчивых сочетаний» (например, составных предлогов) и того, что может рассматриваться как устойчивое сочетание для данной коллекции. В случае со стилистически однородной новостной коллекцией эта мера описывает стилистические особенности данной коллекции, независимо от конкретной тематики сообщений. Выделяемые биграммы относятся к указанию источников информации (напр., *по словам, со ссылкой, РИА Новости)*, места и времени (*в течение, во время, в России*).

Сравнительно многие из рассматриваемых биграмм принято рассматривать как единое слово (напр., составные служебные и дискурсивные слова *в течение, в качестве, может быть*[25]). Интересно, однако, что наряду с ожидаемыми общеязыковыми устойчивыми сочетаниями в списках присутствуют те единицы, которые можно назвать «собственно общеновостными устойчивыми сочетаниями»: напр., *РИА Новости, миллион долларов, миллион рублей, ПО ДАННЫЕ, КАК СООБЩАТЬ, СО ССЫЛКА*[26] (ср. с Таблицей 4).

Выделим несколько основных типов такого рода сочетаний для новостных текстов, маркирующих особенности новостных текстов (см. табл. 4):

- составные служебные и дискурсивные слова, напр., *в течение, в качестве, в ходе, в частности, в результате, пока не, кроме того;*
- сложные номинации, прежде всего, относящиеся к наименованиям источников информации (материал, напр., *РИА Новости*), при переходе к более объемным сочетаниям (три- и более граммам) они входят в состав конструкций «введения источника информации»;

[25]Ср. единицы в Корпусном словаре неоднословных лексических единиц (оборотов) на базе НКРЯ http://www.ruscorpora.ru/obgrams.html

[26] Это, очевидно, составные части более длинных выражений «*как сообщает корреспондент*», «*по данным агенства*», «*со ссылкой на*», которые оказываются среди наиболее частотных три- и более грамм

- колокации-клише (напр., *миллионов долларов*, *миллиарда долларов*), которые при переходе к более объемным сочетаниям могут входить в состав конструкций;
- сочетания, имеющие все показатели конструкций (как правило, компоненты конструкций «введения источника информации»):
  - с глаголом – напр., *сообщает РИА*, *как сообщает*, *это сообщать*,
  - с существительным – напр., *со ссылкой*, *по ссылкам*.

Таблица 4. Биграммы с наиболее высокими значениями меры t-score (в порядке убывания значения меры). Материал портала lenta.ru 2009 года

ОБ ЭТО	об этом
ОДИН ИЗ	по словам
ПО СЛОВО	а также
А ТАКЖЕ	со ссылкой
ПО ДАННЫЕ	ссылкой на
ССЫЛКА НА	по данным
СО ССЫЛКА	кроме того
В РЕЗУЛЬТАТ	РИА Новости
КРОМЕ ТОТ	этом сообщает
РИА НОВОСТЬ	при этом
В ЧАСТНОСТЬ	в том
ЭТО СООБЩАТЬ	в России
МИЛЛИОН ДОЛЛАР	во время
В РОССИЯ	пока не
МИЛЛИАРД ДОЛЛАР	о том
ВО ВРЕМЯ	в результате
ПРИ ЭТО	настоящее время
В КОТОРЫЙ	миллионов долларов
КАК СООБЩАТЬ	связи с
О ТОМ	сообщает РИА
В ХОД	в результате
В ТОТ	в частности
В СВОЙ	миллиарда долларов
ПОКА НЕ	как сообщает

Для научных текстов также выделяется ряд типов t-score-сочетаний, маркирующих научный функциональный стиль (см. табл. 2 и 3):

- составные служебные и дискурсивные слова, напр., *(по) крайней мере, (в) первую очередь, (с) точки зрения, (по) меньшей мере, прежде всего;*
- конструкции и сходные с ними составные обороты: *дает возможность*, *зависит от vs. (в) зависимости от*, *(в) отличие от vs. отличается от* и т.д.

Во введении мы сформулировали – в качестве условного приближения – предположение о том, что производная служебная лексика (напр., предлоги *в течение*, *в качестве*) и дискурсивные слова (напр., *по крайней мере*, *может быть*) расположена в некоторой серединной зоне, равноудаленной и от «ядерных коллокаций», и от «ядерных конструкций». Чем выше предикативность (особенно для дискурсивных слов и наречных образований), тем они оказываются ближе к конструкциям. Другим параметром является степень устойчивости, чем выше она, тем эти единицы оказываются ближе к полюсам сосредоточения коллокаций как

целостных единиц словаря (мы сейчас абстрагируемся от лингвистического анализа процессов фразеологизации).

Соответственно в предлагаемой схеме – в соответствии с признаком предикативности – *в зависимости от* и *в отличие от* находится ближе к середине, а *зависит от* и *отличается от* – чуть ближе к конструкциям.

Степень устойчивости и/или связанности сочетаний уточняется на основании результатов серии экспериментов с информантами и дальнейшей лингвистической интерпретации полученных результатов (подробнее см. [144; 82])[27]. Результаты экспериментов позволили установить дополнительные шкалы, опирающиеся уже не только на значения статистических мер, но и на связность, ощущаемую носителями языка и эксплицируемую в ходе экспериментов. Такой комплексный экспериментальный подход выявил зоны нестабильности в отношении ряда сочетаний (терминологических сочетаний, сложных номинаций, производных служебных слов и т.д.).

В качестве примера зон нестабильности (в соответствии с введением дополнительных шкал, соответствующих результатам экспериментов) приведем некоторые данные по устойчивым сочетаниям (производным служебным словам). Для научных текстов *в частности* и *с помощью* характеризуются большей целостностью и связностью, чем *в качестве*, *за счет*, *на основе*; *с одной стороны*, *с другой стороны*, *по сравнению с* и *в отличие от* характеризуются меньшей целостностью, чем *с точки зрения* и *в соответствии с*. Т.е., напр., морфологическая цельнооформленность *в отличи**е*** *от* не явилось для наивных носителей языка (участников этого эксперимента) решающим признаком для признания высокого уровня целостности и связности.

Аналогично, для новостной коллекции, напр., *этом сообщает*, *в результате* являются менее целостными, чем *как сообщает*, *по данным*; *сообщает РИА Новости*, *об этом сообщается* обладают большей целостностью и связностью, чем *новости со ссылкой*, *по его словам*, *об этом сообщает*.

Данные экспериментов демонстрируют также зависимость от функционального стиля (типа коллекции), напр., *в частности* и *(в) том числе* характеризуется большей целостностью для научных текстов, чем для новостных (подробнее см. [144]). Конечно, окончательный результат будет получен на основании серии взаимодополняющих экспериментов (как по методике, так и по материалу, представленному в анкетах для испытуемых).

На рассматриваемом нами материале типичными представителями конструкций («ядерными конструкциями») являются «конструкции ввода информации» в новостных текстах. В таблице 5 мы привели верхушку списка частотных «пятиграмм» (из рассматриваемого набора только два сочетания не относились к введению источника информации; кроме того, мы не стали исключать слова, написанные латиницей, для иллюстрации того, что в состав этих конструкций в принципе могут входить наименования информационных агентств любого типа). Напомним, что пятиграммы выделялись на основании частоты встречаемости коллокации: для больших n мера t-score как аппроксимация частоты оказывается избыточной.

[27] Надеемся, что в ближайших публикациях мы сможем показать специфику принятия решения испытуемыми при оценке степени устойчивости-связности и дать более тщательную лингвистическую интерпретацию параметров, влияющих на принятие решения.

Таблица 5. Наиболее частотные «пятиграммы», являющиеся «конструкциями ввода информации» в новостных текстах. Материал портала lenta.ru 2009 года (в порядке убывания частоты встречаемости)[28].

«пятиграмма»	Частота (ipm)
РИА Новости **со ссылкой** на	12678
сообщает РИА Новости **со ссылкой**	11048
сообщает Интерфакс **со ссылкой** на	10079
со ссылкой на источник в	9354
Об этом **сообщает** РИА Новости	9149
(об) этом **сообщает** РИА Новости **со**	6845
на источник в правоохранительных органах	6733
(со) ссылкой на источник в правоохранительных	6688
Об этом **сообщает** официальный сайт	6446
Об этом **сообщается** в пресс-релизе	6230
агентство Интерфакс **со ссылкой** на	6083
Об этом **сообщает** Интерфакс **со**	5982
(об) этом **сообщает** Интерфакс **со ссылкой**	5982
сообщает AFP **со ссылкой** на	5880
Об этом **пишет** газета Коммерсант	5841
Новости **со ссылкой** на источник	5683
Об этом **пишет** газета Ведомости	5670
Интерфакс **со ссылкой** на источник	5438
сообщает ИТАР-ТАСС **со ссылкой** на	5002
сообщает агентство Интерфакс **со ссылкой**	4987
Об этом **сообщает** Associated Press	4941
Об этом **сообщается** на сайте	4925
Интерфакс **со ссылкой** на пресс-службу	4895
Об этом **говорится** в официальном	4591
газета Ведомости **со ссылкой** на	4508
Новости **со ссылкой** на пресс-службу	4440
газета Коммерсант **со ссылкой** на	4388

Наиболее частотная схема такой конструкции сводится к:

1 (*об этом*) + 2 глагол (*сообщает*, *сообщается*, *пишет*, *говорится* и др.) + 3 название информационного агентства + 4 *со ссылкой (на)* + 5 источник информации.

В текстах портала «Лента.ру» наиболее часто в состав конструкции входит глагол *сообщает* или *сообщается*, однако это предпочтение носит стилевой характер.

Для того чтобы выяснить это, было проведено дополнительное исследование [162]. Предварительные результаты статистического обследования шести информационных источников свидетельствуют о том, что конструкции «введения источника информации» и особенно глагол, находящийся в вершине такой конструкции, характеризуют информационные источники, прежде всего с точки зрения их главной функции – информационную или публицистическую. Портал «Лента.ру» относится к ярко выраженным информационно насыщенным источникам (новостные ленты и близкие к ним формы подачи материала). Среди рассмотренных информационных источников к информационно насыщенным – ведущим себя в целом аналогично коллекции портала «Лента.ру» – относятся «РИА Новости»,

[28] Среди первых тридцати наиболее частотных «пятиграмм», встретилось двадцать семь конструкций ввода информации.

«РосБизнесКонсалтинг», «Компьюлента». Наиболее яркие свойства публицистической направленности (подчеркнутого внимания к адресату (-ам)) проявляются для «Независимой газеты» [162].

Например, для «Независимой газеты» биграмма *ссылкой на* стоит на 1551 месте, среди словоформных биграмм, упорядоченных по значению меры t-score, а *со ссылкой* – на 1591-м месте. Среди лексем первая биграмма со словом «сообщать» *КАК СООБЩАТЬ* стоит на 967 месте, следующая – *СООБЩАТЬ ИНТЕРФАКС* – на 5096 и т.д. Ср. также с данными «Статистического словаря русской газеты» А.Я. Шайкевича [150] *сообщается* 492, *сообщать* – 1614, *сообщаться* – 29, *сообщение* – 2488, *сообщить* – 8248 (корпус 1997-го года, 15 млн. словоупотреблений).

Для «Независимой газеты» наиболее частотными глаголами в коммуникативной функции оказываются *сказать*, *говорить*, *считать*, *заявить*. Вместо ранее обсуждаемых газетных клише в «Независимой газете» используются более привычные «негазетные» способы передачи информации, эти способы весьма разнообразны, и потому сложно выделить частотные n-граммы, которые можно было бы назвать конструкциями ввода источника информации. В текстах «Независимой газеты» наиболее частотным оказывается то, что характерно для текстов-интервью *отвечать на вопросы* (чуть реже *отвечать на вопрос*), *обратились к X*, где X – это *президенту*, *правительству*, *главе*, *руководству* и т.д. (в порядке убывания частоты встречаемости).

t-score-коллокации

Как уже было сказано, данная мера используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Обычно она считается малопригодной для поиска информационно важных номинаций и терминологических словосочетаний, не используясь для этой цели.

Однако все зависит от контекста, в данном случае от степени монотематичности и однородности коллекции. Так, в процессе данной работы над новостными коллекциями мы обнаружили, что эта мера оказывается полезна при решении задачи о выделении тех единиц, которые характеризуют **все** (или **подавляющее большинство**) текстов коллекции. Основная масса таких сочетаний характеризует скорее особенности стиля текстов коллекции, впрочем, используя минимальный морфологический фильтр из списков t-score-коллокаций, мы могли выделить те сочетания, которые могут рассматриваться как терминологические. Таким образом был получен список терминологических биграмм, общих для **всех** (или **подавляющего большинства**) текстов рассматриваемых коллекций (см. Таблицы 6 и 7).

Таблица 6. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Диалог»

лексемные биграммы	**словоформные биграммы**
РУССКИЙ ЯЗЫК	русского языка
	русском языке
ПРЕДМЕТНЫЙ ОБЛАСТЬ	предметной области

Таблица 7. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Корпусная лингвистика»

лексемные биграммы	**словоформные биграммы**
РУССКИЙ ЯЗЫК	русского языка
	русский язык
КОРПУС ТЕКСТ	корпус текстов
	корпуса текстов
НАЦИОНАЛЬНЫЙ КОРПУС	национального корпуса
	национальный корпус
ЧАСТЬ РЕЧЬ	части речи
	частей речи
АНГЛИЙСКИЙ ЯЗЫК	английского языка
КОРПУС РУССКИЙ	корпус русского
	корпуса русского
МАШИННЫЙ ПЕРЕВОД	машинного перевода
СЕМАНТИЧЕСКИЙ РАЗМЕТКА	семантической разметки
ПРЕДМЕТНЫЙ ОБЛАСТЬ	предметной области
ЛЕКСИЧЕСКИЙ ЕДИНИЦА	лексических единиц
ПАРАЛЛЕЛЬНЫЙ ТЕКСТ	параллельных текстов

Сопоставление списков терминологических биграмм, общих для **всех** (или **подавляющего большинства**) текстов (t-score-биграмм-коллокаций) рассматриваемых коллекций, приводит нас к следующим выводам:

1. Тематика конференции Диалог настолько широка, что на основании общих терминологических сочетаний мы могли бы сделать вывод лишь о том, что, как правило, в качестве основного материала исследований выступает *русский язык*, а также, что в текстах коллекции уделяется внимание *предметной области*.

2. Представляемые на «Корпусной конференции» исследования чаще всего ориентированы на *русский язык* или *английский язык*. В качестве материала (и/или объекта исследования) в большинстве работ выступает *корпус текстов*, что *лексическим единицам* (*частям речи, семантической разметке* лексических единиц) уделяется особое внимание. Что многие исследования ориентированы на решение вопросов *машинного перевода* и связаны с текстами заранее заданной *предметной области*. Таким образом, наши выводы согласуются с традиционной тематикой корпусных исследований, что отражено в наборе «общих» терминологических сочетаний.

Причем именно биграммы (а не триграммы и далее n-граммы) дают на нашем материале наиболее информационно насыщенную картину. Впрочем, возможно, что одна из причин этого лежит в сравнительно небольшом корпусе материалов конференции «Корпусная лингвистика (см. раздел 2.1).

По-видимому, чем выше однородность коллекции, тем более информативным окажется набор подобных t-score-биграмм-коллокаций для описания коллекции как целостного информационного потока (обзор математических моделей информационных потоков см., напр., в [124], о некоторых методах работы с информационными потоками в русле лингвистики текста см. в [87]).

Вместо заключения

Мы постарались обсудить типы коллокаций и конструкций, а главное – разные лингвистические типы шкал «от слова к коллокации и от коллокации к конструкции», которые формируются на основании (1) соотнесенности единицы с «инвентарностью (словарем) vs. конструктивностью (грамматикой)» и (2) с их функционированием в тексте/коллекции, т.е. с «номинативностью vs. предикативностью». Каждая из этих шкал характеризуется нечеткими границами явно выраженной динамической

природы. Положения данной классификации представляются набором гипотез, с одной стороны, уже верифицированных, а с другой – требующих дальнейшей верификации с учетом все большего числа параметров (прежде всего, контекстно-ориентированных параметров). В последнем параграфе четвертой главы про эксперимент на службе анализа текстов мы обсудим возможность введения дополнительных шкал, позволяющих «подключить» интуицию носителей языка (информантов и/или экспертов) и оценить степень целостности интересующих нас единиц.

Наборы рассматриваемых единиц (коллокаций и/или конструкций) характеризуют интересующие нас коллекции, эти наборы можно назвать свертками коллекций по заданным принципам. Именно поэтому мы в своих исследованиях (и даже в примерах) довольно широко варьируем коллекции: с точки зрения представленного функционального стиля, а чаще – гораздо более дробно: с точки зрения тематики, стилевых характеристик (обычно гораздо более точных, чем класс функционального стиля), степени однородности по каждому из этих признаков и т.д. Один из заданных принципов – это статистическая мера и методика обработки полученных списков. Главный заданный принцип заключается в подборе коллекции. Сначала подбирая, а потом описывая коллекцию и/или набор коллекций – через свертку – мы обеспечиваем адекватный контекст для решения задач вычислительного эксперимента: контекст коллекции (а в результате отчасти и текстовый контекст).

Глава 3. Семантическая и информационная структуры при анализе текстов и/или коллекций. Основные элементы этих структур

В третьей главе мы рассмотрим теоретические подходы и приведем примеры, которые были получены в ходе наших экспериментов по изучению текстов, прежде всего, экспериментов с информантами. Как уже было сказано, ключевым для главы является представление о вариативности и неединственности структур текста, извлекаемых при его восприятии (анализе). Часть экспериментов с информантами представляла собой восприятие звучащего текста, однако некоторые результаты этих экспериментов могут быть небезынтересны для наших лекций. Основнной акцент в этой главе делается на исследовании текста.

§ 3.1. Текст. Общие положения

Определим основные характеристики текста, существенные для исследования текста в контексте речевой коммуникации (порождения и восприятия речи):

- развернутость, или «последовательность знаковых единиц» (например, [131]);
- отдельнооформленность [130];
- связность и цельность (например, [130]).

Развернутость соотносится с вопросом о размерности и уровне иерархии такой единицы, как текст, структурными составляющими которого являются слова, синтагмы, фразы, сверхфразовые единства.

Для нас текст – основная конструктивная единица языка и, как уже было сказано, базовый лингвистический контекст, в котором реализуются единицы более низких уровней (слово, коллокация, синтагма, высказывание (фраза), сверхфразовое единство и композиционный фрагмент). Конструктивность и базовость текста кажется очевидной, однако в очередной раз сошлемся на краткую и авторитетную формулировку В.Б.Касевича: «будучи целостной единицей, текст обнаруживает по отношению к своим структурным компонентам (сверхфразовым единствам/абзацам, высказываниям, тем более – словам) свойство ***неаддитивности***: характеристики текста невыводимы полностью из признаков его составляющих; в первую очередь, передаваемое текстом значение несводимо к сумме значений компонентов» [114].

Отдельнооформленность предполагает, с одной стороны, наличие сигналов начала и конца, а с другой – представление о фреймах: знании носителей языка о структуре текстов разных функциональных стилей (текстовой и коммуникативной компетенции) [153]. Выделяют «внешнюю» и «внутреннюю» (смысловую) связность. И. Беллерт определяет связный текст как «такую последовательность высказываний $S_1,\ldots,S_n$, в которой семантическая интерпретация высказывания S_i (при $2<i<n$) зависит от интерпретации высказываний в последовательности $S_1,\ldots,S_{i-1}$» [90: 172]. Можно сказать, что в основе связности и цельности текста – взаимосвязанность и взаимообусловленность его структурных составляющих. Связность реализуется как пространственная (контактно расположенные структурные составляющие), «логическая» и ассоциативная (см., например, [119]).

Цельность и связность текста являются важными, но сложно формализуемыми характеристиками текста. Цельность обычно определяют как наличие единой темы (предметной области, набора ситуаций). Свойство связности (когерентности)

относится к структурной организации текста. При этом различают смысловую (тематическую) и синтаксическую связность (см., например, [130]). Среди формализуемых средств смысловой связности рассматривают, например, связующие слова (союзы, слова с темпоральными и причинно-следственными значениями) и механизмы референции и кореференции (повторяющиеся в тексте слова, другие виды повторной номинации). Синтаксическая связность текста – и высказываний как структурных составляющих текста – выражается, прежде всего, через семантико-синтаксическую структурированность этих единиц.

Исследователи связности текста пользуются разной терминологией. В последних исследованиях все чаще разделяют когезию и когерентность (например, см. [120]). Когезия – связь элементов текста, при которых интерепретация одних элементов зависит от других [120]. Когерентность соотносима с прагматической стороной, она выводит нас за пределы текста в коммуникативную ситуацию и спирается на базу знаний адресата. Когерентность в наибольшей степени связана с презумпцией осмысленности и реализаций (смысловых) ожиданий адресата. Однако в реальных моделях понимания текста носителем языка четко разграничить эти два разных вида связности бывает невозможно[29].

В процедурах речевой деятельности цельность и связность реализуются через механизмы контекстной предсказуемости. Естественно допустить, что если мы возьмем в пределах текста произвольную точку, отвечающую границе между некими языковыми единицами, то характеристики ее правого непосредственного «соседа» будут далеко не случайными. По-видимому, в дополнение к другим структурным характеристикам текст может быть описан как взаимодействие метафорически понимаемых «кривых сил связей между словами» – или между более сложными единицами текста, где некоторые позиции будут оказывать сильное воздействие на то, что может появиться справа, а другие будут предсказывать своих непосредственных «соседей» достаточно слабо. Множественность таких кривых определяется множеством признаков и параметров, по которым осуществляется связывание. Природа этих связей/предсказуемостей может быть различного происхождения: (1) связанной с лексической и семантической сочетаемостью/несочетаемостью, (2) определяющейся правилами синтаксиса, (3) соотносимой с информационной значимостью, (4) задаваемой коммуникативной ситуацией вообще и задачей коммуникации в частности. Предсказуемость может носить и более сложный характер, когда позиции предсказываются не характеристиками непосредственного «соседа» (предшествующего элемента), но на основании знания слушающего о смысловой связности и/или целостности (теме, смысле текста). Силы связей между словами (реже более сложными единицами анализа) хорошо описывается и предсказывается в математических сетевых моделях (напр., [124]). Однако у этих моделей пока существует естественное ограничение в виде уже упоминаемого множества разнотипных по своей лингвистической природе связей, большинство из которых до сих пор плохо изучено. Хочется надеяться, что в ближайшее время будет существенно расширена возможность такого моделирования – с варьированием типов единиц и контекстов – с учетом разнообразных признаков и параметров. Такая работа, по-видимому, может быть осуществлена при подключении

[29] Сейчас мы проводим серию психолингвистических экспериментов по оценке связности между разными единицами текста (словами, предложениями, абзацами).

специально подобранных и лингвистически сбалансированных коллекций, когда каждой задаче соответствует своя коллекция (или набор коллекций).

Естественно, что во время коммуникативного акта человек непрерывно планирует (программирует) свою речь или свое восприятие, осуществляя необходимые регулировки, переключения и т.д. С этой точки зрения, каждая следующая единица должна быть каким-то образом «сверена» и согласована с тем, что уже произнесено (или воспринято) к текущему моменту. Точность прогноза оценивается в прикладном направлении, имеющем до сих пор только английское название “readability” (что соответствует не столько «читабельности», сколько «понимабельности» текста, т.е. правильному извлечению смысла даже при беглом чтении или наличии искажений).

По-видимому, минимальное «окно сверки» («окно анализа») равно одной единице (например, одному высказыванию или одному слову); минимальное необходимое прогнозирование является в то же время как будто типичным, статистически преобладающим (ср. работы по 'cloze tests' или missing-words: [1; 2; 4; 14; 22; 77] и др.); максимальное же прогнозирование определяется текстом и коммуникативной ситуацией в целом. Мы к этому вернемся в последнем параграфе этой главы.

В традиции когнитивных теорий принято рассматривать текст как реализацию некоторого фрейма. Основоположник этого подхода Марвин Минский определяет фрейм как структуру данных, предназначенную для представления некоторой типовой ситуации [129]. Например, существуют фреймы бытовой, деловой и научной коммуникативных ситуаций, позволяющие прогнозировать развитие событий в этой ситуации (в частности, порождение и восприятие текстов разных функциональных стилей). Знание адресатом (слушающим) соответствующего фрейма, по-видимому, соотносится со знанием адресата смысла (цельности) и смысловой связности текста, где текст выступает как реализация этого фрейма.

Существенно противопоставление следующих типов целей и, соответственно, исследовательских процедур исследования текстов:

- понимания и интерпретации текста человеком, чем занимаются в русле традиционного и/или когнитивистского подходов (см., например, работы М.Б. Бергельсон [91-93], а также работы зарубежных авторов (частично рассматриваемые ниже);
- в духе прикладных задач – автоматического понимания текста (или, например, автоматического извлечения информации из текста, задач машинного перевода, автоматического реферирования и пр. (см., например, [127; 146; 125]).

Различие такого рода подходов предполагает помещение в центр исследования разных носителей языка. В случае прикладных исследований в качестве «искусственного носителя языка» выступает автомат. Естественным следствием такого различия является степень вовлеченности того, что можно назвать «базой знаний», позволяющей осуществлять прогнозирование развития событий на основании знания видов коммуникативных ситуаций (внелингвистических данных). Очевидно, что автомат «испытывает затруднения» в формировании некоторой макроструктуры текста, являющейся результатом функционирования в процедурах

восприятия (понимания, интерпретации) не только структурных составляющих текста, но и так называемых фоновых и выводных знаний. Степень вовлеченности фоновых и выводных знаний, по-видимому, зависит от типа фрейма и от знания коммуникантом этого фрейма[30].

§ 3.2. Анализ текста в парадигме когнитивных исследований

Кратко остановимся на наиболее плодотворных положениях современных исследований восприятия и понимания текста[31]. Отправной точкой является то, что «связный текст – больше чем язык сам по себе и гораздо больше, чем последовательность отдельных предложений»[32] (см., например, обзор по [39]). Процедуры восприятия и понимания текста традиционно трактуются как многоуровневые[33]. Однако требуют исследования такие вопросы, как количество и природа уровней, взаимодействие этих уровней и т.д. А.С. Штерн выделяла три уровня восприятия: сенсорный, перцептивный и смысловой [153]. Эти три уровня выделяются, главным образом, на основании психофизиологических критериев восприятия и переработки информации, но не языковых критериев; в частности, сенсорный («нижний») уровень не является языковым. В работе [39], посвященной, впрочем, пониманию письменного текста, выделяется пять следующих уровней: поверхностная структура, текстовая база как система пропозиций, модель ситуации как система референций, контекст коммуникации и функциональный стиль текста (или, может быть, речевой жанр) – the surface code, the propositional textbase, the referential situation model, the communication context, and the discourse genre. Первые три уровня традиционно принимаются большинством психолингвистов, начиная с работы [79]. На уровне ***поверхностной структуры*** адресат работает с такими единицами, как слова (вероятно, даже словоформы) и поверхностная структура клаузы (структурной составляющей текста, характеризующейся смысловой, синтаксической и просодической целостностью, но не превышающей некоторый критический объем[34]). «***База текста***, как правило, представляет собой структурированное множество (систему) пропозиций»[35] [39: 168]; вероятно, при этом речь идет только об ***эксплицитно*** выраженных пропозициях. ***Модель ситуации*** относит адресата к смыслу текста, в ее построении принимают участие как сам текст (explicit text, текст в узком смысле), так и фоновые знания адресата.

Коммуникативный уровень соотносится с прагматическими составляющими коммуникативной ситуации. Уровень ***функционального стиля*** и/или ***речевого жанра***

[30] Тип фрейма, в свою очередь, связан с функциональным типом текста и/или речевыми жанрами. Однако сосуществование разных научных парадигм вводит разную терминологию.

[31] В рамках данной работы используется термин «текст» как синоним терминам «дискурс» и «текст в широком смысле». В большинстве анализируемых теорий, напротив, использовался термин «дискурс». Обзор основных теорий восприятия звучащего текста (дискурса) см. в Gernsbacher 1994; Clark 1993; Levelt 1989; Rubin 1995.

[32] «Connected discourse is more than language per se, and much more than a sequence of individual sentences» [39: 164].

[33] Ср. положение многоуровневой организации деятельности по Бернштейну (Бернштейн 1966). Идеи многоуровневости, более или менее прямолинейно заимствованные из психологии и психофизиологии, оказали значительное влияние на психолингвистические теории.

[34] Ср. приводимое далее положение о том, что для обработки поверхностной структуры задействуется кратковременная память, т.о. объем этой структуры не может превышать психофизиологические возможности данного вида памяти.

[35] «The textbase is normally represented as a structured set of propositions» [39: 168].

(text genre) в зарубежных исследованиях соотносят с различными классами и подклассами, во многом соответствующими выделяемым рядом исследователей (см., например, [6]).

Некоторые положения и термины в предлагаемой Грэссером [39] схеме не бесспорны и требуют уточнений.

В частности, структура, называемая «***базой текста***» (text base) в разных работах понимается двояким образом:

1) как система пропозиций, как правило, соответствующих отдельным высказываниям текста;

2) как макроструктура, для возникновения которой важным (даже необходимым) является использование фоновых и выводных знаний; индивидуальные пропозиции входят в эту макроструктуру на правах членов, вступающих в определенные иерархические отношения [51]. Фоновые знания заполняют смысловые лакуны, неизбежные практически в любом тексте, а выводные знания выводят следствия из пропозиций и их сочетаний и привносят элемент упорядоченности, вносимой адресатом. Таким образом, второе понимание «базы текста» ближе к ***модели ситуации***, причем степень близости определяется степенью вовлеченности фоновых и выводных знаний адресата в восприятие и понимание текста.

Рассмотрим гипотезу Джонсон-Лэйрда, уточняющую место пропозициональной структуры в восприятии и понимании текста (см. обсуждение в [113])[36]. Согласно Ф. Джонсон-Лэйрду, структура пропозиций представляет собой лишь один из видов «семантической записи», которой пользуется человек при восприятии текста и для сохранения результатов этого процесса в памяти. Два других вида – это «ментальные модели» и образы [48]. «Статус последних наименее ясен, хотя утверждается, что имеет место отображение пропозициональных структур на ментальные модели, а последних – на образы. Что же касается ментальных моделей, то такая модель понимается как непосредственное отражение ситуации, описываемой воспринимаемым текстом» [113: 595]. Использующиеся в гипотезе Джонсон-Лэйрда понятия соотносятся с лингвистическими, психолингвистическими и психологическими представлениями. Макроструктура семантики текста как иерархия пропозиций плюс фоновые и выводные знания сопоставима с расширенным представлением о ***системе фреймов***, действительных для данного текста. «*Ментальная модель* – это система собственно-когнитивных фреймов, перцепт, отвечающий семантике текста. <…> «*образ*» в описываемой системе можно понимать как смысл – или некую систему смыслов – в духе Л.С. Выготского и А.Н. Леонтьева» [113: 595].

Семантические представления разных уровней (разной когнитивной глубины), как правило, сосуществуют, представляя собой разные стадии переработки одного и того же текста. Однако в зависимости от коммуникативных целей (стиля текста и личностных установок) человек может задавать уровень понимания (см. обсуждение в [113: 596]). Джонсон-Лэйрд пишет, что в случае высказываний, отражающих стереотипные (determinate) ситуации, глубина существенно выше: испытуемые запоминают лучше смысл высказываний, чем их языковую форму. В обратной ситуации – для высказываний, отражающих нестереотипные, в чем-то необычные ситуации, когнитивная глубина значительно меньше: лучше запоминается языковая

[36] Переиздание монографии в [113].

форма, поверхностная структура высказывания (вплоть до буквального состава) [48: 160–162]. Такого рода экспериментальные данные демонстрируют «отсутствие принудительного набора процедур и операций в процессах восприятия речи (и, соответственно, хранения информации в памяти и извлечения ее из памяти)» [113: 596]. Адресат определяет стратегию восприятия и глубину понимания текста[37].

Выделение «поверхностного восприятия (и понимания)» из «восприятия, понимания и интерпретации текста» соотносимо с попытками выделить различия между «***текстовой базой***» и «***моделью ситуации***» в процедурах восприятия текста (например, [52; 60; 61]. В другой работе – на материале письменных текстов – [83;84] показывается, что соотношение между уровнями – уровнем модели ситуации и более поверхностными уровнями (поверхностных структур и базы текста) – в процедурах понимания может зависеть от функционального стиля текста (публицистический vs. литературно-художественный).

В заключение краткого обзора когнитивистских и речедеятельностных представлений о процедурах понимания текста приведем психофизиологические обоснования возможности функционирования моделируемых процедур (см., например, [31; 54; 78]). Предполагается, что на уровне поверхностной структуры задействуется ***кратковременная*** память[38]. В *рабочей* памяти может удерживаться около двух предложений (с очень большим приближением); там же активируется и функционирует наиболее важная информация, которая может соотноситься с достаточно высокими уровнями представления текста. В экспериментальной работе с носителями языка проявление психофизиологических ограничений – в частности работа различных видов памяти человека – существенным образом определяется условиями коммуникации или (в выше описанной терминологии) коммуникативным контекстом и функциональным стилем текста.

§ 3.3. Анализ текста в парадигмах автоматического понимания текста

В кратком описании особенностей прикладного подхода к пониманию текста будем ориентироваться на книгу Н.Н. Леонтьевой, одного из признанных авторитетов в этой области [125]. Прежде всего, отметим, что автоматическое понимание текстов является необходимой частью разнообразных прикладных задач. Вполне очевидно, что, например, задачи машинного перевода и автоматического аннотирования (или реферирования) суть разные задачи, предполагающие разный результат автоматического понимания текста. Путь учета реальности таких разных подходов Н.Н. Леонтьева видит в последовательном применении идеи «***мягкого» понимания*** текста. «Мягкое» понимание можно трактовать как подстройку работы автомата под

[37] С другой стороны, подобные данные свидетельствуют об относительной автономности семантического компонента языка в процедурах восприятия и понимания. Особое внимание в экспериментальных работах уделяется последовательности/одновременности работы адресата (слушающего и/или читающего) на разных уровнях представления текста. В последние годы все больше исследователей приходит к выводу о том, что функционирование верхних уровней определяет работу синтаксических (например, [49; 59; 66; 80]) и лексических (например, [44; 68]) процедур. Наиболее яркий представитель противоположного подхода Фодор и его модулярная теория [32], в которой предполагается автономность функционирования не только фонологического, но и синтаксического модулей.

[38] Некоторые авторы выделяют ***эхоическую*** (по аналогии с иконической для зрительного восприятия) память, она обеспечивает более или менее точный образ-слепок звукового сигнала, существующий лишь очень недолгое время (около 3 с) после сенсорного восприятия сигнала (см., например, [17]).

разные коммуникативные цели. В отличие от ситуации естественной коммуникации в процедурах автоматического понимания заложено разделение ролей: человек определяет цель и оценивает окончательный результат, на долю автомата приходится понимание текста (в соответствии с поставленной человеком целью) и оценка результата на всех промежуточных уровнях. Результат понимания реализуется в виде построения некоторой семантической структуры. Н.Н. Леонтьева выделяет следующие типы структур[39]:

- «Лингвистические структуры предложений текста (локальное понимание).
- Семантические сети целого текста (глобальное размытое понимание).
- Информационные структуры целого теста (глобальное обобщенное понимание).
- Структуры баз данных и знаний (выборочное специальное понимание)» [125: 22].

1. ***Лингвистические структуры предложений текста*** фиксируют результат «локального» понимания, ограниченного рамками каждого из предложений текста. Наиболее известными лингвистическими структурами, по-видимому, являются структуры, основанные на модели «Смысл ⇔ Текст» [127]. Основой такого семантико-синтаксического представления является синтаксическое дерево предложения, имеющее «семантические» узлы [127] или «семантические» связи [88]. Работа автомата, построенного на базе модели «Смысл ⇔ Текст», опирается на сложные и богатые словари (то есть словари являются компонентами работы автомата), результатом такой работы является представление информации о единицах и связях в пределах предложения. Формализованность такого рода представления является и преимуществом, и недостатком. Главным достоинством такого рода структур является детальность анализа, отражаемого в форме дерева – синтаксического и семантического представления структуры предложения. При наличии словарных статей для всех слов предложения – естественно, при условии правильности структуры с точки зрения законов входного языка – автомат строит правильную синтаксическую структуру, сначала поверхностную, затем глубинную. Если все узлы глубинной синтаксической структуры заменить соответствующими толкованиями из словаря, оставив все связи из глубинной синтаксической структуры, но расширив нотацию[40], будет получена семантическая структура [127]. В системе машинного перевода ЭТАП-2 реализована единая синтаксическая структура, в узлах которой помещены слова исходной фразы. В этой единой синтаксической структуре сохранены подробные связи поверхностной структуры, связи достаточно дифференцированные, поэтому их можно перевести в семантический план; соединяемые ими слова снабжены семантическими характеристиками в комбинаторном словаре [88].

Недостатком жестких древовидных структур является невозможность выхода за пределы предложения, невозможность выборочного восприятия, «выхватывания»

[39] Описываемые далее структуры соответствуют разному положению дел в разработках систем автоматического понимания текста как с точки зрения современности, так и с точки зрения степени реализованности. Задача раздела сводится к сопоставлению работы автомата и человека в процедурах восприятия и понимания текста в разных условиях (цели, задачи, подходы).

[40] Под термином *нотация* принято понимать систему разметки (тэгов, помет). В частности, приводимый тип нотации включает актантные связи (от 1 до 4), атрибутивную и координирующую.

наиболее важной информации[41]. Другое ограничение такого подхода проявляется в слабой корреляции с системами представления знаний. «Пока реально достижимое СемП /семантическое представление/ целого – это последовательность СинСемП /синтактико-семантических представлений/ всех подряд предложений текста» [125: 24].

«*Лингвистические структуры предложений текста*» можно соотнести с уровнем, предшествующим построению *текстовой базы* как системы пропозиций текста. Эти «лингвистические структуры» в какой-то степени соотносятся с рассматриваемыми в предыдущем пункте поверхностными структурами, в которых человек работает с эксплицитными поверхностными структурами (не более одной клаузы или предложения). В то же время «лингвистические структуры» больше по объему: они включают и поверхностное, и глубинное представление о структуре такой единицы, как предложение.

Вероятно, это направление в работах по автоматическому пониманию текста ближе всего к моделированию того, как адресат воспринимает текст, лишенный связности и цельности. Подобного рода эксперименты проводились на текстах, в которых (а) предложения перемешаны случайным образом, (б) клаузы перемешаны случайным образом. С точки зрения задач настоящего исследования, по-видимому, это направление в некоторой степени может быть соотнесено с моделированием того, как адресат воспринимает текст, опираясь на структуру только текущего предложения, если он не может использовать механизмы контекстной предсказуемости, связывающие фрагменты текста за пределами одного предложения[42]. Однако, как уже говорилось выше, смысл текста не есть сумма смыслов составляющих его предложений, и понимание текста человеком принципиально не может быть ограничено лишь этим уровнем.

2. ***Семантическая сеть целого текста***, к построению которой прибегают авторы многих современных работ, представляет собой глобальную размытую структуру понимания. В этой глобальной сети реализуется следующий шаг модели «Смысл ⇔ Текст», включается глубинно-семантический компонент; «смыслом текста» объявляется результат перевода семантико-синтаксических структур (представлений) всех предложений текста на язык более «элементарных» единиц. Для того чтобы выйти в эту сеть, вводятся коммуникативные (информационные) отношения внутри предложений и между предложениями. В качестве таких отношений обычно устанавливают тема-рематические (внутри структур предложений) и референтные связи между структурами соседних предложений. В целом ряде случаев в дополнительном глубинно-семантическом компоненте объединяются собственно языковое и энциклопедическое представления (Перцова 1980).

3. ***Информационные структуры целого текста*** (потоков текстов) в качестве результата фиксируют обобщенное понимание текста (в единицах терминологии выбранной предметной области по классификаторам, тезаурусам, рубрикаторам и пр.). Они используются в информационно-поисковых системах (ИПС). Эти системы работают на материале произвольных текстов (практически без ограничений на

[41] В современных разработках системы ЭТАП-3 возможно интерактивное обращение к человеку с целью снятия семантико-синтаксической неоднозначности [7; 8].

[42] Ср. представления о функционировании контекстной предсказуемости на «окнах сверки» разного объема («от контактно расположенных словоформ до фразы» vs. «от контактно расположенных фраз до текста целиком»).

тематическую область и структуру текста), результатом является поисковый образ документа. Высокая востребованность и масштабность разработок разнообразных ИПС являются безусловным «плюсом» для оценки результативности подобного подхода автоматического понимания. Ограничения этого подхода, вероятно, связаны с небольшим смысловым потенциалом (обобщенное понимание текстов в очень больших масштабах работы систем). В лучшем случае результат такого рода соотносим с результатом классификации текстов человеком по предметной (тематической) области текста или в соответствии с очень грубой моделью ситуации. Конечно, такого рода классификация представляет собой «усеченный» вариант понимания текста. С другой стороны, возможно, такого рода подход соответствует проблематике исследования функционирования в ИПС ключевых слов: определение тематической области текста на основании выделенных из текста ключевых слов (особенно в том случае, если они представляют собой терминологические элементы).

4. ***Структуры баз данных и знаний*** представляют собой выборочное специальное понимание, в максимальной степени учитывающее экстралингвистическое представление, отображение части действительности. Структуры баз данных (БД) – это формальные, жестко фиксированные структуры (например, таблицы с описанием кадрового состава учреждения, все поля такой таблицы заранее заданы), над ними возможны общеизвестные математические операции. Если структуры БД – это формальные, жестко фиксированные структуры, то структуры баз знаний являются полужесткими структурами динамического типа (сценарии, фреймы). Такие структуры получили широкое распространение в системах искусственного интеллекта, они отображают представление целого текста и безразличны к членению на предложения. В системах Р. Шенка такие структуры нацелены на узнавание определенного сюжета в тексте [151]. Задаваемая тема текста может рассматриваться как «квазиденотат», а подход иногда называется денотативным [132; 149]. Главным ограничением такого подхода является «ограничение на мир»[43]. «Экстралингвистические модели … иллюстрируют зависимость понимания текста от предварительных знаний о предмете, но они плохо или никак не моделируют понимание несюжетных текстов, к которым относятся, в частности, научно-технические тексты» [125: 26].

Предлагаемая Н.Н. Леонтьевой идея «мягкого» понимания текста и ее информационно-лингвистическая модель понимания должна примирить методы лингвистического анализа и более грубый информационный анализ; иначе говоря, они должны обеспечить результативное взаимодействие разных уровней обработки текста. Информационный анализ неизбежно сопряжен с потерей части информации (информационный сброс). Определение более или менее информативных составляющих текста может и должно опираться на лингвистические исследования. Реализация процедур понимания «снизу – вверх» (от поверхностных структур к денотативным представлениям) описывается следующим образом: «основное назначение лингвистических структур … состоит в том, чтобы создавать контекст, необходимый и достаточный для вычленения на каждом уровне информативных единиц, которые переходят в структуры следующего уровня» [125: 31]. Лингвистически контролируемый информационный сброс позволяет автомату функционировать в отсутствие идеальных условий: снимать структурные

[43] Так, например, переход к новой предметной области требует построения новой системы автоматического понимания.

ограничения на обрабатываемые тексты (например, автомат может принимать на вход синтаксически неправильные или неполные предложения), допускает работу с неполными словарями и базами знаний. Возможно, исследование функционирования такого рода модели может рассматриваться как моделирование понимания текста «искусственным носителем языка» в разных коммуникативных условиях, для текстов разных функциональных стилей.

Одним из наиболее востребованных механизмов автоматической обработки текста является его компрессия. Задачей такого рода компрессии является получение реферата и/или аннотации: компактной формулировки содержания одного текста или монотематического массива текстов (группы текстов на одну тему). Принципы и степень сжатия определяются, как правило, задачами конкретной системы. Реферат и/или аннотация являются вторичными текстами.

Основным проблемным вопросом, решаемым при моделировании понимания текста автоматом (и построения вторичных текстов), являются цельность и связность текста. Решение такого рода вопроса невозможно без обращения к проблемам референции. Формализуемыми (в разной степени) средствами обеспечения связности являются следующие:

- повторяющиеся в тексте понятия (субъекты, объекты, явления и т.д.) в одном лексическом выражении;
- повторяющиеся в тексте понятия (субъекты, объекты, явления и т.д.) в разных лексических выражениях (например, в виде однокоренных дериватов или слов одного лексико-семантического поля)[44];
- местоимения и местоименные слова (см., например, [134]), чаще всего они также относятся к средствам выражения повторяющихся в тексте понятий;
- «слова-текстопостроители»[45], обозначающие обобщенные логико-композиционные связи между элементами – разного уровня составляющими – текста (например, *итак*, *резюмируя*, *следовательно*, *особо подчеркнем*, *все же*, *так же*, *однако* и т.д.);
- союзные слова характеризующие, главным образом, связи между клаузами, а не предложениями и занимающие промежуточное положение: с одной стороны, они передают синтаксические отношения (как союзы), с другой стороны, их значение соотносятся со значением некоторого полнозначного знаменательного слова (повтор понятия в предложении).

Для повторения одних и тех же понятий вне зависимости от их лексического выражения И.П. Севбо вводит понятие ***нанизывание*** [146]. Для получения компрессированного текста предварительно необходимо осуществить его «развертывание»: в итоге для текста строятся схемы нанизывания через канонические кусты, в которых восстанавливаются все связи[46]. В своей (уже ставшей классической)

[44] Вопрос о частоте встречаемости в тексте таких единиц, как словоформы и лексемы (то есть в одном лексическом выражении) и таких классов, как «класс условной эквивалентности» и «однокоренной класс условной эквивалентности» лексемы (то есть в разном лексическом выражении) рассматривается в главах 4–6. Признак «частота встречаемости в тексте» рассматривается в настоящей работе в контексте исследования коммуникативной структуры текста, формировании наборов ключевых слов текста и процедур «поверхностного понимания» в целом.

[45] И.П. Севбо называет эти слова опорными [146], в настоящей же работе понятие ***опорные слова*** вводится совсем в другом смысле: как наиболее распознающиеся (подробнее см. выше).

[46] Пример записи текста в канонических кустах (из [146]):

1) Боязливые жители вашего города травили меня (Ланцелота) собаками

а 2) собаки у вас (жителей) очень толковые

книге «Структура связного текста и автоматизация реферирования» И.П. Севбо описывает результат своего эксперимента по составлению аннотации текстов разных функциональных жанров на основании особенностей нанизывания: синтаксическая структура упрощенных нормализованных предложений и сведения о повторяемости в тексте понятий и слов.

Обычно для автоматического реферирования используют один из следующих способов (иногда комбинацию способов) (см., например, [146] и многие др.):

1. На основании статистического алгоритма из текста отбираются наиболее существенные предложения. На следующем этапе на основании синтаксического анализа из этих предложений выделяются наиболее значащие фрагменты.
2. На первом этапе применяется алгоритм синтаксического анализа предложений текста, в результате чего выделяются наиболее существенные части этих предложений. На следующем этапе статистическому анализу подвергаются лишь наиболее существенные части предложений текста.
3. «Вес слова» определяется на основании статистического и синтаксического анализа, так, в зависимости от синтаксической роли одно и то же существительное будет иметь разный вес (например, существительное в роли подлежащего более значимо, чем это же существительное в составе предложно-падежной конструкции).

По-видимому, в идеале перечисленные способы автоматического реферирования должны, во-первых, выделять наиболее значимые для понимания текста слова, конструкции и предложения, а во-вторых, характеризовать распределение этих наиболее значимых единиц (а) в структуре текста и (б) в структуре высказываний как составляющих текста. Следовательно, налицо необходимость соотнесения исследований в области автоматического реферирования и моделирования «поверхностного» восприятия и понимания текста человеком, то есть в условиях ограничений на «базу знаний» адресата. Разработано нескольких эффективных алгоритмов реферирования для информационно-аналитических и научно-технических текстов (например, [58; 74]).

В этом отношении большой лингвистический интерес вызывает только что вышедшая монография Н.В.Лукашевич «Тезаурусы в задачах информационного поиска» и те главы, которые посвящены как описанию связности текста, так и созданию по их результатам моделей автоматического реферерирования [126]. Нас больше интересуют экстрактивные аннотации, использующие фрагменты исходного текста (система анализа текста) для порождения текста аннотации (вторичного текста). В работе указываются те лингвистические признаки, которые лежат в основе определения веса (уровня значимости) фрагмента (от слова до предложения): позиция в тексте, частотность слов, именованные сущности и т.д. Одним из новых и наиболее актуальных (во всяком случае в лингвистическом смысле) вопросов является создание аннотации на основе многих текстов (вероятно, в качестве таких наборов

. 3) вот с ними-то (собаками) я (Ланцелот) подружился
. 4) они (собаки) меня (Ланцелота) поняли
, потому что 5) (собаки) любят своих хозяев (жителей)
и 6) (собаки) желают добра им (своим хозяевам, жителям)
. мы (Ланцелот и собаки) болтали почти до рассвета
Исходный текст: Ланцелот. *Боязливые жители вашего города травили меня собаками. А собаки у вас очень толковые. Вот с ними-то я и подружился. Они меня поняли, потому что любят своих хозяев и желают им добра. Мы болтали почти до рассвета* (Шварц 1962: 334)

документов могут выступать кластеры (сюжеты), тексты, организованные в циклы, а, возможно, и более сложные лингвистические информационные объекты). При составлении таких аннотаций (обзорных рефератов) «необходимо решать такие вопросы, как:

- борьба с избыточностью информации,
- идентификация важных различий между документами,
- обеспечение тематической связности текста, что усложняется тем, что предложения могут браться из разных источников» [126: 266].

Проблема модели аннотирования оказывается на стыке не только разных параграфов (этой главы), но и разных глав: текущей и следующей. Лингвистически значимым является анализ композиционной структуры текста (или анализ риторических отношений в терминах теории риторических структур [75]). Даже для научных текстов выделяются разные типы (в разном количестве и с разными весами) композиционных (или риторических) структур. И это возвращает нас к проблеме однородности коллекции или кластера не только в отношении тематической, но стилевой однородности [70]. Как уже было сказано, композиционную структуру мы рассматриваем как одну из стилевых характеристик текста или коллекции. Некоторые стилевые характеристики можно предсказать уже на уровне задачи описания исходных параметров выбора коллекции: событие, череда сходных событий, аналитика, интервью и т.д.

§ 3.4. Коммуникативная и информационная (смысловая) структуры текста

При восприятии речи основной задачей адресата является извлечение смысла (значения) или, вернее, смысловой структуры, которая отвечает тексту как некоторой целостности. Смысловая структура суть «структура содержания» в отличие от рассматриваемой в предыдущей главе просодической структуры. Смысловая структура заведомо многослойна и неоднородна. По-видимому, плодотворно выделять два типа смысловых структур: коммуникативная и собственно смысловая структуры. Далее, каждый из этих типов смысловых структур делится еще на два подтипа:

- коммуникативная структура:
  - тема-рематическая;
  - структура «данное vs. новое»;
- информационная (собственно смысловая) структура[47]:
  - структура пропозиций;
  - структура «ключевые слова vs. неключевые слова».

Остановимся также на таком виде представления коммуникативной и смысловой структур, как «база текста» (text base). «База текста» представляет собой вид иерархической пропозициональной структуры текста: ***макроструктура***, в которой индивидуальные пропозиции (соответствующие отдельным высказываниям текста) вступают в определенные иерархические отношения [53]. В формирование такой макроструктуры, по-видимому, существенный вклад вносят:

[47] Термин «собственно смысловая структура» условен, терминология в этой области еще не устоялась.

- фоновые знания, «настраивающие» адресата на определенную тематическую область и заполняющие смысловые лакуны, присутствующие в подавляющем числе текстов;
- выводные знания (выведение адресатом следствий из пропозиций и их сочетаний).

По-видимому, о «базе текста» имеет смысл говорить в контексте подстройки адресата под структурные особенности текста в процессе его восприятия. Кроме того, в целом ряде случаев представление такого типа о структуре текста может соотноситься с потенциальными путями осуществления контекстной предсказуемости в достаточно широких «окнах сверки», например, от фразы до текста целиком.

Самая глубинная (и в то же время самая грубая) структура, которую можно приписать любому предикативному смысловому образованию – это тема-рематическая структура как сопряжение основных коммуникативных компонентов высказывания (равно и текста)[48]. По-видимому, глубинность этой структуры заключается в глубокой когнитивной природе этого противопоставления. Ч. Хоккет, исследуя языковые универсалии, пишет: «В каждом человеческом языке можно встретить тип предложения двучленной структуры, конституенты которой разумно было бы именовать «тема» – «рема» («topic» vs. «comment») (Хоккет 1970: 70).

В качестве собственно смысловой структуры текста будем рассматривать лишь структуру, задаваемую распределением в тексте ключевых слов (КС) – как основных смысловых вех текста – на фоне неключевых слов (неКС). Структуру подобного рода, возможно, есть основания соотнести с хорошо известным в психологии восприятия противопоставлением фигуры и фона. Намеренно упрощая ситуацию можно сказать, что фигура – это наиболее значимая информация, «смысловые вехи» текста (или его фрагмента). Фон же обеспечивает успешное извлечение этих смысловых вех.

Мы рассматриваем коммуникативное структурирование высказывания (тема и рема, данное и новое); очевидно, однако, что коммуникативная структура высказывания и коммуникативная структура текста существенным образом взаимодействуют в процессах функционирования (порождения и восприятия текста). Подобное взаимодействие проявляется, в частности, в том, что функционирование компонентов структуры высказываний зависит от места расположения в тексте (продвижения от начала к концу текста), что соотносится со структурой «новое vs. данное»[49]. Очевидно, в рамках таких динамических процессов происходит взаимодействие коммуникативной и смысловой структур.

Мы будем говорить о потенциальной значимости коммуникативных и смысловых структур для восприятия не только письменного, но и звучащего текста. Проблемой коммуникативного членения занимались представители разных научных школ, наиболее существенными для данной работы являются следующие положения:

1. Кажется плодотворным выделение Т.Е. Янко двух основных типов коммуникативных значений: конституирующее речевой акт и не-

[48] Представления о том, что из себя представляет тема-рематическая структура и, соответственно, тема и рема (topic & comment, topic & focus, etc.) существенным образом зависят от парадигмы исследования (см., например, обзор «On the notion of topic» [42]. В данной работе тема и рема выделяется на основании результатов эксперимента, в котором эксперты-лингвисты определяют компоненты темы в предъявляемых текстах.

[49] Существуют попытки исследования собственно коммуникативной структуры текста (см. [145]). В настоящей работе предложенная Л.В. Сахарным методика исследования коммуникативной структуры текста в указанном смысле не используется.

конституирующее (подробнее см. [164]). Естественно, конституирующий компонент является обязательным, а не-конституирующий компонент – факультативным. Традиционно в теории актуального членения рассматривают речевой акт как реализацию сообщения (повествовательного предложения); в этом случае рема – это конституирующий компонент, а тема – не-конституирующий. У других типов речевых актов также есть конституирующий и могут быть не-конституирующий компоненты, хотя за ними не всегда закреплена традиционная терминология. В данной работе для любых типов речевых актов конституирующий компонент будем называть ремой, а не-конституирующий – темой.

2. «В языках существует не менее четырех способов маркирования темы, на которые может опираться воспринимающий речь человек. Это позиционный, грамматический, лексический и фонетический способы. Первый и последний из них, можно думать, являются универсальными, в то время как остальные два представлены в одних языках, но отсутствуют в других» [113: 600].
3. Наряду с бинарным противопоставлением тема vs. рема, (согласно теории коммуникативного динамизма) может рассматриваться и «скалярное» деление.

Поясним эти положения на примере противопоставления деловой vs. художественный тексты. В деловом тексте доминирует информативная функция языка, его высказывания рассматриваются как реализация сообщения, то есть материал в наибольшей степени соответствует стандартной «платформе» для исследования коммуникативного членения высказывания («актуального членения предложения» в русскоязычной традиции). Кроме того, от делового текста можно ожидать «нормативности» реализации (наличие обоих компонентов: «тема»+«рема»). Для художественного текста картина существенно другая. Сюжетный текст с элементами диалога не ограничивает множество своих высказываний: кроме сообщений, в нем представлены и вопросы, и императивы, и восклицания (не говоря уж о компонентах, вводящих прямую речь). В данной работе за соответствующими коммуникативными структурами ***всех*** этих высказываний закрепляем терминологию «тема-рематическая», «тема» и «рема». Многообразие этих структур заставляет предположить, что не все из них являются расчлененными, то есть факультативный компонент темы в них может быть эксплицитно не выражен.

Взаимодействие разных способов маркирования темы естественным образом становится одним из центральных для данной работы. Особое значение при извлечении смысла в процессе восприятии речи имеет позиционный критерий: при восприятии звучащего текста, динамически развертывающегося во времени, особенно важно, чтобы сначала у слушающего активировалась тема, то есть чтобы «то, о чем говорится», предшествовало «тому, что говорится». Начальное положение слова (именной группы) в высказывании (вместе с синтаксической немаркированностью) заставляет слушающего выдвигать гипотезу о принадлежности слов к теме. Можно предположить, что в случае каких-либо затруднений именно позиционный критерий становится ведущим. В то же время варианты даже *слабой активации темы* неизбежно должны сказаться на результатах (ср. работы А.А.Кибрика и многих других когнитивистов).

Как известно, теория коммуникативного динамизма в полной мере применима лишь к тем высказываниям, в которых тема предшествует реме, а глагол располагается в центре. Согласно рассматриваемой теории, степень

коммуникативного динамизма у начальной темы минимальна, далее степень коммуникативного динамизма увеличивается с продвижением к концу предложения. Глагол рассматривается как переход от темы к реме, ему приписывается средняя степень коммуникативного динамизма [27][50]. Возможны вариации идеи коммуникативного динамизма. Коммуникативный динамизм по Гаичовой относится к «глубинному» порядку слов, который может на поверхностном уровне реализовываться не только за счет позиционных и синтаксических средств, но и с помощью «размещения интонационного центра». Главное различие между двумя типами шкалирования проявляется в тех случаях, когда фраза несет логическое и/или эмфатическое ударение, то есть наряду с фразовым ударением присутствует второй интонационный центр на непоследнем слове [41]. Впрочем, пражские исследователи утверждают, что появление дополнительного интонационного центра редко встречается в текстах научно-технического функционального стиля[51], то есть далеко не все функциональные стили допускают различие между глубинной и поверхностной структурами (см., например, [43; 44].

Утверждение о преобладании прямого порядка следования *тема – рема* требует подтверждения на репрезентативном массиве текстов. Действительно, преобладание следования ремы за темой (и соответствующего порядка SVO (субъект-объект-предикат)) по корпусным данным реализуется для высказываний текстов разного функционального стиля[52]. Корпусные данные свидетельствуют об увеличении доли подобного порядка слов для научного (научно-технического) стиля (например, [102])[53]. По результатам авторской ручной разметки научно-технических текстов [102] схема «тема предшествует реме» реализуется в 70% случаев; в 16% случаев компонент темы отсутствует (нерасчлененная тема-рематическая структура). По-видимому, тексты литературно-художественного стиля (рассматриваемый художественный текст), напротив, характеризуются максимально свободным порядком слов.

Одним из «непрямых» следствий из положений теории коммуникативного динамизма является общая идея о том, что строгая бинарность оппозиции «тема vs. рема» (или, например, «topic vs. comment») может оказаться нефункциональной: наряду с рассмотрением функционирования подобного противопоставления необходимым становится анализ полевой структуры, а именно выделения ядра и периферии «темы» (а, возможно, и «ремы»).

Несмотря на пристальное внимание, оказываемое проблемам актуального членения предложения, существует «распределенность» между разнообразными подходами. В частности, проблемами актуального членения активно занимаются, с одной стороны, специалисты по автоматической обработке текста, а с другой стороны – когнитивные лингвисты. Очевидно, что это разделение связано с ограничением на

[50] Ср. последнюю формулировку теории коммуникативного динамизма Фирбаса [28].

[51] Именно на материале научно-технических текстов работает целый ряд систем автоматического определения тема-рематических структур.

[52] По данным случайной выборки 1000 предложений, принадлежащих текстам разных функциональных стилей (корпус «Бокренок», корпус-менеджер Бонито) порядок SVO доминирует (37%). Для текстов научно-технического стиля наблюдается увеличение доли SVO до 48%. Доля предложений «объективного» порядка (SVO) в целом составляет более 70% (соответственно, 73% и 77% для всех функциональных стилей и для научно-технического стиля (данные по [102]).

[53] При допущении сходства структурирования текстов научно-технического и официально-делового стилей эти данные, вероятно, могут соответствовать особенностям реализации и текстов официально-делового стиля (рассматриваемого делового текста).

функциональные стили рассматриваемых текстов. Для автоматической обработки текста первостепенной является задача обработки информационных текстов, то есть текстов официально-делового и научного стилей (например, задачи машинного перевода, извлечения знаний, автоматического реферирования). Когнитивные лингвисты в большей степени ориентированы на тексты литературно-художественного стиля (нарратива в широком смысле) и спонтанные тексты (например, тексты «разговорного» стиля). Различие подходов предполагает различие методологической базы. Как уже упоминалось, в случае информационных текстов можно предположить увеличение роли «формально-текстовых» параметров: позиционного и/или синтаксического. При когнитивистском подходе большее внимание, как правило, уделяется принципиально различным параметрам: референции, фокусу внимания (по Чейфу) и т.д. Возможная сложность проведения исследования сопряжена с определением специфики рассматриваемого функционального стиля и конкретного текста. Так, например, для рассматриваемого делового текста – административного уложения – характерно то, что номинация характеризует не сами объекты внешнего мира, а их классы. Таким образом, в случае делового текста можно говорить о дистрибутивном типе референции: референции в переменном релевантном типе пространства [152: 98–100]. Отсутствие традиционных референциальных связей отражается, в частности, в том, что личные местоимения составляют в деловом тексте 0,4% от всех словоупотреблений.

При когнинитивном подходе референцией, как правило, занимаются с позиции говорящего, а не адресата текста, тем не менее полагаем, что основные результаты могут быть применимы и к адресату. Референциальный выбор связывается обычно с активацией референта в сознании говорящего (ср. [9; 40]). В соответствии с универсалией, предложенной в [36], чем больше активация референта, тем больше вероятность того, что он будет выражен более редуцированно. Большинство факторов, предлагаемых в исследованиях референциального выбора, представляют собой лингвистические корреляты когнитивного понятия «активация». Одним из важнейших факторов является фактор линейного (референциального) расстояния, который измеряется в количествах границ предикаций между данным упоминанием референта и его ближайшим антецедентом (ср. [36]). В качестве других значимых факторов рассматриваются фактор риторического (иерархического) расстояния (ср. [33]) и свойства референта: одушевленность и его «дискурсивный вес» (например, является ли референт главным действующим лицом нарративного построения). Модель референциального выбора для русского языка была предложена А.А. Кибриком [50]. Исследование влияния референции на принятие решения о принадлежности словоупотреблений к теме проводится в настоящее время.

Полагаем необходимым подчеркнуть принципиальную ***неединственность*** коммуникативного структурирования текста в целом и высказывания как важнейшей структурной составляющей текста в разных процедурах: порождения / восприятия / анализа текста, разной глубины понимания и/или интерпретации текста и т.д. Таким образом, в общем случае неединственность коммуникативного структурирования определяется структурой коммуникативной ситуации. На такого рода неединственность обратила в свое время внимание Е.В. Падучева, отметив, что различие в описании коммуникативной структуры высказывания – это различие в его функционировании при решении тех или иных задач, например (по [135: 109–112]:

(1) Установить линейно-интонационную структуру при синтезе предложения из его синтаксического представления. «В рамках этой задачи от коммуникативной структуры требуется, чтобы в ней содержалась вся <…> дополнительная семантическая или прагматическая информация, которая должна быть добавлена к синтаксическому представлению предложения».
(2) Выяснить, какие значения передаются в данном языке варьированием линейно-интонационной структуры высказывания.
(3) Выразить оптимальным образом (например, при переходе от семантического представления предложения к синтаксическому) заданное содержание – смысл и разного рода акценты, контрасты, меняющиеся фокусировки и проч. и т.д.

Очевидно, что решение каждой из перечисленных задач задает свой тип коммуникативного структурирования.

Мы будем отталкиваться от общих «психолингвистических» представлений о ключевых словах (КС)[54] на основании работ, выполненных в рамках научной школы Л.В. Сахарного и А.С. Штерн, а также на основании работ А.И. Новикова (например, [130; 131]). КС определяются в ходе эксперимента с информантами, которые должны прослушать текст, подумать над его содержанием и выписать 10-15 слов, наиболее важных с точки зрения его содержания.

Наиболее функциональными, таким образом, могли бы стать основные положения, изложенные в работах Л.В. Сахарного и А.С. Штерн и их учеников. Кратко их можно сформулировать следующим образом:

- ✓ КС отражают тему текста;
- ✓ их упорядоченность – в наборе ключевых слов (НКС) – может трактоваться как эксплицитно невыраженная рема текста;
- ✓ при допущении того, что рема в тексте может быть не выражена эксплицитно (но лишь за счет ассоциативных связей), НКС рассматривается как один из минимальных вариантов «текста»;
- ✓ такого типа «текст» характеризуется «ядерной» цельностью и минимальной связностью (см., например, [18; 146]).

Первое положение кажется максимально обоснованным. Оставшиеся требуют дополнительного обсуждения. По-видимому, расширение понятия «текст», позволяющее включить НКС в множество возможных текстов, не является для нас необходимым. В то же время возможность развертывания НКС в процедурах порождения текста является экспериментально доказанной. Для порождения такого рода текстов необходима инструкция, запускающая механизм: например, «напишите связный осмысленный текст, употребив слова…» [133]. Вероятно, упорядоченность слов в НКС активирует ассоциативные связи, необходимые для существования любого текста [133]. Понимание ремы как «связующего элемента», соединяющего темы и подтемы, оказывается противоречащим выше описываемому подходу к исследованию тема-рематической структуры и представлению о реме как о конституирующем компоненте (как высказывания, так и текста).

Немного о результатах, полученных в экспериментах по восприятию звучащего текста в шуме (сигнал/шум 0 дБ), с одной стороны, иллюстрирующих некоторые из

[54] Впрочем, вычислительный подход к оценке КС как наиболее значимых для текста (коллекции) иногда выступал в качестве самостоятельной задачи, а не только для сопоставления результатов двух разных экспериментов.

положений об информационной и коммуникативной структуре текста, с другой – позволяющие оценить осуществляемую текущим образом подстройку под текст (процедуры анализа и понимания).

Коммуникативная структура

- ✓ Для ***делового*** текста (1) элементы темы распознаются лучше, чем элементы ремы (особенно на конечном фрагменте); (2) от начального к конечному фрагменту текста происходит ***улучшение*** распознаваемости каждого из элементов; (3) перцептивно значимой является позиция перед паузой.
- ✓ Для ***художественного*** текста (1) элементы ремы распознаются лучше, чем элементы темы (особенно на конечном фрагменте); (2) от начального к конечному фрагменту текста происходит ***ухудшение*** распознаваемости элементов ***темы*** и ***улучшение*** элементов ***ремы***; (3) мелодика (прежде всего, понижение частоты основного тона) представляет собой перцептивно наиболее значимый фонетический признак для коммуникативного членения художественного текста; этот признак маркирует ***новое*** (не только для структуры «тема vs. рема», но и «данное vs. новое»).

Собственно смысловая структура

- Для ***делового*** текста (1) КС распознаются лучше, чем неКС (особенно на конечном фрагменте); (2) от начального к конечному фрагменту текста происходит ***улучшение*** распознаваемости каждого из элементов; (3) перцептивно значимой является позиция перед паузой.
- Для ***художественного*** текста (1) неКС распознаются лучше, чем КС (на начальном и конечном фрагментах); (2) наилучшей распознаваемостью обладает ***середина*** текста (здесь происходит нейтрализация противопоставления КС vs. неКС); (3) мелодика (прежде всего, понижение частоты основного тона) – перцептивно наиболее значимый фонетический признак маркирования КС.

При распознавании слов делового текста наиболее существенным является фактор знакомства с текстом (его темой, структурой и наиболее частотными словами), КС и элементы темы (как им и положено) распознаются сравнительно неплохо, конец текста предсказуем и хорошо распознается. Для художественного текста бóльшая «опорность» приходится на начальный (преамбула) и срединный (развитие сюжета) композиционные фрагменты и по-разному соотносится с компонентами коммуникативного и смыслового членения: с темой для преамбулы, с диалогом (особенно неКС или ремой) для срединного фрагмента. Таким образом, говоря о структурах текста и процедурах анализа, мы должны учитывать разнообразные виды контекста, в частности, функциональный стиль, композиционную структуру и риторическую связность текста.

Рассуждая о текущем словаре и ключевых словах в главе 2, мы уже упомянули, что выходя на более высокий уровень анализа и сопоставляя – (1) «текущие словари», принадлежащие тексту и коллекциям разной степени однородности («уровня вложенности»), (2) ключевые слова, характеризующие текст и коллекции разной степени вложенности – можем получить полноценную информацию не только для информационно насыщенных текстов, но и для художественных текстов. Естественно, такого рода анализ предполагает сопоставление эксперимента с информантами и вычислительного эксперимента (для последнего особую значимость приобретает формирование контрастивной коллекции). Для научных текстов наличие

пересечений как характеристика степени тематической однородности коллекций и центральном/периферийном положении текста в информационном пространстве коллекций не противоречит общей методике анализа (Пивоварова, Ягунова 2011). Поэтому остановимся на более сложном примере анализа информационных структур циклов Н.В.Гоголя:

- «Петербургские повести» – максимальная компактность и прозрачность информационной структуры; достаточно большие наборы ключевых слов, выделяемых на основании и эксперимента с информантами, и вычислительного эксперимента; два набора ключевых слов – выделяемых на основании вычислительного эксперимента[55] и эксперимента с информантами – хорошо демонстрируют различия между двумя типами информационных структур: извлекаемой человеком в процессе понимании текстов vs. выделяемой автоматом при реализации процедур информационного поиска.
- «Украинская тематика» (Цикл «Миргород» и «Вечера на хуторе близ Диканьки») – максимальная неоднородность структуры; списки ключевых слов, выделяемые в ходе вычислительного эксперимента, интуитивно кажутся наиболее адекватными для понимания текстов носителем языка.
- поэма «Мертвые души» – демонстрирует промежуточную картину.

Сопоставление с данными о распределении потенциально ключевого слова в пространстве текста позволяет произвести формализованную классификацию типов КС (типы действующих лиц, ключевые слова, аккумулирующие содержательные вехи описания и/или рассуждения и т.д.) [154-156].

Исследуя наборы ключевых слов мы – в зависимости от методики интерпретации полученных данных – можем создать свертку разной степени компрессии. Это определяется выбором единицы анализа. В выше приведенных рассуждениях (результатах) была использована традиционная для КС единица анализа: лексема. Словоформы из анкет информантов нормализовались и представлялись в виде списка лексем. Однако в то же время для решения ряда других задач нужно представлять данные в виде более компактных лексико-семантических единиц (с точностью до синонима, с точностью до ближайших дериватов (объединяя, напр., *статистика* и *статистический*), с точностью до тематического класса[56]. Однако без использования тезауруса эта задача решается неоднозначно, мнения экспертов часто не совпадают. В [126] приводятся крайне интересные решения выделения-построения-использования «лексических цепочек в построении тематического представления текста (там же, глава 19: 375-393). Конечно же, предлагаемые решения базируются на использовании тезаурусного представления РуТез [126], и заставляют серьезно задуматься лингвистов.

[55] Ср. КС, выделенные на основании tf-idf: *Акакиевич, рука, шинель, Ковалев, лицо, ростовщик, Акакий, медж, асессор, Яковлевич, пуф, коллежский, маиор, нос, титулярный, Шиллер, квартальный, коломна, Чартков, бакенбарды, лорнет, Пискарев, время, прыщик, проспект, департамент, Рафаэль, Чертокуцкий, голова, Фидель, чорт, комната, Психея, портрет, художник, происшествие, человек, слово, чиновник, Невский, Испания, дама, глаза, штаб-офицерша, казаться, Гофман, беспрестанный.*

[56] Устоявшейся терминологии нет, в [158] я использовала термины «классов эквивалентности» и «классов условной эквивалентности».

§ 3.5. Избыточность. Компрессия текста. Свертки текста

Мы обозначим контуры парадигмы исследования восприятия и понимания текста – делового и художественного – на материале экспериментально полученных компрессированных текстов (в результате лингвистической компрессии), то есть эти тексты можно охарактеризовать как имеющие и адаптационную, и лингвистическую компрессию. В качестве компрессированных вариантов представления текста рассматривается два типа: лакунарный текст и наборы опорных слов (НОС). Под опорными словами мы здесь понимаем слова, характеризующиеся максимальной разборчивостью в одном из экспериментальных режимов. Для примера ограничимся только одним: восприятием текста в шуме. Предполагается, что высокая разборчивость (не менее 30% информантов) отражает высокую информационную значимость этих слов, что проявляется и в их реализации, и в результатах восприятия (понимания). НОС обладают существенно большей степенью компрессии по разным условиям; степень компрессии существенно зависит от функционального стиля текста. НОС – последовательности лучше всего распознающихся словоупотреблений текста – рассматриваются как свертки текстов, то есть как варианты вторичных текстов (полученных в результате понимания исходного)[57].

Лакунарные тексты представляют собой формализованный вариант исключения каждого четвертого слова (замены каждого четвертого слова амплитудно-модулированным белым шумом). Естественно, лакунарные эксперименты можно и нужно проводить как для письменных, так и для звучащих текстов (исследуя “readability” и контекстную предсказуемость) [111-113]. Лакунарные тексты оказываются подвергнутыми незначительной компрессии, принцип удаления слов из текста абсолютно формален – каждое четвертое; таким образом, среди удаленных

[57] За пределами формата глав этого учебника, ориентированного, главным образом, на письменный текст, остались очень интересные результаты:

- функциональный стиль текста определяет распределение опорных слов[57] в пределах текста;
- функциональный стиль текста определяет выбор процедур идентификации слов текста; разные модели идентификации через обращение к словарю (lexical access) – когортные vs. сетевые модели восприятия – отражают особенности разных процедур идентификации слов текста[57].

При распознавании слов художественного текста класс поиска, формируемый с опорой на начальный слог (с учетом дополнительных факторов), основывается на сравнительно небольшом числе гипотез и сравнительно «коротком пути» их верификации. Наилучшая распознаваемость – у двусложных слов, они же являются наиболее частотными ритмическими структурами в рассматриваемом тексте. Для распознавании слов делового текста приходится допустить более широкое использование других – «некогортных» – стратегий восприятия, дополняющих / замещающих собственно когортные. Эта часть наших данных скорее согласуется с представлениями модели SHORTLIST, в рамках которой не акцентируется значимость начальной позиции слова (см., например, [71]).

Различие в сегменте, преимущественно запускающем процедуру поиска слов в словаре («lexical access»), не является единственным различием. Главное различие в последовательности / параллельности процедур обработки. Главное преимущество сетевых моделей в том, что они предполагают **параллельную** обработку информации по разным путям проверки (например, сегмент, число слогов, место ударения и т.д.). Для более простых ситуаций коммуникации (известная тема, короткие слова, короткие синтагмы и фразы) и функциональных стилей художественного текста или тем более бытового разговора может оказаться достаточной когортной модели распознавания слов в тексте. При усложнении ситуации коммуникации с необходимостью включаются более сложные модели. Сетевые модели позволяют параллельную обработку, достигают необходимого результата (слова-кандидата) с большей вероятностью при больших искажениях – как субъективных (например, сложная предметная область или незнание ее адресатом), так и объективных (внесения помех в объект (текст) или передачу объекта. Более того – сетевые модели и параллельный анализ входной информации приближают нас к учету психофизиологической природе распознавания и понимания текста. Именно эти модели подходят для наиболее востребованных в компьютерной лингвистике информационно насыщенных текстов (научном, официально деловом, тексте новостных сообщений).

слов оказываются слова с разными фонетическими и/или внефонетическими характеристиками (служебные и знаменательные, полноударные и клитики, элементы темы и ремы и т.д.).

Эксперименты по восстановлению текста на основании обоих вариантов компрессированных текстов предоставляют возможность исследования контекстной предсказуемости и шире – понимания текста. Лакунарный текст и НОС существенно различаются по принципу сжатия и степени компрессии. Использование этих методик позволило по-новому взглянуть на роль общеязыковой частотности / частоты встречаемости для указанных функциональных стилей (в словарях С.Шарова и О.Ляшевской), собственно смысловых и коммуникативных структур с точки зрения контекстной предсказуемости. Выводы, полученные на материале двух экспериментальных режимов – зашумленный и лакунарный текст – подтверждают и/или взаимодополняют друг друга.

Общеязыковая частота встречаемости

Деловой лакунарный текст

- При восприятии лакунарного делового текста (в отличие от восприятия текста в шуме) общеязыковая частотность словоформ играет существенную роль.
- Для этого текста возможно переструктурирование словаря в результате подстройки слушающего под особенности этого текста.
- На конечном фрагменте текста редкие лексемы восстанавливаются значимо лучше, чем на начальном или конечном.

Художественный лакунарный текст

- Для художественного текста общеязыковая частотность словоформ играет существенную роль.
- Для художественного текста общеязыковая частотность лексем играет существенную роль.
- Не происходит переструктурирования словаря по мере продвижения слушающего по тексту (по мере понимания смысла текста [тема и набор подтем текста]).
- Нет значимых различий в восстанавливаемости редких словоформ (и редких лексем) на конечном фрагменте и на начальном (или конечном) фрагменте текста.

Общим свойством, характеризующим восстановление и делового, и художественного лакунарных текстов, является значимость роли общеязыковой частотности словоформ (как основных единиц перцептивного словаря): чем выше частота встречаемости, чем лучше точная восстанавливаемость. При восприятии текста в шуме этот признак имел значение лишь для художественного текста. Для художественного лакунарного текста значимым признаком оказалась также частота встречаемости лексемы, что может отражать то, что задача восстановления отдельных (лакунаризируемых) единиц может задействовать в стратегии восприятия и более высокий уровень – уровень понимания сюжета. Восприятие (и восстановление) лакунарного делового текста в существенно меньшей степени включает процедуру понимания: испытуемые незнакомы с предметной областью текста.

Собственно смысловая структура

Деловой лакунарный текст

- Имеет место подстройка слушающего под особенности текста; в результате этой подстройки частотность слова (или более крупной структурной

составляющей текста) по тексту имеет существенное значение при оценке степени предсказуемости.

- На конечном фрагменте КС восстанавливаются значительно лучше, чем неКС.
- По мере продвижения слушающего от начального фрагмента текста к конечному происходит улучшение предсказуемости КС, в результате чего на конечном фрагменте КС восстанавливаются несколько лучше, чем неКС.
- Данные о роли собственно смысловых структур на материале лакунарного эксперимента аналогичны данным эксперимента по восприятию текста в шуме.

Художественный лакунарный текст

- Подстройка слушающего под особенности текста не сопровождается изменением частотности элементов текста по сравнению с общеязыковой частотой встречаемости.
- На начальном фрагменте КС восстанавливаются значительно хуже, чем неКС.
- Данные о роли собственно смысловых структур на материале лакунарного эксперимента и на материале эксперимента по восприятию текста в шуме дополняют друг друга.

Общим для восприятия делового и художественного лакунарных текстов является то, что от начального к конечному фрагменту текста происходит увеличение предсказуемости позиций КС, то есть понимание текста сопровождается формированием смысловых вех.

В процедурах восприятия текста конкурируют разные «окна сверки» контекстной предсказуемости, каждому из которых присваиваются разные веса. Наибольший вес минимальное «окно сверки» имеет для восстановления компонентов неоднословных целостностей, например, для фразеологизмов *и так далее*, *во все горло, от всей души, чтобы духу твоего не было*; сложных номинаций *внутренний рынок, транспортные средства* и т.д. (т.е. для коллокаций и конструкций). Текстовое «окно сверки» является максимальным; формирование смысловых вех характеризует именно такое «окно сверки». Для художественного текста, по-видимому, характерно взаимодействие «окна сверки», равного всему художественному тексту и равного смысловому блоку текста.

Коммуникативная структура

Деловой лакунарный текст

- От начального к конечному фрагменту текста происходит увеличение предсказуемости позиций темы.
- На конечном фрагменте элементы темы восстанавливаются значительно лучше, чем элементы ремы.
- Данные о роли коммуникативных структур на материале лакунарного эксперимента аналогичны данным эксперимента по восприятию текста в шуме.

Художественный лакунарный текст

- В условиях этого экспериментального режима нейтрализуется различие в предсказуемости между элементами темы и ремы.
- Нейтрализуется зависимость предсказуемости элементов темы и ремы от продвижения по тексту (от смыслового блока).

- Данные о роли собственно смысловых структур на материале лакунарного эксперимента и на материале эксперимента по восприятию текста в шуме взаимодополняют друг друга.

Можно ли НОС представить в виде свертки текста, а затем в «обратном» эксперименте восстановить текст на основании этих наборов, полученных на разных текстах (функциональных стилях) и разных фрагментах? НОС – упорядоченные последовательности (фонетических) словоформ, где опорные слова являются наиболее распознаваемыми при восприятии текста в шуме[58]. Более традиционным видом сверток является набор ключевых слов (НКС), в котором задан порядок – порядок введения КС в текст. КС отражают тему текста, а упорядоченность слов в НКС активирует ассоциативные связи, необходимые для существования любого текста (что проявляется в возможности развертывания НКС в цельный и связный текст). НКС и НОС можно рассматривать как разные виды сверток текста, отражающих разные виды смыслового структурирования. НОС – представляет свертку более динамичного характера (отражающего процедуры он-лайн-понимания текста), НКС – свертку статичного характера, понимание текста в целом предшествует определению КС, в свертке представлены результаты уже свершившегося действия.

Особенности НОС как сверток текста исследовались в двух сериях эксперимента по восстановление текста на основании (1) полного НОС и (2) фрагмента НОС, соответствующего начальному фрагменту текста. Эксперимент проводился в письменно-письменной форме[59].

В основу эксперимента легли следующие ***гипотезы***:

- НОС позволяют осуществить построение целостного связного текста;
- НОС задают функциональный стиль развертываемого текста;
- НОС определяют предметную (тематическую) область развертываемого текста;
- развертывание НОС позволяет определить позиции слов и/или конструкций, обладающих максимальной контекстной предсказуемостью.

Определение того, принадлежат ли восстановленные тексты тому же функциональному стилю, что и исходный, производилось на основании двух критериев:

- экспертная оценка:
  - заключение эксперта о принадлежности восстановленного текста к данному функциональному стилю (деловому или художественному),
  - степень статичности vs. динамичности смены описываемых ситуаций (как дополнительный признак);
- количественные (формальные) критерии:
  - коэффициент лексического разнообразия текстов (КЛР), что отражает степень разнообразия лексических средств при построении текста и соотносится с функциональным стилем текста (разнообразие лексики (высокий КЛР) характеризует художественный текст, а клишированность (низкий КЛР) – деловой текст),

[58] Результаты аналогичных экспериментов, в которых НОС были опорными словами при распознавании текстов в других режимах искажения описаны в [158].

[59] В письменной инструкции было указано: «*Перед Вами* ***последовательность*** *слов, извлеченных из текста. Попробуйте на их основе восстановить текст*».

- длина текстов в словах (как дополнительный признак).

Существенное значение для рассматриваемого собственно смыслового структурирования текста имеет степень динамичности текста, которая определяется количеством описываемых ситуаций:

- Пересечение НОС и НКС существенно выше для статичного делового текста по сравнению с динамическим художественным (52% vs. 13%).
- Смысловая структура, заложенная в НОС, является более динамичной, чем смысловая структура, представленная в виде НКС. Это проявляется в увеличении доли глагольной лексики.

Деловой текст, НОС для всего текста

1. Все тексты, восстановленные на основе НОС, извлеченных из исходного текста делового функционального стиля, воспроизводят этот стиль. Большинство развернутых текстов можно отнести к жанру нормативных актов. Все развернутые тексты – как и исходный текст – относятся к статическому варианту: как правило, предписание, регулирующее положение дел, или – в ряде случаев – описание некоторого положения дел.
2. Высокая степень повторяемости словоформ (КЛР 0,46) в предъявляемом НОС провоцирует высокую повторяемость слов в восстанавливаемых испытуемыми текстах.
3. При восстановлении текста испытуемыми использовались КС (как присутствующие в НОС, так и отсутствующие в наборе) как представляющие смысловую (тематическую) область данного текста. Подавляющее число используемых слов относятся к деловому функциональному стилю (при отнесении лексики к деловому функциональному стилю использовался разработанный С.А. Шаровым на основе НКРЯ набор частотных списков по жанрам (http://corpus.leeds.ac.uk/serge/frqlist/), в котором представлен и «Частотный словарь административных текстов, значимая лексика»[60]).
4. Результаты восстановления текста – развертывания НОС в текст – позволили определить позиции, обладающие максимальной контекстной предсказуемостью.

Деловой текст, НОС для начального фрагмента

1. Развертывание НОС в текст можно считать успешным как с точки зрения построения цельных связных текстов, так и с точки зрения восстановления предметной области исходного текста. Все развернутые тексты – как и исходный текст – относятся к деловому функциональному стилю. Большинство развернутых текстов можно отнести к жанру нормативного акта.
2. Значения КЛР для начального фрагмента исходного делового текста показывает сравнительно высокую степень повторяемости лексических единиц, однако степень клишированности исходного делового текста на всем тексте существенно выше. Клишированность текста лишь отчасти задается через НОС в инструкции для развертывания текстов испытуемыми. КЛР для лексем восстановленных текстов в среднем близок к соответствующему показателю исходного текста. КЛР для словоформ восстановленных текстов в среднем ниже, чем для исходного текста.

[60] На момент проведения эксперимента этот частотный список предлагал более ожидаемые значения частот, чем частотные словари О.Ляшевской и С.Шарова на сайте ИРЯ РАН.

3. Результаты восстановления текста не позволили определить позиции, обладающие высокой контекстной предсказуемостью.

Различия между КЛР существенны по всем параметрам: между восстановленными деловыми текстами – на основании «НОС для всего текста vs. НОС для начального фрагмента» – как для словоформ, так и для лексем. По-видимому, это различие соотносится с ранее сделанными выводами о важности текстового «окна сверки» для восприятия делового текста (в экспериментальных режимах восприятия текста в шуме и восприятия лакунарного текста). Для сравнения приведем данные по ***художественному тексту***. ***НОС для всего текста.***

1. Все развернутые тексты – как и исходный текст – относятся к художественному функциональному стилю, характеризуются элементами динамизма (в разной степени). При разворачивании НОС обычно восстанавливается сюжет с двумя действующими лицами, в тексте присутствуют диалоговые фрагменты; характеристики действующих лиц (как и место действия) могут быть самыми разными. Общее содержание развернутых текстов может быть охарактеризовано как «конфликт» (сопровождающий выпрашивание денег или какое-либо другое требование).
2. Художественный текст с динамически развивающимся сюжетом характеризуется разнообразием знаменательной лексики. Повторяемыми являются, главным образом, местоименная лексика и, в какой-то степени – наименования действующих лиц.
3. КЛР для словоформ и лексем восстановленных текстов несколько выше, чем для исходного текста. Вероятно, это незначительное расхождение связано с индивидуальными стратегиями порождения текста (богатство сюжетной линии, использование синонимических ресурсов и местоименной лексики).

Художественный текст, НОС для начального фрагмента (преамбула и завязка).

1. Все развернутые тексты – как и исходный текст – относятся к художественному функциональному стилю. Большинство развернутых художественных текстов характеризуется статичностью (отсутствием смены ситуаций). Как правило, при разворачивании НОС для начального фрагмента текста восстанавливается статичный текст, а не сюжет с двумя действующими лицами и фрагментами диалога.
2. Степень разнообразия лексем может быть связана со степенью статичности текста: восстановленные тексты являются более статичными, чем исходные, и КЛР для лексем восстановленных текстов ниже, чем для исходного текста. Возможным подтверждением статичности восстанавливаемого фрагмента художественного текста служит то, что КЛР для лексем восстановленных художественных и деловых текстов (начальные фрагменты) различаются лишь на уровне тенденции.
3. Результаты восстановления текста – развертывания НОС в текст – позволили определить позиции слов, обладающих максимальной контекстной предсказуемостью.

Главным результатом является подтверждение гипотезы о том, что НОС (для всего текста) задают функциональный стиль тех текстов, что восстанавливаются испытуемыми в эксперименте. То, что восстановленные тексты принадлежат тому же функциональному стилю, что и исходные, подтверждается не только качественными, но и количественными критериями. Значимость различий между КЛР (для словоформ

и лексем) восстановленных деловых и художественных текстов является количественным показателем принадлежности текстов к разным функциональным стилям.

Полученные данные не противоречат гипотезе о том, что НОС (для начального фрагмента текст) задают функциональный стиль тех текстов, что восстанавливаются испытуемыми в эксперименте. Важным результатом является взаимосвязь композиционного начального фрагмента со степенью «статичности vs. динамичности» всего текста: для статичного делового текста статичность сохраняется в восстановленных текстах. Для динамичного художественного текста степень статичности повышается: фрагменты «преамбула и завязывание сюжета» имеют бóльшую динамичность, чем восстанавливаемые тексты.

Глава 4. Объект исследования современной лингвистики текста. Текст vs. информационный поток

В четвертой главе мы рассмотрим, главным образом, общие подходы и планы на будущее; также приведем конкретные примеры и те данные, которые были получены в ходе наших экспериментов с информантами и/или вычислительных экспериментов. Ключевым для этой главы является представление о вариативности информационной структуры в соотношении таких единиц анализа как текст vs. информационный поток. Одним из пунктов было исследование типологии текстов с точки зрения способа реализации информационной структуры

§ 4.1. Объекты исследования современной лингвистики текста. Информационный поток

Изменившиеся условия существования человека коренным образом перестроили процедуру анализа информации. Развитие технологий информационного и фактографического поиска открывает ***новое*** поле деятельности для специалистов в области компьютерной лингвистики текста. Раньше основным и единственным объектом лингвистического исследования был ***текст*** (его анализ, понимание). Но для того, чтобы полноценно жить в информационном обществе, человек должен обрабатывать огромное количество информации. Лавина информации, содержащаяся в информационных потоках, не может быть воспринята и проанализирована человеком в силу его психофизиологических ограничений. Новый информационный объект – ***информационный поток*** – требует использования новых технологий, которые выступают в качестве посредника при извлечении адресатом коммуницируемого смысла. В нашей «лингвистической» работе ***информационный поток*** понимается, прежде всего, как ***множество текстов, выступающих как единый объект***: адресатов интересует смысл, заключенный сразу в сотнях и даже тысячах текстов.

Гораздо подробнее – всесторонне и модельно – тема информационных потоков рассматривается Д.В.Ландэ в части 6 данного пособия. Однако мы, лингвисты, умеем работать, главным образом, с теми объектами, которые имеют лингвистическую природу. Тематические информационные потоки гораздо ближе к сфере интересов и возможностей лингвистики, именно их мы в своих работах чаще всего называем информационными потоками, учитывая структурные связи между текстами (документами) и внутри самих текстов.

«В самом естественном языке устойчивость частот слов (существование ансамбля статистически однородных текстов) вызывает сомнение. Любой целостный текст обладает индивидуальностью. Попытка найти реальные статистически однородные ансамбли текстов никому еще не удавалось. Точнее говоря, не удавалось наблюдать такой набор текстов, в которых слова встречались с одинаковым спектром частот. В то же время словник любого текста, который по разумным содержательным соображениям удается считать замкнутым, можно упорядочить» (Часть VI.Глава 2 «Самоподобие в информационном пространстве» данного пособия).

Использование принцип самоподобия в интернетике по самым разным причинам соотносится с задачей выбора контекста. Одним из вариантов информационных потоков является коллекция текстов. В качестве такого рода коллекций могут выступать самые разные коллекции, с точки зрения, как структуры коллекции, так и

структуры текста (или подколлекций этой коллекции). Это плодотворная и увлекательная тема, которую в этом учебном пособии нам удалось лишь затронуть (глава 2 и 4).

Что такое информационное пространство? Является ли общее информационное пространство видом контекста? Вероятно, да. Но пока еще трудно нащупать лингвистические принципы организации такого рода контекста. И явно информационное пространство – в современном информационном обществе – выходит за рамки привычных лингвистических контекстов (наподобие, скажем, Национального корпуса). Идеи так называемого Semantic WEB уже ближе к информационному пространству, хотя и не покрывают всей сложности и многообразия связей, сосуществующих в сети. Может ли математическое исследование информационного пространство приблизить нас к пониманию природы лингвистических объектов? Думаю, что на этот вопрос должен быть положительный ответ. В результате мы поймем информационную и лингвистическую природу таких объектов как текст, кластер (сюжет), коллекция, тематическая коллекция,… можем продолжить, и назвать в качестве примера еще полнотекстовую базу результатов однотипных запросов поисковых машин.

Одна из основных практических особенностей с коллекциями состоит еще в том, что это система коммуникации «автомат→человек», а часто и в необходимости компрессированной выдачи информации человеку: например, набора ключевых слов или даже аннотации (или обзорного реферата). Набор ключевых слов – свертка исходного текста, проблемы формирования набора ключевых слов связаны исключительно с **анализом** текстового материала коллекции, требования к выбору единицы анализа гораздо менее четкие, чем при аннотировании.

В «качестве информационного портрета темы, соответствующей запросу, можно рассматривать множество ключевых слов, наиболее точно (по статистическим и смысловым алгоритмам) отражающее информацию, получаемую в результате поиска по данному запросу. Построение информационных портретов в реально функционирующих системах выполняется на основе эмпирических и статистических методов, основу которых, как и в случае автореферирования, составляют частотно-лингвистические алгоритмы». Например, «информационный портрет может быть реализован как отдельная семантическая карта или как таблица на экране с результатами поиска» [123: 167]. Да, действительно, информационные портреты часто «живут» в ИПС, помогая уточнять систему запроса. Однако это далеко не единственное применение наборов ключевых слов как информационных портретов темы (коллекции того или иного вида). Позволю себе заметить также, что существует обилие уже упомянутых статистических и смысловых алгоритмов для получения информационных портретов. В ряде случаев для получения таких портретов используют элементы Information Extraction (например, для извлечения наименований персон, организаций, географических наименований), в результате элементами анализа становятся как слова, так и коллокации, что сближает наборы ключевых слов (или словосочетаний), выделяемых автоматически и в ходе эксперимента с информантами.

При создании (обзорной) аннотации осуществляется и **анализ** исходного текстового материала коллекции, и **синтез** текста аннотации. Все это налагает гораздо более жесткие требования к выбору единиц, к последовательности их размещения и реализации связности (тематической и семантико-синтаксической).

«На сегодня существует множество путей решения задачи, которые достаточно четко подразделяются на два направления – квазиреферирования и краткого изложения содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов, – выделении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала основывается на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее существенной информации и порождении новых текстов, содержательно обобщающих первичные документы» [123: 158].

Представлял бы крайний интерес лингвистический анализ аннотаций в сопоставлении со структурой исходного объекта: степень информационной насыщенности (vs. воздействия на адресата, напр., в интервью и даже некоторых видах аналитики), статичность vs. динамичность (событие vs. сюжет со сменяющимися ситуациями vs. череда повторяющихся событий), компактность vs. диффузность информационной структуры и т.д.

§ 4.2. Коллокации и конструкции как составляющие текстов

В предыдущей главе выборка анализируемых **текстов** – текстов в условном отрыве от коллекций как баз текстов – была ограничена возможностями экспериментов с информантами, т.е. объектом исследования становились отдельные тексты (см. [158]). Попробуем реализовать следующий виток, когда объектом исследования становятся большая текстовая коллекция объемом в миллионы словоупотреблений и тематически однородные кластеры (подколекции). В результате различных вычислительных экспериментов на основе таких коллекций мы получаем данные, с одной стороны, позволяющие соотнести особенности структуры двух разных объектов (коллекции vs. единичные тексты), с другой – определить интересующие нас типы текстов (структур текстов) и, тем самым, сузить материал для экспериментальной работы с информантами. В результате мы имеем возможность наиболее тщательно исследовать роль контекста: большой коллекции текстов → тематически однородной подколлекции текстов (сюжет или кластер) → единичного текста и → минимального синтаксического контекста (подробнее см. [158; 162]). Мы в своем исследовании языка и речи идем от реализации, от имеющегося в нашем распоряжении материала.

Рассматриваем **все** связанные сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и/или экспериментов с информантами. Выделяемые единицы представляют собой неоднородное множество, требующее интерпретации (см. главу 2). Возвращаемся к теме «единица и контекст» уже на витке, приближающемся к конкретным текстам (своего рода связка между главой 2 и 4):

- минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;
- текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;
- контекст, предполагающий учет текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т.д.)

Неоднословнные связанные сегменты выступают, прежде всего, как структурные составляющие текста или однородных коллекций (например, сюжетов). Анализ этих структурных составляющих позволяет исследовать структуру текста и/или текстов. Единицы и контекст(-ы) анализируются во взаимодействии: контекст и коммуникативная задача определяют выбор единиц анализа. Тематически однородная коллекция (сюжет) изучается методами, пришедшими из лингвистики текста (дискурса).

Нами оценивались следующие данные:

- полученные в ходе вычислительных экспериментов:
  - список наиболее связанных n-грамм по коллекции;
  - список наиболее связанных n-грамм по подколлекции (подколлекция является тематически более однородной, чем исходная коллекция);
  - отдельные тексты, представленные в виде последовательности связанных сочетаний («сегментов» в терминологии автора программы).
- полученные в ходе эксперимента с информантами отдельные тексты, представленные в виде последовательности связанных сочетаний.

Подтвердились следующие **гипотезы**:

- с увеличением степени однородности (коллекция→ однородная коллекция→текст) характерными становятся более длинные n-граммы;
- с увеличением степени однородности (коллекция→ однородная коллекция→текст) увеличивается число конструкций (в соотношении конструкция vs. типовая коллокация), увеличивается число предикативных сочетаний;
- набор связанных сочетаний, подсчитанных для каждого текста отдельно в ходе вычислительного эксперимента, сходен с набором сочетаний, полученных в ходе экспериментов с информантами,
- набор связанных сочетаний, выделенный в ходе экспериментов с информантами, содержит несколько больше предикативных сочетаний, чем набор связанных сочетаний, сформированный в ходе вычислительного эксперимента.

Такое исследование предполагает сочетание вычислительного эксперимента и эксперимента с информантами. В ходе вычислительного эксперимента меры совместной встречаемости определяется на основании видоизмененной меры Дайса (Dice) [19]:

$$Dice'(x, y) = \log_2\left(\frac{2 * f(x, y)}{f(x) + f(y)}\right),$$

где *f(x)* и *f(y)* – частота встречаемости слов x и y в коллекции, а *f(x,y)* – частота совместной встречаемости слов *x* и *y*.

Процесс вычислительного эксперимента можно коротко описать следующим алгоритмом. Сначала для всех пар слов по всей коллекции считается коэффициент Дайса. Затем для каждого конкретного текста, представляющего собой цепочку слов или, вернее, цепочку пересекающихся пар (слово *x* с предшествующим словом и слово *x* с последующим словом), осуществляется «сборка» связанных сегментов. При последовательном прохождении от слова к слову в каждом тексте уже известны соответствующие значения меры Дайса для всех пересекающихся пар. На основании значений этой статистической меры слова объединяются в связанные группы с учетом ближайшего контекста (принимается решение о том, надо ли присоединить

текущее слово к предыдущему). Слово не присоединяется к предыдущему, если значение коэффициента Дайса для данной пары ниже порогового, или если оно ниже, чем среднее арифметическое того же коэффициента для левой и правой пары. Во всех остальных случаях слово присоединяется. Связанный сегмент может включать не более семи слов (мы ни разу не приблизились к этому порогу). В результате такого вычислительного эксперимента мы получаем набор связанных сочетаний, подсчитанных для каждого текста отдельно, а затем объединенный в некое подобие частотного словаря связанных сочетаний. Программа, реализующая этот алгоритм, доступна для скачивания с сайта ее создателя: http://donelaitis.vdu.lt/~vidas/tools.htm.

Используемая мера выделяет связанные сегменты (как коллокации, так и конструкции), характеризующиеся информационной ценностью на материале однородной коллекции текстов (ср. [20; 21]). Свое предположение мы проверили через сопоставление с результатами, полученными с помощью стандартных статистических мер MI и t-score, с ключевыми словами, выделяемыми на основании коэффициента важности tf-idf (этот коэффициент позволяет оценить степень важности слова по отношению к той или иной коллекции (подколлекции)) и рядом дополнительных методик. Выдвинутое предположение об информационной значимости связанных сегментов, выделяемых с помощью меры Дайса на материале тематически однородной коллекций текстов, подтверждается в ходе предыдущих исследований с использованием меры MI (напр., [159; 161]). При рассмотрении указанных сегментов в рамках единичных текстов (по результатам вычислительного эксперимента и эксперимента с информантами) будем называть их значимыми структурными составляющими текста (значимыми для анализа текстов).

Материалом послужили тексты и/или коллекции:

- тексты портала Лента.ру за 2010 год - 40000 текстов общим объемом около 9,5 млн. токенов (т.е. словоупотреблений и знаков препинания);
- два сюжета (или кластера), т.е. две небольших коллекции тематически однородных текстов, полученных с помощью ресурса «Галактика Зум»[61]:
  - приезд А. Шварцнеггера в Москву - 360 текстов, около110 тыс. токенов,
  - назначение С. Собянина - 660 текстов, 170 тыс. токенов,
  - все тексты кластеров берутся из новостного потока, они близки по времени появления и посвящены одному событию;
- три текста о А. Шварцнеггере (из Лента.ру, РИАН, Газета.ру) и два текста о Собянине (Лента.ру, РИАН) для экспериментов с информантами.

Конечно, эти тексты (наряду со прочими текстами кластеров) использовались и в вычислительных экспериментах, т.ч. задача состояла в сопоставлении результатов этих двух экспериментов для каждого рассматриваемого текста.

Крайне важен этап выбора конкретных новостных сюжетов (кластеров), а далее среди них – наиболее представительных текстов. Конечно, все мы знаем, что результаты кластеризации текстов не всегда нас полностью удовлетворяют. Для исследования выбирается «чистый и компактный» кластер сравнительно большого объема, состоящий из максимально тематически однородных текстов. Отбирались кластеры с информационно значимым сюжетом (по субъективной оценке), имеющие четко выстроенный сюжет (основное действующее лицо (или лица), основное

61 Этот материал любезно предоставлен нам Александром Антоновым и Станиславом Баглеем, Галактика-Zoom: galaktika-zoom.ru, http://www.webground.su

действие, сопровождающие действующие лица и/или организации, сопровождающие действия, время, место и т.д.). О других характеристиках скажем чуть ниже.

В эксперименте с информантами – эксперименте по шкалированию – приняло участие около 20 студентов СПбГУ и РГПУ им. А.И.Герцена, получающих гуманитарное образование[62]. Эксперимент с информантами представлял собой оценку связности между текстоформами (пробельными словами) в тексте в шкале от 0 до 5, где 5 – соответствовало максимальной, а 0 – минимальной степени связности. В анкете информанту предлагался текст с «пробелами для заполнения» и инструкция, требующая оценить «*степень связности между словами или словом и знаком препинания в шкале от 0 до 5 баллов. «0» соответствует минимальной силе связности, а «5» – максимальной силе связности. Проставьте эти баллы (от 0 до 5) во ВСЕ позиции, между ВСЕМИ словами и/или словами и знаками препинания*». Информантам отдельно не объяснялся принцип оценки связности, они должны были действовать, опираясь на интуитивные представления о связности и, конечно, на свою текстовую базу знаний. Экспериментатор не навязывает информанту предпочтение, например, синтаксического или лексико-семантического подхода, однако полученные данные позволяют судить о том, что информанты в целом справляются с поставленной задачей. Усредненные данные по группе информантов, представили непротиворечивую оценку степени связности между словами. На основании этих данных можно выстраивать сколь угодно длинные цепочки слов в соответствии с устанавливаемым пороговым значением связности. Эмпирически мы подобрали пороговое значение, равное 3,7 баллам. Если полученное число было больше, чем 3,7, пару слов рассматривали как связную, если меньше – как не связную.

Носитель языка имеет интуитивные представления о неслучайно встречающихся сочетаниях слов: текстовые базы по текстам разных функциональных стилей, по текстам разных тематик или по текстам, посвященным определенной теме. На основании этого знания адресат воспринимает каждый конкретный текст как непротиворечащий некоторой текстовой базе адресата (в качестве ее аналога при вычислительном эксперименте выступают коллекции и подколлекции текстов разной степени однородности). Тематически однородные кластеры представляли достаточно обсуждаемые события, поэтому нельзя было предположить, что информанты не знакомы с этими темами. Эксперимент проводился примерно через месяц после описываемых событий, так что эти темы не могли быть забыты.

Наибольший интерес представляет анализ данных, полученных на материале кластеров для словоформ. При интерпретации данных по рассматриваемым сюжетам мы опирались на данные, полученные на материале двух сюжетов и пяти указанных текстов, однако для иллюстрации возможностей предлагаемого метода приведем результаты только двух текстов: одного текста о А. Шварценергере и одного текста о С. Собянине из «Лента.ру» 2010 года[63].

Сюжет в целом и анализируемый текст о А. Шварценергере гораздо более динамичные (последовательность нескольких ситуаций) и более сложные по тематической структуре: реализующий, например, темы а) приезд известного киноактера, б) приезд губернатора Калифорнии, в) встреча с президентом

[62] Пользуясь случаем, хотим поблагодарить Галину Доброву за помощь в проведении эксперимента.

[63] В статье мы ограниваемся новостными текстами, однако при интерпретации данных частично учитывались также результаты, полученные на материале научных текстов (тематически однородная коллекция материалов конференции «Корпусная лингвистика» и 4 текста из этой коллекции)

Д. Медведевым, г) активное использование твиттеров, д) инвестирование проекта «Сколково». Носитель языка (или автомат) «вправе» сам устанавливать значимость каждой из тем. Сюжет в целом и анализируемый текст о С. Собянине гораздо более статичные и имеют сравнительно простую тематическую структуру («выборы» представляют собой вариант частотного фрейма).

Различие в рассматриваемых сюжетах непосредственно отразилось на результатах эксперимента. Для иллюстрации в таблицах 1 и 2 представлены сегменты, состоящие не менее чем из трех текстоформ (слов, разделителем между которыми служат пробелы и/или знаки препинания) – данные вычислительного эксперимента и эксперимента с информантами – на материале сюжета и текста о А. Шварценеггере (табл. 1) и о С.Собянине (табл. 2). Полужирным шрифтом выделены сегменты или их фрагменты, которые присутствуют как в списке, полученном в ходе вычислительного эксперимента, так и в эксперименте с информантами[64].

Таблица 1. *Связанные сегменты, состоящие не менее чем из трех текстоформ*

Вычислительный эксперимент			Эксперимент с информантами, единичный текст о А. Шварценеггера
Коллекция (Лента.ру 2010 г)	Сюжет о Шварценеггере (однородная коллекция)	Единичный текст о А. Шварценеггера	
тем не менее	глобальное инновационное партнерство	только что приземлился	Губернатор Калифорнии Арнольд Шварценеггер
в связи с	представителей ведущих компаний	**могу дождаться встречи**	прилетел в Москву.
в 2009 году	с губернатором калифорнии	вскоре после этого	в российскую столицу
то же время	могу дождаться встречи	**ответил** калифорнийскому губернатору	Не **могу дождаться встречи** с президентом Медведевым
в настоящее время	во главе делегации	англоязычная версия твита	российский президент Дмитрий Медведев **ответил**
со ссылкой на	создать настоящий технологический бум	ответил ему взаимностью	в своем микроблоге
возбуждено уголовное дело	сфере высоких технологий	это же время	добро пожаловать в Москву
по сравнению с	только что приземлился		Жду встречи с вами
в 2008 году	тогда вам сказал		Медведев добавил микроблог
и т.д.	которые занимаются инновационными разработками		с делегацией представителей
	их российскими партнерами		он встретится с российскими министрами
	российская венчурная компания		во время посещения Медведевым
	стать мэром москвы		российский президент завел себе
	Global Technology Symposium		
	главами американских инвестиционных компаний		
	видение дальнейшего развития		
	Silicon Valley Bank		
	пост мэра москвы		
	самых разных событий происходит		
	июне этого года		
	после непродолжительной беседы		
	и т.д.		

[64] В графу «Сюжет о Шварценеггере (однородная коллекция)» попала верхушка наиболее частотных связанных сегментов, упорядоченных по частоте, остальные графы (наборы) представлены полностью.

Предложенная нами методика учитывает различные виды контекстов: «тематический» (сюжет) и «стилистический» (Лента.ру) (см. табл. 1). В «стилистическом» контексте существенными оказывались характерные для СМИ конструкции и обороты (например, *в настоящее время*, *со ссылкой на*), из которых нельзя сделать выводы о конкретном содержании текстов, но можно составить общее впечатление об их стилистической направленности (см. табл. 1). В «тематическом» контексте наиболее значимыми оказывались сложные номинации (*глобальное инновационное партнерство*) и предикативные конструкции, описывающие ситуацию (*только что приземлился*) (см. табл. 1). Структурные составляющие сюжета дали более полное и объективное представление о сюжете, чем структурные составляющие единичного текста. Информанты в целом выделяли более длинные сегменты, чем программа. Информанты были нацелены на описание ситуаций, они выделяли большее число предикативных сочетаний – длинные конструкции в целом более типичны, чем длинные коллокации.

Число пересекающихся длинных связанных сегментов, выделяемых программой и информантами, в существенной степени зависит от типа текста. Для более динамичных сюжетов и текстов (включающих описание последовательности событий) число пересечений меньше, для более статичных – больше[65]. Это один из параметров, позволяющих оценить структуру единичного текста и текстов сюжета в целом. Мы ни в коей мере не рассматривали эксперимент с информантами как вариант оценки работы программы; вычислительный эксперимент и эксперимент с информантами имели одинаковый статус.

Набор длинных связанных сегментов, выделяемых информантами, на наш взгляд, может считаться самоценным для анализа структуры текста, т.к. вполне вероятно, что они отражают расстановку структурных составляющих текста, важных для восприятия (ср. идею о том, что при восприятии адресат стремится оперировать наиболее крупными оперативными единицами, в главе 3). Продемонстрируем это на примере текста, в котором длинные связанные сегменты интерпретировались в духе гештальтпсихологии в качестве фигуры (они выделены полужирным шрифтом), а все остальные фрагменты текста рассматриваются как фон:

Губернатор Калифорнии Арнольд Шварценеггер 10 октября **прилетел в Москву.** / После прибытия **в российскую столицу** он сделал в своем микроблоге на Twitter соответствующую запись (Только что приземлился в Москве. Прекрасный день. **Не могу дождаться встречи с президентом Медведевым**), а также разместил фотографию, сделанную по дороге из аэропорта.

Вскоре после этого **российский президент Дмитрий Медведев ответил** калифорнийскому губернатору **в своем микроблоге**: @Schwarzenegger, **добро пожаловать в Москву**. Англоязычная версия твита Медведева также содержала слова "**Жду встречи с вами** и вашей делегацией в @skolkovo".

Кроме того, **Медведев добавил микроблог** Шварценеггера в друзья. Губернатор Калифорнии ответил ему взаимностью.

Как сообщает РИА Новости, Шварценеггер приехал в Россию **с делегацией представителей** венчурных фондов и инновационных компаний Кремниевой долины. Планируется, что помимо президента Медведева, **он встретится с российскими министрами**.

Президент России и губернатор Калифорнии **в этом году** уже встречались - **это произошло в июне / во время посещения Медведевым** США. В это же время **российский президент завел себе** микроблог.

[65] По нашим предварительным данным, для научных текстов такого рода пересечений гораздо больше, чем для новостных текстов.

Объединение набора двухсловных и длинных связанных сегментов увеличивает вес темы «значимость визита А. Шварцнеггера для развития высоких технологий», а насколько эта тема важна – решать адресату, т.е. тому, кто анализирует и понимает этот текст. Возможно, причина невыделения сегментов, несущих такую информацию, в том, что большинство информантов – гуманитарии, однако структура рассматриваемых текстов как минимум позволяет прочтение, в котором «*развитие высоких технологий*» является второстепенным фактом.

На материале результатов вычислительных экспериментов картина более неоднозначная. Если для кластера в целом длинные связанные сегменты информативны, то в случае единичного текста в указанном примере длинных связанных сегментов мало, мы не можем извлечь ценную информацию (понять текст) из их набора (во всяком случае до включения в «расширенный набор» связанных сегментов, состоящих из 2 текстоформ) (подробнее см. [161]).

Почему, если рассматривать каждый из текстов из кластера про Шварцнеггера, то длинных связанных сегментов, полученных в результате вычислительного эксперимента, практически никогда не оказывается достаточно для анализа информационной структуры этого текста? Почему для этого материала столь велико различие между набором длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента?

Основные причины лежат динамичности текста и в особенностях семантико-синтаксической структуры анализируемого в примере текста. Телетайпный, отрывочный стиль написания большинства текстов кластера про А. Шварцнеггера (возможно, обыгрывающий общение в твиттере) характеризуется короткими структурами и навязывает короткие связанные сегменты. Характеристику анализируемого текста можно дополнить отсутствием четко выраженной композиционной структуры сюжета и уже упоминающимся разнообразием тем. Выбор примера – и кластера (сюжета), и текста как его наиболее яркого представителя – обусловил резкое различие между результатами эксперимента с информантами и вычислительного эксперимента.

В качестве контрпримера приведем кластер текстов о С. Собянине и один из них (также из Лента.ру). Наблюдаем значительное сходство между наборами длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента. Длинные связанные сегменты, полученные в результате эксперимента с информантами, рассмотрим в силу нашего допущения как достаточные для анализа (понимания) текста.

Длинные связанные сегменты, полученные в результате вычислительного эксперимента, обладают, главным образом, одним «недостатком»: в их состав не попадают наименования персон, действующих лиц этого сюжета. Если бы мы добавили к этому набору набор двухсловных связанных сегментов или наименования персон (с элементами Ф.И.О.), то вся информация, необходимая для восстановления текста, присутствовала бы в объединенном наборе. Для рассматриваемого текста набор двухсловных связанных сегментов с элементами ФИО следующий: *Собянин утвержден, Сергей Собянин, за Собянина, Юрий Лужков, Дмитрия Медведева, помимо Собянина, Игорь Левитин, соратник Лужкова, Валерий Шанцев, Людмила Швецова, Медведев объявил, Сергее Собянине, Дмитрия Медведева, избрать Собянина, Сергей Собянин, Владимира Путина, Дмитрия Медведева, Владимира Путина*.

Таблица 2. *Связанные сегменты из текста про* С. *Собянина , состоящие не менее, чем из 3 текстоформ*[66]

Кластер про С. Собянина (однородная коллекция)	Вычислительный эксперимент	Эксперимент с информантами
на пост мэра	**Московской городской думы**	Сергей Собянин утвержден
Московской городской думы	**проголосовали 32 депутата**	на посту мэра Москвы
проголосовали 32 депутата	участвовали 34 человека	**Московской городской думы**
тот же день	присяга нового мэра	**проголосовали 32 депутата**
губернатор Нижегородской области	тот же день	против высказались двое
нового мэра Москвы	**Как сообщалось ранее**	голосование в Мосгордуме
из 35 депутатов	**18 : 00**	**Как сообщалось ранее**
инаугурация нового мэра	избрании нового градоначальника	торжественное мероприятие планируется провести
центральном Федеральном округе	руководивший исполнительной властью	в **18:00**
кандидатуру Сергея Собянина	9 октября партия	21 октября 2010 года
на посту мэра	представила президенту четыре кандидатуры	нового градоначальника Москвы
добросовестно исполнять возложенные	список единоросов попали	исполнительной властью столицы
благополучию его жителей	**губернатор Нижегородской области**	с утратой доверия президента
участвовали 34 человека	прошлом - вице-мэр	Соответствующий указ Дмитрия Медведева
губернатором Тюменской области	**исполняющая обязанности вице-мэра**	на пост мэра Москвы
остановил свой выбор	остановил свой выбор	**губернатор Нижегородской области**
по его словам	после этого фракция	**исполняющая обязанности вице-мэра** Москвы
присяга нового мэра	из 35 мест	президент Медведев объявил
Московская городская дума	органах власти начался	аппарата правительства РФ
руководивший исполнительной властью	**городе Когалым Ханты-мансийский округа**	пообещала поддержать выбор Дмитрия Медведева
9 октября партия	**ответственные государственные посты**	в **городе Когалым Ханты-Мансийский округа**
избрании нового градоначальника	губернатором Тюменской области	в разные годы
до 2008 года	до 2008 года	занимал **ответственные государственные посты**
из 35 мест		
органах власти начался		
ответственные государственные посты		

Результаты вычислительного эксперимента и эксперимента с информантами эксплицируют разные информационные структуры одного и того же текста: разные варианты извлечения информации в соответствии с намерениями и возможностями адресата. Адресат (носитель языка или автомат) выделяет важные вехи в тексте на основании коммуникативной ситуации, собственных целей и задач. Разные возможности и задачи соответствуют разным коллекциям (в соответствии тематической областью коллекции и/или разной степенью однородности) или разным базам знаний информантов (степени компетентности информантов). Главное – мы проиллюстрировали то, что получаемые результаты в существенной степени зависят от лингвистической природы моделируемого объекта: в первую очередь, сюжета

[66] Полужирным шрифтом выделены те сегменты или их фрагменты, которые присутствуют в списках, полученных как в ходе вычислительного эксперимента, так и эксперимента с информантами.

(кластера), а во вторую – конкретного текста как представителя этого кластера. Следовательно, лингвистический анализ объекта (набора объектов) может и – надеюсь, во многих случаях должен – предшествовать вычислительным процедурам, выделяя те закономерности, которые можно предсказать на начальном этапе («постановке» задачи коммуникации, формулировке гипотез методами лингвистики текста).

§ 4.3. Свертки для описания разных информационных объектов: от текстов до информационных потоков

При всем различии рассматриваемых информационных объектов – текст и информационный поток – нас интересует то, что они обладают информационной (смысловой) структурой и могут быть свернуты до набора слов и словосочетаний. Этот набор может выступать представителем (носителем) информационной структуры объекта (и текста, и информационного потока). Эту тему мы поднимали в первом параграфе этой главы.

Напомним, что ключевыми словами (или аналогами ключевых слов) в разных контекстах называют, напр.:

1 выписанные группой информантов слова, наиболее важные для решения поставленных в инструкции задач (обычно – понимания текста),

уровень значимости слова определяется как относительная частота его встречаемости в протоколах информантов,

2 автоматически выделяемые неслучайно встречающиеся в документах слова и словосочетания, важные для рассматриваемой выборки (выдачи) в рамках общего массива документов,

уровень значимости слова рассчитывается на основании некоего алгоритма.

Чтобы осуществить свертывание текста в виде КС, этот текст нужно понять. Поэтому естественно считать, что свертки представляют собой результат понимания текста или, иначе говоря, извлечения смысла из текста. Рассмотрим пример экспериментального исследования информационной значимости сверток (с точки зрения той задачи, которая стояла перед информантами). С помощью такого эксперимента изучалась возможность восстановления исходного смысла или информационной структуры текста.

Ресурс Галактика-Зум (http://galaktika-zoom.ru/, также см. http://webground.su) предоставляет возможности для проведения исследования на материале сверток (наборов) автоматически определяемых ключевых слов. Для каждой выдачи (в соответствии с запросом) этот ресурс вычисляет и предоставляет пользователю ***Информационный портрет*** (или Инфорпотрет), т.е. набор автоматически определяемых слов и словосочетаний, важных для рассматриваемой выборки (среза) в рамках общего массива документов. Инфопортрет как сверка множества текстов является основной возможностью для извлечения адресатом ***целостной информационной структуры***: большой объем не позволяет человеку оперировать непосредственно с каждым текстом.

Основной задачей данного эксперимента было определить, является ли Инфопортрет реальной сверткой текста, т.е. сможет ли информант восстановить по нему информацию об объекте, описанном в данном тексте, в частности, информацию процедурно-временного характера. Для этого перед информантами ставится задача

определения временного периода, к которому относится группа текстов. При этом из свертки должны быть удалены все непосредственные указания на временной период (месяц, квартал, конкретные даты).

Нами анализировались новостные тексты: их достаточное количество по выбранной нами тематике, они, в основном, компактны и ограничены лексически. В качестве запросов были выбраны запросы «ЕГЭ» и «единый государственный экзамен», т.е. выбирались тексты, содержащие данные слово или словосочетание. Основания для выбора именно таких запросов были следующие:

- «ЕГЭ» («единый государственный экзамен») может быть по праву названо одним из «ключевых слов» 2009 года. Актуальность и востребованность этой темы позволила получить в выдаче большое количество текстов (см. табл. 1). Причем выборочный анализ текстов выдачи показывает, что тематически они достаточно однородны.
- Тема «ЕГЭ» (или «единый государственный экзамен») была выбрана из-за того, что в самой природе рассматриваемого объекта (и текстов, его описывающих) заключена периодизация и хорошо знакомый лингвистам принцип построения сюжета. Причем эти периоды несут особую информационную нагруженность (подготовка – проведение – подведение итогов), что позволяет в процедуре проведения эксперимента с информантами через определение интервала эксплицировать основную информацию, содержащуюся в предъявляемых Инфопортретах.

На вход системе «Галактика-Зум» были посланы запросы: (1) «ЕГЭ» и (2) «единый государственный экзамен». Результаты этих запросов система распределила по 9 выборкам, где каждая из выборок содержит документы, относящиеся к одному из прошедших месяцев 2009 года (от января по сентябрь включительно). В общей сложности 9 выборок содержало 7768 текстов для запроса «ЕГЭ» и 2232 для «единый государственный экзамен». Два эксперимента – эксперимент 1 («ЕГЭ») и эксперимент 2 («единый государственный экзамен») – содержал по 9 информационных портретов (для каждого по 9 выборок своего запроса). Каждому информанту выдавалась инструкция:

«Каждый из 9 листов соответствует выборке одного месяца 2009года.

Ваша задача оценить и отметить на каждом листе свой выбор:

1. *предположительный период: подготовка к экзамену – проведение экзамена – подведение итогов;*
2. *месяц: от января до сентября 2009 года;*
3. *критерии, особенности, комментарии и т.д.»*[67]

[67] Информанты – 16 (17) студентов и аспирантов СПбГУ гуманитарных специальностей. Они не являлись специалистами в предметной области (переход на систему ЕГЭ) ни в силу профессиональной деятельности, ни в силу жизненного опыта (т.к. сами сдавали традиционные экзамены). Процедуры принятия решения и используемые ими критерии не связаны со специальными знаниями и навыками (напр., аналитической работой с информационными потоками). Смысловая структура текстов (выборок текстов) данной предметной области в большинстве случаев неоднородна и предполагает конкуренцию критериев, т.к. включает в себя в качестве подтем как минимум три: окончание школы – сдача ЕГЭ – поступление в вуз. Второй эксперимент проводился через 2,5 месяца после первого с той же бригадой информантов (добавился один новый). Методика проведения эксперимента должна была минимизировать влияние индивидуальных ассоциативных связей. Собранная бригада участвовала в двух экспериментах. В промежутке результаты эксперимента с информантами не обсуждались. Сопоставительный анализ протоколов второго эксперимента исключают возможность влияния на его результаты первого эксперимента; таким образом, мы считаем, что экспериментальный дизайн удовлетворяет требованиям чистоты эксперимента.

На основании результатов определения испытуемыми периода в эксперименте 1 («ЕГЭ») можно выделить четыре класса (по убыванию числа правильных и согласованных ответов информантов):

1. Февраль, март, сентябрь (подготовка экзамена и подведение итогов);
2. Январь (подготовка экзамена);
3. Апрель, август (подготовка экзамена и подведение итогов);
4. Май, июнь, июль (неопределенность проведение экзамена/подведение итогов).

Свертки, предъявленные испытуемым в ходе Эксперимента 2 («единый государственный экзамен»), дали другое распределение правильных и согласованных ответов испытуемых:

1. июль, февраль (подведение итогов и подготовка, соответственно);
2. апрель (проведение экзаменов вместо подготовки);
3. январь, май, сентябрь (подготовка, проведение, и подведение итогов, соответственно);
4. июнь (подготовка экзамена вместо проведения, но сравнительно высокая согласованность);
5. август, март (подведение итогов и подготовка, соответственно).

Анализ тем текстов разных выдач показывает, что однозначное определение периода и месяца не обязательно должны соответствовать друг другу. Сроки проведения ЕГЭ колеблются от апреля до июля (согласно приказу «Об утверждении сроков и единого расписания проведения …» (http://www1.ege.edu.ru/content/view/475/36/):

- досрочное проведение – апрель,
- для основной массы выпускников 2009 года – июнь (а также 26 и 29 мая),
- для выпускников прошлых лет – июль.

Выборочный анализ текстов выдач по рассматриваем запросам (месяцы апрель-июль) показывает, что выдача на запрос «единый государственный экзамен» в большей степени ориентированы на «проблемные» случаи, а на запрос «ЕГЭ» – на типичные. Для «апреля» (Эксперимент 2) проблемным является досрочное проведение ЕГЭ (высокая согласованность и сравнительно неплохое восстановление месяца). Для июня и июля – сдача ЕГЭ выпускниками прошлых лет (неравное положение выпускников 2009 года и прошлых лет, т.е. более сложные условия для последних). Поэтому «июнь» дает большее внимание к подготовке, а июль к подведению итогов.

Задача краткого обсуждения результатов сверток по коллекциям документов, сформированным двумя сходными запросами, позволил дать анализ, прежде всего, лингвистической природе коллекции. Например, проследить (1) роль шкалы информационная насыщенность vs. воздействие на адресата, (2) взаимодействие и переплетение тем и подтем сложного сюжетного объекта, (3) композиционную структуру сюжета, выстраивающуюся по законам нарратива. Все эти параметры являются, прежде всего, лингвистическими и информационными. Отметим, что шкала от информационно насыщенного текста до текста, реализующего воздействие на адресата, была (в данной предметной области) задана, прежде всего, на уровне запроса.

Применяя нарративную метафору, можно рассмотреть девять сверток (для каждого из периодов, которому соответствовала одна выборка) как компоненты

единой смысловой структуры высокого уровня, характеризующейся динамичной сменой ситуаций (при том, что каждая из этих ситуаций сама имеет сложную смысловую структуру). Тогда свертку «январь» можно описать как ***преамбулу*** (фазу ориентации), «февраль» как основу ***завязывания сюжета***, «сентябрь» – как ***коду*** (мораль всей истории). Именно эти компоненты нарратива ведут себя сходным образом и для запроса «ЕГЭ», и для запроса «единый государственный экзамен». Наиболее сюжетными и неоднозначными оказались свертки «апрель–июль», на которых происходит развитие сюжета. Анализ результатов экспериментов демонстрирует разные сюжетные линии. Степень «воздействия на адресата» (напр., убеждения) задает разные направления: типичное положение дел (для «информационных текстов») или проблемные случаи (для «текстов воздействия на адресата»).

Методы экспериментальной лингвистики – психолингвистики и лингвистики текста – находят свое применение в исследовании нового для лингвистики объекта.

Список используемой литературы

1. *Aborn M. et al.* Sources of contextual constraint upon words in sentences // J. of Exp. Psychology. 1959, vol. 57.
2. *Alderson J. C.* Native and non-native speaker performance on cloze tests // Language Learning. 1980, vol.30.
3. *McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W.*. Are good text always better? Text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 1996, 14, 1-43.
4. *Bachman L. F.* Performance on cloze tests with fixed-ratio and rational deletions // TESOL Quarterly. 1985, vol. 19.
5. *Barlow M., Kemmer S.* (eds) Usage-based models of language. Stanford, Calif.: CSLI Publications. 2000.
6. *Biber D.* Variation Across Speech and Writing – Cambridge:Cambridge Univ. Press, 1988.
7. *Boguslavsky I., Iomdin L., Sizov V.* Multilinguality in ETAP-3. Reuse of Linguistic Resources // Proceedings of the Workshop "Multilingual Linguistic Resources. 20th International Conference on Computational Linguistics» – Geneva, 2004
8. *Boguslavsky I., L. Iomdin, V. Sizov.* Interactive enconversion by means of the ETAP-3 system // Proceedings of the International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies – Alexandria, 2003
9. *Chafe W.* Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing – Chicago:University of Chicago Press, 1994
10. *Church K., Hanks, P.*, 'Word association norms, mutual information, and lexicogra-phy', Computational Linguistics, 16(1), 22–29. 1990
11. *Church, K., W. Gale, P. Hanks and D. Hindle* 1991 Using statistics in lexical analysis. In U. Zernik ed Lexical Acquisition. Englewood Cliff, NJ: Erlbaum. 115-64.
12. *Church, K., W. Gale, P. Hanks and D. Hindle* Using statistics in lexical analysis. In U. Zernik ed Lexical Acquisition. Englewood Cliff, NJ: Erlbaum. 115-64. 1991
13. *Clark H. H.* Arenas of Language Use – Chicago: Univ. ChicagoPress, 1993
14. Cloze tests in English, Thai, and Vietnamese. Native and non-native performance 1972;
15. *Croft W.* Radical Construction Grammar. Syntactic theory in typological perspective. Oxford: Oxford University Press. 2001.
16. *Croft W., Cruse A.* Cognitive Linguistics. Cambridge: Cambridge University Press. 2004.
17. *Crowder R. G., Morton J.* Precategorical Acoustic Storage // Perception and Psychophysics. 1969, №5
18. *Сиротко-Сибирский С. А*. О проблеме понимания текста в лингвистике и психолингвистике // … Слово отзовется: памяти А. С. Штерн и Л. В. Сахарного – Пермь, 2006
19. *Daudaravicius V.* Automatic identification of lexical units. // Computational Linguistics and Intelligent text processing CICling-2009, Meksikas, Meksika. 2010a.

20. *Daudaravičius V.* The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance. In proceedings of Computational Linguistics and Intelligent text processing CICling-2010, Iasi, Romania. Lecture Notes in Computer Science. Springer-Verlag.648–660. 2010б.
21. *Daudaravicius, V.*, Marcinkeviciene, R.: Gravity Counts for the Boundaries of Collocations. International Journal of Corpus Linguistics 9(2), 321–348. 2004
22. *Fillenbaum S. et al.* The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript // J. of Verbal Learning and Verbal Behavior. 1963, vol. 2
23. *Fillmore Ch. J.* Inversion and constructional inheritance.// Webelhuth G., Koenig J., Kathol A. (eds.). Lexical and constructional aspects of linguistic explanation. Stanford, Ca: CSLI. 113-128. 1999.
24. *Fillmore Ch. J., Kay, P.*. Construction Grammar Coursebook. Manuscript, University of California at Berkeley Department of linguistics. 1993.
25. *Fillmore Ch. J., Lee-Goldman R. R., and Rhodes R.* The FrameNet Constructicon // Sign-Based Construction Grammar. Stanford: CSLI Publications / H.C. Boas & I.A. Sag (eds.) 2011
26. *Fillmore Ch., J., Kay P., O'Connor M. C.* Regularity and idiomaticity in grammatical constructions: The case of Let alone. // Language 64, 3. 501-538. 1988.
27. *Firbas J.* On defining the theme in functional sentence analysis // Travaux linguistiques de Prague, vol. 1. 1966
28. *Firbas, J.* Functional Sentence Perspective in Written and Spoken Communication. Studies in English Language – Cambridge: CambridgeUniversity Press, 1992
29. *Firth, J.R.:* 1957. Papers in Linguistics 1934–1951. London.
30. *Firth, J.R.:* 1968.Selected Papers of J.R. Firth, 1952–1959. London.
31. *Fletcher C. R., Bloom C. P.* Causal reasoning in the comprehension of simple narrative texts // *J. Mem. Lang.* 1988, 19
32. *Fodor J. A.* The Modularity of Mind – Cambridge (Mass.), 1983
33. *Fox B.* Discourse structure and anaphora – Cambridge: Cambridge Univ. Press, 1987
34. *Fried M., Östman J.* 2004. Construction Grammar: a thumbnail sketch.// Fried M., Östman J. (eds.). Construction Grammar in a cross-language perspective. 11-86. Amsterdam: John Benjamins.
35. *Gernsbacher, M. A.* Handbook of Psycholinguistics / M. A. Gernsbacher. – New York : Academic Press, 1994. – 1174 p
36. *Givón, Talmy*. Topic continuity in discourse: An Introduction / Topic continuity in discourse: a quantitative cross-linguistic study (Typological studies in language, vol. 3) – Amsterdam: J. Benjamins, 1983.
37. *Goldberg A.*. Constructions. A Construction Grammar approach to argument structure. Chicago: University of Chicago Press. 1995
38. *Goldberg A.*. Constructions at Work: the nature of generalization in language. Oxford University Press 2006
39. *Graesser A. C. et al.* Discourse comprehension // Ann. Rev. Psychol. 1997, 48
40. *Gundel J. K., Hedberg N., Zacharsky R.* Cognitive status and the form of referring expressions in discourse // Language. 1993, 69.
41. *Hajičová E.* Surface and underlying word order // Prague Linguistic Circle Papers, V.1 – Amsterdam: J. Benjamins, 1995.
42. *Hajičová E., Korbayová I.* On the notion of topic // Prague Linguistic Circle Papers –Amsterdam: J. Benjamins, in print.
43. *Hajičová E., Sgall P.* Towards an automatic identification of topic and focus/ // Proceedings of the second conference on European chapter of the Association for Computational Linguistics – Geneva, 1985.
44. *Hajičová E., Sgall P., Skoumalová H.* Identifying Topic and Focus by an Automatic Procedure // *Hess DJ, Foss DJ, Carroll P.* Effects of global and local context on lexical processing during language comprehension // J. Exp. Psychol. Gen. 1995, 124.
45. *Halliday M.* Lexis as a Linguistic Level. // Bazell, C., Catford, J., Halliday, M., and Robins, R. (eds.). In Memory of J. R. Firth. Longman, London 1966.
46. Ресурс и библиография по Грамматике конструкций/ http://constructiongrammar.org/
47. *Iordanskaja, L., Paperno, S.* A Russian-English Collocational Dictionary of the Human Body, Columbus/Ohio. 1996.
48. *Johnson-Laird P.* Mentëal models: Towards a cognitive science of language, inference and consciousness – Cambrige, 1983

49. *Just M. A., Carpenter P. A.* A capacity theory of comprehension: Individual differences in working memory // Psychological Review. 1992, 98
50. *Kibrik A. A.* Anaphora in Russian narrative discourse: A cognitive calculative account // Studies in Anaphora / B. Fox (Ed.) – Amsterdam: J. Benjamins, 1996
51. *Kintsch W. K.* The Representation of Meaning in Memory – Hillsdale (NJ), 1974
52. *Kintsch W.* Text comprehension, memory, and learning // Am. Psychol. 1994, 49
53. *Kintsch W. K.* The Representation of Meaning in Memory – Hillsdale (NJ), 1974.
54. *Kintsch, W. K.* Toward a model of text comprehension and production / / W. K. Kintsch, T. A. van Dijk // Psychological Review, 1978, 85. – P. 363-394.
55. *Lagus K., Kohonenand O., Virpioja S.* Towards unsupervised learning of constructions from text // Proceedings of the Workshop on extracting and using constructions in NLP / Sahlgren M. and Knutsson O. NODALIDA 2009
56. *Lagus K., Kohonenand O., Virpioja S.* Towards unsupervised learning of constructions from text // Proceedings of the Workshop on extracting and using constructions in NLP / Sahlgren M. and Knutsson O. NODALIDA 2009
57. *Levelt W.* Speaking: From Intention to Articulation – Cambridge: MIT Press, 1989
58. *Liang S.F., Devlin S., Tait J.* Can automatic abstracting improve on current extracting techniques in aiding users to judge the relevance of pages in search engine results? // 7th Annual CLUK Research Colloquium – Birmingham, 2004
59. *MacDonald M. C., Pearlmutter N. J., Seidenberg M. S.* The lexical nature of syntactic ambiguity resolution // Psychol. Rev. 1994, 101
60. *McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W.* Are good text always better? Text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 1996 14, 1-43.
61. *Mannes S.* Strategic processing of text // J. Educ. Psychol. 1994, 86
62. *Manning C., Schutze H.* Collocations // Manning C., Schutze H. Foundations of Statictical Natural Language Processing, 2002, pp.151-189
63. *Manning C., Schutze H.* Collocations // Manning C., Schutze H. *Foundations of Statictical Natural Language Processing*, 2002, pp.151-189
64. *Manning Ch.* and *Schutze H*. Foundations of Statistical Natural Language Processing . MIT Press, Cambridge, MA, 1999
65. *Masini F.* Multi-word Expressions between Syntax and the Lexicon: the case of Italian Verb-particle Constructions. SKY // Journal of Linguistics 2005. 18. 145-173
66. *McClelland J. L., Rumelhart D. E. et al.* Parallel Distributed Processing: Explorations in the Microstructure of Cognition – Cambridge, MA: MIT Press 1986.
67. *Mel'chuk I.A.* 1995 Phrasemes in Language and Phraseology in Linguistics // Idioms: Structural and Psychological perspectives. Hillsdale, New Jersy 1995, 167-232.
68. *Morris R. K.* Lexical and message-level sentence context effects on fixation times in reading // J. Exp. Psychol. Learn. Mem. Cogn. 1994, 20
69. *Muischnek K., Sahkai H.* Using collocation-finding methods to extract constructions and to estimate their productivity // Proceedings of the Workshop on extracting and using constructions in NLP / Sahlgren M. and Knutsson O. NODALIDA 2009
70. Nenkova A., Gravano A., Hirschberg J. High frequency word entrainment in spoken dialogue // Proceedings of ACL/HLT 2008 (short paper), pages 169-172. Columbus, OH, June 2008.
71. *Norris, D.* Shortlist: a connectionist model of continuous speech recognition // Cognition. 1994, 52
72. *Petrovic S., Snajder J., Basic B.D., Kolar M.* Comparison of collocation extraction for document indexing // Journal of Computing and information technology 14, 4. 321-327. 2006
73. *Rubin, J.* The contribution of video to the development on competence in listening / J. Rubin // A guide for the teaching of second language listening / D. Mendelsohn, J. Rubin (Eds.). – San Diego, CA : Dominie Press, 1995. – P. 151–165.
74. *Seki Y.* Automatic summarization focusing on document genre and text structure. Doctoral abstract // ACM SIGIR Forum, №39(1) – New York, 2005
75. *Soricut* R., *Marcu D.*: Sentence Level Discourse Parsing using Syntactic and Lexical Information // Proceedings of HLT-NAACL 2003.
76. *Stubbs M.*. Collocations and semantic profiles: on the case of the trouble with quantitative studies. // *Functions of language 2:11*, 23-55, Benjamins, 1995.

77. *Taylor W. L.* Cloze procedure: a new tool for measuring readability // Journalism Quarterly. 1953, vol. 30
78. *Trabasso T., Magliano J.P.* Conscious understanding during comprehension // Discourse Processes. 1996, 21
79. *van Dijk T. A., Kintsch W.* Strategies of discourse comprehension – NewYork etc., 1983
80. *Whitney P., Budd D., Bramucci R. S.* Crane RS. On babies, bathwater, and schemata: a reconsideration of top-down processes in comprehension // Discourse Process. 1995, 20
81. *Хоккет Ч.* Проблема языковых универсалий // Новое в лингвистике. Вып. V – М.: Иностр. лит., 1970
82. *Yagunova E., Savina A.* The Experimental Study of Terminology Collocations: Calculations and Experiments with Informants // Proceedings of the workshop on creation, harmonization and application of terminology resources: CHAT 2011. NEALT Proceedings Series. May 11, 2011. / Eds. Tatiana Gornostay, Andrejs Vasiļjevs and Inguna Skadiņa
83. *Zwaan R. A.* Effects of genre expectations on text comprehension // J. Exp. Psychol. Learn. Mem. Cogn. 1994, 20
84. *Zwaan, R.A.* Towards a model of literary comprehension // Models of understanding text / B.K. Britton, A.C. Graesser (Eds) – Hillsdale, NJ: Erlbaum, 1996
85. *Анохин П. К.* Теория функциональных систем. — Успехи физиологических наук. 1970, т. 1, № 1.
86. *Анохин П. К.* Узловые вопросы теории функциональных систем. М., 1980.
87. *Антонов А.В., Ягунова Е.В.* Охват содержимого информационных потоков путем анализа сверток текстов // Материалы XII Всероссийской научной конференции RCDL'2010 «Электронные библиотеки : перспективные методы, технологии, электронные коллекции» (Казань, 13 –17 октября 2010 года) Казань, 2010
88. *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др.* Лингвистическое обеспечение системы ЭТАП-2 – М.: Наука, 1989
89. *Баранов А. Н.* Введение в прикладную лингвистику: Учебное пособие. Изд. 2-е, исправленное. М.: Едиториал УРСС, 2003. - 360 с.
90. *Беллерт И.* Об одном условии связности текста // Новое в зарубежной лингвистике. Вып.VIII. Лингвистика текста – М.: Прогресс, 1978
91. *Бергельсон М.Б.* THEORY OF MIND, или ФАКТОР АДРЕСАТА // Материалы Международной научной конференции MegaLing'2007 Горизонты прикладной лингвистики и лингвистических технологий (24 - 28 сентября 2007 г. Украина, Крым,Партенит) – Симферополь, 2007.
92. *Бергельсон М.Б.* Моделирование культурно обусловленной коммуникативной компетентности с помощью когнитивных категорий: анализ повседневных рассказов и представление знаний // Первая российская конференция по когнитивной науке: Тезисы докладов – Казань: Изд-во КГУ, 2004.
93. *Бергельсон М.Б.* Прагматическая и социокультурная мотивированность языковой формы – М., 2007.
94. *Бернштейн Н.А.* Очерки по физиологии движений и физиологии активности – М., 1966
95. *Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю.* 2008. Словарь глагольной сочетаемости непредметных имен русского языка (электронный ресурс). Доступен для скачивания по адресу: ttp://dict.ruslang.ru/abstr_noun.php
96. *Богданов С. И., Рыжова Ю. В.* 1997. Русская служебная лексика. Сводные таблицы. СПб.
97. *Бондарко Л.В., Вербицкая Л.А., Гордина М.В., Зиндер Л.Р., Касевич В.Б.* Стили произношения и типы произнесения // Вопросы языкознания, 1974, №2
98. *Венцов А.В., Касевич В.Б.* Проблемы восприятия речи – СПб., 1994.
99. *Венцов А.В., Касевич В.Б.* Проблемы восприятия речи (2е изд.) – М.: УРСС, 2003
100. Вероятностное прогнозирование в речи. Отв. ред. Р.М. Фрумкина. – М. 1971.
101. *Глазанова Е.В.* Типы связей в ментальном лексиконе и экспериментальные методы их исследования: дис…канд. филол. наук – СПб., 2001.
102. *Гордеев С.С., Азарова И.В.* Характер корреляции между порядком слов и коммуникативной перспективой в научно-технических текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая-3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея – М.: Изд-во РГГУ, 2007
103. *Грановская Р.М.* Восприятие и модели памяти. Л., 1974.
104. *Даль Э.* Возникновение и сохранение языковой сложности. М., 2009.

105. *Ерофеева Е.В.* Вероятностная структура идиомов: социолингвистический аспект – Пермь: Изд-во Пермского ун-та, 2005
106. *Иорданская Л. Н., Мельчук И. А.* Смысл и сочетаемость в словаре. М.: Языки славянских культур. 2007.
107. *Касевич В. Б.* Морфонология // Касевич В. Б. Труды по языкознанию. – Т.1. – СПб, 2006а.
108. *Касевич В. Б.* Семантика. Синтаксис. Морфология // Касевич В. Б. Труды по языкознанию. – Т.1. – СПб, 2006б
109. *Касевич В. Б.* Элементы общей лингвистики М., 1977
110. *Касевич В. Б., Ягунова Е.В.* Корпуса письменных текстов и моделирование восприятия речи // Вестник СПбГУ. Серия 2. 2006, Вып.3, 2006б.
111. *Касевич В.Б., Ягунова Е.В.* Еще к вопросу о перцептивной значимости начала слова // Лингвистическая полифония: Сборник статей в честь юбилея профессора Р. К. Потаповой / Отв. ред. чл.-корр. РАН В. А. Виноградов – М.: Языки славянских культур, 2007.
112. *Касевич В.Б., Ягунова Е.В.* Контекстная предсказуемость слов в тексте (на материале русского и французского языков) // Вестник Пермского ун-та, вып. 3. – 2006а. – С. 61-70.
113. *Касевич В.Б., Ягунова Е.В.* Контекстная предсказуемость слов в тексте (на материале русского и французского языков) // Вестник Пермского университета, вып.3, 2006б.
114. *Касевич В. Б.* Текст как целостность // Материалы XXIV Международной научной конференции Восточного факультета «Источниковедение и историография стран Азии и Африки» (СПбГУ, 10-12 апреля 2007) – СПб., 2007.
115. *Касевич В.Б., Ягунова Е.В.* Перцептивный словарь взрослых и детей // Проблемы социо- и психолингвистики. Вып.6. Вариативность речевого онтогенеза – Пермь, 2004.
116. *Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю.* Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ. С. 181-185.
117. *Кожина М.Н.* Речеведение и функциональная стилистика: вопросы теории. Пермь, 2002.
118. *Копотев М. В.*. Принципы синтаксической идиоматизации. Хельсинки: Helsinki University Press. 2008
119. *Котюрова М. П.* О понятии связности и средствах ее выражения в русской научной речи // Язык и стиль научной литературы. Теоретические и прикладные проблемы – М., 1977
120. *Кронгауз* М.А. Семантика// М., 2001
121. *Кустова Г. И.* 2008. Словарь русской идиоматики. Сочетания слов со значением высокой степени (электронный документ). Доступен для скачивания по адресу: http://dict.ruslang.ru/magn.php
122. *Ландэ Д.В.* Основы интеграции информационных потоков: Монография. – М., 2006
123. *Ландэ Д.В.* Поиск знаний в Internet. - М.:Диалектика, 2005
124. *Ландэ Д.В., Снарский А.А., Безсуднов И.В.* Интернетика. Навигация в сложных сетях: модели и алгоритмы. М., 2009.
125. *Леонтьева Н.Н.* Автоматическое понимание текста: системы, модели, ресурсы: учебное пособие – М.: Издательский центр «Академия», 2006
126. *Лукашевич* Н.В. Тезаурусы в задачах информационного поиска М., МГУ 2011.
127. *Мельчук И.А.* Опыт теории лингвистических моделей «Смысл↔Текст» – М., 1999
128. *Мельчук И. А.* О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания 4. С. 73-80. 1960
129. *Минский М.* Фреймы для представления знаний – М., 1979
130. *Мурзин Л. Н., Штерн А. С.* Текст и его восприятие – Свердловск, 1991
131. *Николаева Т.М.* Лингвистика текста: современное состояние и перспективы // Новое в зарубежной лингвистике.Вып. VIII. Под ред. Т. М.Николаевой – М.: Прогресс, 1978
132. *Новиков А.И.* Семантика текста и ее формализация, – М.: Наука, 1983
133. *Овчинникова И.Г.* Ассоциативный механизм в речемыслительной деятельности: дис. … докт. филол. наук – СПб. 2002
134. *Откупщикова, М. И.* Синтаксис связного текста: учебное пособие / М. И. Откупщикова. – Л., 1982. – 103 с.
135. *Падучева Е.В.* Высказывание и его соотнесенность с действительностью – М., 2001
136. *Пивоварова Л.М., Ягунова Е.В.* Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов. Предварительные наблюдения // Материалы второго

Международного симпозиума «Терминология и знание» М., 2010
137. *Пивоварова Л. М.* Устойчивые конструкции, характеризующие тексты СМИ // Прикладная и математическая лингвистика: Материалы секции XXXIX Международной филологической конференции, СПб. 2010.
138. *Л.М. Пивоварова, Е.В. Ягунова* Информационная структура научного текста. Текст в контексте коллекции // *Труды международной конференции "Корпусная лингвистика – 2011"* – СПб., 2011 (в печати)
139. *Пиотровский Р.Г.* Лингвистическая синергетика: исходные положения, первые результаты, перспективы СПб. 2006
140. *Пиотровский Р.Г.* Информационные измерения языка. Л., 1968
141. *Пиотровский Р.Г.* Лингвистический автомат (в исследовании и непрерывном обучении СПб., 1999
142. Прогноз в речевой деятельности. М. / Авт. колл.: Р.М. Фрумкина (отв. ред.), А.П. Василевич, П.Ф. Андрукович, Е.Н. Герганов. 1974.
143. *Рогожникова Р. П.* Толковый словарь сочетаний, эквивалентных слову. М., 2003.
144. *Савина А.В., Ягунова Е.В.* Исследование коллокаций с помощью экспериментов с информантами // Труды международной конференции "Корпусная лингвистика – 2011" – СПб., 2011 (в печати)
145. *Сахарный Л.В.* Тема-рематическая структура текста: основные понятия // Язык и речевая деятельность. 1998, №1
146. *Сахарный Л.В., Сибирский С. А., Штерн А. С.* Набор ключевых слов как текст // Психолого-педагогические и лингвистические проблемы исследования текста – Пермь, 1984
Севбо И.П. Структура связного текста и автоматизация реферирования – М.: Наука, 1969
147. *Соколова Е. Г., Семенова С. Ю., Кононенко И. С., Загорулько Ю. А., Кривнова О. Ф., Захаров В. П.* Особенности подготовки терминов для русско-английского тезауруса по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.). Вып. 10 (17).- М.: Изд-во РГГУ, 2011.
148. СЭСРЯ – Стилистический энциклопедический словарь русского языка / Под. ред. М.Н. Кожиной – М.: Флинта: Наука, 2003.
149. Файн В.С. Распознавание образов и машинное понимание естественного языка – М.: *Наука, 1987*
150. *Шайкевич А.Я., Андрющенко В.М., Ребецкая* Н.А., Статистический словарь языка русской газеты (1990-е годы). М. 2008
151. *Шенк Р.* Обработка концептуальной информации – М., 1980
152. *Шмелев А.Д.* Русский язык и внеязыковая действительность – М.: Языки славянских культур, 2002
153. *Штерн А. С.* Перцептивный аспект речевой деятельности – Л., 1992
154. *Ягунова Е.В.* Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов // Русский язык: исторические судьбы и современность: IVМеждународный конгресс исследователей русского языка (Москва, МГУ им. М.В.Ломоносова, филологический факультет 20-23 марта 2010 г.): Труды и материалы / Составители М.Л. Ремнева,, А.А Поликарпов. – М.: Изд-во Моск. ун-та. 2010а
155. *Ягунова Е.В.* Эксперимент и вычисления в анализе ключевых слов художественного текста // Философия языка. Лингвистика. Лингводидактика №1 Пермь 2010б
156. *Ягунова Е.В.* Ключевые слова в исследовании текстов Н.В. Гоголя // Проблемы социо- и психолингвистики. Пермь, 2011
157. *Ягунова Е. В., Пивоварова Л. М.* От коллокаций к конструкциям // Русский язык: конструкционные и лексико-семантические подходы / Отв. ред. С.С.Сай. СПб, 2011 (ACTA LIINGUIISTIICA PETROPOLIITANA. Труды Института лингвистических исследований РАН. Отв. редактор / Н. Н. Казанский) (в печати)
158. *Ягунова Е.В.* Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь. 2008.
159. *Ягунова Е.В., Пивоварова Л.М.* Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Научно-техническая информация, Сер.2, №6. М. 2010а. с.30-40

160. Ягунова Е.В., Пивоварова Л.М.. Извлечение и классификация коллокаций на материале научных текстов. предварительные наблюдения // V Международная научно-практическая конференция "Прикладная лингвистика в науке и образовании" памяти Р.Г. *Пиотровского (1922-2009) : Материалы. СПб.* 2010б *С. 356-364*

161. *Ягунова Е.В., Пивоварова Л.М.* Исследование структуры новостного текста как последовательности связных сегментов // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. Вып. 10 (17).. М., 2011.

162. *Ягунова Е.В., Пивоварова Л.М., Клышинский Э.С.* Коммуникативная функция глаголов в газетных и научных текстах // Материалы конференции «Понимание в коммуникации-5». Сб. работ. – М. 2011 С. 206-209.

163. *Савина А.В., Ягунова Е.В.* Исследование коллокаций с помощью экспериментов с информантами // *Труды международной конференции "Корпусная лингвистика – 2011"* – СПб., 2011 (в печати)

164. *Янко Т. Е.* Коммуникативные стратегии русской речи – М., 2001

ЧАСТЬ II. КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: МЕТОДЫ, РЕСУРСЫ, ПРИЛОЖЕНИЯ *(БОЛЬШАКОВА Е.И.)*

Глава 1. Введение

Термин *компьютерная лингвистика* (КЛ) в последние годы все чаще встречается в связи с разработкой различных прикладных программных систем, в том числе – коммерческих программных продуктов. Связано это бурным ростом в обществе текстовой информации, в том числе в сети Интернет, и необходимостью автоматической обработки текстов на естественном языке (ЕЯ). Указанное обстоятельство стимулирует развитие компьютерной лингвистики как области науки и разработку новых информационных и лингвистических технологий.

В рамках компьютерной лингвистики, существующей уже более 50 лет (и известной также под названиями *машинная лингвистика*, *автоматическая обработка текстов на ЕЯ*) предложено много перспективных методов и идей, но далеко не все они еще нашли свое выражение в программных продуктах, используемых на практике. Наша цель – охарактеризовать специфику этой области исследований, сформулировать ее основные задачи, указать ее связи с другими науками, дать краткий обзор основных подходов и используемых ресурсов, а также кратко охарактеризовать существующие приложения КЛ. Для более подробного ознакомления с этими вопросам можно рекомендовать книги [4, 16, 37].

Глава 2. Задачи компьютерной лингвистики

Компьютерная лингвистика возникла на стыке таких наук, как лингвистика, математика, информатика (Computer Science) и искусственный интеллект. Истоки КЛ восходят к исследованиям известного американского ученого Н. Хомского в области формализации структуры естественного языка [7]; ее развитие опирается на результаты в области общей лингвистики (языкознания) [36]. Языкознание изучает общие законы естественного языка – его структуру и функционирование, и включает такие области:

- *Фонология* – изучает звуки речи и правила их соединения при формировании речи;
- *Морфология* – занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории;
- *Синтаксис* – изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка.
- *Семантика и прагматика* – тесно связанные области: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика – особенностями выражения этого смысла в связи с конкретными целями общения;
- *Лексикография* описывает лексикон конкретного ЕЯ – его отдельные слова и их грамматические свойства, а также методы создания словарей.

Результаты Н. Хомского, полученные на стыке лингвистики и математики, заложили основу для теории формальных языков и грамматик (часто называемых *генеративными*, или *порождающими* грамматиками). Эта теория относится ныне к *математической лингвистике* и применяется для обработки не столько ЕЯ, но искусственных языков, в первую очередь – языков программирования. По своему характеру это вполне математическая дисциплина.

К математической лингвистике относят также и *квантитативную лингвистику*, изучающую частотные характеристики языка – слов, их комбинаций, синтаксических конструкций и т.п., При этом используется математические методы статистики, так что можно назвать этот раздел науки статистической лингвистикой [14].

КЛ тесно связана и с такой междисциплинарной научной областью, как искусственный интеллект (ИИ) [40], в рамках которого разрабатываются компьютерные модели отдельных интеллектуальных функций. Одна из первых работающих программ в области ИИ и КЛ – это известная программа Т. Винограда, которая понимала простейшие приказы человека по изменению мира кубиков, сформулированные на ограниченном подмножестве ЕЯ [32]. Отметим, что несмотря на очевидное пересечение исследований в области КЛ и ИИ (поскольку владение языком относится к интеллектуальным функциям), ИИ не поглощает всю КЛ, поскольку она имеет свой теоретический базис и методологию. Общим для указанных наук является компьютерное моделирование как основной метод и итоговая цель исследований.

Таким образом, задача КЛ может быть сформулирована как разработка компьютерных программ для автоматической обработки текстов на ЕЯ. И хотя при этом обработка понимается достаточно широко, далеко не все виды обработки могут быть названы лингвистическими, а соответствующие процессоры – лингвистическими. *Лингвистический процессор* должен использовать ту или иную формальную модель языка (пусть даже очень простую), а значит, быть так или иначе языково-зависимым (т.е. зависеть от конкретного ЕЯ). Так, например, текстовый редактор Microsoft Word может быть назван лингвистическим (хотя бы потому, что использует словари), а редактор NotePad – нет.

Сложность задач КЛ связана с тем, что ЕЯ – сложная многоуровневая система знаков, возникшая для обмена информацией между людьми, выработанная в процессе практической деятельности человека, и постоянно изменяющаяся в связи с этой деятельностью [36, 38]. Другая сложность разработки методов КЛ (и сложность изучения ЕЯ в рамках языкознания) связана с многообразием естественных языков, существенными отличиями их лексики, морфологии, синтаксиса, разные языки предоставляют разные способы выражения одного и того же смысла.

Глава 3. Особенности системы ЕЯ: уровни и связи

Объектом лингвистических процессоров являются тексты ЕЯ. Под текстами понимаются любые образцы речи – устной и письменной, любого жанра, но в основном КЛ рассматривает письменные тексты. Текст имеет одномерную, линейную структуру, а также несет определенный смысл, язык же выступает как средство преобразования передаваемого смысла в тексты (синтез речи) и наоборот (анализ речи). Текст составлен из более мелких единиц, и возможно несколько способов разбиения (членения) текста на единицы, относящихся к разным уровням.

Общепризнано существование следующих уровней [36, 38]:

- уровень предложений (высказываний) – *синтаксический уровень*;
- уровень слов (словоформ – слов в определенной грамматической форме, например, столом, дружбы) – *морфологический уровень*;
- уровень фонем (отдельных звуков, с помощью которых формируются и различаются слова) – *фонологический уровень*.

Фонологический уровень выделяется для устной речи, для письменных текстов в языках с алфавитным способом записи (в частности, в европейских языках) он соответствует *уровню символов* (т.к. фонемы примерно соответствуют буквам алфавита).

Уровни, по сути, есть подсистемы общей системы ЕЯ (взаимосвязанные, но в достаточной степени автономные), и в них самих могут быть выделены подсистемы [36]. Так, морфологический уровень включает также *подуровень морфем*. *Морфема* – это минимальная значащая часть слова (корень, приставка, суффикс, окончание, постфикс).

Отметим, что единицы всех перечисленных уровней, кроме фонологического, являются знаками в смысле *семиотики* (общей науки о знаках), поскольку имеют значение (а отдельно взятая фонема или буква смысла не имеет). Иерархия уровней проявляется в том, что более высокий уровень в большой степени обуславливает организацию нижележащего уровня – так, синтаксическая структура предложения в значительной мере определяет, какие должны быть выбраны словоформы.

Вопрос о количестве уровней и их перечне до сих пор остается открытым в лингвистике. Как отдельный может быть выделен *лексический уровень* – уровень лексем. *Лексема* – это слово как совокупность всех его конкретных грамматических форм (к примеру, лексему *стол* образуют формы *стол*, *стола*, *столу*, *столом*). В тексте встречаются *словоформы* (лексемы в определенной форме), а в словаре ЕЯ – лексемы, точнее, в словаре записывается каноническая словоформа лексемы, называемая также *леммой* (например, для существительных это форма именительного падежа единственного числа: *стол*).

Относительно синтаксического уровня может быть выделен *подуровень словосочетаний* – синтаксически связанных групп слов (*купил книгу*, *новый год)*, и *надуровень сложного синтаксического целого*, которому примерно соответствует абзац текста. Сложное синтаксическое целое, или *сверхфразовое единство* – это последовательность предложений (высказываний), объединенных смыслом и лексико-грамматическими средствами [38]. К таким средствам относятся в первую очередь лексические повторы и *анафорические ссылки* – ссылки на предшествующие слова текста, реализуемые при помощи местоимений и местоименных слов (*они*, *этот* и т.д.).

Можно также говорить еще об одном уровне – *уровне дискурса*, под которым понимается связный текст в его коммуникативной направленности. Под дискурсом понимается последовательность взаимосвязанных друг с другом предложений текста, обладающая определенной смысловой целостностью, за счет чего он выполняет определенную прагматическую задачу [45]. Во многих типах связных текстов проявляется традиционная схематическая (*дискурсивная*) структура, организующая их общее содержание, например, определенную структуру имеют описания сложных технических систем, патентные формулы, научные статьи, деловые письма и др.

Отдельным является вопрос об *уровне семантики*. В принципе, она присутствует всюду, где есть знаковые единицы языка (морфемы, слова, предложения). Однако наличие именно уровня зависит от существования некоторого универсального набора семантических единиц, при помощи которых можно было бы выразить смысл любого высказывания. Подтверждением самостоятельности уровня семантики считается то, что человек обычно запоминает смысл высказывания, а не

его конкретную языковую форму. Элементарные единицы этого уровня называются *семами*, и в ряде исследований считается, что таких единиц в ЕЯ не более 2 тысяч.

Если сравнивать ЕЯ и искусственные языки, в частности, языки программирования, которые наиболее близки к ЕЯ по выполняемым лингвистическим функциям и успешно обрабатываются автоматически, то в первую очередь следует указать следующие их отличия, связанные с тем, что искусственные языки есть результат целенаправленной деятельности человека, а ЕЯ – продукт долгого исторического, и в определенной степени стихийного развития.

1) Открытость системы ЕЯ: язык постоянно изменяется (это не очень заметно в пределах нескольких лет, но ощутимо по прошествии одного-двух десятилетий). Изменения касаются не только словарного запаса языка (новые слова и новые смыслы старых), но также его синтаксического и фонетического уровней. Следствие открытости – принципиальная невозможность единожды описать конкретный ЕЯ и построить соответствующий лингвистический процессор. Необходимо пополнение знаний о языке на всех его уровнях, а, следовательно, КЛ должна разрабатывать средства автоматизации пополнения этих знаний.
2) Нестандартная сочетаемость (*синтактика*) единиц на каждом уровне ЕЯ. В частности, если в искусственных языках синтаксическая сочетаемость знаков диктуется их семантикой, то в ЕЯ соединение слов на уровне предложений лишь частично может быть описана законами грамматики. В любом языке достаточно большое количество грамматически правильных сочетаний реально не употребляется, например, в русском языке правильным сочетанием является *крепкий чай*, но не *тяжелый чай* (как в английском *heavy tea*). Тем самым, КЛ должна вырабатывать представления нестандартной сочетаемости единиц языка.
3) Б***о***льшая системность ЕЯ, т.е. в нем больше число уровней, четче границы между ними, а также более выражена *ассиметрия* связи между единицами языка и выражаемыми ими смыслами, проявляющаяся на всех уровнях языковой системы. Под ассиметрией понимаются нарушения регулярности этих связей, что выражается в таких явлениях как п*олисемия* (многозначность) – наличие у одной единицы языка нескольких связанных между собой значений (например, полисемия слов, например: *земля* – суша, почва, конкретная планета); с*инонимия* – полное или частичное совпадение значений разных единиц (например, синонимия слов: *негодяй* и *подлец*), *омонимия* – совпадение по форме двух разных по смыслу единиц. Таким образом, КЛ должна иметь средства решения проблем неоднозначности, связанной с этими явлениями.

Добавим, что омонимия существенно проявляется на всех уровнях ЕЯ, укажем некоторые ее виды:

- *Лексическая омонимия* означает одинаково звучащие и пишущиеся слова, не имеющие общих элементов смысла, например, *рожа* – лицо и вид болезни.
- *Морфологическая омонимия* – совпадение форм одного и того же слова (лексемы), например, словоформа *круг* соответствует именительному и винительному падежам.
- *Лексико-морфологическая* омонимия (наиболее частый вид) возникает при совпадении словоформ двух разных лексем, например, *стих* – глагол в единственном числе мужского рода и существительное в единственном числе, именительном падеже),

- *Синтаксическая омонимия* означает неоднозначность синтаксической структуры, что приводит к нескольким интерпретациям: *Студенты из Львова поехали в Киев, Flying planes can be dangerous* (известный пример Хомского) и др.

Глава 4. Моделирование в компьютерной лингвистике

Разработка лингвистического процессора (ЛП) предполагает описание лингвистических свойств обрабатываемого текста ЕЯ, и это описание организуется как *модель языка*. Как и при моделировании в математике и программировании, под моделью понимается некоторая система, отображающая ряд существенных свойств моделируемого явления (т.е. ЕЯ) и обладающая поэтому структурным или функциональным подобием.

Используемые в КЛ модели языка обычно строятся на основе теорий, создаваемых лингвистами путем изучения различных текстов и на основе своей лингвистической интуиции (интроспекции). В чем же специфика именно моделей КЛ? Можно выделить следующие их особенности [4]:

- Формальность и, в конечном счете, алгоритмизируемость;
- Функциональность (цель моделирования – воспроизведение функций языка как «черного ящика», без построения точной модели синтеза и анализа речи человеком);
- Общность модели, т.е. учет ею довольно большого множества текстов;
- Экспериментальная обоснованность, предполагающая тестирование модели на разных текстах;
- Опора на словари как обязательную составляющую модели.

Сложность ЕЯ, его описания и обработки ведет к разбиению этого процесса на отдельные этапы, соответствующие уровням языка, Большинство современных ЛП относятся к модульному типу, при котором каждому уровню лингвистического анализа или синтеза соответствует отдельный модуль процессора. В частности, в случае анализа текста отдельные модули ЛП выполняют:

- Графематический анализ, т.е. выделение в тексте словоформ (переход от символов к словам);
- Морфологический анализ – переход от словоформ к их *леммам* (словарным формам лексем) или *основам* (ядерным частям слова, за вычетом словоизменительных морфем);
- Синтаксический анализ, т.е. выявление грамматической структуры предложений текста;
- Семантический и прагматический анализ, при котором определяется смысл фраз и соответствующая реакция системы, в рамках которой работает ЛП.

Возможны разные схемы взаимодействия указанных модулей (последовательная работа или параллельный перемежающийся анализ), однако отдельные уровни – морфология, синтаксис и семантика все же обрабатываются разными механизмами.

Таким образом, ЛП можно рассматривать как многоэтапный преобразователь, переводящий в случае анализа текста каждое его предложение во внутреннее представление его смысла и наоборот в случае синтеза. Соответствующая модель языка может называться *структурной*.

Хотя полные модели КЛ требуют учета всех основных уровней языка и наличия соответствующих модулей, при решении некоторых прикладных задач можно обойтись без представления в ЛП отдельных уровней. К примеру, в ранних

экспериментальных программах КЛ, обрабатываемые тексты относились к очень узким проблемным областям (с ограниченным набором слов и строгим их порядком), так что для распознавания слов можно было использовать их начальные буквы, опуская этапы морфологического и синтаксического анализа.

Еще одним примером редуцированной модели, ныне достаточно часто используемой, является языковая модель частотности символов и их сочетаний (биграмм, триграмм и пр.) в текстах конкретного ЕЯ [19]. Такая *статистическая модель* отображает лингвистическую информацию на уровне символов (букв) текста, и ее достаточно, например, для выявления опечаток в тексте или для распознавания его языковой принадлежности. Аналогичная модель на базе статистики отдельных слов и их совместной встречаемости в текстах (биграмм, триграмм слов) применяется, например, для разрешения лексической неоднозначности [18] или определения части речи слова (в языках типа английского).

Отметим, что возможны *структурно-статистические модели*, в которых при представлении отдельных уровней ЕЯ учитывается та или иная статистика – слов, синтаксических конструкций и т.п.

В ЛП модульного типа на каждом этапе анализа или синтеза текста используется соответствующая модель (морфологии, синтасиса и т.п.).

Существующие в КЛ морфологические модели анализа словоформ различаются в основном по следующим параметрам:

- результату работы – лемма или основа с набором морфологических характеристик (род, число, падеж, вид, лицо и т.п.) заданной словоформы;
- методу анализа – с опорой на словарь словоформ языка или на словарь основ, либо же бессловарный метод;
- возможности обработки словоформы лексемы, не включенной в словарь.

При морфологическом синтезе исходными данными являются лексема и конкретные морфологические характеристики запрашиваемой словоформы данной лексемы, возможен и запрос на синтез всех форм заданной лексемы. Результат как морфологического анализа, так и синтеза в общем случае неоднозначен.

Для моделирования синтаксиса в рамках КЛ предложено большое число разных идей и методов, отличающихся способом описания синтаксиса языка, способом использования этой информации при анализе или синтезе предложения ЕЯ, а также способом представления синтаксической структуры предложения [6]. Весьма условно можно выделить три основных подхода к созданию моделей: генеративный подход, восходящий к идеям Хомского [7], подход, восходящий к идеям И. Мельчука и представленный моделью «Смысл⇔Текст» [42], а также подход, в рамках которого делаются те или иные попытки преодолеть ограничения первых двух подходов, в частности, теория синтаксических групп [33].

В рамках генеративного подхода синтаксический анализ производится, как правило, на основе формальной контекстно-свободной грамматики, описывающей фразовую структуру предложения, или же на основе некоторого расширения контекстно-свободной грамматики. Эти грамматики исходят из последовательного линейного членения предложения на фразы (синтаксические конструкции, например, именные группы) и отражают поэтому одновременно как его синтаксическую, так и линейную структуры. Полученная в результате анализа иерархическая синтаксическая структура предложения ЕЯ описывается *деревом составляющих*, в листьях которого находятся слова предложения, поддеревья соответствуют

входящим в предложение синтаксическим конструкциям (фразам), а дуги выражают отношения вложения конструкций.

К рассматриваемому подходу могут быть отнесены сетевые грамматики, представляющие собой одновременно аппарат для описания системы языка и для задания процедуры анализа предложений на основе понятия конечного автомата, например, расширенная сеть переходов ATN [23].

В рамках второго подхода для представления синтаксической структуры предложения используется более наглядный и распространенный способ – *деревья зависимостей*. В узлах дерева расположены слова предложения (в корне обычно глагол-сказуемое), а каждая дуга дерева, связывающая пару узлов, интерпретируется как синтаксическая *подчинительная связь* между ними, причем направление связи соответствует направлению данной дуги. Поскольку при этом синтаксические связи слов и порядок слов в предложении отделены, то на основе деревьев подчинения могут быть описаны разорванные и *непроективные* конструкции [36], достаточно часто возникающие в языках со свободным порядком слов.

Деревья составляющих больше подходят для описания языков в жестким порядком слов, представление с их помощью разорванных и непроективных конструкций требует расширения используемого грамматического формализма. Зато в рамках этого подхода более естественно описываются конструкции с неподчинительными отношениями. В то же время общая трудность для обоих подходов – представление однородных членов предложения.

Синтаксические модели во всех подходах пытаются учесть ограничения, накладываемые на соединение языковых единиц в речи, при этом так или иначе используется понятие валентности [38]. *Валентность* – это способность слова или другой единицы языка присоединять другие единицы определенным синтаксическим способом; *актант* – это слово или синтаксическая конструкция, заполняющая эту валентность. Например, русский глагол *передать* имеет три основные валентности, которые можно выразить следующими вопросительными словами: *кто? кому? что?* В рамках генеративного подхода валентности слов (прежде всего, глаголов) описываются преимущественно в виде специальных фреймов (*subcategorization frames*) [4], а в рамках подхода, основанного на деревьях зависимостей – как *модели управления*.

Модели семантики языка наименее проработаны в рамках КЛ. Для семантического анализа предложений были предложены так называемые падежные грамматики и *семантические падежи* (валентности), на базе которых семантика предложения описывается как через связи главного слова (глагола) с его семантическими актантами, т.е. через семантические падежи [4]. Например, глагол *передать* описывается семантическими падежами *дающего* (агенса), *адресата* и *объекта передачи*.

Для представления семантики всего текста обычно используются два логически эквивалентных формализма (оба они детально описаны в рамках ИИ [40]):

- Формулы исчисления предикатов, выражающих свойства, состояния, процессы, действия и отношения;
- Семантические сети – размеченные графы, в которых вершины соответствуют понятиям, а вершины – отношениям между ними.

Что касается моделей прагматики и дискурса, позволяющих обрабатывать не только отдельные предложения, но и текст в целом, то в основном для их построения

используются идеи Ван Дейка [30]. Одна из редких и удачных моделей – модель дискурсивного синтеза связных текстов [41]. В подобных моделях должны учитываться анафорические ссылки и другие явления уровня дискурса.

Завершая характеристику моделей языка в рамках КЛ, остановимся чуть подробнее на теории лингвистических моделей «Смысл⇔Текст» [42], и в рамках которой появилось много плодотворных идей, опередивших свое время и актуальных до сих пор.

В соответствии с этой теорией ЕЯ рассматривается как особого рода преобразователь, выполняющий переработку заданных смыслов в соответствующие им тексты и заданных текстов в соответствующие им смыслы. Под смыслом понимается инвариант всех синонимичных преобразований текста. Содержание связного фрагмента речи без расчленения на фразы и словоформы отображается в виде специального семантического представления, состоящего из двух компонент: *семантического графа* и сведений о *коммуникативной организации смысла*.

Как отличительные особенности теории следует указать:

- ориентацию на синтез текстов (способность порождать правильные тексты рассматривается как основной критерий языковой компетенции);
- многоуровневый, модульный характер модели, причем основные уровни языка разделяются на поверхностный и глубинный уровень: различаются, к примеру, *глубинный* (семантизированный) и *поверхностный* («чистый») синтаксис, а также поверхностно-морфологический и глубинно-морфологический уровни;
- интегральный характер модели языка; сохранение информации, представленной на каждом уровне, соответствующим модулем, выполняющими переход с этого уровня на следующий;
- специальные средства описания синтактики (правил соединения единиц) на каждом из уровней; для описания лексической сочетаемости был предложен набор *лексических функций*, при помощи которых сформулированы правила синтаксического перифразирования;
- упор на словарь, а не на грамматику; в словаре хранится информация, относящаяся к разным уровням языка; в частности, для синтаксического анализа используются модели управления слов, описывающие их синтаксические и семантические валентности.

Эта теория и модель языка нашли свое воплощение в системе машинного перевода ЭТАП [26].

Глава 5. Лингвистические ресурсы

Разработка лингвистических процессоров требует соответствующего представления лингвистической информации об обрабатываемом ЕЯ. Эта информация отображается в разнообразных компьютерных словарях и грамматиках.

Словари являются наиболее традиционной формой представления лексической информации; они различаются своими единицами (обычно слова или словосочетания), структурой, охватом лексики (словари терминов конкретной проблемной области, словари общей лексики и т.п.). Единица словаря называется *словарной статьей*, в ней представляется информация о лексеме. Лексические омонимы обычно представляются в разных словарных статьях.

Наиболее распространены в КЛ морфологические словари, используемые для морфологического анализа, в их словарной статье представлена морфологическая

информация о соответствующем слове – часть речи, словоизменительный класс (для флективных языков), перечень значений слова и т.п. В зависимости от организации лингвистического процессора в словарь может быть добавлена и грамматическая информация, например, модели управления слова.

Существуют словари, в которых представлена и более широкая информация о словах. Например, лингвистическая модель «Смысл⇔Текст» существенно опирается на *толково-комбинаторный словарь*, в словарной статье которого помимо морфологической, синтаксической и семантической информации (синтаксические и семантические валентности) представлены сведения о лексической сочетаемости этого слова.

В ряде лингвистических процессоров используются **словари синонимов**. Сравнительно новый вид словарей – **словари паронимов**, т.е. внешне схожих слов, различающихся по смыслу, например, *чужой* и *чуждый*, *правка* и *справка* [34].

Еще один вид лексических ресурсов – **базы словосочетаний**, в которые отбираются наиболее типичные словосочетания конкретного языка. Такая база словосочетаний русского языка (около миллиона единиц) составляет ядро системы КроссЛексика [28].

Более сложными видами лексических ресурсов являются **тезаурусы и онтологии**. Тезаурус – это семантический словарь, т.е. словарь, в котором представлены смысловые связи слов – синонимические, отношения род-вид (иногда называемые отношением выше-ниже), часть-целое, ассоциации. Распространение тезаурусов связано с решением задач информационного поиска [39].

С понятием тезауруса тесно связано понятие онтологии [11]. Онтология – набор понятий, сущностей определенной области знаний, ориентированный на многократное использование для различных задач.Онтологии могут создаваться на базе существующей в языке лексики – в этом случае они называются *лингвистическим*и.

Подобной лингвистической онтологией считается система WordNet [24] – большой лексический ресурс, в котором собраны слова английского языка: существительные, прилагательные, глаголы и наречия, и представлены их смысловые связи нескольких типов. Для каждой из указанных частей речи слова сгруппированы в группы синонимов (*синсеты*), между которыми установлены отношения антонимии, гипонимии (отношение род-вид), меронимии (отношение часть-целое). Ресурс содержит примерно 25 тыс. слов, число уровней иерархии для отношения род-вид в среднем равно 6-7, достигая порою 15. Верхний уровень иерархии формирует общую онтологию – систему основных понятий о мире.

По схеме английского WordNet были построены аналогичные лексические ресурсы для других европейских языков, объединенные под общим названием EuroWordNet.

Совершенно другой вид лингвистических ресурсов – это **грамматики ЕЯ**, тип которых зависит от используемой в процессоре модели синтаксиса. В первом приближении грамматика представляет собой набор правил, выражающих общие синтаксические свойства слов и групп слов. Общее число правил грамматики также зависит от модели синтаксиса, изменяясь от нескольких десятков до нескольких сотен. По существу, здесь проявляется такая проблема, как соотношение в модели языка грамматики и лексики: чем больше информации представлено в словаре, тем короче может быть грамматика и наоборот.

Отметим, что построение компьютерных словарей, тезаурусов и грамматик – объемная и трудоемкая работа, иногда даже более трудоемкая, чем разработка лингвистической модели и соответствующего процессора. Поэтому одной из подчиненных задач КЛ является автоматизация построения лингвистических ресурсов [10, 15].

Компьютерные словари часто формируются конвертацией обычных текстовых словарей, однако нередко для их построения требуется гораздо более сложная и кропотливая работа. Обычно это бывает при построении словарей и тезаурусов для быстро развивающися научных областей – молекулярной биологии, информатики и др. Исходным материалом для извлечения необходимой лингвистической информации могут быть **коллекции и корпуса текстов**.

Корпус текстов – это коллекция текстов, собранная по определенному принципу представительности (по жанру, авторской принадлежности и т.п.), в которой все тексты размечены, т.е. снабжены некоторой лингвистической разметкой (аннотациями) – морфологической, акцентной, синтаксической и т.п. [3].В настоящее время существует не менее сотни различных корпусов – для разных ЕЯ и с различной разметкой, в России наиболее известным является Национальный корпус русского языка [43].

Размеченные корпуса создаются лингвистами и используются как для лингвистических исследований, так и для настройки (обучения) используемых в КЛ моделей и процессоров с помощью известных математических методов машинного обучения. Так, машинное обучение применяется для настройки методов разрешения лексической неоднозначности, распознавания части речи, разрешения анафорических ссылок.

Поскольку корпуса и коллекции текстов всегда ограничены по представленным в них языковым явлениям (а корпуса, ко всему прочему, создаются довольно долго), в последнее время все чаще в качестве более полного лингвистического ресурса рассматриваются тексты сети Интернет [13, 35]. Безусловно, Интернет является самым представительным источником образцов современной речи, однако его использование как корпуса требует разработки специальных технологий.

Глава 6. Приложения компьютерной лингвистики

Область приложений компьютерной лингвистики постоянно расширяется, так что охарактеризуем здесь наиболее известные прикладные задачи, решаемые ее инструментами.

Машинный перевод [21] – самое раннее приложение КЛ, вместе с которым возникла и развивалась сама эта область. Первые программы перевода были построены более 50 лет назад и были основаны на простейшей стратегии пословного перевода. Однако довольно быстро было осознано, что машинный перевод требует полной лингвистической модели, учитывающей все уровни языка, вплоть до семантики и прагматики, что неоднократно тормозило развитие этого направления. Достаточно полная модель использована в отечественной системе ЭТАП [26], выполняющей перевод научных текстов с французского на русский язык.

Заметим, однако, что в случае перевода на родственный язык, например, при переводе с испанского на португальский или же с русского на украинский (у которых много общего в синтаксисе и морфологии), процессор может быть реализован на

основе упрощенной модели, например, на основе все той же стратегией пословного перевода.

В настоящее время существует целый спектр компьютерных систем перевода (разного качества), от больших интернациональных исследовательских проектов до коммерческих автоматических переводчиков. Существенный интерес представляют проекты многоязыкового перевода, с использованием промежуточного языка, на котором кодируется смысл переводимых фраз. Другое современное направление – статистическая трансляция [5], опирающаяся на статистику перевода слов и словосочетаний (эти идеи, к примеру, реализованы в переводчике поисковика Google).

Но несмотря на многие десятилетия развития всего этого направления, в целом задача машинного перевода еще весьма далека до полного решения.

Еще одно довольно старое приложение компьютерной лингвистики – это **информационный поиск** и связанные с ним задачи индексирования, реферирования, классификации и рубрикации документов [1, 20, 22].

Полнотекстовый поиск документов в больших базах документов (в первую очередь – научно-технических, деловых), проводится обычно на основе их *поисковых образов*, под которыми понимается набор *ключевых слов* – слов, отражающих основную тему документа. Сначала в качестве ключевых слов рассматривались только отдельные слова ЕЯ, а поиск производился без учета их словоизменения, что некритично для слабофлективных языков типа английском. Для флективных языков, например, для русского потребовалось использование морфологической модели, учитывающей словоизменение.

Запрос на поиск также представлялся в виде набора слов, подходящие (релевантные) документы определялись на основе похожести запроса и поискового образа документа. Создание поискового образа документа предполагает **индексирование** его текста, т.е. выделение в нем ключевых слов [12] . Поскольку очень часто гораздо точнее тему и содержание документа отображают не отдельные слова, а словосочетания, в качестве ключевых слов стали рассматриваться словосочетания. Это существенно усложнило процедуру индексирования документов, поскольку для отбора значимых словосочетаний текста потребовалось использовать различные комбинации статистических и лингвистических критериев.

По сути, в информационном поиске в основном используется *векторная модель текста* (называемая иногда *bag of words* – мешок слов), при которой документ представляется вектором (набором) своих ключевых слов. Современные интернет-поисковики также используют эту модель, выполняя индексирование текстов по употребляемым в них словам (в то же время для выдачи релевантных документов они используют весьма изощренные процедуры ранжирования).

Указанная модель текста (с некоторыми усложнениями) применяется и в рассматриваемых ниже смежных задачах информационного поиска.

Реферирование текста – сокращение его объема и получение его краткого изложения – реферата (свернутого содержания), что делает более быстрым поиск в коллекциях документов. Общий реферат может составляться также для нескольких близких по теме документов.

Основным методом автоматического реферирования до сих пор является отбор наиболее значимых предложений реферируемого текста, для чего обычно сначала вычисляются ключевые слова текста и рассчитывается коэффициент значимости

предложений текста. Выбор значимых предложений осложняется анафорическими связями предложений, разрыв которых нежелателен – для решения этой проблемы разрабатываются определенные стратегии отбора предложений.

Близкая к реферированию задача – **аннотирование** текста документа, т.е. составление его аннотации. В простейшей форме аннотация представляет собой перечень основных тем текста, для выделения которых могут использоваться процедуры индексирования.

При создании больших коллекций документов актуальны задачи **классификации** и **кластеризации** текстов с целью создания классов близких по теме документов [31]. Классификация означает отнесение каждого документа к определенному классу с заранее известными параметрами, а кластеризация – разбиение множества документов на кластеры, т.е. подмножества тематически близких документов. Для решения этих задач применяются методы машинного обучения, в связи с чем эти прикладные задачи называют Text Mining и относят к научному направлению, известному как Data Mining, или интеллектуальный анализ данных [27].

Очень близка к классификации задача **рубрицирования** текста – его отнесение к одной из заранее известных тематических рубрик (обычно рубрики образуют иерархическое дерево тематик).

Задача классификации получает все большее распространение, она решается, например, при распознавании спама, а сравнительно новое приложение – классификация SMS-сообщений в мобильных устройствах. Новое и актуальное направление исследований для общей задачи информационного поиска – многоязыковой поиск по документам.

Еще одна относительно новая задача, связанная с информационным поиском – **формирование ответов на вопросы** (Question Answering) [9]. Эта задача решается путем определения типа вопроса, поиском текстов, потенциально содержащих ответ на этот вопрос, и извлечением ответа из этих текстов.

Совершенно иное прикладное направление, которое развивается хотя и медленно, но устойчиво – это **автоматизация подготовки и редактирования** текстов на ЕЯ. Одним из первых приложений в этом направлении были программы автоматической определения переносов слов и программы орфографической проверки текста (спеллеры, или автокорректоры). Несмотря на кажущуюся простоту задачи переносов, ее корректное решение для многих ЕЯ (например, английского) требует знания морфемной структуры слов соответствующего языка, а значит, соответствующего словаря.

Проверка орфографии уже давно реализована в коммерческих системах и опирается на соответствующий словарь и модель морфологии. Используется также неполная модель синтаксиса, на основе которой выявляются достаточно частотные все синтаксические ошибки (например, ошибки согласования слов). В то же время в автокорректорах не реализовано пока выявление более сложных ошибок, к примеру, неправильное употребление предлогов. Не обнаруживаются и многие лексические ошибки, в частности, ошибки, возникающие в результате опечаток или неверного использования схожих слов (например, *весовой* вместо весомый). В современных исследованиях КЛ предлагаются методы автоматизированного выявления и исправления подобных ошибок, а также некоторых других видов стилистических

ошибок [25, 29]. В этих методах используется статистика встречаемости слов и словосочетаний.

Близкой к поддержке подготовки текстов прикладной задачей является **обучение естественному языку**, в рамках этого направления часто разрабатываются компьютерные системы обучения языку – английскому, русскому и др. (подобные системы можно найти в Интернете). Обычно эти системы поддерживают изучение отдельных аспектов языка (морфологии, лексики, синтаксиса) и опираются на соответствующие модели, например, модель морфологии.

Что касается изучения лексики, то для этого также используются электронные аналоги текстовых словарей (в которых по сути нет языковых моделей). Однако разрабатываются также многофукциональные компьютерные словари, не имеющие текстовых аналогов и ориентированные на широкий круг пользователей – например, словарь русских словосочетаний Кросслексика [28]. Эта система охватывает широкий круг лексики – слов и допустимых их словосочетаний, а также предоставляет справки по моделям управления слов, синонимам, антонимам и другим смысловым коррелятам слов, что явно полезно не только для тех, кто изучает русский язык, но и носителям языка.

Следующее прикладное направление, которое стоит упомянуть – это **автоматическая генерация** текстов на ЕЯ [2]. В принципе, эту задачу можно считать подзадачей уже рассмотренной выше задачи машинного перевода, однако в рамках направления есть ряд специфических задач. Такой задачей является многоязыковая генерация, т.е. автоматическое построение на нескольких языках специальных документов – патентных формул, инструкций по эксплуатации технических изделий или программных систем, исходя из их спецификации на формальном языке. Для решения этой задачи применяются довольно подробные модели языка.

Все более актуальная прикладная задача, часто относимая к направлению Text Mining – это **извлечение информации** из текстов, или Information Extraction [8], что требуется при решении задач экономической и производственной аналитики. Для этого осуществляется выделение в тесте ЕЯ определенных объектов – именованных сущностей (имен, персоналий, географических названий), их отношений и связанных с ними событий. Как правило, это реализуется на основе частичного синтаксического анализа текста, позволяющего выполнять обработку потоков новостей от информационных агентств. Поскольку задача достаточно сложна не только теоретически, но и технологически, создание значимых систем извлечения информации из текстов осуществимо в рамках коммерческих компаний [44].

К направлению Text Mining относятся и две другие близкие задачи – выделение мнений (Opinion Mining) и оценка тональности текстов (Sentiment Analysis), привлекающие внимание все большего числа исследователей. В первой задаче происходит поиск (в блогах, форумах, интернет-магазинах и пр.) мнений пользователей о товарах и других объектах, а также производится анализ этих мнений. Вторая задача близка к классической задаче контент-анализа текстов массовой коммуникации, в ней оценивается общая тональность высказываний.

Еще одно приложение, которое стоит упомянуть – **поддержка диалога** с пользователем на ЕЯ в рамках какой-либо информационной программной системы. Наиболее часто эта задача решалась для специализированных баз данных – в этом случае язык запросов достаточно ограничен (лексически и грамматически), что позволяет использовать упрощенные модели языка. Запросы к базе,

сформулированные на ЕЯ, переводятся на формальный язык, после чего выполняется поиск нужной информации и строится соответствующая фраза ответа.

В качестве последнего в нашем перечне приложений КЛ (но не по важности) укажем **распознавание и синтез звучащей речи**. Неизбежно возникающие в этих задачах ошибки распознавания исправляются автоматическими методами на основе словарей и лингвистических знаний о морфологии. В этой области также применятся машинное обучение.

Глава 7. Заключение

Компьютерная лингвистика демонстрирует вполне осязаемые результаты в различных приложениях по автоматической обработке текстов на ЕЯ. Дальнейшее ее развитие зависит как от появления новых приложений, так и независимой разработки различных моделей языка, в которых пока не решены многие проблемы. Наиболее проработанными являются модели морфологического анализа и синтеза. Модели синтаксиса еще не доведены до уровня устойчиво и эффективно работающих модулей, несмотря на большое число предложенных формализмов и методов. Еще менее изучены и формализованы модели уровня семантики и прагматики, хотя автоматическая обработка дискурса уже требуется в ряде приложений. Отметим, что уже существующие инструменты самой компьютерной лингвистики, использование машинного обучения и корпусов текстов, может существенно продвинуть решение этих проблем.

Список использованной литературы

1. Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, Adison Wesley, 1999.
2. Bateman, J., Zock M. Natural Language Generation. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p.304.
3. Biber, D., Conrad S., and Reppen D. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge, 1998.
4. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
5. Brown P., Pietra S., Mercer R., Pietra V. The Mathematics of Statistical Machine Translation. // Computational Linguistics, Vol. 19(2): 263-311. 1993.
6. Carroll J R. Parsing. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 233-248.
7. Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957.
8. Grishman R. Information extraction. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 545-559.
9. Harabagiu, S., Moldovan D. Question Answering. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 560-582.
10. Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998, p.131-151.
11. Hirst, G. Ontology and the Lexicon. In.: Handbook on Ontologies in Niformation Systems. Berlin, Springer, 2003.
12. Jacquemin C., Bourigault D. Term extraction and automatic indexing // Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 599-615.

13. Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. Computational linguistics, V. 29, No. 3, 2003, p. 333-347.
14. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
15. Matsumoto Y. Lexical Knowledge Acquisition. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 395-413.
16. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.
17. Oakes, M., Paice C. D. Term extraction for automatic abstracting. Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M. L'Homme (Eds), John Benjamins Publishing Company, Amsterdam, 2001, p.353-370.
18. Pedersen, T. A decision tree of bigrams is an accurate predictor of word senses. Proc. 2nd Annual Meeting of NAC ACL, Pittsburgh, PA, 2001, p. 79-86.
19. Samuelsson C. Statistical Methods. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 358-375.
20. Salton, G. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA: Addison-Wesley, 1988.
21. Somers, H. Machine Translation: Latest Developments. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 512-528.
22. Strzalkowski, T. (ed.) Natural Language Information Retrieval. Kluwer,1999. 385 p.
23. Woods W.A. Transition Network Grammers forNatural language Analysis/ Communications of the ACM, V. 13, 1970, N 10, p. 591-606.
24. Word Net: an Electronic Lexical Database. /Edit. by Christiane Fellbaum. Cambridge, MIT Press, 1998.
25. Wu J., Yu-Chia Chang Y., Teruko Mitamura T., Chang J. Automatic Collocation Suggestion in Academic Writing // Proceedings of the ACL 2010 Conference Short Papers, 2010.
26. Апресян Ю.Д. и др. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.
27. Барсегян А.А. и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – 2-е изд. – СПб.: БХВ-Петербург, 2008.
28. Большаков, И.А. КроссЛексика – большой электронный словарь сочетаний и смысловых связей русских слов. // Комп. лингвистика и интеллект. технологии: Труды межд. Конф. «Диалог 2009». Вып. 8 (15) М.: РГГУ, 2009, с.. 45-50.
29. Большакова Е.И., Большаков И.А. Автоматическое обнаружение и автоматизированное исправление русских малапропизмов // НТИ. Сер. 2, № 5, 2007, с.27-40.
30. Ван Дейк Т.А., Кинч В. Стратегия понимания связного текста.// Новое в зарубежной лингвистике. Вып. XXIII– М., Прогресс, 1988, с. 153-211.
31. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
32. Виноград Т. Программа, понимающая естественный язык – М., мир, 1976.
33. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М., Наука, 1985.
34. Гусев, В.Д., Саломатина Н.В. Электронный словарь паронимов: версия 2. // НТИ, Сер. 2, № 7, 2001, с. 26-33.
35. Захаров В.П. Веб-пространство как языковой корпус// Компьютерная лингвистика

и интеллектуальные технологии: Труды Межд. конференции Диалог ‘2005 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П.Селегея – М.: Наука, 2005, с. 166-171.
36. Касевич В.Б. Элементы общей лингвистики. — М., Наука, 1977.
37. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
38. Лингвистический энциклопедический словарь /Под ред. В. Н. Ярцевой, М.: Советская энциклопедия, 1990, 685 с.
39. Лукашевич Н.В., Салий А.Д. Тезаурус для автоматического индексирования и рубрицирования: разработка, структура, ведение. // НТИ, Сер. 2, №1, 1996.
40. Люгер Дж. Искусственный интеллект: стратегии и методы решения сложных проблем. М., 2005.
41. Маккьюин К. Дискурсивные стратегии для синтеза текста на естественном языке // Новое в зарубежной лингвистике. Вып. XXIV. М.: Прогресс, 1989, с.311-356.
42. Мельчук И.А. Опыт теории лингвистических моделей «СМЫСЛ ↔ ТЕКСТ». — М., Наука, 1974.
43. Национальный Корпус Русского Языка. http://ruscorpora.ru
44. Хорошевский В.Ф. OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Т. 2. – М.: Физматлит, 2004, с.573-581.
45. Шевченко Н.В. Основы лингвистики текста: учебное пособие – М.: Приор-издат, 2003.

ЧАСТЬ III. НАЧАЛЬНЫЕ ЭТАПЫ АНАЛИЗА ТЕКСТА *(КЛЫШИНСКИЙ Э.С.)*

Глава 1. Этапы анализа текста

Самые большие возможности и высокое качество анализа текстов можно получить, проведя его полный анализ. Однако сложности, возникающие при создании подобного анализа таковы, что на практике до сих пор не реализованы все теоретические положения, разработанные на данный момент. Основными проблемами здесь являются сложность синтаксического анализа текста и сложность создания полноценной экспертной системы, реализующей полноценную модель окружающего мира. Сложность анализа текста заключается в том, что текст эллиптичен, неполон и насквозь пронизан умолчаниями. Ярким примером может служить китайский театр, в котором человек, который при ходьбе выбрасывает в стороны несгибающиеся ноги и поглаживает бороду, воспринимается как положительный гражданский герой, тогда как «обольстительная красавица» должна переступать плотно сжав колени. Аналогично и в тексте встречаются конструкции, предназначенные скорее для живого воображения, чем для формальной обработки: «давить мух», «сделать ноги». Анализ подобных текстов может составить серьезную проблему не только для ЭВМ, но и для человека, так как большинство ситуаций имело под собой какую-то реальную или вымышленную основу и вставка их в текст служит как бы ссылкой на такую ситуацию (хотя зачастую большинство может уже и не помнить о чем идет речь, а просто восстанавливает истинный смысл фразы).

Для полноценной работы система анализа текста должна иметь возможность проанализировать тест, поданный пользователем на вход, с точки зрения синтаксиса (структуры предложений), семантики (понятий, применяемых в тексте) и прагматики (правильности употребления понятий и целей их употребления). Далее система должна сгенерировать свой отклик во внутреннем представлении, пригодном для логического вывода, и просинтезировать свой отклик на естественном языке.

В целом система, поддерживающая полный анализ, должна содержать в себе следующие модули.

Графематический анализ – обеспечивает выделение синтаксических или структурных единиц из входного текста, который может представлять собой линейную структуру, содержащую единый фрагмент текста. Однако в более общем случае текст может состоять из многих структурных единиц: основного текста, заголовков, вставок, врезок, комментариев и т.д. При машинном переводе ставится задача сохранить подобную структуру текста. Однако в случае диалоговых систем обычно используется первый вариант (без вставок). Но и в этом случае графематический анализ должен выделять синтаксические единицы: абзацы, предложения, отдельные слова и знаки препинания. В ряде случаев здесь же проводится предморфологический анализ – объединение неразрывных неизменяемых словосочетаний в одну единицу: «_что_-_то_», «_таким_образом_», «_и_так_далее_», … .

Морфологический анализ – обеспечивает определение нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе. Это делается для того, чтобы ориентироваться в дальнейшем только на нормальную форму, а не на все словоформы, использовать параметры, например, для проверки согласования слов.

Предсинтаксический анализ отвечает за две противоположные задачи: объединение отдельных лексических единиц в одну синтаксическую или, наоборот, ее разделение на несколько. В одну синтаксическую единицу объединяются изменяемые неразрывные словосочетания (например, «бить баклуши»). Делением слов особенно необходимо заниматься, например, в немецком языке, где несколько произвольных связанных между собой слов могут объединяться в одно сложное «на лету», а помещать в морфологический анализ все подобные сочетания не представляется возможным. Еще одной задачей предсинтаксического анализа является проведение синтаксической сегментации. Её задачей является разметка линейного текста на фрагменты, привязанные правилам следующего этапа – синтаксического анализа, который является задачей с экспоненциальным ростом сложности. В связи с этим любая помощь при его проведении может привести к существенному ускорению его работы.

Синтаксический анализ – самая сложная часть анализа текста. Здесь необходимо определить роли слов и их связи между собой. Результатом этого этапа является набор деревьев, показывающих такие связи. Выполнение задачи осложняется огромным количеством альтернативных вариантов, возникающих в ходе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от различных нормальных форм), так и неоднозначностью самих правил разбора.

Постсинтасический анализ служит двум целям. С одной стороны нам необходимо уточнить смысл, заложенный в слова и выраженный при помощи различных средств языка: предлогов, префиксов или аффиксов, создающих ту или иную словоформу. С другой стороны, одна и та же мысль может быть выражена различными конструкциями языка. В случае с многоязыковой диалоговой системой, одну и ту же мысль можно выразить различными синтаксическими конструкциями. В связи с этим дерево необходимо нормализовать, т.е. конструкция, выражающая некоторое действие различным образом для различных языков или ситуаций, должна быть сведена к одному и тому же нормализованному дереву. Кроме того, на этом же этапе может проводиться обработка разрывных изменяемых словосочетаний, в которых слова словосочетания могут изменяться и могут быть разделены другими словами («белый офицер» vs «белый корниловский офицер»).

Семантический анализ проводит анализ текста «по смыслу». С одной стороны, семантический анализ уточняет связи, которые не смог уточнить постсинтаксический анализ, так как многие роли выражаются не только при помощи средств языка, но и с учетом значения слова. С другой стороны, семантический анализ позволяет отфильтровать некоторые значения слов или даже целые варианты разбора как «семантически несвязные».

Этапом семантического анализа заканчивается анализ входного текста. Последующие этапы требуются для генерации отклика, например, в ходе диалога с пользователем или при переводе документов с иностранного языка для их дальнейшей обработки аналитиком. Сам отклик может, например, выбираться из некоторого корпуса текстов или генерироваться «на лету». В случае генерации ответа необходимо провести следующие этапы синтеза.

Генерация внутреннего представления отклика. Прежде, чем давать какой-либо отклик, диалоговая система должна сформулировать ответ. Для этого ей, например, может потребоваться собрать и проанализировать какую-то информацию. Отклик

системы будет зависеть от состояния диалога и других параметров. После этого необходимо определить форму ответа (или вопроса), подставить в него конкретные слова и значения и лишь затем приступать к синтаксическому синтезу текста отклика.

Предсинтаксический синтез. Задачи данного этапа прямо противоположны задачам постсинтаксического анализа. Здесь мы обязаны вернуть в предложение языкозависимые конструкции, пытаясь раскрыть роль слов средствами языка. В зависимости от контекста необходимо выбрать ту или иную форму выражения роли слов и основных идей предложения, расшифровать словосочетания, развернуть нормализованное дерево.

Синтаксический синтез превращает дерево предложения в линейный порядок слов. При этом осуществляется согласование параметров слов между собой.

Предморфологический синтез разъединяет слова, объединенные в целях экономии смысла в единую лексическую единицу. Здесь же может осуществляться обратная задача: слияния отдельных слов в одно, если того требуют правила языка.

Морфологический синтез по нормальной форме слова и его параметрам находит соответствующую словоформу.

Графематический синтез объединяет слова в единый текст, следит за соответствием фрагментов входного текста фрагментам выходного. На этом синтез отклика заканчивается.

Генерация отклика в разной мере присуща всем видам диалоговых систем, некоторым видам систем составления рефератов текста, статистического анализа текста, генерации текстов. Вопросно-ответные системы могут генерировать отклик как результат обработки запроса пользователя, системы общения обязаны делать это по определению, исполнительные системы могут комментировать происходящее или генерировать ответ на запрос пользователя. Но действия систем не ограничиваются только генерацией ответов. Вопросно-ответные системы должны сконвертировать запрос пользователя в какой-либо запрос на формальном языке (например, SQL при поиске в базе данных) и на основании полученных результатов решить, какой вид ответа необходимо выбрать. Исполнительная система должна определить алгоритм выполнения запроса пользователя и реализовать его. Но эти вопросы не входят сейчас в наше рассмотрение.

Кратко изложив последовательность этапов, в той или иной степени необходимых для обработки текста, рассмотрим теперь более подробно особенности реализации каждого из этих этапов.

Глава 2. Морфологический анализ и синтез

§ 2.1. Словарный морфологический анализ и синтез

Для того чтобы подчеркнуть различия в употреблении слов люди придумали формы слов или словоформы. Однако какова бы ни была словоформа, она выражает одно и то же понятие. Обсуждая понятие само по себе, принято использовать его нормальную форму – просто одну из словоформ, выделенную для обозначения понятия. Т.е., если у нас есть слово «мама», то для него существует несколько форм: мамы, маме, маму и т.д. К каждой форме приписывается ряд характеристик или параметров (род, падеж, число), характеризующих данную словоформу. Также каждому слову приписывается часть речи, показывающая, какого рода понятием мы оперируем. В речи мы привыкли к тому, что в данном месте должно стоять слово с заданной частью речи в определенной форме, но при машинной обработке подобные интуитивные рассуждения должны быть формализованы. Кроме того, подобное разнообразие вносит известные проблемы при анализе текста. Вместо того, чтобы работать с единственным словом, мы вынуждены обрабатывать все его словоформы. Для того чтобы избежать подобной ситуации были введены этапы морфологического анализа и синтеза.

Задачей *морфологического анализа* является определение по словоформе нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных к данной словоформе. При этом может оказаться, что одной словоформе может быть сопоставлено несколько таких пар.

Задача *морфологического синтеза* прямо противоположная. Здесь необходимо по нормальной форме и набору параметров получить словоформу.

Дадим более формальные определения, необходимые для рассмотрения этих этапов.

Нормальная форма слова – это форма слова (строка), принятая для обозначения понятия, связанного с данным словом. Обычно считается, что от нормальной формы образуются все остальные формы слова. Однако в таких случаях, как «идти – шел», связь между нормальной формой и словоформой не прослеживается. В связи с этим будем считать, что нормальная форма всего лишь одна из словоформ данного слова, выделенная согласно традиции данного языка. Словоформа – это форма слова (строка), связанная с нормальной формой слова и указывающая на особенности употребления данного слова. Будем считать, что словоформа характеризуется пятеркой – строкой словоформы; частью речи; нормальной формой, от которой была образована данная словоформа; частью речи нормальной формы; набором морфологических параметров, приписываемых к данной словоформе. Часть речи нормальной формы нам необходима, так как, например, деепричастие удобно считать формой глагола, а не выводить в отдельное слово.

Список основных частей речи в целом уже устоялся, хотя различные исследователи всё еще спорят о составе служебных частей речи. При реализации конкретного морфологического словаря важно с самого начала определиться с их списком, так как его изменение потом может оказаться дорогостоящей операцией. Для практических задач удобна любая из имеющихся логически обоснованных систем деления слов на части речи. В связи с этим мы не будем обсуждать здесь различные подходы к классификации слов.

Морфологический параметр – это пара <имя параметра, значение параметра>. Именем параметра может служить род, число, время, склонение, краткость формы прилагательного и другие признаки слов, принятые в данном языке. Значение параметра – это конкретное значение, которое может принимать данный признак. Так, например, падеж может быть именительным, родительным, местным, аккузативным; род может быть мужским, женским, средним; число – единственным, множественным, двойственным и т.д.

Параметры равны между собой, если равны их имена и значения. Параметры равны по имени, если совпадают их имена.

В ряде случаев значение параметра определить невозможно или в этом нет необходимости. Например, в русском языке существительным во множественном числе не приписывают род. Также существуют слова, которые имеют только форму множественного числа. Если словам, обладающим единственным числом значение рода может быть приписано из единственного числа, то слова, не обладающие единственным числом (очки, часы), такой информации лишены полностью. В этом случае будем считать, что значение параметра принимает фиксированное значение, обозначаемое «0». Примем, что параметр со значением «0» равен другому параметру, если равны их имена.

Подобный подход при хранении параметров может быть хорош в целом ряде случаев. Так, например, мы можем просто проверять наличие параметра с каким-либо значением. Это может пригодиться для того, чтобы убедиться, что параметр принял значение, отличное от заданного. Кроме того, мы можем приписать параметр слову для того, чтобы как-то выделить его среди других слов. В этом случае само наличие параметра у слова будет нести важную информацию.

Однако если нам необходимо просто провести морфологический анализ слов в тексте, может использоваться другой подход. Мы можем составить полный перечень всех значений параметров и дать им уникальные имена. В этом случае мы можем сэкономить место, так как хранится только имя параметра, сохраняя при этом различительную силу параметров. Но так как имена обычно даются символьные, то степень экономии зависит от фантазии разработчиков.

Вместо символьных имен параметров может использоваться цифровое представление. В этом случае мы можем создать справочник, в котором каждому символьному имени параметра будет сопоставлено некоторое уникальное число. При машинной обработке подобный подход позволит сэкономить место в памяти и ускорит процесс выдачи результатов. Заодно он объединит оба подхода, оставив людям уникальные и понятные для них символьные имена и предоставив компьютеру иметь дело с более удобным и компактным представлением.

При различных подходах слово «мама» может быть записано следующим образом.

мама	сущ., женск., одуш., единств., именит.
мама	сущ., род=женск., одуш=одуш., число=единств., падеж=именит.
12345	01 0202 0501 1101 1201

В последнем случае должен иметься справочник, указывающий, что существительному соответствует код 01, параметр род имеет код 02, а для него женский род кодируется числом 02 и т.д. Сопоставлением нормальной формы и ее кода занимается сам морфологический словарь. Так, если на анализ подается строка,

то на выходе будет числовой идентификатор, тогда как в синтезе информация преобразуется в противоположную сторону: по идентификатору можно получить строковую запись слова.

Набор параметров для частей речи фиксируется. Среди параметров слова выделяют словообразовательные и формообразовательные. Словообразовательные параметры не изменяются при изменении слова по формам. Так, например, слово «мама» остается женского рода в любой своей форме. Формообразовательные параметры изменяются при изменении слова по формам. Для приведенного примера падеж будет являться формообразовательным параметром. Обычно разделение на словообразовательные и формообразовательные параметры задается для всех слов, принадлежащих одной части речи. При этом словообразовательные параметры для одних частей речи могут являться формообразовательными для других. Например, параметр рода не меняется у существительных, однако будет образовывать формы у прилагательных и глаголов. Отнесение части параметров является спорной. Например, переходность глаголов может относиться как к словообразовательным, так и к формообразовательным параметрам, в зависимости от предпочтений разработчиков и их целей.

Формально морфологическая омонимия – это ситуация, когда одной словоформе можно приписать несколько кортежей, содержащих нормальную форму, части речи и набор параметров. Ниже приведены примеры таких ситуаций.

мамы	Мама	сущ., ж.р., ед. ч., род. п., одуш.
	Мама	сущ., ж.р., мн. ч., им. п., одуш.
стекло	Стекло	сущ., ср.р., ед. ч., им. п., неодуш.
	Стекать	гл., ср. р., ед. ч., 3 л., нн.ф., пр. вр.
дракон	дракон (животное)	сущ., м.р., ед. ч., им. п., одуш.
	дракон (корабль)	сущ., м.р., ед. ч., им. п., неодуш.
замок	замок (за́мок)	сущ., м.р., ед. ч., им. п., неодуш.
	замок (замо́к)	сущ., м.р., ед. ч., им. п., неодуш.

Приведенные примеры показывают различные виды омонимии. Подобная ситуация весьма распространена во многих языках, хотя и в разной мере. Отсутствие данного явления изрядно сократило бы количество шуток. В [7] приводится следующее определение омонимии: «Омонимами (гр. homos - одинаковый + onyma - имя) называются слова, разные по значению, но одинаковые по звучанию и написанию». При этом различают полные и неполные (частичные) омонимы. При полной омонимии слова принадлежат к одному грамматическому классу (у них одна часть речи) и все их формы совпадают. Примерами полной омонимии являются слова «дракон» (животное vs корабль), «кошка» (одушевленная vs неодушевленная), «коса» (прическа, отмель, инструмент). К неполной омонимии можно отнести слова, у которых совпадают лишь некоторые формы. Это, например, «закапывать», происходящее от слов «закапать» и «закопать», или приведенное выше «стекло». Выделяют и более частные явления. Так, полисемией называется свойство слов употребляться в разных значениях. В теоретической лингвистике считается, что слова полисемичны, если они сохранили некоторую логическую связь между собой: существуют некоторые аналогии, ассоциации, не потеряны исторические корни. Так, например, слово «ядро», как нечто центрообразующее, считается полисемичным

(пушечное ядро, ядро ореха, ядро армии), тогда как «коса» будет омонимичной (хотя и здесь можно найти общее – нечто тонкое и вытянутое). Кроме того, слова могут быть омоформами друг друга, если у них совпадают одно или несколько написаний, или омографами, когда у них совпадают написания, но различаются произношения («за́мок» vs «замо́к»).

Рассмотрев то, над чем работает морфология, перейдем к тем методам, с помощью которых она реализуется. Различают два вида морфологических словарей: словарные и бессловарные. Словарная морфология предполагает наличие словаря, в котором каждой словоформе сопоставлены нормальная форма и набор параметров. Т.е. у нас хранится полный словарь слов, и мы не можем проанализировать или просинтезировать слова, отсутствующие в словаре. Самым простым решением проблемы создания морфологического словаря является таблица, в которой в первой колонке будет записана словоформа, во второй – нормальная форма, а в третьей – набор параметров.

Для морфологического анализа в таком словаре необходимо просто найти все соответствующие словоформы и выдать найденные результаты. Для синтеза требуется найти заданную нормальную форму с требуемым набором параметров и выдать словоформу, находящуюся в той же строке (Рис. 2.1).

Однако при подобном подходе очень велики накладные расходы. Если предположить, что среднее слово будет занимать 8 байт, среднее количество параметров положить равным 4 и для обозначения параметров использовать 8 байт, то с частью речи одна запись в среднем будет занимать 48 байт. Предположим, что для каждой нормальной формы у нас имеется 8 словоформ. Тогда морфологический словарь объемом 100 тыс. слов будет занимать 36,6 Мб. Однако это весьма оптимистические предположения. Так, 8 словоформ были взяты просто потому, что это круглая цифра. В русском языке существительные имеют 12 форм, а прилагательные – 24; с другой стороны, наречия и предлоги имеют всего одну форму, но их количество уступает количеству тех же существительных. Максимальное количество словоформ может превышать 300 (если мы считаем деепричастия и причастия формами глагола), и среднее количество в этом случае составит около 25 форм на парадигму. На практике потребуется также место для индексов, в случае реляционной таблицы придется зарезервировать место не под среднюю длину слова, а максимальную и т.д. В итоге словарь в 100 тыс. слов (чего не достаточно для анализа текстов широкой тематики) будет содержать порядка 2–2,5 млн. входов против 800 тыс., которые были взяты для расчета. В связи с этим объем занимаемой памяти может вырасти на порядок – до 256 Мб без учета индексов. Кроме того, и время поиска будет пропорционально логарифму от объема базы, умноженному на среднюю длину слова.

Выходом может служить переход к дереву. Анализ словоформы в таком случае проводится побуквенно, начиная от корня дерева. В каждом узле хранится массив указателей на следующую вершину, причем каждый указатель отвечает за свою букву. Решение в лоб заключается в том, чтобы в каждом узле хранить массив длиной в размерность алфавита языка. Однако для русского языка для хранения всех последовательностей из 8 букв потребовалось бы более 46 млрд указателей. Большая часть таких вариантов будет отсеяна, так как, например, в русском языке нет слов, начинающихся на твердый или мягкий знак. Кроме того, массивы будут заполнены плотно только близко к корню дерева. Ближе к листовым вершинам массивы

становятся сильно разрежены. Это свойство можно использовать и после 2-4 уровней хранить два массива. Первый массив хранит список букв, для которых имеются указатели, а второй массив – собственно указатели в соответствующих ячейках. Кроме того, часть постфиксов слов будут уникальны для данной ветви и будут представлять собой линейную цепочку, то есть их можно хранить в виде строки, а не последовательности вершин дерева. На Рис. 2.2 показан пример дерева префиксов.

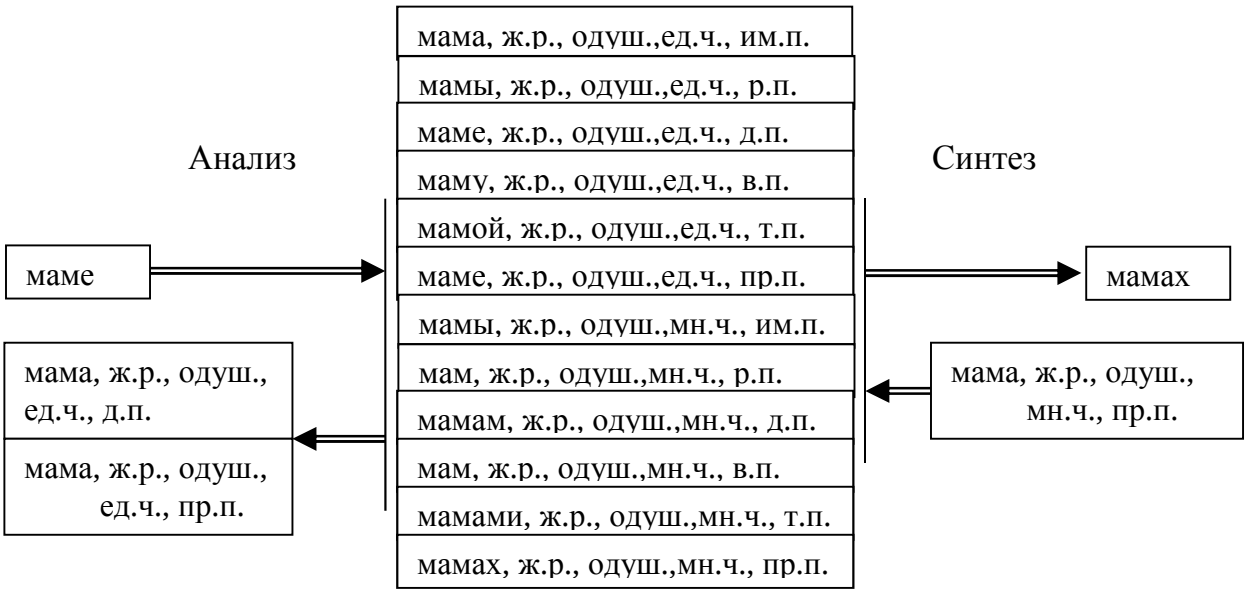


Рис. 2.1. Пример табличного морфологического анализа и синтеза

Кроме того, традиционно раздельно хранят деревья префиксов и постфиксов. Дело в том, что для большинства слов можно выделить неизменяемую часть – префикс – и набор постфиксов с привязанными к ним параметрами. Подобный набор постфиксов будем называть парадигмой изменения слова. Заметим, что мы не употребляем здесь слова «окончания», а именно постфикс, так как при изменении слова у него может появляться, исчезать или меняться не только окончание, но и суффикс. Частым явлением является изменение корневой буквы. Ярким примером является слово «идти», которое целиком попадет в изменяемый постфикс и будет иметь пустой неизменяемый префикс («идти» → «шел»). Кроме того, во многих сложносоставных словах меняется не только первое, но и второе слово. В этом случае в постфикс попадает вся часть слова, начиная с изменяемой части первого слова. Так, например, в фамилии Римский-Корсаков в постфикс попадет часть «ий-Корсаков».

Можно заметить, что слова группируются по парадигмам. Парадигма – это множество всех постфиксов и связанных с ними параметров для всех словоформ данного слова. Так, например, слова «лектор» и «завлаб» имеют одну парадигму. Мы можем хранить единственный набор ветвей в дереве постфиксов, сокращая тем самым занимаемый объем памяти.

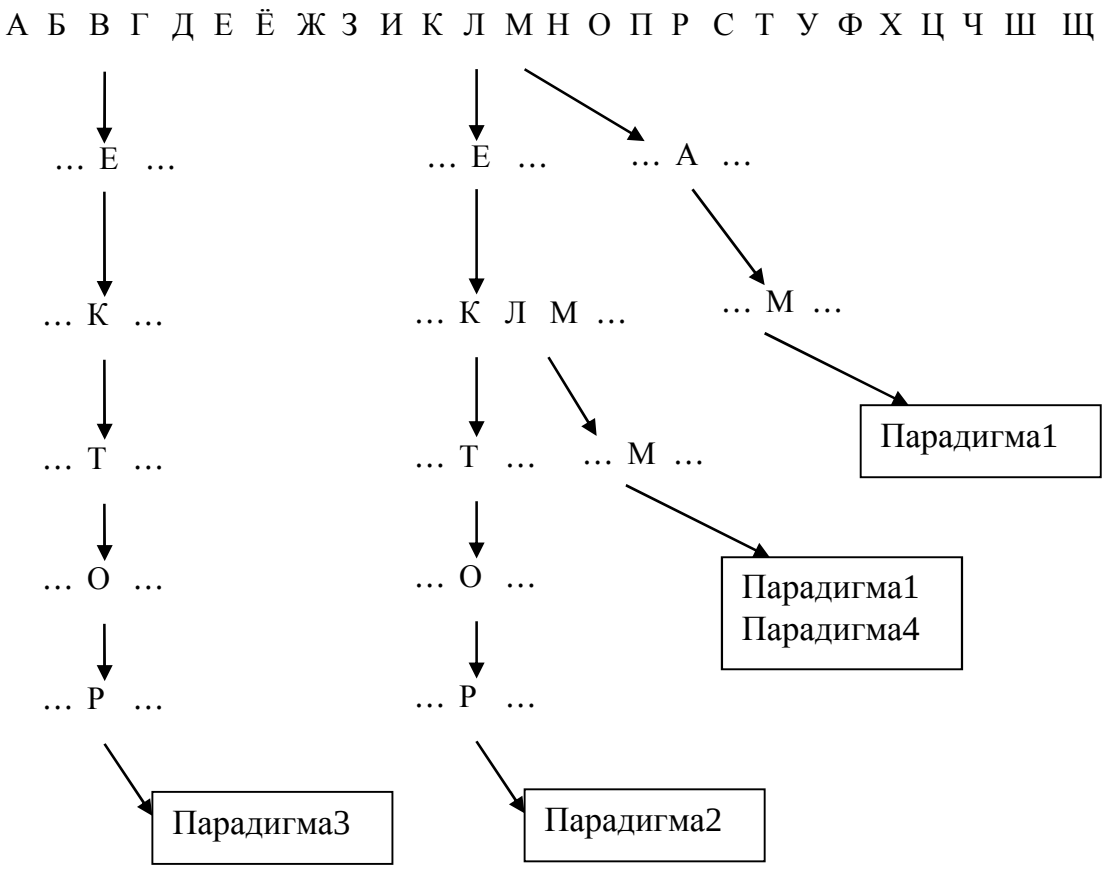


Рис. 2.2. Фрагменты дерева префиксов для слов «вектор», «лектор», «мама», «лемма»

Заметим, что вопрос о совпадении парадигм зависит от воли проектировщика. Так, слова «лектор» и «вектор» ни в коем случае не попадут в одну парадигму, так как будут иметь различные формы в винительном падеже. Однако слова «мама» и «лемма» попадут или не попадут в одну парадигму в зависимости от того, учитываем ли мы их одушевленность или нет. В первом случае парадигмы будут различными, так как для каждой формы слова «мама» будет прописан параметр «одушевленная» (хотя может существовать еще одно слово в морфологии, являющееся неодушевленным и означающее материнскую плату – его парадигма совпадет с парадигмой слова «лемма»), тогда как для слова «лемма» будет приписан параметр «неодушевленная». Во втором случае этих различий не будет, и слова будут принадлежать одной парадигме.

Конечная вершина дерева префиксов должна содержат информацию о том, какой постфикс или набор постфиксов соответствует данному префиксу. При этом указатель должен показывать на листовую вершину постфикса, в связи с тем, что развернуть его от корня будет несколько затруднительно. Заметим, что конечная вершина дерева префиксов не обязана быть листовой, так как в том месте, где закончилась основа одного слова, может продолжиться другое (например, «лук-ом» и «луков-ый»).

Заметим, что листовые вершины представленного дерева сходятся в парадигмах. Кроме того, часть путей в дереве могут совпадать, в связи с чем можно перейти от дерева к графу, объединив совпадающие части в одну ветвь. Такой граф уже можно назвать конечным автоматом, так как для каждого перехода в нем определен символ,

по которому осуществляется переход. Подобные и другие ухищрения позволяют существенно сократить объем памяти, занимаемой морфологическим словарем.

Морфологический анализ будет проходить следующим образом. Мы двигаемся побуквенно по строке, содержащей слово, перемещаясь при этом по дереву префиксов. Изначально в качестве текущей вершины выбирается корень дерева префиксов. Если переход из текущей вершины по очередной букве строки отсутствует, то разбор заканчивается. Если достигнута вершина, помеченная как конечная, то проводится проверка постфикса. При этом мы двигаемся по дереву постфиксов от листовой вершины к корню. Если корень был успешно достигнут, то информация из парадигмы изменения слова переносится в результат. Если в дереве префиксов не была достигнута листовая вершина, то движение по нему продолжается. В случае, когда множество результатов оказалось пустым, сообщаем о неуспешном анализе. В противном случае возвращаем множество результатов. Скорость анализа будет пропорциональна длине слова, а не объему словаря.

Для дерева постфикса возможен и другой вариант хранения. В этом случае мы храним префикс не с конца, а с его начала (движение будет производиться от корня дерева к листовым вершинам). При этом вершины деревьев префиксов и постфиксов хранят номер парадигмы. В этом случае, проведя анализ префикса, мы начинаем поиск постфикса от корня дерева. Если в конце слова была достигнута вершина с тем же номером парадигмы, что и в дереве префиксов, то слово считается успешно найденным.

Следует заметить, что скорость работы морфологического анализа будет сильно зависеть от задач, которые перед ним ставятся. Так, например, если у нас имеется поисковая система, задача которой найти все вхождения данного слова в документах, вне зависимости от формы слова, то нам вполне достаточно вернуть лишь нормальную форму слова. Если мы хотим ранжировать поиск в зависимости от совпадений морфологических параметров, или, как в нашем случае, морфологические параметры будут входить в критерий оценки, то для добавления параметров к результату и их обработки потребуется еще некоторое время. Кроме того, морфологический анализ может хранить и семантическую информацию, добавление которой к результату еще снизит скорость работы.

Морфологический синтез будет осуществляться следующим образом. В индексе нормальных форм находим все нормальные формы, для которых будет производиться синтез. Одной строке нормальной формы может соответствовать несколько слов со своими парадигмами, например, слово «кошка», имеющее одушевленную и неодушевленную формы, будет иметь разные постфиксы в винительном падеже множественного числа: «кошек» vs «кошки». Однако в данном случае ища, например, одушевленную кошку мы не обнаружим параметра «одушевленность» со значением «одушевленная» среди параметров неодушевленной кошки. Следовательно, она не попадет в результат. Аналогичные проблемы возникают и в других словах: «лист» – «листы» vs «листья»; «язык» – «языки» (часть тела) vs «языков» (язычник (древнерусское) или пленник).

Далее мы берем парадигму, соответствующую выбранной нормальной форме, находим нужный нам набор параметров, берем неизменяемую часть слова, присоединяем к ней постфикс, получая тем самым искомую словоформу. Помещаем ее в множество результатов.

При сравнении параметров может получиться так, что успешно сравнятся несколько наборов параметров. Это происходит потому, что как в множестве параметров, хранимых в парадигме, так и в множестве параметров, поступивших на вход, могут содержаться параметры с нулевым значением. Здесь следует помнить, что предпочтение следует отдавать полному совпадению параметров, т.е. желательно, чтобы значения параметров, имеющих на входе нулевое значение, в парадигме также имели нулевое значение. При наличии альтернативы лучше выбирать набор параметров, в котором большее количество параметров сравнилось точно. Еще одной проблемой при синтезе является неполный набор параметров, поступивший на вход. Это связано с тем, что мы не сумели выяснить полный набор параметров на предыдущих этапах. Такой вариант также необходимо предусматривать при реализации системы морфологического синтеза.

Так, например, если мы попытаемся сгенерировать родительный падеж единственного числа от слова «чай», то мы получим два варианта: «чая» и «чаю», которые оба являются морфологически верными и употребимыми. А попытка получить прошедшее время глагола, не уточнив предварительно его род, приведет к тому, что мы получим как минимум три варианта синтеза, так как прошедшее время глагола в русском языке не различается по лицам.

§ 2.2. Автоматизированное пополнение морфологического словаря

Автоматическое порождение гипотез о парадигмах изменения незнакомых слов является хорошей возможностью автоматизировать процесс заполнения баз. При переходе к новой предметной области встает вопрос о неполноте морфологического словаря. Каждая предметная область использует собственную лексику. В связи с этим встает вопрос о пополнении ею словарей. Данный процесс может быть автоматизирован, если имеющийся модуль морфологического анализа позволяет проводить предсказание лексических параметров незнакомых слов. Для этого необходимо выделить все слова, отсутствующие в имеющемся морфологическом словаре, и подвергнуть их анализу с предсказанием. Результатам анализа, как это отмечалось в соответствующем разделе, является кортеж словоформы $<\mathbf{f}_{nf}, r, \mathbf{P}_{const}(r,s) \cup \mathbf{P}_{var}(r,s)>$, где $\mathbf{f}_{nf}=<s_{nf}, r, \mathbf{P}_{var}(r,s)>$ - лексема нормальной формы, r – часть речи словоформы, s и s_{nf} – анализируемый токен (строка слова) и токен нормальной формы, а **P** – наборы параметров. По результатам анализа мы можем объединить все слова, обладающие одинаковыми токенами нормальной формы в единые гипотезы.

В ходе выдвижения гипотез можно использовать несколько сильных, но интуитивно верных положений.

1. Гипотезы, порожденные на основе редковстречающихся парадигм, в рассмотрение не брались. Под редковстречающейся понимается парадигма, по которой изменяется количество лексем не выше заданного порога.

2. Для словарных слов, принадлежащей одной парадигме, определяется список букв, заканчивающих их псевдоосновы. В случае если для словоформы выдвигается гипотеза о ее принадлежности к данной парадигме, и если при этом ее псевдооснова не оканчивается ни на одну из полученных букв, то такая гипотеза отвергается. Использование двух букв псевдоосновы позволяет проводить выбор с весьма высокой точностью.

3. Можно отсеивать гипотезы, образованные от словоформы, встретившейся единственный раз в исследуемом корпусе и являющиеся единственной словоформой,

использованной в данной парадигме, так как подобная словоформа скорее всего написана с ошибкой. Исключение можно делать для парадигм не изменяющихся слов (т.е. содержащих единственную позицию в парадигме).

4. Псевдоосновы несловарных словоформ, объединяемых в рамках одной парадигмы, должны содержать хотя бы один символ.

После кластеризации проводится отсеивание полученных лексем по критерию максимальной встречаемости словоформ, вошедших в лексему. Т.е. для каждого слова определяется, сколько раз оно встретилось в тексте. Далее эти значения суммируются по парадигмам и оставляются лишь парадигмы с максимальной сумой.

Получаемые результаты будут существенно зависеть от типа используемой морфологии. Для стемминга будут объединяться все слова, обладающие одной псевдоосновой. Так, в одну парадигму в зависимости от алгоритма выделения псевдоосновы могут попасть слова «компьютер», «компьютерный», «компьютеризация» и т.д. При использовании лемматизации результаты зависят от списка используемых параметров. При полном отсутствии таковых слова объединяются без образования альтернатив. Однако полный набор параметров создает проблемы. Среди прочего, это связано с тем, что в русском языке встречаются парадигмы, объединяющие один и тот же набор флексий, однако приписывающие им различные наборы параметров. Так, для слова «админ» можно породить лексемы, показанные на Рис. 3.3. Здесь «-» означает пустой постфикс. В скобках написаны словарные представители парадигмы. Из приведенных примеров видно, что даже один и тот же набор словоформ может быть различным образом размещен в различных парадигмах.

Единственное число	им.	род.	вин.	дат.	тв.	пр.
АДМИН (ТЕЛЕФОН) м.р., неодуш	-	А	-	У	ОМ	Е
АДМИН (ТОН) м.р., неодуш	-	А	-	У	ОМ	Е
АДМИН (БАЛ) м.р., неодуш	-	А	-	У	ОМ	Е/У
АДМИН (АКТИВИСТ) м.р., одуш	-	А	А	У	ОМ	Е
АДМИН (ОПЕР) м.р., одуш	-	А	А	У	ОМ	Е
Множественное число	им.	род.	вин.	дат.	тв.	пр.
АДМИН (ТЕЛЕФОН) м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (ТОН) м.р., неодуш	А/Ы	ОВ	А/Ы	АМ	АМИ	АХ
АДМИН м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (АКТИВИСТ) м.р., одуш	Ы	ОВ	ОВ	АМ	АМИ	АХ
АДМИН (ОПЕР) м.р., одуш	А/Ы	ОВ	ОВ	АМ	АМИ	АХ

Рис. 3.3. Пример неоднозначности предсказания слова по всем его словоформам

Большое количество ошибок, встречающихся в любых текстах, зашумляет выход системы лемматизации и требует длительного ручного труда по отделению корректных вариантов от ошибочных. К счастью, возможностей для ошибки предоставляется очень много, и поэтому большинство ошибок встречается один или два раза и отсеиваются на этапах фильтрации или кластеризации. Однако некоторые ошибочные словоформы могут войти в состав других парадигм, изменив тем самым результаты кластеризации не в лучшую сторону. Кроме того, у многих авторов существуют так сказать «любимые» ошибки, когда одна и та же ошибка допускается многократно в различных словоформах.

Однако даже небольшая автоматизация процесса предсказания парадигм несловарных слов позволяет существенно повысить производительность труда лингвистов в ходе формирования словарей. Кроме того, в ряде задач может использоваться не лемматизация, а, например, обсуждаемый ниже стемминг, в ходе которого лексические параметры указываться не будут. В этом случае необходимо сгенерировать (в том или ином виде) нормальную форму слова и указанные выше проблемы окажутся неактуальны. В этом случае возможно создание полностью автоматической процедуры пополнения словарей.

§ 2.3. Методы бессловарного морфологического анализа

Бессловарные морфологические словари появились во времена, когда оперативная память была существенно ограничена. Однако на данный момент несколько мегабайт или даже десятков мегабайт оперативной памяти не составляют проблемы, в связи с чем наибольшее распространение получили словарные морфологии. Существенным плюсом бессловарных морфологий является то, что ни могут предсказать морфологические характеристики практически любого слова, если его парадигма изменения попадает под одну из хранимых. Классическим примером здесь является предложение «Глокая куздра штеко будланула бокра и курдячит бокрёнка», предложенное одним из основоположников отечетственного языкознания академиком Л.В. Щербой еще около 1930 года при чтении курса лекций «Основы языкознания». Из этого предложения мы можем понять, что «куздра» имеет женский род, единственное число, именительный падеж и т.д. и разобрать синтаксис предложения, при этом совершенно не понимая, о чем идет речь. С другой стороны, наша уверенность в том, что куздра имеет женский род во многом основывается на результатах неявно проводимого синтаксического анализа, который говорит о том, что «глокая» является прилагательным в женском роде и согласуется со словом «куздра».

Бессловарные морфологии хранят парадигмы слов. При этом в парадигме в качестве постфикса может храниться только окончание. Часто в бессловарной морфологии может храниться набор приставок и суффиксов и привязанная к ним семантическая информация. Например, про суффиксы «-онок-» и «-ёнок-» будет написано, что их добавление обозначает детеныша животного, а приставка «при-» означает присоединение или приближение.

Анализ и синтез в бессловарных морфологиях ведется так же, как и в словарных, но без поиска по дереву префиксов и с учетом возможности выделения нескольких последовательно идущих постфиксов.

Однако такой подход часто приводит к существенным ошибкам. Так, например, «октябрёнок» может стать детенышем «октября» (что формально верно), «припевать» будет трактоваться как «приближение + певать» или «присоединение + певать», а «перебиваться» - возвратной несовершенной формой от «перебить», причем не ясно от какого из значений: прервать, уничтожить или прибить заново. При этом также не совсем понятно, какой вид возвратной формы будет иметься ввиду: перебивать себя или перебивать самому. При этом на самом деле возвратная форма от «перебить» будет подразумевать совсем иное значение – «обойтись без чего-либо», хотя вариант «перебить себя» представляется маловероятным, но не невозможным.

Для бессловарных морфологий существовали алгоритмы, позволявшие избегать этих ошибок. Так, например, для сочетания «пере+бивать+ся» можно в явном виде

прописать значение слова. Но в итоге мы получаем словарную морфологию со специальным алгоритмом архивирования базовых понятий, для которых нет исключений.

Также выделяют системы на основе стемминга. В случае стемминга зачастую отбрасывается вся морфологическая информация, а в качестве нормальной формы берется неизменяемая псевдооснова, называемая стем. Так, для слова «мама» стемом будет являться строка «мам». Именно эта основа и используется в дальнейшем для идентификации слова во всех его формах. Неудобство состоит в том, что для различных слов может порождаться один и тот же стем, например, «люб-овь» и «люб-ить». В случаях, когда необходимо различать эти понятия (например, при поиске слов в тексте), возможен единственный вариант – хранить информацию о части речи. В прямо противоположном случае, когда различать слова не обязательно, подобное совпадение может сослужить добрую службу. Однокоренные слова чаще всего относятся примерно к одному и тому же понятию («любить» означает «продуцировать любовь»). В связи с этим при сравнении текстов целиком такие понятия не будут размываться, а скорее наоборот – будут давать совместный вклад в результат сравнения. Однако в случае стемминга весьма вероятно смешение различных понятий. Так к стему глагола «люб-ить» будет отнесен и глагол «люб-оваться» (ведь у него есть форма «люб-уюсь»), что приводит к смешению различных понятий.

Для слов, подверженных флексии, т.е. замене букв в корне слова, берется несколько стемов. Так, например, для слова «шов» будет образовано два стема: «шов» и «шв», а для слова «идти» - «ид», «ше» и «шл». Это не позволяет идентифицировать их как одно со всеми вытекающими последствиями. Для решения этой проблемы создаются сложные парадигмы, объединяющие несколько стемов.

Собственно анализ в подобных системах будет проводиться аналогичным образом с лексической морфологией. Однако здесь возможны два варианта. В первом случае мы храним как стемы, так и парадигмы изменения и алгоритм анализа и синтеза не претерпевает никаких изменений. Во втором случае хранится только набор парадигм. В этом случае оставшаяся основа и будет являться искомым стемом. Из множества полученных стемов выбирается, например, самый короткий или самый длинный. Также применяется вариант, когда проводится анализ последних нескольких букв стема: наречия, имеющие пустое окончание, заканчиваются на «-о» или «-е», некоторые глаголы на «-ова-» или «-ева-» и т.д. Это также помогает отсеять ряд результатов.

Морфология на основе стемминга обладает рядом достоинств. Так, например, за счет упрощения алгоритма и уменьшения объема выдаваемой информации существенно (до нескольких раз) возрастает скорость анализа, а при использовании лишь массива парадигм сокращается объем хранимых баз. Главным достоинством морфологии на основе стемминга является тот факт, что при отсутствии словаря основ мы фактически получаем морфологическую базу неограниченного объема, настраиваемую непосредственно на имеющийся текст. Это очень удобно при создании информационно-поисковых систем с нефиксированной лексикой. В этом случае при индексировании текстов мы получаем некоторый набор стемов, которые и заносим в индекс. При этом морфология никогда не сообщает нам, что такого слова нет в словаре.

Однако подобный подход не лишен недостатков. Первым из них является невысокая точность метода. Так, например, стем «шл» будет соответствовать и глаголу «слать» («шлют» и омонимичное «шли»). Соответственно, при анализе мы объединим два этих разных слова в один «куст». В результате на информационный запрос пользователя о слове «шлют» будет выдана информация и о глаголе «идти». В зависимости от применяемого алгоритма могут быть выданы все формы для обоих глаголов. А в зависимости от слова в один «куст» могут быть объединены и слова различных частей речи. В ряде случаев это может оказаться весьма полезным, так как однокоренные слова обычно относятся к одним понятиям («грузчик» – «грузить») и, например, при информационном поиске начинают за счет этого объединяться в кластеры. Однако для исполнительных систем такой подход неприменим.

Следующим недостатком является невозможность морфологического синтеза на базе без основ. Справедливости ради следует заметить, что подобная задача необходима не во всех практических приложениях. И, как это было замечено выше, стемминговый подход не применим к таким приложениям, как исполнительные системы. Лишившись морфологической информации мы перестаем понимать, является ли слово действием, которое мы должны выполнить, или объектом этого действия, каким именно объектом действия является слово и т.д. Без подобной информации качественное выполнение действий невозможно.

Следует заметить, что грань между стемминговой морфологией, базирующейся на неизменяемой псевдооснове, и лексической морфологией, выдающей полный набор морфологических параметров и оперирующей с нормальными формами слова, довольно тонка. С одной стороны, лексическая морфология использует неизменяемую основу, т.е. стем. С другой стороны, при хранении полного набора лексической информации стемминг отличается от лемматизации лишь выдаваемой строкой нормальной формы. Так, система морфологического анализа MyStem компании Яндекс (http://company.yandex.ru/technology/mystem) хотя и называется стеммером (точнее – парсером), однако выдает полный набор лексической информации о слове. Аналогичный по объемам и выдаваемым характеристикам морфологический словарь «Диалинг» (http://www.aot.ru/) является полноценным лемматизатором и ни в коем случае не заявляется как стеммер.

Одним из современных вариантов реализации бессловарной морфологии в чистом виде является стеммер Портера (http://snowball.tartarus.org/). В нем пришедшая на вход строка проверяется на наличие заданных постфиксов, причем постфиксы проверяются в определенном порядке, а часть постфиксов может комбинироваться. Так, после выделения постфикса прилагательного может остаться постфикс причастия. Все, что осталось после их последовательного «откусывания», объявляется стемом. В зависимости от найденного постфикса слову может быть приписана та или иная часть речи, хотя в подавляющем большинстве задач этого не требуется. Алгоритм предельно прост, обладает очень высокой скоростью, однако дает большой процент ошибок. Так, например, если требуется подсчитать частотность слов, то для уже разбиравшихся слов «идти» и «шов» будут сгенерированы несколько никак не связанных стемов. В результате частотность данных слов будет существенно понижена за счет «размазывания» ее по нескольким группам. Кроме того, деление на постфиксы является в значительной мере спорным. Скажем, постфикс «-ев» относится к существительным, тогда как слово «ошалев» таковым не является. Также алгоритм выдает единственный вариант разбора,

полностью скрывая омонимию слов. Заметим также, что исходный алгоритм был несколько дополнен его отечественными пользователями, что несколько сократило процент ошибок.

Алгоритм Портера очень слабо учитывает тот факт, что для различных частей речи и даже для различных парадигм перед постфиксом могут стоять различные буквы. Этот факт используется в системе морфологического анализа Stemka (http://www.keva.ru/stemka/stemka.html), где хранятся не только сами постфиксы, но и еще две предшествующие буквы псевдоосновы. Сами комбинации букв и постфиксов хранятся в виде конечного автомата справа налево.

Существенным плюсом бессловарных морфологий является то, что они могут выдать результат для любых слов, встречающихся в тексте, что очень удобно при анализе текстов из незнакомой предметной области или содержащих много нелитературных или редко употребляемых слов. Однако корректность выдаваемой информации находится на уровне 90-95%. Это привело к отказу от бессловарных морфологий в задачах, когда точность анализа должна превалировать над его полнотой, и к переходу к словарным морфологиям в таких задачах, как машинный перевод и диалоговые системы. Однако на практике существует большое количество задач, решаемых статистическими методами, в которых вполне достаточно приблизительного знания о связях между словами. Это задачи рубрикации, информационного поиска, частично – задачи реферирования, ряд других задач.

Методы бессловарных морфологий активно используются в словарных морфологиях для предсказания нормальной формы и набора параметров слов, которые отсутствуют в морфологическом словаре. Для этого необходимо проанализировать постфиксы слова и попытаться образовать нормальную форму исходя из полученного префикса и парадигмы, приписываемой постфиксу. Для этого по найденным постфиксам определяются постфиксы нормальной формы, которые присоединяются к полученным префиксам, и наборы морфологических параметров. Существенным недостатком является большое количество предсказанных вариантов. Так, например, слово «кони» может быть предсказано как существительное мужского рода («ог*ни*» – «ко*ни*» → «ог*онь*» – «ко*онь*»), женского рода одушевленное или неодушевленное («кошк*и*» – «кон*и*» → «кошк*а*» – «кон*а*»), обладающее только множественным числом («сан*и*» – «кон*и*»), глаголы «конить» и «кнать» в повествовательном наклонении («г*они*» – «к*они*» → «г*нать*» – «к*нать*»; «юл*и*» – «кон*и*» → «юл*ить*» – «кон*ить*») и т.д.

Количество таких вариантов может быть существенно сокращено за счет фильтрации. Так, например, при наличии морфологического словаря достаточно большого объема можно утверждать, что в нем находятся все местоимения, предлоги, союзы и некоторые другие части речи. Отсев можно произвести и с точки зрения статистики. Достаточно большое количество парадигм содержит всего по нескольку слов. Также много парадигм, созданных для единственного слова. Например, парадигму составляют все формы слова «идти», так как в словарной морфологии оно будет обладать пустой основой. Эти парадигмы в большинстве своем являются закрытыми, т.е. добавление новых слов в них уже невозможно. В связи с этим можно отсечь подобные парадигмы, запретив выдвижение гипотез на их основе. Для этого достаточно подсчитать количество слов, относящихся к каждой из гипотез, и выдвигать гипотезы только на основе парадигм, к которым принадлежит количество слов, большее заданного порога.

Еще один вариант отсеивания основывается на том, что у слов, принадлежащих одной парадигме, совпадает не только изменяемая часть, но и последние несколько символов неизменяемой. Так, например, у слов «шествовать», «повествовать», «любопытствовать» псевдооснова заканчивается на «-ств-». Соответственно, новое слово, обладающее постфиксом «-ств-овать» (или в более общем случае «-ств+постфикс парадигмы») должно быть предсказано именно по этой парадигме. Для предсказания может быть использовано от одной до трех букв с конца псевдоосновы. Три буквы позволяют получить статистически значимые результаты для часто повторяющихся длинных последовательностей. Но так как одни и те же трехбуквенные последовательности встречаются гораздо реже, чем единственная буква, то использование трех букв дает гораздо больше вариантов таких последовательностей, используемых для проверки.

На практике подобный метод используется для некоторого разрешения неоднозначности выдаваемых результатов, когда кроме стема требуется определить и хотя бы часть речи. В такой ситуации три предыдущие буквы позволяют с достаточно высокой точностью определить, к какой части речи должно относится слово. Для этого помимо псевдоокончания хранят и несколько последних букв. Большее количество букв псевдоосновы позволяет более точно определить параметры слова. При этом меньшее количество букв позволяет достичь большей скорости разбора и сокращает объем словаря.

Методы бессловарных морфологий используются также и для расширения словаря. Так, например, очень многие существительные могут употребляться с «не-» в начале. Введение всех существительных с «не-» значительно увеличит объем словаря и снизит скорость его работы. В связи с этим при морфологическом анализе можно предварительно проверить наличие одного из префиксов («пере-», «при-» и т.д.) или аффиксов (например, «-ся» и «-сь»), а оставшуюся часть слова попытаться найти в словаре. При этом, например, для слова «оставшийся» после удаления аффикса основы найдено не будет, а слово «делавшийся» будет успешно разобрано (если предположить, что ни того, ни другого нет в словаре). Заметим, что запускать подобный алгоритм имеет смысл лишь после того, как выяснится, что пришедшее слово целиком отсутствует в словаре. Заметим также, что подобная операция может быть проведена средствами предсинтаксического анализа, в котором могут быть предусмотрены соответствующие средства.

§ 2.4. Коррекция орфографических ошибок

Небольшая модификация алгоритма морфологического анализа может помочь распознавать слова, написанные с ошибками. Для этого используется расстояние Левенштейна – минимальное количество ошибок, исправление которых приводит одно слово к другому. Считается, что существуют ошибки трех видов: ошибка вставки, ошибка замены и ошибка пропуска. Ошибка вставки – это случай, когда в слово вставлена лишняя буква; ошибка пропуска – когда буква была пропущена; и ошибка замены – когда одна буква заменена другой. При поиске слова с ошибкой нам придется поочередно пытаться пропускать буквы то во входном слове (в случае ошибки вставки), то в дереве (в случае ошибки пропуска), или и там, и там (в случае ошибки замены). При этом если пропуск буквы во входной строке является действием тривиальным, то пропуск вершины в дереве ведет к тому, что мы не знаем,

какого именно потомка данной вершины следует выбрать. В связи с этим приходится осуществлять перебор всех потомков и отбирать все успешные варианты.

Ошибка в ходе морфологического анализа может проявляться один из двух способов. При анализе слова мы можем дойти до места, когда переход по текущей букве будет отсутствовать. Ошибка может быть совершена в одной из предыдущих букв, поэтому разбор слова следует начать сначала. Еще одним вариантом определения ошибки в слове будет тот факт, что множество парадигм, привязанных к найденным префиксам, не будет пересекаться с множеством парадигм, найденных для постфиксов.

Кроме того, следует вводить ограничение на количество ошибок, которые мы позволяем допустить. Как говорит поговорка: «Если в слове хлеб допустить всего четыре ошибки, то получится слово пиво». Если мы фиксируем число ошибок, то для коротких слов оно может оказаться избыточным. Так, с одной ошибкой слово «быть» сводится к словам «быт», «мыть», «выть», «бить». С двумя ошибками мы уже получаем «плыть», «забыть», «ять», «зять» и множество других слов, уже гораздо менее похожих на оригинал. Аналогично в длинных словах нельзя применять процентное соотношение, так как, например, 25% букв от слова «великолепнейший» дает нам 3 ошибки, которые позволяют свести данное слово ко всем его словоформам. В связи с этим верхнюю границу числа ошибок обычно ограничивают как процентным соотношением, так и фиксированным числом. Например, не более 30% букв входного слова, но не более 3. При этом все равно стараются найти слова с минимальным количеством ошибок. Т.е., если для слова «великолепнейший» (предположим, что именно этой словоформы у нас нет в словаре) мы найдем слово «великолепнейшая» (2 ошибки), то слово «великолепнейшего» (3 ошибки) рассматриваться уже не будет.

Также могут вводиться специфические варианты ошибок. Так, при ручном наборе текста вариантом ошибки замены является смена порядка следования двух букв (транспозиция): «рпивет» вместо «привет». При распознавании текста могут рассматриваться специфичные замены: «ю» → «io» и т.д.

Компанией «Диктум» был предложен следующий вариант алгоритма коррекции ошибок при анализе текстов. В качестве основных берутся ошибки вставки, пропуска, замены и транспозиции. Кроме того рассматриваются ошибки следующего вида.

1. клавиатурная близость клавиш: *«анеудот – анекдот»*;
2. ошибки в безударных гласных: *«аностасия – анастасия»*;
3. фонетическая похожесть букв: *«брюпетка – брюнетка»*;
4. парные буквы: *«автограв – автограф»*;
5. вставка лишнего пробела: *«сло во – слово»*;
6. отсутствие пробела или дефиса: *«футбольныйклуб – футбольный клуб»*;
7. схожесть написания цифр и букв (ч-4, о-0, з-3): *«4естно – честно»*;
8. идентичное написание букв в разных раскладках: «XP0M0C0MA – хромосома»;
9. буквы и символы в разных раскладках: *«<лизнец – близнец»*;
10. ошибки после шипящих и ц: *«жолтый – желтый»*;
11. перепутывание и смещение рук при слепой печати: *«инвнжае – телефон»*;
12. перевод транслитерации на русское написание: *«kartinki – картинки»*;
13. исправление неправильной раскладки клавиатуры как для целого, так и для части слова: *«jlyjrkfccybrb – одноклассники»*.

Для указанных ошибок расстояние Левенштейна берется из интервала [0;1], так как их вероятность выше замен другими буквами. Для хранения весов расстояния Левенштейна используются квадратные матрицы, определенные для всего алфавита.

Формально задача ставится следующим образом. Пусть дан алфавит A, на данном алфавите определен словарь языка $L \subset A^+$. Тогда для слова $\mathbf{w} \in A^+$ требуется найти множество слов $\{\mathbf{w}'\}$, таких, что $dist(\mathbf{w}, \mathbf{w}') \leq \sigma$ и $F(\mathbf{w}') = \max(v)$, где $\mathbf{v} \in L$ и $dist(\mathbf{w}, \mathbf{v}) \leq \sigma$. При вычислении функции *dist* используются значения из матриц. Если занчение не найдено, вес ошибки принимается равным единице.

Однако для устранения части ошибок требуется учесть лексический контекст. Так, например, одной из широко распространенных ошибок является вставка мягкого знака в глаголы в третьем лице (в этом случае получается нормальная форма глагола): «нравиться» vs «нравится». Подобное слово содержится в словаре, но написано с ошибкой, которая не позволяет провести синтаксический анализ. Кроме того, коррекция ошибок в слове может дать несколько вариантов написания. Для устранения подобных ситуаций следует учитывать контекст слова. Например, «белый грип» – «белый гриб», но «птичий грип» – «птичий грипп», или «очень нравиться телефон» при правильном «очень нравится телефон».

Для исправления подобных ошибок могут использоваться синтаксический анализ с коррекцией или метод n-грамм. При ошибке, изменяющей форму слова, в их базе n-грамм не будет обнаружено соответствующего сочетания. При этом придется предложить варианты замены для всех слов в n-грамме. Так, например, фраза «очень нравиться телефон» может быть истолкована и как «очень нравиться телефону». Однако в полной фразе «Мне очень нравиться телефон» последний вариант представляется маловероятным и будет отсеян. Отсев будет проведен либо на этапе устранения омонимии, либо на синтаксическом анализе, так как фраза не содержит субъекта.

Глава 3. Постморфологический и предсинтаксический анализ

§ 3.1. Автоматизированное снятие омонимии

В некоторых задачах оказывается возможным автоматическое обучение по большому корпусу текстов. Т.е. имеется возможность собрать статистическую информацию, которая потом будет использоваться для вероятностного определения характеристик текстов. К подобным подходам относится метод n-грамм. Суть метода состоит в следующем. На большом размеченном корпусе текстов со снятой омонимией подсчитывается частота встречаемости для всех имеющихся в тексте комбинаций из n последовательно идущих слов. При этом обычно опускается нормальная форма слова, т.е. в расчет принимаются лишь часть речи и лексические параметры. В ходе разбора текста подобная информация может использоваться для вероятностного снятия омонимии. Мы можем предположить, что первые n-1 слово в n-грамме определяют нам вероятность появления n-го слова. Таким образом, зная первые n-1 слово мы можем выбрать наиболее вероятный вариант последнего. В ряде случаев используют линейную комбинацию вероятностей нескольких сокращающихся последовательностей слов.

Другим способом применения данного метода является отсечение наименее вероятных вариантов, т.е. частичное снятие омонимии. Так, если в корпусе текстов нам ни разу не встретилась определенная комбинация слов, то можно считать, что она вообще не должна встречаться в текстах. Таким образом, если мы встретили ее в реальном тексте, то часть входящих в нее омонимов может быть отброшена.

Значение n обычно принимается равным трем, так как биграммы обладают слишком малой историей и в дальнейшем не дают хороших результатов, а 4-граммы порождают слишком большое количество вариантов. При порождении n-грамм количество вариантов может быть уменьшено последующей их обработкой. Так, например, может оказаться, что для некоторой пары слов, стоящей в начале триграммы, существует полный набор вариантов для третьего слова, причем вероятность их появления примерно равна. Это будет означать, что первые два слова не определяют третье, и весь набор троек может быть удален как неинформативный. Подобный подход работает лишь в ситуации, когда мы выбираем наиболее вероятный вариант. При отсеивании наименее вероятных омонимов сам факт наличия подобных троек в базе будет означать, что у третьего слова не может быть удалено ни одного омонима.

Приведенный метод позволяет учитывать согласование параметров слов. Для каждой триграммы можно хранить лишь те значения лексических параметров, которые совпадают у отдельных комбинаций или всех слов в n-грамме. В методе следует различать знаки препинания, не сводя их, например, к одной части речи, так как роли, например, запятой и двоеточия в предложениях существенно различны.

Описанный метод позволяет получить точность снятия омонимии до 95% при выборе единственного наиболее вероятного варианта (при полном снятии омонимии) и порядка 99% (в зависимости от степени снятия) при отсеивании наименее вероятных вариантов (при частичном снятии омонимии).

Существует несколько методов, применяющих информацию об n-граммах. Наивный классификатор Байеса – это наиболее простой вид теггера (программы морфологической разметки, возможно, совмещенной со снятием омонимии), обучающегося на эталонном корпусе, который применяется для снятия омонимии с

помощью лексических параметров соседних слов, используя варьируемое окно контекста. Классификатор Байеса основывается на том предположении, что все параметры статистистически не зависимы между собой. В задаче снятия омонимии контекст, в котором появляется омонимичное слово представляется набором параметров (F_1;F_2; … ; F_n), а значение самого омонимичного слова представляется классом (S). Параметры F_i могут быть бинарными и представлять, появляется или нет какое-либо омонимичное слово с некоторым набором слов слева и справа от него. Для наивного Байесовского классификатора суммарная вероятность появления комбинации контекстных параметров с данным словом описывается следующим образом (более подробно теория по данному вопросу изложена в Часть V.§ 1.3):

$$P(F_1, F_2, ..., F_n, S) = P(S)\prod_{i=1}^{n} P(F_i \mid S) \qquad (3.1)$$

Любой из параметров, который равен нулю, говорит о том, что наше слово никогда не появляется с определенным значением. Забегая несколько вперёд, скажем, что такие значения сглаживаются путём присвоения им по умолчанию очень маленькой вероятности. В общем случае, каждый параметр F_i может входить с соответствующим весовым коэффициентом в выражение 3.1. Знаки препинания могут учитываться или нет, в зависимости от конкретной реализации системы автоматической обработки текста. В системах автоматической обработки теста капитализация слов, как правило, никогда не учитывается. Окно контекста может охватывать только левых, только правых или сразу левых и правых соседей слова. Выбор размера окна контекста оптимальным образом - это отдельная задача.

Как уже было отмечено, статистическое моделирование естественного языка предназначено для морфологической разметки текста с помощью закономерностей, которые не могут быть получены в явной аналитической форме. Здесь возникает проблема выбора наиболее подходящей статистической модели q(x), которая бы учитывала все особенности обучающей выборки. Таким образом, сама обучающая выборка является ограничениями, которые накладываются на q(x). Обратимся к такому понятию как энтропия, которое является основным для теории информации. Энтропия - это мера априорной неопределенности системы. Энтропия обладает следующими полезными свойствами: обращается в ноль, когда одно состояние системы достоверно, а другие невозможны; при заданном числе состояний обращается в максимум, когда эти состояния равновероятны[68]. Согласно принципу максимальной энтропии, вид модели q(x) подбирается таким образом, чтобы максимизировать предмет энтропии $H(q)$, не делая никаких дополнительных предположений для последовательности из N слов, не представленных в обучающей выборке. Принцип максимальной энтропии записывается в следующем виде:

$$H(q) = -\sum_x q(x)\log q(x) \qquad (3.2)$$

Средний показатель энтропии для английских текстов составляет 6-10 бит на слово, который может зависеть от вида N-граммной модели и жанра текста. В рамках задачи по разметке текста, энтропия – это среднее число бит, нужное, чтобы определить значение слова в данной обучающей выборке. Показатель связанности (perplexity) – это среднее геометрическое количество слов, которое может оказывать

[68] В случае N-граммной модели это означает, что вероятность появления N-граммы вычисляется по методу максимального правдоподобия.

влияние на неизвестное слово. Это еще одна стандартная мера для сравнения моделей языка, которая выражается следующей формулой:

$$PP = 2^{H(q)} \quad (3.3)$$

Следует подчеркнуть, что энтропия это некоторая функция, которая характеризует как саму модель естественного языка, так и имеющуюся обучающую выборку. Среди двух вероятностных моделей, имеющих одинаковый уровень ошибок предпочтительнее та, у которой энтропия меньше. Вообще же, количество ошибок и энтропия не однозначно связаны между собой.

В соответствии с принципом максимальной энтропии у нас есть возможность выбрать наиболее оптимальную базовую вероятностную модель естественного языка. Но эта базовая модель основана на принципе максимального правдоподобия, который не позволяет учитывать неравномерности в обучающей выборке и делать разметку при неполной информации.[69] Разумеется, что от системы по автоматической обработке текста требуется, чтобы она обрабатывала как можно более широкий круг текстов, а не только тот, что был представлен в обучающей выборке[70]. Таким образом, сглаживание используется, а зачастую просто необходимо, в том случае, когда для обучения доступен небольшой корпус, и есть возможность получить нулевые вероятности для последовательности слов, не представленных в обучающей выборке. Цель сглаживания сделать распределение более равномерным, другими словами, повысить вероятности для последовательностей слов, которые встречаются редко или вообще не встречаются и, соответственно, несколько снизить вероятности для комбинаций слов, которые часто встречаются. Как правило, методы сглаживания позволяют повысить качество работы триграммных тэггеров и, тем более, тэггеров на основе скрытых марковских моделей высоких порядков.

Различные методы сглаживания N-граммной вероятностной модели позволяют подобрать оптимальную[71] статистическую модель естественного языка. Проблему выбора оптимального тэггера попробуем обрисовать на следующем примере. Предположим, что у нас есть обучающее множество $X = \{x_1,x_2,\ldots,x_N\}$ и нам нужно получить распределение вероятностей $q(x)$. В самом простом случае, мы используем максимум правдоподобия и полагаем, что $q(x)$ совпадает с эмпирическим распределением $p`(x) = c(x) / N$, где $c(x)$ – число раз, которое встречалось слово x, а N – размерность обучающей выборки. Но в таком случае, мы придём к переобучению модели и не сможем разбирать N-граммы, не представленные в обучающей выборке. Другими словами, необходимо чтобы $q(x)$ соответствовала только наиболее значимым свойствам распределения $p`(x)$.

Для наглядности, приведем примеры. Предположим, что $x = (w_1; w_2)$, где w_1 и w_2 английские слова, которые появляются в некотором большом корпусе английских текстов. Таким образом, задача сводится к оценке частоты появления биграмм, в данном случае, в английском языке. Предположим биграмму, которая не появляется в обучающем множестве – «**PIG DOG**». Имеем, p(**PIG DOG**) = 0 , но интуитивно мы хотим, чтобы p(**PIG DOG**) > 0, т.к. эта биграмма имеет некий шанс появиться. Еще

[69] Такая модель плохо размечает или вообще не размечает последовать из N слов, не представленных в обучающей выборке.

[70] Теоретически, при наличии представительной обучающей выборки, по предельной теореме перейдем от частот появления N-грамм к вероятности их появления, и с помощью принципа максимального правдоподобия получим наиболее оптимальную вероятностную модель автоматической обработки текста. Имеющимися современными средствами на практике такое не достижимо.

[71] Строго говоря, на практике обычно подбирается субоптимальная модель естественного языка.

один пример, предположим, что слово «**Mateo**» может появляться только после слова «**San**» в обучающем корпусе (биграмма «**San Mateo**»). Таким образом, имеем, что p (w_1 Mateo) = 0 для всех $w_1 \neq$ San, но интуитивно, мы хотим, чтобы $p(x) > 0$ для всех w_1, а не только для w_1 = San. Мы хотим, чтобы наша модель максимально хорошо разбирала случаи, представленные в обучающей выборке и, также максимально хорошо разбирала неизвестные случаи. Другими словами, требуется, чтобы система принимала решения при неполной информации.

Прежде чем привести вид сглаженной Марковской модели напомним, что такое оценка максимального правдоподобия для биграммной модели:

$$P_{ML}(w_i \mid w_{i-1}) = \frac{P(w_i \mid w_{i-1})}{P(w_i)} = \frac{c(w_{i-1} \mid w_i)/Nw}{c(w_{i-1})/Nw} = \frac{c(w_{i-1} \mid w_i)}{c(w_{i-1})} \quad (3.4)$$

, где Nw – число слов в обучающей выборке, $c(w_{i-1})$ и $c(w_{i-1} \mid w_i)$ – число раз, которое встречается строка w_{i-1} и $w_{i-1} \mid w_i$ в обучающем корпусе. Нулевая вероятность биграммы может привести к ошибкам при распознавании речи. При снятии неоднозначности с помощью N-граммных моделей высокого порядка можно получить относительно высокий процент ошибок при низком покрытии текста. Таким образом, чтобы получать более точные оценки вероятностей применяются различные виды сглаживания моделей. Пример самого простого сглаживания это прибавить единицу к частоте появления биграммы:

$$P_{ML+1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1} \mid w_i)+1}{c(w_{i-1})+V} \quad (3.5)$$

, где V – размер словаря.

Сглаживание биграммной модели с помощью униграммной выглядит следующим образом:

$$P_{\text{int}\,erp}(w_i \mid w_{i-1}) = \lambda\ P_{ML}(w_i \mid w_{i-1}) + (1-\lambda\ )P_{ML}(w_i) \quad (3.6)$$

, где λ – положительной весовой коэффициент.

В общем виде выражение для сглаженной Марковской модели N-го порядка можно записать в следующей форме:

$$P_{\text{int}\,erp}(w_i \mid w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} P_{ML}(w_i \mid w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}})P_{\text{int}\,erp}(w_i \mid w_{i-n+2}^{i-1}) \quad (3.7)$$

, где P_{ML} – оценка максимального правдоподобия для модели предыдущего порядка (порядка N-1), λ – положительные весовые коэффициенты. Таким образом, сглаженная модель N-го порядка определяется рекурсивно как линейная интерполяция между моделью максимального правдоподобия и сглаженной моделью порядка N-1. Чтобы закончить рекурсию, можно взять в качестве сглаженной модели первого порядка оценку максимального правдоподобия (выражение 3.4), простое сглаживание (выражение 3.5) или предположить равномерное распределение вероятности появления каждого слова:

$$P_{unif}(w_i) = \frac{1}{V} \quad (3.8)$$

Для каждой последовательности слов w_{i-n+1}^{i-1} есть свой набор весовых коэффициентов $\lambda_{w_{i-n+1}^{i-1}}$, вычисляемых оптимальным образом. Последовательности слов,

которые наблюдались много раз в обучающем корпусе и те, которые встречались лишь несколько раз будут иметь разные коэффициенты λ. Чтобы не хранить набор параметров для каждой последовательности слов и не затягивать процесс обучения, N-граммы объединяются в относительно небольшое число классов[72], а коэффициенты λ вычисляются отдельно уже для этих классов. В выражении (3.8) интерполяция может быть нелинейной и тогда общий вид выражения несколько изменится. Есть методы сглаживания вероятностей, которые больше подходят для большой обучающей выборки, а есть которые подходят для относительно маленькой выборки. Для биграммной модели, обученной на большом корпусе метод сглаживания Church-Gale предпочтительнее, в то время как, сглаживание по методу Katz лучше применять для биграмм, полученных с небольшого обучающего корпуса.

Для широкопопулярной триграммной модели, впервые описание эксперимента по применению взвешенной суммы вероятностей моделей первого и второго порядка применил Frederick Jelinek в 1980 году. Сглаженная триграммная модель содержит линейные комбинации триграммных, биграммных и униграммных байесовских вероятностей:

$$Psmooth(w_i \mid w_{i-2} * w_{i-1}) = \lambda_3 * P(w_i \mid w_{i-2} * w_{i-1}) + \lambda_2 * P(w_i \mid w_{i-1}) + \lambda_1 * P(w_i) \qquad (3.9)$$

, где сумма коэффициентов $\lambda_1 + \lambda_2 + \lambda_3 = 1$, причем $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 > 0$. Значения для λ_1, λ_2, λ_3, получены решением системы линейных уравнений. В публикации Чешских исследователей за 1998 год была представлена точность около 93% для обычного HMM тэггера, а с использованием сглаженной триграммной модели точность разметки возросла до более чем 95%.

Еще одним вариантом устранения омонимии является выделение словосочетаний. Входные правила для данного этапа также могут быть порождены автоматически. Для каждого слова на основе большого корпуса текстов (не обязательно даже размеченного) можно посчитать частоту его встречаемости. Далее мы можем посчитать частоту встречаемости, например, пар слов. Для тех пар, вероятность встретить которые выше, чем их совместная вероятность, можно высказать предположение, что они являются словосочетаниями. На практике, если полученная по анализу корпуса частота встречаемости превосходит в два раза теоретическую, то данная пара гарантированно является словосочетанием.

Для словосочетаний большей длины не обязательно учитывать соответствующие комбинации. Дело в том, что пара слов может оказаться началом многословного словосочетания. Таким образом, обнаружив все двухсловные комбинации мы также выделили и кандидатов на многословные. Если сохранить их позиции в тексте, то имеется возможность проверить следующие несколько слов на предмет совпадения. За счет этого снимается необходимость анализировать все возможные комбинации слов, количество которых будет существенно расти с ростом длины комбинаций.

Применение неразмеченного корпуса требует дальнейшего проведения работ по приписыванию отдельным словам их лексических характеристик. Во многом разметка будет определяться стоящими перед системой задачами. В некоторых случаях будет достаточно выделить сами словосочетания, которые потом будут использоваться как единые лексические единицы, например, для определения тематики текста. Для задач синтаксического анализа можно приписать каждому словосочетанию готовое дерево зависимостей. При машинном переводе необходимо

[72] В английском языке это называется bucketing (от слова bucket – корзина).

сопоставить ему эквивалент в другом языке. И если первая задача может быть решена автоматически на размеченном корпусе со снятой омонимией, то вторая и третья обычно решаются вручную.

Приведенные методы позволяют сравнительно быстро получить высокие результаты при условии, что мы имеем доступ к размеченному корпусу текстов со снятой омонимией. В связи с этим создание подобных корпусов является важной научной и практической задачей. Наиболее полным размеченным корпусом в русском языке является Национальный корпус русского языка (www.ruscorpora.ru). Он содержит в себе тексты различной направленности, для которых проведен морфологический анализ, т.е. каждому слову приписаны его лексические характеристики. Для части корпуса снята омонимия, т.е. проделана большая работа по выбору корректных наборов лексических параметров. Вообще, создание представительного корпуса является важной национальной задачей, подхлестывающей развитие исследований в области данного языка, его использование. С практической точки зрения корпусом могут пользоваться, например, переводчики. При наличии сомнений о том, каким образом следует перевести ту или иную фразу, можно задать запрос поисковой системе корпуса и сравнить частоту употреблений имеющихся вариантов. При наличии разметки в корпусе можно проверить корректность фразы с точки зрения синтаксиса. С научной точки зрения корпус дает богатый материал для изучения структуры языка, используемых в нем слов и фраз, построения предложений, истории развития. Для этого корпус должен включать в себя тексты, написанные в разное время в различных жанрах и относящихся к разным предметным областям. Обычно в корпуса включают не только письменные источники, но и живую речь.

Но вернемся к разбору словосочетаний. Многие сочетания слов могут иметь несколько вариантов употребления, в которых они будут или не будут являться словосочетаниями. Так, например, сочетание «так сказать» в различных контекстах трактуется по-разному. В фразе вида «это, так сказать, явление» оно будет являться словосочетанием, определяющим наше отношение или вариативность произношения, но в предложении «Так сказать нельзя!» оно указывает на невозможность произнесения определенного текста. В связи с этим метод требует обязательной ручной доработки результатов, полученных автоматически.

Заметим, что в разобранном случае словосочетание будет трех- или даже четырехсловным, включающим в себя еще одну или две запятые. Та же самая фраза, но с внесенной в нее запятой («Так сказать, нельзя!») будет показывать, что мы пытаемся сформулировать запрет в другом виде («Всё это сделать весьма и весьма затруднительно. Так сказать, нельзя!»). Таким образом, после выделения пар слов, претендующих на роль словосочетаний, необходимо анализировать контекст не только справа, но и слева от них. Кроме того, расширение контекста может привести к получению более однозначных словосочетаний за счет возможной потери более коротких и чаще встречающихся.

§ 3.2. Постморфологический анализ

Предсинтаксический анализ необходим для выделения элементов текста, морфологического анализа этих элементов, разделения сложносоставных элементов на части, объединение простых связанных по смыслу элементов в группы, выделение фрагментов текста, которые могут разбираться самостоятельно. Его название

показывает место предсинтаксического анализа в общей системе: перед синтаксическим анализом. Задачей предсинтаксического анализа является подготовка данных для синтаксического анализа в наиболее удобной форме, максимально облегчающей выполнение задачи последнему.

На вход системы поступает текст. В первую очередь необходимо определить единицы этого текста: абзацы, предложения, отдельные слова и знаки препинания. В отличие от систем машинного перевода, диалоговым системам нет необходимости выделять заголовки, сноски, комментарии, врезки и прочие элементы текста, необходимые для сохранения форматирования. Подобное форматирование текста может понадобиться диалоговой системе только для приобретения новых знаний из существующих текстов, однако создание систем, способных на подобные экзерсизы – дело будущего. Выделение всех описанных элементов текста (как слов, так и врезок) является задачей графематического анализа.

Выделение абзацев в современных редакторах является тривиальной задачей. В них уже существует разметка на абзацы. При полностью текстовом вводе абзацы зачастую отмечаются символом перевода строки. В начале абзаца часто ставят два и более пробела или пробельную строку. В случае, когда каждая строка текста оканчивается символами конца строки, задача выделения абзаца может потребовать специальных знаний о структуре данного текста.

Задача выделения предложений менее тривиальна. Обычно предложение заканчивается точкой, вопросительным или восклицательным знаком, иногда – многоточием. Однако на практике те же знаки препинания используются и для других целей. Точка часто применяется в сокращениях. При этом если сокращение приходится на конец предложения, то ставится только одна точка, относящаяся как к сокращению, так и к концу предложения. Восклицательный и вопросительный знаки часто используются в выразительных вставках в тексте: «… и он, Великая удача!, …», «… он смотрел, А что он мог сделать?,...». Таким образом, знаки препинания не являются стопроцентной гарантией окончания предложения. К счастью, при общении с диалоговыми системами выразительные вставки обычно не используются, однако проблема точки остается. Еще одна проблема на данном этапе – это слова, написанные с большой буквы, после точки. Так в предложении «Мишка очень любит мёд…» без использования прагматики и контекста не совсем понятно о ком идет речь: о ласково называемом медведе или личности по имени Михаил, также называемом уменьшительно-ласкательным именем. В первом случае однозначно рассматривается начало предложения, во втором есть шанс, что предшествующая точка могла стоять после сокращения.

Еще одну проблему представляют собой цитаты и прямая речь, также нечасто используемые при общении с диалоговыми системами. В состав цитаты может входить как несколько слов, так и несколько предложений. Прямая речь обычно содержит некоторый связанный фрагмент текста. В связи с этим и разбирать их лучше как отдельный текст. Однако, например, в английских текстах, прямую речь принято выделять не кавычками, а апострофами. Те же самые апострофы используются для обозначения притяжательного падежа и сокращений: «man’s», «mans’», «it’s», «I’ll», «’cause». В текстах, пытающихся подчеркнуть простонародность речи, сокращения (и как результат апострофы) встречаются очень часто. В связи с этим проблема выделения начала и конца прямой речи стоит остро.

Следует заметить, что притяжательный падеж в английском языке может образовываться не только от существительных, но даже, например, от глаголов: «last gone's daughter». В этом случае не совсем понятно, что делать с притяжательностью: формально глаголы не обладают параметром «притяжательность». Более того, приведенный пример может сам по себе рассматриваться как отдельное предложение: «last gone is (was) daughter». Заметим, что знаки препинания могут встречаться в слове несколько раз: «a-number-one's».

Отдельной проблемой являются тире и дефис. В двухбайтных кодировках они различаются, и системы редактирования текстов имеют возможность ставить их в нужных местах. Однако однобайтные кодировки знают только один символ – минус. В связи с этим в однобайтной кодировке (или в случае ошибок пользователя) отличить «кто-то» от «кто то» в следующих двух предложениях можно только в результате синтаксического анализа: «он знает, что кто-то пришел к другу …» от «он знает кто – то пришел к другу …». Ориентироваться на наличие пробелов в такой ситуации можно, но и в этом случае мы не имеем стопроцентной гарантии.

Для решения этих и других проблем, возникающих при членении текста на составляющие, используется графематический анализ. Для работы графематического анализа нам среди прочего потребуется этап деления сложносоставных слов. Данный этап занимается тем, что делит сложное слово, составленное из нескольких, на составляющие его, например, «сине-зелено-красный». Данная проблема особенно актуальна в немецком и турецком языках. В немецком языке разрешено формировать произвольные сложносоставные слова, если образующие его слова относятся к одному понятию. Часть из них, например «Siecherheitdienst» (Siecherheit + dienst), уже устоялись и считаются одним словом, но большинство подобных слов образуется «на лету» и заносить их в морфологический словарь нет никакой возможности.

Однако примеры из немецкого языка меркнут перед примерами из более редких языков. **Mamihlapinatapai** - слово из яганского языка Племя Яган, (Огненная Земля), указано в книге рекордов Гиннесса в качестве «наиболее сжатого слова» и считается одним из самых трудных для перевода слов. Оно означает «Взгляд между двумя людьми, в котором выражается желание каждого в том, что другой станет инициатором того, чего хотят оба, но ни один не хочет быть первым». Слово состоит из рефлексивного / пассивного префикса ма-(МАМ-перед гласным), корень ihlapi, что значит быть в недоумении, как то, что делать дальше, то stative суффикса-н, достижение Суффикс-ate, и двойной суффикс-apai, который в составе с рефлексивным mam- есть взаимные чувства.

Пример восьмипорядковой деривации в эскимосском языке: *igdlo-ssua-tsia-lior-fi-gssa-liar-qu-gamiuk* (дом-большой-довольно-изготавливать-место-быть-идти-велеть-когда.он.его), «Велев ему пойти туда, где строился довольно большой дом».

К великому счастью, подобные языки обычно не анализируются автоматизированными методами.

В отличие от европейских языков, турецкий язык (и не только он) является агглютинативным. В нем часть информации об употреблении слова добавляется в конец слова в виде аффиксов. Так, например, в одно слово можно сказать kitap_lar_ım_da_ki_ler_i: «те (вин. падеж), что лежат на моих книгах». За счет добавления аффиксов у одного существительного может появиться несколько тысяч словоформ, хранить которые в морфологическом словаре также будет просто невозможно, так как аффиксы будут добавляться после любой имеющейся

словоформы, не содержащей аффиксов, хотя и в строго определенном порядке. В связи с этим было бы проще ввести ряд дополнительных параметров и «откусив» соответствующий аффикс, добавить оставшейся словоформе параметр с заданным для данного аффикса значением.

Остальные сложные слова должны разделяться по определенным правилам. В противном случае только слово «поляк» будет иметь более 20 вариантов деления. Так, все прилагательные, кроме последнего, в примере, приведенном для русского языка («сине-зелено-красный»), должны находиться в краткой форме (и как минимум присутствовать в словаре).

Таким образом, этап деления сложносоставных слов имеет довольно сложную структуру и управляется языкозависимыми правилами. Работа графематического анализа будет выглядеть следующим образом. Сперва, графематический анализ по заданным критериям выделяет абзацы. Далее выделяется строка до первого разделителя (пробела, перевода строки, иного знака препинания). Если строка состоит из одних цифр, то она помечается частью речи «числительное» и отправляется в промежуточный массив. В противном случае строка подается на этап деления сложносоставных слов. Если после выделенной строки стоит единственный знак препинания, разрешенный для присутствия в словах, представленных в морфологическом словаре, то мы выделяем следующую строку, объединяем с предыдущей, вновь проверяем на наличие разрешенного разделителя до тех пор, пока такой разделитель присутствует. После этого подаем полученную строку на этап деления сложносоставных слов. Если слово не представлено в морфологическом словаре и не может быть разделено, этап должен вернуть ошибку. При этом в массив выделенных слов будет помещен специальным образом помеченный кортеж, содержащий выделенную строку, с неизвестным словом. В противном случае мы возвращаем выделенный набор слов и помещаем его в массив выделенных слов. Выделенные знаки препинания также подаются в массив выделенных слов.

Для борьбы с неоднозначной расстановкой точек можно вводить правила анализа сокращений. Можно выделить несколько видов сокращений. Фиксированные сокращения не изменяют своей формы записи. Это, к примеру, «и т.д.», «и т.п.», «т.о.». Найдя подобные обороты, мы можем заменить их отдельными словами. При этом приведенные примеры лучше заменить одним сложным словом, так как они являются неразрывными неизменяемыми словосочетаниями и обрабатываются как одна лексическая единица. Заменив их на несколько слов, мы можем получить неоднозначность анализа структуры предложения.

Изменяемые сокращения представляют собой сокращение и некоторую обязательную, но произвольно изменяющуюся часть. Это, например, сокращения в датах, именах, названии населенных пунктов и географических мест: «о. Врангеля», «г. Москва», «1904 г.», «г-н Герострат», «А.С. Пушкин». Часть из сокращений имеет смысл развернуть в полное слово. При этом может возникнуть проблема с приписыванием параметров развертываемому слову. «1904 г.» может обозначать «год», «году», «года» и т.д.

Такие сокращения, как «и т.д.», «1904 г.» и прочие, могут встречаться в конце предложения. При этом точка будет обозначать как конец предложения, так и точку в сокращении (для примера см. последнее предложение предыдущего абзаца). В связи с этим следует ввести дополнительный параметр в правила – может ли данное сокращение заканчивать предложение.

Устранив все «лишние» точки и расставив маркеры о возможном окончании предложений, мы можем считать, что все точки, восклицательные и вопросительные знаки или маркеры, после которых идет слово с большой буквы, заканчивают предложение.

По окончании выделения отдельных слов имеет смысл провести свертку части таких слов. Это необходимо сделать, чтобы облегчить работу синтаксическому анализу.

Во-первых, можно свернуть числительные, написанные в виде слов, превратив их в числовое написание. Во многих языках мира словесное написание числительных выглядит следующим образом. Изначально присваиваем текущей и итоговой сумме нулевые значения. При этом итоговая сумма будет хранить окончательный результат, а текущая сумма – промежуточное значение при подсчете количества сотен, тысяч, миллионов и т.д. Если следующее за текущим числительным слово обозначает числительное меньшее, чем текущее, то его следует прибавить к текущей сумме. Если следующее слово обозначает числительное большее, чем текущее, то текущую сумму необходимо умножить на это большее числительное и прибавить к итоговой сумме. Текущая сумма при этом обнуляется. По окончании числа текущая сумма прибавляется к итоговой.

Сто	пятнадцать	тысяч	двести	один	
100	+15	*1000	200	+1	=115000+201=115201

Для дробных числительных следует различать рациональные и десятичные дроби. Для дробных числительных основная проблема состоит в том, что они состоят из двух чисел, первое из которых выражается счетным числительным, а второе — порядковым. При этом в общем случае довольно сложно определить их границу. Основным критерием может служить тот факт, что первое число может быть меньше второго (семь восьмых, пятьсот шесть девятьсот одиннадцатых). Однако когда числитель больше знаменателя, задача может не иметь однозначного решения (один миллион двести восемнадцать тысяч трехсотых может быть эквивалентно 1000000/218300 и 1218000/300). К счастью, большинство подобных дробей легко сокращаются и приводятся к более очевидному варианту.

Для десятичных дробей на этапе анализа числительных необходимо ввести разделитель, показывающий место, с которого начинается дробная часть (целых в русском языке, point в английском). Сама дробная часть в различных языках выражается по-разному. Так, в русском языке после разделителя будет идти рациональная дробь, вторая часть которой выражается единственным словом (десятых, тысячных, стомиллионных. ...). Также возможна такая же запись дроби, как и в английском языке: после разделителя перечисляются цифры в том порядке, в котором они идут после запятой (точки). Но такая форма записи редко встречается на письме, особенно в человеко-машинном диалоге.

Для немецкого языка указанный алгоритм работать не будет в связи с тем, что порядок слов в немецких числительных несколько иной. Так, единицы в нем ставятся перед десятками, например, «fünf und zwanzig» - «пять и двадцать». Аналогично и в английском языке могут встречаться конструкции, включающие в себя слово «and»: «twenty and five». В связи с этим в алгоритм необходимо вводить слова, обозначающие сложение и даже умножение текущей и последующей сумм. Следует

помнить, что подобные слова сами могут вносить некоторую неоднозначность. Так, например, во фразе «Twenty and five vehicle» в качестве ответа на вопрос «How many peoples goes?» будет иметься в виду двадцать человек и пять машин.

Кроме того, слово «один» (и аналогичные ему в других языках) имеет собственную семантику. С одной стороны, оно может употребляться для обозначения некоторого объекта: «Дай одну», «В одном зоопарке, не помню каком,…». С другой стороны, оно может опускаться в числительных: обычно говориться «тысяча пять», а не «одна тысяча пять», хотя второй вариант также употребим, особенно в формальных документах.

Аналогично слова «десяток», «сотня» и «тысяча» могут выступать в роли существительных: «Во главе пяти сотен воинов с осадными орудиями он подошел к стенам города и начал правильную осаду». В данном примере слово «сотня» означает воинское подразделение, а не количество людей.

При обработке числительных следует помнить, что в тексте может встречаться запись групп числительных: телефоны, IP-адреса и т.п., записанные не числами, а словами. Т.е. по ходу выполнения алгоритма необходимо отслеживать, что следующее число не принадлежит одной группе (десятки, сотни…) с предыдущим и текущая сумма не больше итоговой. Например: «Сто девяносто два сто шестьдесят восемь ноль один». Однако подобная запись может обозначать как IP-адрес 192.168.0.1, так и число 19216801, хотя наиболее вероятным с точки зрения прагматики в такой ситуации представляется первый вариант. Второй вариант более вероятен при группировке цифр тройками для выделения тысяч, миллионов и т.д., например, 108.891.452. В этом примере не может быть записан IP-адрес, так как он должен содержать в себе четыре группы цифр от 0 до 255. Приведенная запись чисел характерна для финансовых расчетов и может предполагать до или после себя обозначения денежной единицы ($108.891.452 или 108.891.452 руб.).

Семантически нагруженные группы цифр: даты, номера телефонов, IP-адреса и т.д. также имеет смысл объединять в одно слово. Их можно объединять по шаблонам. Например, xx/xx/xxxx или xx.xx.xxxx для дат, xxx-xxxx, +x(xxx)xxx-xxxx, xxx-xx-xx для телефонов и т.д. Это также уменьшит количество слов, поступающих на синтаксический анализ и, как следствие, упростит его работу.

Превращая слова в цифры, следует иметь в виду, что слова могут подчеркивать роль цифр в слове. Они могут играть роль порядковых или счетных числительных. В связи с этим следует приписывать различную часть речи в зависимости от того, какое слово идет последним в группе. Так, «сто один» будет счетным числительным, а «сто первый» - порядковым.

Еще одной задачей предсинтаксического анализа является обработка словосочетаний. Можно выделить следующие виды словосочетаний: неразрывные неизменяемые, неразрывные изменяемые и разрывные. Неразрывные неизменяемые словосочетания состоят из одних и тех же словоформ, идущих одна за другой. Например, «таким образом», «так сказать», «не взирая на» и т.д. В состав неразрывных неизменяемых словосочетаний могут входить и знаки препинания: ср. «для того чтобы» в начале предложения и «для того, чтобы» в середине. Для поиска подобных словосочетаний необходимо просто проанализировать входной текст. Найденные словосочетания имеет смысл обрабатывать дальше как единую словоформу.

Кроме того, словосочетания можно разделить на открытые и закрытые. В закрытых словосочетаниях отдельные слова теряют собственный смысл и могут трактоваться только в составе словосочетания. При этом слова в открытых словосочетаниях сохраняют все лексические связи и, как следствие, могут подчинять себе другие слова. При этом может происходить разрыв словосочетания.

Следует опасаться неоднозначности у словосочетаний. Так, например, сочетание «так сказать» может встретиться в фразе вида «так сказать нельзя», хотя как вводная конструкция оно должно быть выделено запятыми.

Неразрывные изменяемые словосочетания состоят из идущих подряд словоформ, образованных от фиксированных нормальных форм, но в зависимости от контекста обладающих различными параметрами. Например, «бить баклуши» («бил баклуши», …), «наивная модель мира», «искусственный интеллект» и т.д. Обычно слова в неразрывном изменяемом словосочетании согласуются. При желании можно выделить главное и зависимые слова, однако свертка подобных словосочетаний позволяет, как и в предыдущем случае, сократить количество синтаксических единиц и тем самым упростить задачу синтаксического анализа. Заметим, что грань между изменяемым словосочетанием и простой синтаксической конструкцией довольно тонка. Так, если «Черный квадрат» можно отнести к словосочетаниям, то «красный квадрат» таковым являться не будет. Дело в том, что словосочетания имеет смысл выделять лишь в тех случаях, когда сдвигается семантика или прагматика предметов, указанных в словосочетании. «Черный квадрат» обозначает название картины и является именем собственным (кстати, в такой роли он должен быть обрамлен кавычками), а красный квадрат является просто квадратом, одно из свойств которого – цвет – обладает значением «красный». Аналогично, если искусственный интеллект является специализированной отраслью знаний, то, например, обширный интеллект – это интеллект, обладающий свойством обширности. Как всегда и здесь не обходится без неоднозначности. Так, в фразе «несколько искусственный интеллект» будет иметься в виду интеллект, обладающий свойством заполненности искусственно придуманными знаниями. Приведенный пример и сам является искусственным, однако на практике подобная ситуация встречается довольно широко.

Для анализа неразрывных неизменяемых словосочетаний необходимо предварительно провести морфологический анализ. Далее мы ищем требуемые словоформы в заданном порядке. При этом может быть необходимо проверить согласование слов по параметрам: заданные параметры должны обладать одними и теми же значениями. Как уже упоминалось, прилагательное и существительное в русском языке согласуются по ряду параметров. Следовательно, для словосочетаний, составленных из существительного и подчиненных ему прилагательных, необходимо проверить подобное согласование.

Разрывные словосочетания – это связанные слова, между которыми могут вклиниваться другие слова. Как и в предыдущем случае, связка слов дает несколько иное значение, чем просто сумма значений слов. В случае с разрывными словосочетаниям связь между словами либо очевидна, либо используемый вид согласования не требует того, чтобы слова стояли рядом. Более того, кроме подчиненного слова, образующего словосочетание, к главному слову могут присоединяться и другие зависимые слова, являющиеся его неоднородными членами. В итоге все подчиненные члены имеют право идти вперемешку. Например, «отправка самолетов», «прибытие самолетов» → «отправка и прибытие самолетов».

Так как слова в разрывных словосочетаниях могут быть разнесены по предложению, то сперва требуется провести синтаксический анализ предложения. Следовательно, работа с разрывными словосочетаниями не может быть отнесена к предсинтаксическому анализу.

§ 3.3. Синтаксическая сегментация

Еще до проведения синтаксического анализа есть возможность выдвинуть некоторые предположения о структуре разбираемого предложения, выделить его фрагменты (сегменты), которые можно разобрать независимым образом. Дальнейший синтаксический анализ будет опираться на эти предположения и получит возможность сразу отбросить часть вариантов. Для использования этих возможностей вводится этап синтаксической сегментации.

Первой задачей *синтаксической сегментации* является уменьшение количества омонимов, соответствующих каждой словоформе. Так, например, если слово может являться как наречием, так и прилагательным, то следует проанализировать следующее за ним слово. Если оно является однозначным глаголом, то слово будет наречием. Если следующее слово является однозначным прилагательным или существительным, то данное слово будет прилагательным. Подобных правил достаточно много, но очень часто они могут служить лишь предположениями, так как существует альтернативный вариант прочтения данного фрагмента. Однако зачастую даже выделение нескольких альтернатив помогает сократить количество возможных вариантов.

Вторая задача – выделение синтаксических конструкций. Например, мы можем выделить начало сложноподчиненного предложения, обнаружив ключевые слова «, который», «, потому что», «, когда» и т.д. Можно выделить деепричастные обороты, найдя место, где за существительным после запятой идет деепричастие. Кроме того, можно попытаться определить связность и подчинение фрагментов. Так, например, во фразе «...**когда на столе**, покрытом скатертью**, они расставили тарелки**...» жирным выделен единый фрагмент, одному из слов которого (стол) подчинен вставленный в него в фрагмент.

Для поиска фрагментов нам потребуется понятие шаблона поиска. Под шаблоном поиска слова будем понимать кортеж <N,S,P>, где N – нормальная форма искомого слова, S – часть речи и P={p} множество искомых параметров искомого слова. Нормальная форма слова может представлять собой строку с искомой нормальной формой, либо пустую строку, если нормальная форма нас не интересует. Аналогично описывается и часть речи. Множество параметров может быть пустым, если параметры при поиске нас не интересуют. Искомые параметры, в отличие от обычных параметров, будут содержать дополнительный флажок, указывающий на тип поиска. Предполагаются следующие типы поиска:

- точный, когда сравнивается как имя, так и значение параметра;
- по имени, когда проверяется наличие параметра у данного слова вне зависимости от его значения;
- совпадающий, когда значение параметра должно совпадать со значениями таких же параметров у других шаблонов;
- несовпадающий, когда значение параметра может принимать любое значение, кроме указанного.

Под шаблоном поиска фрагмента будем понимать упорядоченное множество шаблонов поиска слов.

Приведем пример шаблона поиска фрагмента.

<‘’,’прилагательное’, {[‘род’,совп.,’’],[‘число’,совп.,’’],[‘падеж’,точн.,’им’]}>
<‘’,’существительное’, {[‘род’,совп.,’’],[‘число’,совп.,’’],[‘падеж’,совп.,’’]}>

Здесь мы пытаемся найти прилагательное и следующее за ним существительное вне зависимости от их нормальной формы, при этом у найденных слов должны совпадать род, число и падеж, причем падеж должен быть именительным.

Как уже упоминалось, нам может потребоваться найти слово, имеющее единственное значение. В связи с этим в шаблоне поиска слова необходимо добавить флаг, который будет показывать, должно ли значение быть единственным.

Следует заметить, что распространенной является ситуация, когда к одному и тому же месту в предложении подходят сразу несколько имеющихся шаблонов поиска фрагментов. При этом шаблоны могут пересекаться или один может включать в себя другой. Подобную ситуацию следует учитывать при работе с шаблонами.

Шаблон поиска фрагмента только ищет необходимый нам фрагмент. Второй частью задачи является преобразование найденного фрагмента. Для этого потребуется шаблон формирования фрагмента. Данный шаблон будет показывать, какие найденные слова требуется включить в выходной фрагмент и какие новые слова в него вставить. Так, например, нам может потребоваться найти несколько слов и слить их в одну лексическую единицу. В этом случае шаблон формирования фрагмента будет включать единственный элемент, полностью формирующий новое слово.

Элементы шаблона формирования фрагмента – шаблоны формирования слов – будут показывать, откуда следует взять нормальную форму слова и его часть речи, параметры слова (ввести новое, взять из слова входного предложения, из какого именно слова). Кроме того, нам могут потребоваться метки для синтаксического анализа. Формат меток будет существенно зависеть от методики проведения синтаксического анализа. Не нарушая общности рассуждений, возьмем метки начала и конца правил синтаксического анализа, разбирающих некоторые синтаксические конструкции.

Правило синтаксической сегментации будет состоять из шаблона поиска фрагмента, шаблона формирования фрагмента и списка исключений. Исключения показывают, какие правила необходимо исключить из исполнения в случае, если их шаблоны сравнятся с уже найденным фрагментом. Такая ситуация возможна, например, когда у нас есть два шаблона, один из которых описывает более частную ситуацию.

Синтаксическая сегментация будет работать по следующему алгоритму. Во входном предложении ищутся места, попадающие под хранимые в правилах шаблоны поиска фрагментов. Далее анализируется список исключений, и при нахождении пересекающихся исключаемых правил мы выбрасываем их из рассмотрения. Если после этого остались пересекающиеся фрагменты, необходимо создать несколько копий входных данных и разнести найденные шаблоны по копиям таким образом, чтобы исключить пересечения. При этом непересекающиеся шаблоны должны быть включены во все копии. Далее для всех полученных копий по шаблонам

формирования фрагмента производится формирование новых слов и замена слов, подошедших под шаблон, на вновь сформированные.

Приведем пример работы синтаксической сегментации.

Пусть у нас имеется следующий набор правил.

<«»; прил.; род +, число +, падеж +><«»; однозначно сущ.; род +, число +, падеж +> ⇒ <однозначно прилагательное><не изменяется>

<«,»; знак; ><«который»; мест.;> ⇒ с первого слова начинается сложноподчиненное предложение

Пусть на вход поступает следующее предложение: «Я увидел человека с красным лицом, который быстро бежал по улице».

Слово «красный» может быть как прилагательным, так и существительным («Борьба красных и белых в ходе гражданской войны…»). По первому правилу вариант существительного будет отсеян. По второму правилу мы заранее определим, что фрагмент «…, который бежал по улице» является сложноподчиненным предложением, и не будем разбирать другие варианты, например, перечисление «… лицом, … улице».

Уровень омонимии может быть снижен на этапе синтаксической сегментации за счет поиска часто употребляющихся конструкций, оборотов и словосочетаний. Так, например, в научной литературе часто употребляются такие конструкции, как «Под … будем понимать …», «Допустим, что …» и т.д. При этом в первой фразе определяемое слово должно находиться в творительном падеже, а основное слово определения и согласуемые с ним слова (обычно прилагательные и местоимения) – в винительном. Найдя подобную конструкцию можно с большой долей вероятности утверждать, что определяемое слово является существительным, а определяющая конструкция отвечает конкретным требованиям. Исходя из этого, можно отбросить часть омонимов, не подходящих под указанный шаблон.

Используя подобные положения, на ВМК МГУ был разработан язык лексико-синтаксических шаблонов (http://lspl.ru/), служащий, правда, для несколько иных целей. Язык шаблонов позволяет на основе отдельных слов и отношений между ними описывать целые конструкции. Отдельное слово описывается следующим образом: часть речи <нормальная форма; список параметров через запятую>. Параметры записываются в формате имя=значение. При необходимости нормальная форма и параметры могут опускаться. Приведем несколько примеров записи отдельных слов.

A<важный; c=nom, g=fem> - описывает формы «важная» и «важна», так как показатель формы не указан.

A<важный> - описывает все формы слова «важный».

V<t=pres, p=3, n=plur> - описывает любой глагол в настоящем времени, третьем лице множественного числа.

Каждый шаблон обладает именем и в него может входить несколько шаблонов для слов. При совпадении части речи у слов шаблона проводится их нумерация.

N1 N2 – два последовательно идущих произвольных прилагательных.

Полная итерация (ноль и более употреблений) некоторой конструкции обозначается при помощи фигурных скобок. При этом имеется возможность указать в треугольных скобках количество повторений этой конструкции.

{A}<1,3> N - от одного до трех прилагательных, после которых идет существительное.

Конструкция, заключенная в квадратные скобки, считается факультативной.

{A} N ["не"] V – существительное, перед которым может идти произвольное количество прилагательных, за которым следует глагол, перед которым может стоять «не».

Шаблон позволяет задавать альтернативы одной конструкции с использованием символа |. Также язык позволяет задавать согласование параметров между отдельными словами конструкции.

A<тяжелый> N <A.g=N.g, A.n=N.n, A.c=N.c>

A<тяжелый> N <A=N> - прилагательное «тяжелый» в произвольной форме, за которым идет существительное, согласующееся с данным прилагательным.

Для шаблонов задается главное слово, от которого берутся все лексические параметры конструкции. Имя конструкции может быть использовано в прочих шаблонах.

AP = A|Pa – шаблон AP задает последовательность из одного прилагательного или причастия.

AN = {AP} N <AP=N> (N) – шаблон AN задает полную итерацию AP (прилагательных или причастий), за которой следует существительное, являющееся главным словом конструкции, согласующееся с первой конструкцией.

ACT = AN V <AN=V> - шаблон ACT задает последовательность из конструкции AN и согласующего с ней глагола.

NP = AN1 {AN2<c=gen>} (AN1) – существительное с подчиненной ей группой существительных.

DT = NP1<c=acc> ["мы"] «назовем» NP2<c=ins> <NP1.n = NP2.n> - конструкция вида «… мы назовем …».

Используя подобные шаблоны можно как проводить упрощенный синтаксический анализ, так и снимать омонимию в определенных конструкциях.

ЧАСТЬ IV. ИНСТРУМЕНТАЛЬНЫЕ СИСТЕМЫ РАЗРАБОТКИ ПРИЛОЖЕНИЙ ПО АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ *(НОСКОВ А.А.)*

Глава 1. Введение

В настоящее время в связи с быстрым ростом количества текстовой информации, в том числе в сети Интернет, возникает необходимость в быстрой разработке прикладных программных систем (приложений) для автоматической или автоматизированной обработки текстов на естественном языке (ЕЯ-текстов). Примерами такой обработки является сбор и фильтрация данных из различных источников, извлечение знаний [33], реферирование, аннотирование и т.п.

Разработка приложений для решения такого рода задач имеет ряд сложностей, в первую очередь - необходимость интеграции большого числа программных компонентов, реализующих алгоритмы обработки ЕЯ-текста, работающие на различных его уровнях (обработка слов, предложений, абзацев и т.п.).

Например, задача выделения сущностей и отношений [27] из ЕЯ-текстов требует графематического анализа (разбиения текста на слова и предложения), морфологического анализа, выделения определенных конструкций естественного языка на синтаксическом уровне, последующей семантической обработки результатов, а также преобразования текстовых данных между различными форматами и представлениями.

Другую сложность составляет организация внутреннего представления обрабатываемого ЕЯ-текста, поскольку оно должно поддерживать две его важные особенности: естественную иерархию составляющих текст элементов (абзацев, предложений, слов) и свойственную естественному языку множественность интерпретаций на каждом уровне (например, словоформа “мыла” на морфологическом уровне имеет 2 интерпретации: существительное или как глагол).

Кроме того, сложность обеспечивается тем, что для достижения успешного выполнения поставленных задач системы по обработке ЕЯ-текстов должны решать следующие проблемы [18]:

1. Точность (accuracy) — алгоритмы обработки естественного языка не гарантируют получение только корректных результатов, так что дизайн системы должен учитывать это и предоставлять возможности для повышения точности за счет, например, отката к использованию других алгоритмов;
2. Эффективность (efficiency) — в большинстве исследований вопросы эффективности практически рассматриваются как Ђдетали реализацииЛ, однако при интерактивной работе задержки в ответе системы более чем на несколько секунд могут быть неприемлемыми для пользователей;
3. Продуктивность (productivity) — усилия, затрачиваемые на разработку ЕЯ-приложений часто выше, чем для многих других областей разработки ПО в связи с отсутствием возможности знаний о существующих ресурсах, невозможности интегрировать сторонние компоненты;
4. Гибкость (flexibility) — как и другое ПО ЕЯ-системы должны быть гибкими, они должны поддерживать различные форматы, источники данных и

возможность применения для различных задач;

5. Устойчивость (robustness) — системы должны сохранять работоспособность в различных условиях, содержать обработку различных исключительных ситуаций и провалов алгоритмов;
6. Масштабируемость (scalability) — для возможности обработки больших объемов данных системы должны иметь возможность увеличения производительности например, за счет использования распределенной схемы;
7. Многомодальность (multimodality) — различные лингвистические компоненты используют при анализе различные аспекты информации, соответственно система должна поддерживать возможность доступа к ним;
8. Разреженность данных (data sparseness) — поскольку многие алгоритмы обработки ЕЯ-текстов требуют наличия подготовленных данных для обучения, их построение (особенно для многоязычных случаев) может представлять проблему;
9. Многоязычность (multilinguality) — пользователи говорят на разных языках и, соответственно, как в ПО необходима интернационализация, поддержка соответствующих кодировок, однако в для ЕЯ-систем ситуация значительно усложняется тем, что для каждого языка необходимы свои наборы тренировочных данных, правил обработки, а кроме того, могут различаться используемые алгоритмы.

Налицо необходимость в инструментах, позволяющих упростить создание приложений по обработке текстов на естественном языке (ЕЯ-приложений) за счет повторного использования тех или иных программных компонентов и сборки приложения из них - это неоднократно обсуждается в различных работах, в том числе [18]). Эти же инструменты должны предоставлять гибкую модель данных, позволяющую учесть указанные свойства ЕЯ-текстов.

В дальнейшем рассматриваются такие системы и различные аспекты их построения. В частности, в главе 2 рассматриваются различные аспекты, связанные с повторным использованием тех или иных средств, в 3 – подходы к представлению данных в системах, в 4 – архитектурные решения, используемые при их построении. Наконец, в 5 рассматриваются примеры таких систем, относящиеся к различным направлениям развития.

Глава 2. Программные средства лингвистической обработки

Повторное использование тех или иных средств играет важную роль при разработке программного обеспечения вообще. Высокий уровень повторного использования позволяет уменьшить количество ресурсов, необходимых для построения успешного ПО и обеспечивает общую концептуальную базу для построения различных систем, что положительно сказывается на возможности их совместного использования.

В случае ЕЯ-систем уровень повторного использования достаточно низок [18], чему способствует сложность написания компонентов, отсутствие стандартизованных интерфейсов, а также несовместимость различных подходов и компонентов.

При этом в ЕЯ-системах есть несколько классов средств которые могут быть успешно повторно использованы:

- Данные, необходимые для работы прикладной системы - обучающие выборки, корпуса текстов, словари, онтологии;

- Концептуальные элементы системы - алгоритмы обработки текстовых данных, форматы их представления, а также шаблоны проектирования систем;
- Программный код, реализующий алгоритмы обработки данных и преобразование данных между различными форматами, а также связующий код, обеспечивающий взаимодействие компонентов и обмен данными между ними.

При движении от данных к программному коду сложность повторного использования средств увеличивается, однако вместе с тем возрастает положительный эффект от их повторного использования.

В качестве примеров средств первой категории можно расмотреть языковые корпуса, такие как Национальный Корпус Русского Языка (НКРЯ) [35] (http://ruscorpora.ru/index.html) или Хельсинский Аннотированный Корпус (ХАНКО) [31] (http://www.ling.helsinki.fi/projects/hanco/). На сайтах этих корпусов предоставлены средства для поиска размеченных текстов по заданным параметрам. Кроме того, для НКРЯ предоставляется случайная выборка предложений объемом 180 тыс. словоупотреблений, которая может быть загружена со страницы http://ruscorpora.ru/corpora-usage.html.

Программные средства могут поддерживать повторное использование средств, используемых при обработке текстов различными путями, в соответствии с чем их можно разбить на три основных группы:

- Системы, разработанные для решения конкретных задач. Эти системы, как правило позволяют переиспользование спецификаций и элементов проектирования (алгоритмов, моделей и подходов), однако переиспользование кода в них как правило достаточно мало;
- Программные библиотеки, реализующие те или иные алгоритмы обработки данных. Они позволяют повторно использовать код программных компонентов, но часто ограничивают разработчика приложения в выборе модели представления данных.
- Программные среды, предназначенные для разработки некоторого класса приложений. В их случае возможно повторное использование как программных компонентов и моделей данных, так и кода, обеспечивающего взаимодействие различных компонентов. Однако использование программных сред часто накладывает ограничения на архитектуру приложения и модель данных, а также иногда на класс используемых алгоритмов.

Таким образом, при движении от системы к фреймворку возрастает количество переиспользуемых сущностей, но также возрастают и ограничения, накладываемые на разрабатываемое приложение.

Рассмотрим некоторые примеры программных библиотек:

Для европейских языков наиболее известны программные библиотеки OpenNLP(http://incubator.apache.org/opennlp/) и LingPipe(http://alias-i.com/lingpipe/), предоставляющие наборы компонентов для разбиения на лексемы, выделения предложений, синтаксического анализа, определения языка, выделения сущностей и других.

Для русского языка одной из наиболее широких библиотек для лингвистической обработки является набор модулей на сайте aot.ru [32], который включает графематический, морфологический, синтаксический и семантический анализаторы.

Другими известными морфологическими анализаторами являются Mystem [26] (http://company.yandex.ru/technology/mystem/), разработанный в компании Яндекс и PyMorphy (http://packages.python.org/pymorphy/). Оба из них доступны для загрузки и использования в научных целях.

В качестве примера еще одной программной библиотеки стоит рассмотреть модули поддержки языка лексико-синтаксических шаблонов LSPL [34], позволяющие выделять из текстов на русском языке синтаксические конструкции по их формальному описанию.

Далее будут рассмотрены различные программные среды.

Глава 3. Представление лингвистических данных

Все ЕЯ-системы так или иначе сталкиваются с проблемой представления лингвистической информации. Они вынуждены тем или иным образом представлять, хранить и интерпретировать обрабатываемые в системе данные. В данном случае представление данных включает в себя средства и формализмы, используемых для представления данных, методов их хранения в процессе обработки и вовне системы, а также интепретацию данных системой, в частности, ограничения, накладываемые на онтологию, объекты которой могут быть представлены в системе.

§ 3.1. Подходы к представлению данных

В соответствии с [12] подходы к представлению данными в ЕЯ-системах могут быть распределены по четырем категориям:

- **Основанные на разметке** (markup-based), в которых дополнительная информация хранится непосредственно в тексте в форме дополнительной разметки (например, на SGML или XML); Примерами систем использующих такой подход могут служить LT-NSL, Wraetlic и др. (см. в §4.5.1);
- **Основанные на аннотациях** (annotation-based), в которых информация хранится отдельно и содержит ссылки на исходный текст; Это в первую очередь системы, использующие идеи проекта TIPSTER: GATE, Ellogon и др. (см. в §4.5.2);
- **Основанные на абстракции** (abstraction-based), в которых текст хранится только как часть некоторой структуры данных, которая представляет всю информацию в некотором основанном на конкретном формализме виде;
- Отсутствие ограничений на представление данных.

Подходы, основанные на абстракции обычно основываются на формализмах, описывающих организацию представляемой информации в унифицированной форме, не содержащей непосредственно текста. Их использование ограничивают переиспользование компонентов не совместимых с определенным формализмом. Например, ALEP, основанный на нейтральном к грамматике формализме HPSG не позволяет использовать множество компонентов, не совместимых с этим формализмом, например, основанных на статистических методах. Аналогично, никакой другой формализм не совместим со всем множеством существующих компонентов и алгоритмов.

§ 3.2. Лингвистическая разметка

Лингвистическая разметка представляет из себя задание информации о лингвистических единицах непосредственно в тексте в форме разметки на специальном языке (например, SGML или XML). Ниже приведен пример такой разметки:

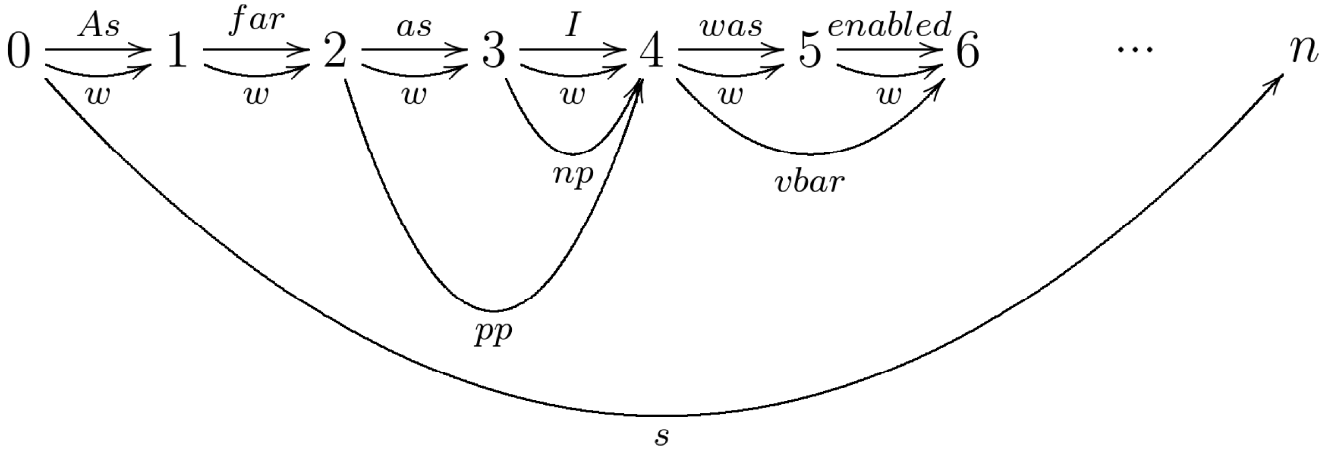
```
<s id="6293">
 <w c="w" pos="IN" id="624">As</w>
 <w c="w" pos="RB" id="625">far</w>
 <pp id="21236">
  <w c="w" pos="IN" id="626" head="yes">as</w>
  <np number="singular" person="1" id="627">
   <w c="w" pos="PRP" head="yes" id="628">I</w>
  </np>
 </pp>
 <vbar voice="passive" time="past" id="629" args="+6302">
  <w c="w" pos="VBD" stem="be" head="yes" id="630">was</w>
  <w c="w" pos="VBN" stem="enable" id="631">enabled</w>
 </vbar>
```

Использование специальной разметки для представления лингвистической информации в документе выглядят весьма удобными для задач обработки естественного языка. Они позволяют легко анализировать результаты обработки пользователем или разработчиком, игнорировать не относящуюся к задаче разметку и использовать стандартные инструменты для обработки. Но основная проблема этого подхода – представление сложных и пересекающихся структур. Например, такие структуры могут возникнуть вследствие неоднозначности анализа текста на одном из этапов обработки. Представление таких структур значительно усложняет используемую схему разметки, что сводит на нет преимущества от использования стандартных программ и анализируемости человеком.

§ 3.3. Лингвистические аннотации

Аннотации представляют из себя информацию о лингвистических единицах, хранящуюся отдельно от текста и ссылающуюся на его участки. Каждая аннотация как правило содержит тип и набор атрибутов, описывающих ее характеристики. Ниже приведен пример набора аннотаций, соответствующего вышеприведенной разметке:

Использование аннотаций не имеет недостатков, связанных с представлением пересекающихся лингвистических единиц и ограничений на поверхностную модель обработки. Кроме того, информация из аннотаций может быть преобразована в/из разметку, что, например, успешно реализовано в системе GATE. Однако, стоит заметить, что как подходы, привязанные к конкретному формализму ограничивают использование компонентов, основанных на других формализмах, подходы, основанные на аннотациях усложняют интеграцию компонентов глубокого анализа, требующих некоторых более сложных структур данных. Это ограничение приводит к необходимости разработки специальных средств интеграции, например как в системе Whiteboard.



Начиная с проекта TIPSTER (см. в §4.5.2) аннотации использовались достаточно широко в проектах по извлечению информации, однако используемые для их представления форматы были несовместимы, что привело к попыткам создания формализмов и стандартов, позволяющих обобщить различные подходы к их представлению.

Графы аннотаций

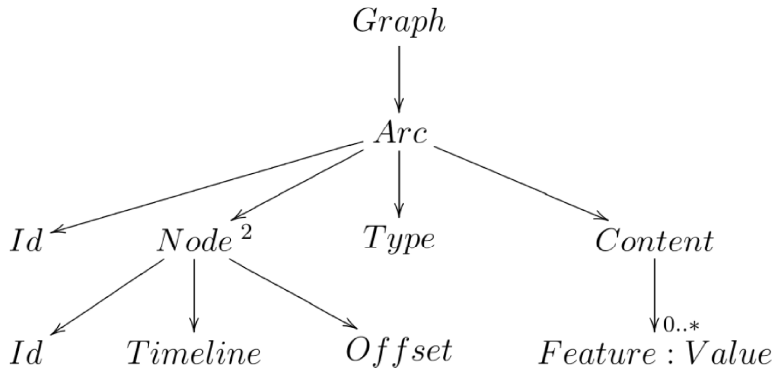
Графы аннотаций [4; 5] представляют из себя формальную систему, предложенную в 1999 году для унификации различных форматов представления аннотаций. Система позволяет аннотировать различные сигналы (не обязательно текстовые) в рамках одного набора данных, имеющие независимые шкалы времени.

Основу системы составляет понятие временной шкалы, соответствующей сигналу на которой могут быть отмечены позиции. На каждой временной шкале отмечается некоторое множество узлов, представляющих неаннотируемые участки. Каждая аннотация имеет начальный и конечный узел и аннотирует промежуток между ними.

Таким образом, аннотации и узлы образуют ориентированный граф в котором аннотации можно направленны в сторону узлов с большей позицией. Такой граф является ациклическими по построению, кроме того, система требует, чтобы в графе аннотаций не существовало узлов, не инцидентных каким-либо аннотациям.

При этом, каждая аннотация имеет тип и содержимое, представленное в виде набора пар имя-значение.

Взаимосвязи объектов, формирующих систему можно изобразить следующим образом:

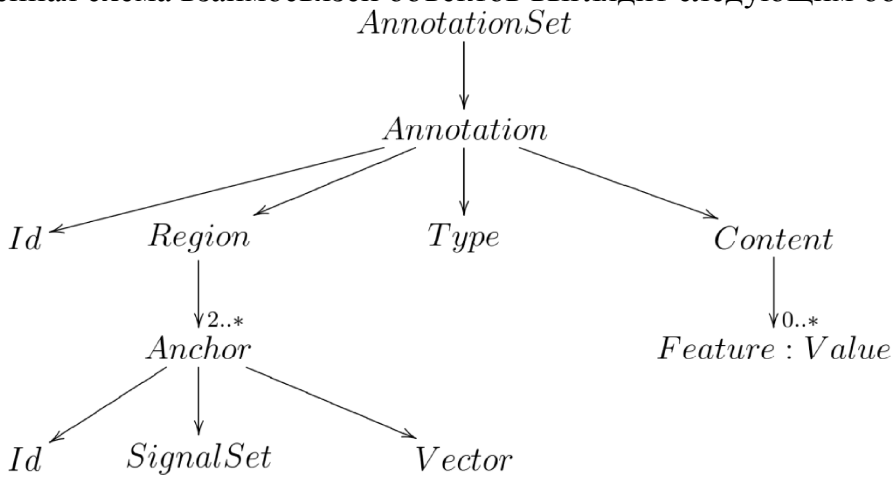


Модель ATLAS

Модель ATLAS [4] (Architecture and Tools for Linguistics Analysis Systems) возникла в 2000 году как обобщение графов аннотаций в двух основных направлениях:

- Возможность наличия нескольких размерностей в пространстве аннотируемого сигнала - например, при аннотировании участков изображения или видео;
- Представление нескольких наборов аннотаций для одного и того же сигнала - например, для разделения аннотаций, создаваемых компонентами различного уровня анализа;

Расширенная схема взаимосвязей объектов выглядит следующим образом:



В новой модели разделены понятия региона, который представляет из себя просто некоторый участок сигнала (в одномерном случае - отрезок) и аннотации, которая представляет из себя информацию, приписанную региону и состоящую из типа и набора признаков. Авторы представляют аннотации, как «отношения между регионами и (структурными) метками».

§ 3.4. Представления, основанные на абстракции

Структуры признаков являются распространенным средством для представления лингвистической информации и ведут свою историю от фреймов. Многие формализмы для осуществления анализа базируются на их использовании.

Структуры признаков

[28]

Структура признаков представляет собой набор пар «признак-значение», где значение может быть атомарным или сложным - другой структурой признаков, списком или множеством. Структуры признаков обычно записываются в виде матриц следующего вида:

$$\begin{bmatrix} f : \begin{bmatrix} q : a \end{bmatrix} \\ g : c \\ h : d \end{bmatrix}$$

Приведенная матрица описывает структуру, имеющую три признака: *f*, *g* и *h*, причем значением первого является структура $\begin{bmatrix} q : a \end{bmatrix}$, имеющая признак *q* со

значением *a*, в то время как оставшиеся два признака имеют атомарные значения *c* и *d* соответственно.

Между структурами признаков может быть задано отношение частичного порядка, называемое отношением категоризации (subsumption) и связывающее менее общие структуры (несущие меньше информации) с более общими (несущими больше информации).

Структура признаков α считается более общей, чем β если они идентичны или множество признаков α является подмножеством множества признаков β, а значения соответствующих признаков в α более общие. Например, среди следующих структур признаков *A* и *B* являются более общими, чем *E*, однако не связаны друг с другом отношением категоризации.

A

$$\begin{bmatrix} a : \begin{bmatrix} b : c \end{bmatrix} \\ k : m \end{bmatrix}$$

B

$$\begin{bmatrix} a : \begin{bmatrix} b : c \\ o : p \end{bmatrix} \end{bmatrix}$$

E

$$\begin{bmatrix} a : \begin{bmatrix} b : c \end{bmatrix} \end{bmatrix}$$

Для структур признаков определена операция унификации, которая возвращает наиболее общую структуру, обладающую всеми признаками как первой, так и второй, причем для каждого из признаков его значение так же является наиболее общим. Например, результатом унификации вышеприведенной структуры и $\begin{bmatrix} f : \begin{bmatrix} f : e \end{bmatrix} \end{bmatrix}$ будет являться структура:

$$\begin{bmatrix} f : \begin{bmatrix} q : a \\ f : e \end{bmatrix} \\ g : c \\ h : d \end{bmatrix}$$

Другой важной особенностью формализма является возможность представления ссылок на один и тот же элемент с помощью введение переменных. Более того, переменные могут ссылаться на элемент, значение которого не определено (полностью или частично). Например:

$$\begin{bmatrix} f : \boxed{1} \begin{bmatrix} q : \boxed{2} \end{bmatrix} \\ g : \boxed{1} \\ h : \boxed{2} \end{bmatrix}$$

Здесь признаки *f* и *g* ссылаются на одно и то же значение - структуру $\begin{bmatrix} q : \boxed{2} \end{bmatrix}$. При этом признак *h* и признак *q* во вложенной структуре так же ссылаются на один и тот же элемент, хотя его значение и не определено. Возможность представления ссылок на неопределенные значения позволяет значительно расширить возможности унификации структур признаков. Например, результатом унификации вышеописанной структуры и $\begin{bmatrix} h : e \end{bmatrix}$ является структура:

$$\begin{bmatrix} f : \boxed{1} \begin{bmatrix} q : e \end{bmatrix} \\ g : \boxed{1} \\ h : e \end{bmatrix}$$

На основе унификации для структур признаков были разработаны механизмы логичского вывода, аналогичные используемым для утверждений логики первого порядка (например, метод резолюций[14]).

Структуры признаков с переменными могут быть представлены как ациклические ориентированные графы, где ребра помечены признаками, а вершины представляют значения.

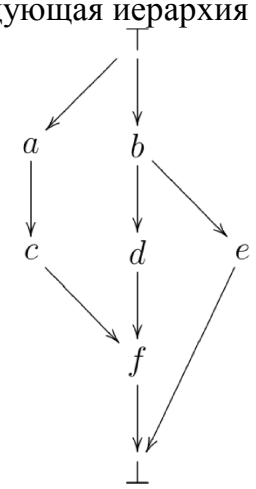
Для структур признаков часто используется запись значения, получаемого по пути признаков, например ⟨*Xfq*⟩, обозначающая значения признака *q* признака *f* структуры *X*.

Типизированные структуры признаков

Типизированные структуры признаков получены как расширение формализма структур признаков за счет приписывания признакам типов и использования информации об отношениях между типами во время унификации.

Отношения между типами обычно задаются в виде отношения частичного порядка “тип-подтип”.

Например, если имеется следующая иерархия типов:



То структуры $[\ q:a\ ]$ и $[\ q:b\ ]$ могут быть унифицированы - результатом будт структура $[\ q:f\ ]$, поскольку *f* является наиболее общим подтипом *a* и *b*.

§ 3.5. Недоспецифицированные представления

Анализ текста на естественном языке в рамках одного их этапов обработки часто не может быть выполнен однозначно в силу того, что информация, необходимая для разрешения неоднозначности может находится на более глубоких уровнях. Одним из подходов обработки таких ситуаций является выдача этапом обработки множества наиболее вероятных результатов анализа. Однако, использование такого подхода приводит к значительным вычислительным затратам, а оптимизация путем отбрасывания части результатов приводит к потере части информации и возможности отсутствия допустимых интерпретаций на одном из последующих этапов.

Другим подходом является использование недоспецифицированных представлений, в рамках которых информация представляется таким образом, чтобы избежать выбора между различными вариантами. В основе недоспецифицированных формализмов лежит идея того, что каждый этап обработки текста на естественном языке предоставляет в качестве результата семантическую информацию в неполной форме.

Minimal Recursion Semantic

В принципе, использование структур признаков уже позволяет представлять информацию в недоспецифицированной форме, путем наличия признаков без значений, с которыми могут быть связаны переменные. То есть такое представление позволяет **недоспецифицировать значения признаков**. Однако, во многих случаях неоднозначность имеет другую форму: например, имеются структуры признаков A и B и неоднозначность состоит в том вложена ли A в B или наоборот. Такая неоднозначность структуры не может быть представлена на основе структур признаков.

В качестве решения, позволяющего представлять **недоспецифицированность структуры** был предложен формализм MRS [9] (Minimal Recursion Semantic). Основная идея формализма состоит в преобразовании вложенной структуры в плоскую. Таким образом, вложенная структура признаков (или предиктов) преобразуется в множество структур (которые могут быть объединены символами коньюнкции).

Robust Minimal Recursion Semantic

Формализм RMRS [8] (Robust Minimal Recursion Semantic) является развитием MRS, основное отличие которого состоит в том, что структуры из нескольких признаков (многоаргументные предикаты) разбиваются на однопризнаковые структуры (бинарные предикаты).

Если рассматривать представление структур признаков в виде ориентированных графов, такое представление соответствует хранению множества ребер графа, где для каждого ребра указаны начальная и конечная вершина, причем такие указания могут быть представлены как константами, так и переменными. При этом в представлении могут быть заданы дополнительные ограничения, например требования различности значения некоторых переменных.

Глава 4. Архитектура инструментальных ЕЯ-систем

Различные ЕЯ-системы имеют различную архитектуру, однако практически все в той или иной мере предполагают разбиение на независимые модули (часть из которых может быть создана сторонними разработчиками), которые можно в общем назвать компонентами. Это позволяет говорит об общих свойствах систем, связанных с их компонентной организацией, таких как: особенности понимания сущности компонентов, задачи, которые выполняются ими, схема и средства взаимодействия между ними, порядок работы различных компонентов при функционировании системы в целом.

Кроме того, в соответствии с [30], архитектура ЕЯ-системы должна иметь три основных слоя:

- Слой взаимодействия (communication layer), который описывает взаимодействия между различными компонентами в системе для решения задачи обработки текста;
- Слой данных (data layer), включающий различные форматы представления данных и правила их преобразования между компонентами;
- Слой интепретации (interpretation layer), описывающий то, как компоненты интерпретируют те или иные данные в процессе своей работы.

Слои данных и интерпретации соответствуют представлению данных, которые были рассмотрены в 4.3, здесь же наиболее важен слой взаимодействия, содержащий отношения между различными компонентами.

§ 4.1. Компонентная организация

С точки зрения организации взаимодействия компонентов стоит начать с классификации, предложенной в [30], которая подразделяет системы на три категории:

- Каждый компонент взаимодействует непосредственно с другим компонентом. В этом случае все компоненты должны иметь информацию о среде выполнения, формате данных и их интерпретации;
- Компоненты взаимодействуют через центральный координатор, который берет на себя ответственность за распределение задач и преобразование данных между различными форматами;
- Компоненты работают с некоторым общим хранилищем данных, не имея информации о среде выполнения.

Однако, приведенная классификация весьма приблизительна, каждая из категорий должна быть уточнена. Для начала, о непосредственном взаимодействии.

Компоненты, взаимодействующие независимо друг с другом могут совместно работать как минимум одним из нескольких способов:

- Подход, основанный на потоках данных (data-flow), в котором компоненты организуют некоторый граф и обмениваются данными друг с другом по соединениям, являющимся ребрами этого графа. Примером использования такого подхода является архитектура Catalyst [2]. Одним из преимуществ подобного подхода является возможность осуществлять обработку различными компонентами одновременно, возможно в распределенной среде.
- Другим вариантом организации компонентов является схема “каналов и фильтров”, в которой компоненты передают данные между собой

последовательно. Примером реализации такой схемы является архитектура LT-NSL. В этом случае одновременность обработки ограничена локальностью анализа конкретных компонентов. Например, tokenizer и стеммер обабатывают данные локально и могут успешно работать параллельно, однако компоненты, требующие анализа более протяженных сущностей могут блокировать работу последующих компонентов.

Также, знание компонентов друг о друге вовсе не является обязательным при этом подходе. Использование некоторого стандартизированного представления данных в системе позволяет компонентам взаимодействовать без наличия информации об устройстве системы. Такая информация необходима только на этапе сборки системы, осуществляемой разработчиком.

Подходы, основанные на взаимодействии через центрального координатора также имеют несколько разновидностей, но они имеют больше специфики с точки зрения реализации, поэтому здесь не имеет смысла их рассматривать.

При использовании общего хранилища данных значительным аспектом является порядок доступа компонентов к данным. В частности, в рассматриваемых приложениях наиболее распространены две модели:

- Линейная схема активации, в которой компоненты работают с данными последовательно, реализуя различные этапы обработки. При этом последовательность обработки задается на этапе сборки системы. Подобный подход используется в системе GATE. С точки зрения одновременности обработки подход эквивалентен схеме “каналы и фильтры”.
- Другая разновидность схема представлена в системе Ellogon, где порядок активации компонентов определяется на основе пред- и постусловий. Эта схема является более гибкой и некоторые варианты ее реализации позволяют построить обработку, аналогичную dataflow подходу. Кроме того, задание предусловий позволяет организовать автоматическое формирование порядка обработки.

Обе модели данных значительно ограничены в возможности параллельной работы компонентов, однако при некоторых дополнительных ограничениях, накладываемых на компоненты (например, требование последовательной выдачи обработанных данных) могут достигать уровня, аналогичного подходам группы «каждый-с-каждым».

§ 4.2. Процессы обработки текста

Фиксированный процесс

Один из базовых подходов к обработке текстов в ЕЯ-системах состоит в последовательном применении к тексту нескольких фиксированных этапов обработки [19]. Простейшим примером такого подхода применительно к системе переводя является следующая цепочка этапов:

- Графематический анализ;
- Морфологический анализ;
- Синтаксический анализ;
- Преобразование;
- Генерация текста.

Такой подход не является очень гибким - достаточно часто необходимо в зависимости от тех или иных условий использовать различные этапы обработки.

Динамический процесс

Более гибким вариантом является возможность использования в процессе обработки различных этапов в зависимости от выполнения каких-то условий. Такой подход называется динамическим процессом [19] и часто реализуется с помощью специального компонента, инкапсулирующего знания о необходимом порядке выполнения этапов и условиях, влияющего на него.

В начале обработки текст передается этому компоненту, который выбирает первый этап обработки и передает текст на него. По окончании этапа текст возвращается центральному компоненту, который выбирает следующий этап или возвращает его в качестве результата.

Вложенные процессы

Часто полный текст не нужен для выполнения этапа обработки (например, синтаксический анализ может производится в рамках одного предложения). В таких случаях используется вложенный процесс обработки текста [19].

Во вложенном процессе один или несколько из этапов обработки могут осуществлять разбиение текста (или его части) на более мелкие участки с тем, чтобы применить некоторый новый процесс к каждому из них. Например, разбить текст на предложения с тем, чтобы произвести синтаксический анализ и последующий перевод каждого предложения по отдельности.

При этом каждый подпроцесс может включать другие вложенные процессы (например, разбиение предложения на слова и их обработка).

Итерационные процессы

Другим интересным расширением процесса обработки является итеративное применение некоторой его части [17]. Например, если есть два компонента синтаксического анализа, первый из которых осуществляет поверхностный анализ, а второй - глубокий, используя некоторый анализ в качестве первого приближения, то можно используя поверхностный анализ в качестве первого приближения итеративно применять компонент глубокого синтаксического анализа к результату предыдущей итерации (или же к взвешенной сумме предыдущей и первой итерации).

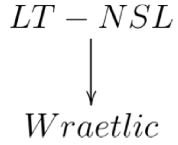
В некоторых случаях использование такого подхода позволяет увеличить точность производимого анализа.

Глава 5. Системы обработки ЕЯ-текстов

Ниже рассмотрены некоторые системы обработки ЕЯ-текстов в рамках направлений, в которых они развивались.

§ 5.1. Системы на базе разметки

Одна из крупнейших ветвей систем обработки ЕЯ-текстов основана на представлении лингвистической информации в форме разметки в одном из стандартных форматов, например SGML и XML. При этом, различные компоненты системы обычно обмениваются размеченными документами в текстовой форме.



LT-NSL

Архитектура LT-NSL [21] была разработан в 1996 году как система для обработки больших корпусов текста, возможно содержащих сложную лингвистическую информацию.

Для представления лингвистической информации в системе LT-NSL используется разметка в документе на языке SGML. Такое решение мотивировано тем, что:

- SGML – четко определенный язык, позволяющий задавать структурную информацию в тексте;
- Доступ приложения к тексту в SGML может быть осуществлен с выбором необходимого уровня абстракции. В частности, приложение может просто игнорировать узлы, не затрагиваемые конкретным видом анализа;
- SGML способствует формальному описанию используемой нотации и предоставляет утилиты для проверки соответствия документов этому формальному описанию.

В процессе разработки системы авторы столкнулись с тем, что использование полного SGML как внутреннего представления данных неэффективно, поскольку требует больших затрат на его анализ и проверку корректности. Для решения этой проблемы было использовано решение в котором документ сначала проверяется на корректность и приводится к специальной упрощенной форме. При дальнейшей обработке предполагается, что документ находится в этой форме и заведомо корректен, что позволяет значительно упростить его обработку. Для разбора документов в упрощенной форме реализован специальный программный интерфейс, позволяющий осуществлять это более эффективно чем обычные парсеры.

Использование разметки в тексте для представления лингвистической информации приводит к проблеме перекрывающихся элементов разметки. В LT-NSL эта проблема решается путем использования ссылок на разметку, находящуюся вне документа.

Одной из важных особенностей архитектуры является то, что текст вместе с лингвистической информацией рассматривается как поток данных, обрабатываемый с использование компонентов, соединяемых через каналы (pipe) UNIX. Каждый компонент выполняет свою специфическую задачу и работает как фильтр, т.е.

принимает на вход поток данных и выдает модифицированный поток на выходе. Такой подход позволяет не хранить документ в памяти полностью, что важно при обработке больших корпусов текстов.

В процессе обработки текста используется специальный язык запросов для выделения конструкций текста, аналогичный современному XPath, позволяющий извлекать элементы по описанию пути к ним от корня документа и задавать дополнительные требования, такие как необходимость наличия подэлементов или текста, удовлетворяющего регулярному выражению.

Wraetlic

Другой системой, использующей разметку для представления лингвистической информации является Wraetlic NLP suite [1], разрабатываемая в 2000-2006 годах. Использование разметки мотивировано возможностью легко объединять компоненты в систему используя стандартные средства операционной системы, такие как каналы.

Лингвистическая разметка представляется в форме XML, при этом проблема представления пересекающейся разметки решена путем использования внешних XML-узлы и ссылок в документе. Для обработки такой разметки реализован специальный модуль, позволяющий извлекать участки документа по ссылкам.

Преимуществом использования XML является возможность простого создания инструментов для виузализации результатов в виде XSLT-преобразований, генерирующих данные для браузера.

Для увеличения эффективности, модули могут быть объединены в один процесс и использовать программный интерфейс на Java для доступа к обрабатываемым данным, что исключает накладные расходы, связанные с генерацией и разбором XML.

Система имеет модульную архитектуру, предоставляя разработчику возможность расширять его новой функциональностью. Большинство модулей реализовано на Java, но некоторые написаны на C из соображений производительности или же использования платформо-зависимых библиотек.

Система содержит модули для графематического и морфологического анализа, извлечения именованных сущностей, классификации и поверхностного синтаксического анализа.

§ 5.2. Системы на базе аннотаций

Значительное влияние на ЕЯ-системы оказал проект TIPSTER [15], проводимый в 1991-1998 годах. Целью проекта было проведение исследований и разработок в области извлечения информации. Одним из направлений проекта было создание стандартной архитектуры для систем, осуществляющих извлечение информации из текстов на естественном языке, позволяющей решить следующие задачи:

- Предоставление программного интерфейса для задач управления документами в системе и извлечения информации из них;
- Поддержка одноязычных и многоязычных приложений;
- Обмен модулями различных разработчиков;
- Работоспособность в различном аппаратном и программном окружении;
- Масштабируемость на различные объемы документов и потоки их обработки;
- Небольшое время отклика приложений;
- Возможность использования многоуровневой системы безопасности;

В модели данных TIPSTER центральным объектом является документ, который является:

- Хранилищем информации о тексте в виде аттрибутов и аннотаций;
- Атомарным элементом коллекций;
- Атомарным элементом в операциях выделения;

Документы организованы в коллекции, которые также могут иметь аттрибуты. Коллекции представляют хранилище для документов в рамках архтектуры TIPSTER. Аннотации хранят информацию о фрагментах документов. Каждый фрагмент представляется в виде набора отрезков, каждый из которых состоит из пары чисел, указывающих позиции его начала и конца в документе. Хотя большинство аннотаций привязываются только к одному отрезку текста, поддержка набора отрезков позволяет ссылаться на разрывные фрагменты текста.

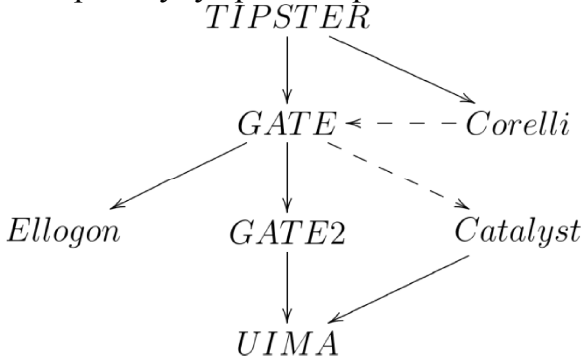
Архитектура TIPSTER не позволяет аннотациям менять текст (например, вставлять символы), поскольку это требует представления позиций в тексте, позволяющих вставки.

Каждая аннотация имеет тип (например, ЋлексемаЛ, ЋпредложениеЛ и т.п.) и набор аттрибутов. Аттрибуты могут содержать как простые значения (например, часть речи), так и более сложные, например, ссылки на другие аннотации.

В рамках архитектуры TIPSTER предполагается, что приложение осуществляет доступ к аннотациям очень часто. Для того, чтобы упростить такой доступ введено понятие набора аннотаций (AnnotationSet), который позволяет искать аннотации по позиции или типы, последовательно перебирать их, добавлять, удалять и т.п. Каждый документ содержит набор аннотаций, представляющий все аннотации этого документа.

Основные процедуры обработки текстов реализуются в приложениях в виде компонентов, которые генерируют новую разметку в форме аннотаций. Компонент принимает на обработку документ (с некоторой разметкой, сгенерированной предыдущими компонентами или без нее) и добавляет новые аннотации в разметку.

Идеи, заложенные в TIPSTER были успешно использованы во многих ЕЯ-системах, некоторые из которых будут рассмотрены ниже:



GATE

Одной из наиболее известных систем для автоматической обработки текстов на естественном языке, основанных на идеях проекта TIPSTER, является GATE [12] (General Architecture for Text Engineering). GATE предоставляет для приложений общую модель представления, хранения и обмена данными между компонентами

приложения, а также графические инструменты для управления данными, их визуализации и анализа.

Первая версия GATE была представлена в 1996 году и предназначалась для решения следующих задач:

- Обмен информацией между компонентами в максимально общей теоретико-нейтральной форме;
- Интеграция компонентов, написанных на различных языках программирования на различных платформах;
- Тестирование и доработка лингвистических компонентов и систем, построенных из них через единообразный графический интерфейс.

GATE использует модель данных, разработанную в рамках проекта TIPSTER, однако, позволяет осуществлять преобразование из аннотаций в XML-разметку и обратно, что позволяет интегрировать приложения со стандартными инструментами по обработке XML-документов.

Система состоит из трех слоев:

- GDM (Gate Document Manager) - хранилище текстов и сопутствующей лингвистической информации, предоставляющее компонентам единообразный интерфейс для доступа и манипуляций с данными.
- CREOLE (Collection of REusable Objects for Language Engineering) - набор переиспользуемых лингвистических компонентов для различных задач анализа текстов. Компоненты используют GDM для доступа к данным. Некоторые из этих компонентов были специально разработаны для использования в GATE, другие являлись обертками вокруг уже существующих.
- GGI (Gate Graphical Interface) - графический интерфейс, который позволяет использовать ресурсы GDM и CREOLE для интерактивного создания и тестирования компонентов и приложений. Интерфейс позволяет создание, просмотр и редактирование коллекций документов, которые управляются GDM, а также отображение результатов работы компонентов в форме аннотаций.

Использование GATE в различных проектах по обработке текстов на естественном языке и извлечению информации показало [20] преимущества ее использований, выраженные в том, что GATE:

- Способствует переиспользованию лингвистических компонентов, что уменьшает усилия, необходимые для интеграции и разработки приложений;
- Способствует объединению усилий в лингвистических исследований за счет представления общей базы для разработки компонентов и приложений;
- Облегчает сравнение алгоритмов и их реализаций в виде компонентов;
- Объединяет лучшие элементы подхода TIPSTER с возможностью экспорта аннотаций в XML (а также импорта из него);
- Предоставляет удобный графический интерфейс.

Так же были выявлены некоторые недостатки GATE, в частности:

- Не поддерживаются компоненты, представляющих источники данных;
- Не поддерживается генерация текста;
- Реализация базы данных неэффективна;
- Графический интерфейс иногда сложен и неочевиден;
- Модель обработки не масштабируется на большое число компонентов;
- Невозможно добавить новые компоненты для виузализации данных;

- Для интеграции различных компонентов необходима совместимость их схем аннотаций.

Ellogon

Ellogon [22] (http://www.ellogon.org/) – многоязыковая платформа для создания приложений обработки текстов на естественном языке, использующая идеи из проектов TIPSTER и GATE. Платформа была разработана в 2002 году как имеющая низкое потребление ресурсов и поддерживающая обработку текстов на различных языках (за счет использования Unicode). Ellogon предоставляет инфраструктуру для:

- Управления, хранения и обмена текстами вместе с лингвистической информацией;
- Создания, управления и интеграции лингвистических компонентов;
- Поддержки взаимодействия между различными компонентами;
- Визуализации текстовой и лингвистической информации.

Ellogon состоит из трех подсистем: ядра, графического интерфейса и подключаемых компонентов. Ядро реализует всю базовую функциональность по управлению текстовой и лингвистической информацией, используемой в других частях системы. Подключаемые компоненты выполняют конкретные задачи лингвистической обработки. Ядро Ellogon написано на C++ из соображений эффективности, однако имеет программный интерфейс для других языков, таких как Tcl и Java.

В соответствии с архитектурой TIPSTER, лингвистическая информация представляется в форме аннотаций, причем одна аннотация может быть связана с несколькими непрерывными промежутками в тексте.

Каждый компонент в Ellogon имеет набор предусловий и постусловий. Предусловия описывают лингвистическую информацию, которая должна быть представлена в документе для того, чтобы компонент мог осуществить его обработку, а постусловия описывают то, какая информация добавляется компонентом в текст в процессе обработки. Эта информация используется для определения порядка обработки документа. Также, каждый компонент определяет набор параметров, которые определяют его поведение и могут быть отредактированы пользователем в графическом интерфейсе и набор визуализаторов позволяющих представлять обрабатываемую информацию в графическом интерфейсе.

Для быстрой разработки компонентов Ellogon предоставляет специальные средства в графическом интерфейсе, позволяющие генерировать заготовки для новых компонентов, перезагружать компоненты после их редактирования и т.п. Компоненты могут быть реализованы как на C++ и работать напрямую с ядром так и на других языках с использованием программного интерфейса для дополнительных языков.

Ключевыми возможностями Ellogon являются:

- Полная поддержка Unicode и различных языков в ядре и графическом интерфейсе;
- Портируемость на большинство платформ;
- Совместимость с другими ЕЯ-системами, такими как GATE за счет предоставления специальных оберток для их программного интерфейса;
- Сжатие данных в памяти, позволяющее значительно уменьшить использование памяти приложением;
- Возможность работы в качестве веб-сервера, предоставляющего функциональность приложения через протокол HTTP.

Система Ellogon доступна для свободного использования и может быть загружена с сайта системы.

GATE 2

Для исправления недостатков первой версии GATE, был произведен полный редизайн системы и в 2001 году разработана новая версия GATE [6], (http://gate.ac.uk/) реализованная на языке Java. Переход на Java позволил использовать представление текста в Unicode, что обеспечило поддержку обработки текстов на различных языках.

Одним из важнейших отличий стала поддержка компонентов трех типов:

- Языковые ресурсы, которые хранят и предоставляют сервисы для доступа к различным типам лингвистических ресурсов, таким как документы, корпуса и онтологии; Хотя предоставляемые по умолчанию ресурсы предназначены для доступа к текстовой информации, система не содержит ограничений на использование ресурсов другого типа, что позволяет реализовать ресурсы для обработки информации других типов;
- Обрабатывающие ресурсы — лингвистические компоненты, которые осуществляют различные лингвистические задачи, такие как выделение лексем, морфологический анализ, и т.п;
- Визуальные ресурсы — графические компоненты, предоставляющие возможности визуализации и редактирования обрабатываемой информации или редактирования процесса выполнения приложения.

Каждый компонент GATE 2 имеет набор свойств, которые управляют его функционированием, например, для большинства обрабатывающих компонентов одним из таких свойств является ссылка на обрабатываемый языковой ресурс. Выполнение приложения состоит в последовательном применении всех обрабатывающих компонентов к соответствующим им языковым ресурсам.

Модель данных из проекта TIPSTER была модифицирована в целях совместимости с форматом Atlas.

Также, в GATE 2 был введен специальный язык JAPE [11], позволяющий описывать процесс обработки разметки в компонентах в форме набора правил, основанных на регулярных выражениях. Использование этого языка значительно упростило создание многих компонентов для обработки естественного языка.

С использованием JAPE процесс обработки документа описывается в форме набора правил, каждое из которых состоит из двух частей, одна из которых определяет условия применения, а другая совершаемые действия.

Левая часть правила задает так называемый шаблон, который используется для выделения последовательности аннотаций. Шаблон может содержать элементы, сопоставляемые с аннотациями, а также операторы регулярных выражений (?,*,+,|). При этом в шаблоне могут быть заданы ограничения на свойства аннотаций, входящих в выделяемую последовательность, например Token.kind == number.

Правая часть Jape-правил описывает преобразования, применяемые к разметке в точке сопоставления левой части правила. Обычно такими преобразованиями является добавление новых аннотаций, с заданием им тех или иных свойств. Новые аннотации добавляются к фрагментам текста, соответствующим меткам, указанным в левой части правила.

Правила Jape применяются следующим образом: производится поиск в тексте последовательности аннотаций, соответствующей левой части правила, затем меткам

ставятся в соответствие фрагменты текста, к которым в последствии добавляются аннотации правой части.

Система GATE 2 доступна для свободного использования и может быть загружена с сайта системы.

Catalyst

Catalyst [2] – это архитектура, выделенная из вопросно-ответной системы Qanda в 2002 году и предназначенная для решения задач обработки текстов на естественном языке и извлечения информации. Исходная система Qanda впоследствии послужила прототипом для разработки других систем на базе архитектуры.

Основными причинами для разработки архитектуры послужил тот факт, что существующие системы плохо масштабировались при росте объемов обрабатываемой информации и требований к скорости ее обработки.

Для представления данных Catalyst использует модель, основанную на аннотациях TIPSTER.

Важным отличием системы Catalyst от других, является использование концепции потоков данных для интеграции компонентов. Каждый компонент соединяется с другими каналами, по которым передаются аннотации, упорядоченные в соответствии с их позициями в документе (а для совпадающих позиций - по типам аннотаций).

Каждый компонент объявляет какие типы аннотаций необходимы ему для обработки и какие аннотации он генерирует на выходе. Такая информация позволяет не передавать между компонентами аннотации, которые не будут использован и значительно уменьшить кол-во информации, передаваемой по каналам, что важно при построении хорошо масштабируемых распределенных системы.

Такой подход к интеграции компонентов имеет ряд преимуществ:

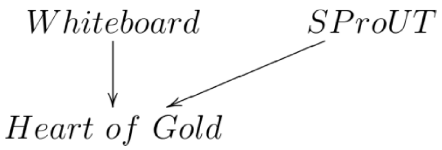
- Многие ошибки, связанные с неправильной организацией приложения могут быть выявлены на этапе сборки;
- Отсутствуют накладные расходы на преобразование разметки в какой-либо стандартный формат (например, на базе XML);
- Разработчики компонентов могут работать непосредственно с аннотациями, а не с разметкой в каком-либо формате;
- Система может работать как на одной машине, так и распределенно. Компоненты могут быть реплицированы для увеличения производительности;
- Код компонентов упрощается поскольку в них не требуется проверка данных на корректность (она выполняется на этапе сборки приложения).

Поскольку отладка и мониторинг распределенной системы может представлять значительные сложности Catalyst предоставляет возможности для распределенного логирования и мониторинга. Использование распределенного логирования позволяет собирать информацию со всех или указанных процессов в системе. Собранная информация включает события начала и завершения обработки, пользовательские сообщения, сообщения об ошибках и т.п. Логирование может быть включено при запуске системы или в процессе ее функционирования. Система мониторинга позволяет отслеживать конфигурацию и состояние приложения, включая потоки передаваемых данных в процессе его функционирования. Предоставляется информация об активных компонентах, потоках данных, количестве буферизованных сообщений и т.п.

Работа с Catalyst состоит из написаний компонентов для фреймворка и сборке приложений из компонентов. Catalyst разработан так, чтобы в первую очередь упростить вторую задачу - сборка приложения представляет из себя размещение компонентов и соединение их каналами.

Компоненты для фреймворка могут быть разработаны непосредственно с использованием модели аннотаций или представлять из себя обертки вокруг уже существующих компонентов.

§ 5.3. Системы интеграции поверхностной и глубокой обработки



SProUT

Система SProUT [13; 25] (Shallow Processing with Unification and Typed Feature Structures) была разработана для задач поверхностной обработки текста в 2002-2004 годах. Мотивацией для разработки системы послужила необходимость в системе, которая позволяла бы гибко интегрировать различных обрабатывающих компонентов и в то же время представляла бы хорошее соотношение между производительностью и выразительностью используемого формализма.

Идеей системы было объединение формализмов конечных преобразователей, для которых существуют эффективные алгоритмы, и унификационных грамматик, позволяющих естественным образом выражать синтаксические и семантические ограничения. В качестве средства для такого объединения использовалась машина с конечным числом состояний, работающая над типизированными структурами признаков. Таким образом правила преобразований в правой части содержат регулярное выражения над структурами признаков, а правая часть представляет выходную структуру признаков. При этом ограничения на равенство признаков заменяются их унифицируемостью.

Описанный формализм был расширен путем добавления функциональных операторов и возможности вызова дополнительных правил в процессе сопоставления. Функциональные операторы позволяют расширять формализм путем подключения новых функций, используемых для вычисления значений в результирующей структуре. Использование вызова дополнительных правил позволяет вызывать в левой части правил сопоставление других правил (или, возможно, того же самого правила) тем самым расширяя выразительность формализма до контекстно-свободного ценой небольшого снижения эффективности сопоставления (поскольку такой вызов приводит к порождению нового процесса сопоставления).

Ядро системы состоит из четырех основных компонентов - инструментария для обработки конечных машин, компилятора регулярных выражений, интерпретатора формализма XTDL и пакета типизированных структур признаков. С использованием этих компонентов разработаны переиспользуемые компоненты для обработки лингвистической информации. Компоненты легко интегрируются внутри системы поскольку имеют унифицированное представление данных в виде типизированных структур признаков.

Компоненты осуществляют обработку последовательно, но возможна более сложная конфигурация путем использования специально разработанного языка описания процесса обработки.

Whiteboard

Система Whiteboard [10; 25] была разработана в 2000-2002 годах и предполагала возможность интеграции лингвистических компонентов для поверхностной и глубокой обработки текста.

Такая интеграция компонентов для поверхностного и глубокого анализа проблематична в связи с различиями в их производительности и точности. Одно из возможных решений состоит в том, чтобы выполнять анализ параллельно, используя результаты глубокого анализа при их наличии. Однако, для больших наборов данных такой подход приводит к рассинхронизации работы компонентов. Авторы системы предложили решение, основанное на анализе данных с использованием компонентов поверхностного анализа для определения участков, которые необходимо обработать с помощью компонентнов глубокого анализа. Кроме того, использование результатов поверхностного анализа на таких участках может быть использовано в качестве дополнительной эвристики для компонентов глубокого анализа, увеличивая производительность их работы.

Для представления лингвистической информации на поверхностном уровне в системе используется XML-разметка. При этом более сложные структуры, которые не могут быть выражены в XML хранятся отдельно и доступны компонентам, выполняющим глубокую обработку. Это позволяет, с одной стороны компонентам поверхностного анализа работать с хорошо известным представлением, а компонентам глубокого анализа не быть ограниченными этим представлением.

Heart of Gold

Архитектура Heart of Gold [7; 24; 25] была разработана в 2004-2005 как развитие Whiteboard исправляющее ряд ее основных недостатков. Задача архитектуры состоит в том, чтобы сохранить преимущества поверхностной обработки текста (в первую очередь устойчивость и эффективность), но при этом увеличить точность и глубину анализа в тех местах, где это необходимо.

В качестве единого формата для представления данных выбран RMRS. Для компонентов, которые используют другие представления, основанные на XML, применяется преобразование данных, описываемое на языке XSLT.

Платформа работает как медиатор между приложениями и набором компонентов. Приложения посылают запросы об анализе документов центральному модулю, который рассылает запросы различным компонентам и затем осуществляет слияние результатов их обработки. При этом, результаты запросов сохраняются в базе данных, что позволяет избежать повторной обработки при получении тех же запросов. Параметры запросов указывают на анализируемый документ, участок в тексте, который необходимо проанализировать, а также необходимую глубину анализа.

В соответствии с идеей, заложенной в систему результат обработки каждого компонента представляет из себя недоспецифицированную семантическую информацию, которая может быть углублена в процессе дальнейшей обработки. В соответствии с этим, стратегия обработки запроса состоит в том, чтобы обработать запрос с помощью всех компонентов начиная с минимальной глубины до заданной в

запросе, откатываясь к результату работы предыдущего компонента в случае, если какой-либо компонент не обработал запрос. Для возможности последующего анализа условий при которых был получен тот или иной результат, вместе с ним сохраняется метаинформация о переданых параметрах, времени обработки запроса и т.п.

Ядро системы написано на Java, однако компоненты и приложения могут быть написаны на любых языках и осуществлять взаимодействие с ядром через протокол XML-RPC.

§ 5.4. Системы, развивающие отдельные аспекты обработки текста

Здесь представлены другие системы, осуществляющие обработку текстов на естественном языке так или иначе интересные идеями, заложенными в них.

Fastus

Система Fastus [16] была разработана в 1994 как средство для эффективного и точного извлечения информации из текстов и интересна в первую очередь тем, что являлась одной из первых систем, использующих регулярные шаблоны для решения задач извлечения информации.

При разработке системы были учтены такие особенности извлечения информации из текстов, как релевантность только небольшой части анализируемого текста, необходимость отображения информации в заранее определенное относительно простроен представление и незначительность многих аспектов, связанных со значением и целью написания текста. При этом рассмотрение извлечения информации как подзадачи понимания текста и применение систем, предназначенных для понимания текстов приводило к относительно низкому качеству результатов при низкой эффективности обработки.

Для преодоления этой проблемы авторами было предложено использовать для извлечения информации более простой регулярный формализм вместо обычно используемых контекстно-свободных с тем, чтобы увеличить эффективность работы системы.

Работа системы FASTUS состояла из четырех этапов:

1. Обнаружение ключевых слов, свидетельствующих о наличии в предложениях релевантной информации;
2. Выделение именных групп, глагольных групп и важных классов слов, таких как предлоги, союзы, и т.п;
3. Выделение необходимой информации по шаблонам. Шаблоны для выделения представлялись в виде конечных автоматов, где переходы соответствовали ключевым словам или словосочетаниям заданного типа, выделенным на предыдущем этапе. При этом в шаблонах использовались конструкции, позволяющие выделять более сложные языковые конструкции (например, сложные именные группы) для представления более полной информации в результирующей структуре.
4. Структуры, выделенные для одного и того же предложения сливались в одну для получения наиболее полной информации об описываемом событии.

Проведенные эксперименты показали эффективность выбранного подхода для извлечения информации.

CAFE

Система CAFE [29] (Cascading, Asynchonous, Feedback Envrionment) разрабатывалась в 2001 году и предназначалась для решения задач распознавания речи.

CAFE интересна в первую очередь принципами взаимодействия компонентов, основанных на идеях так называемого Ћнепрерывного пониманияЛ, в котором компоненты, обрабатывающие данные не дожидаются полных результатов предыдущего этапа, а производят анализ параллельно, при этом предоставляя компонентам предыдущего информацию, корректирующую их функционирование.

Традиционно взаимодействие компонентов, составляющих систему строится на основе предоставления каждым компонентом единственного наилучшего результата анализа на основе информации на своем уровне. Однако, во многих случаях выбор различных вариантов на определенном уровне анализа не может быть осуществлен только на основе информации этого уровня, а требует использования информации с более глубоких уровней анализа текста.

Для решения этой проблемы в CAFE используется следующая мождель взаимодействия между компонентами, реализующими анализ на различных уровнях:

- Каждый компонент предоставляет следующему уровню результаты своего анализа постепенно, по мере их получения не дожидаясь окончания обработки всего документа или предложения; Соответственно, информацию от предыдущего уровня компонент также получает постепенно и может возникать ситуация в которой новой информации еще не поступило;
- Компонент передает на следующий уровень несколько наилучших вариантов анализа на своем уровне, при этом в случае отсутствии информации от предыдущего уровня компонент может продолжать передавать следующему уровню оставшиеся варианты для текущей порции данных;
- Компонент передает на предыдущий уровень информацию об оценке полученных вариантов анализа, что позволяет скорректировать направление анализа на предыдущем этапе.

Таким образом в системе CAFE во-первых достигается больший уровень параллелизма работы компонентов, а во-вторых существует обратная связь между этапами, которая помогает скорректировать получаемые результаты.

LinguaStream

Система LinguaStream [3] разработана в 2003 году и основывается на использовании декларативного описания процесса обработки, который может быть представлен в виде графа.

Приложение разабатывается путем выбора компонентов, каждый из которых имеет набор параметров, входов и выходов и их соединении. Платформа основана на использовании XML и может обрабатывать любой XML-файл, сохраняя его изначальную структуру.

В платформе используются декларативные механизмы описания процесса обработки.

Система использует идею, согласно которой различные модели анализа дополняют друг друга, и, соответственно, не отдает предпочтения какому-либо из них. Для обеспечения совместимости между различными компонентами используется унифицированное представление разметки и аннотаций в виде наборов признаков (feature sets).

Важным аспектом является возможность использования различных минимальных единиц на различных этапах анализа. Когда какая-либо модель требует наличия минимальной единицы анализа (например, лексемы), то эта единица может быть определна локально, только для соответствующего компонента. Кроме того, каждый компонент отмечает какие элементы разметки он обрабатывает. Описанные возможности позволяют определить различные точки зрения на обратываемый документ для кажого этапа.

Каждое приложение может быть переиспользовано в качестве компонента для создания более сложного процесса обработки.

Помимо компонентов, выполняющих конкретные задачи обработки текста, платформа включает компоненты, представляющие различные формализмы реализации задач обработки, например:

- Унификационных грамматик (на базе Prolog);
- Преобразований с конечным числом состояний;
- Грамматик, основанных на ограничениях;
- Регулярных выражений.

Learning Based Java

Система Learning Based Java [23] представляет средства для интеграции и обучения различных статистических компонентов.

В основе системы лежит представление задачи анализа текста в виде поиска набора выходных данных, максимизирующего некоторые оценочные функции и при этом удовлетворяющего заданным ограничениям.

Приложение в системе представляется как набор моделей, описывающих признаки, передаваемые на вход статистической модели. В качестве примера рассмотрим простую модель, осуществляющую выделение в качестве признаков множества слов новостной статьи:

```
/** This feature generating classifier "senses" all the
 * words in the document that begin with an alphabet
 * letter. The result is a bag-of-words representation
 * of the document. */
discrete% BagOfWords(Post post) <- {
  for (int i = 0; i < post.bodySize(); ++i)
    for (int j = 0; j < post.lineSize(i); ++j) {
      String word = post.getBodyWord(i, j);
      if (word.length() > 0 &&
          word.substring(0, 1).matches("[A-Za-z]"))
        sense word;
    }
}
/** The label of the document. */
discrete NewsgroupLabel(Post post) <-
  { return post.getNewsgroup(); }
```

В приложении задается используемый компонент машинного обучения и модели, признаки из которых используются для обучения и распознавания объектов:

```
/** Here, we train averaged Perceptron for many
 * rounds of the training data. **/
discrete NewsgroupClassifierAP(Post post) <-
learn NewsgroupLabel
  using BagOfWords
  from new NewsgroupParser("data/20news.train.shuffled")
    40 rounds
  with SparseNetworkLearner {
    SparseAveragedPerceptron.Parameters p =
      new SparseAveragedPerceptron.Parameters();
    p.learningRate = .1;
    p.thickness = 3;
    baseLTU = new SparseAveragedPerceptron(p);
  }
  progressOutput 20000
  testFrom new NewsgroupParser("data/20news.test")
end
```

§ 5.5. Прочие системы

Corelli

Система Corelli [30] была разработана в 1996 году и предназначена для интеграции распределенных лингвистических компонентов, реализованных на различных языках.

В системе используется модель данных проекта TIPSTER, при этом тип представляемой информации не ограничивается. Для решения проблем совместимости в системе предоставляется библиотека, осуществляющая преобразование данных.

Система состоит из центрального сервера, реализованного на Java, который отвечает за хранение документов и сопутствующей лингвистической информации и различных обрабатывающих компонентов, которые получают необходимые данные от сервера документов. Компоненты взаимодействуют с центральным сервером напрямую с использованием программного интерфейса на Java или удаленно на базе CORBA. Кроме того, компоненты система поддерживает подключение и отключение компонентов в процессе выполнения.

Поскольку взаимодействие с центральным сервером осуществляется по фиксированному протоколу, его реализация может быть заменена в соответствии с нуждами приложения. В частности, предоставляется три основных версии центрального сервера использующие для хранения данных файловую систему, специальное объектное хранилище или реляционную базу данных. В последнем случае центральный сервер предоставляет возможности для транзакционного взаимодействия.

UIMA

Система UIMA (http://uima.apache.org/) разрабатывается с 2004 года по настоящее время. Для представления данных используется модель TIPSTER.

Обработка документов осуществляется последовательно, каждый компонент добавляет аннотации в представление документа.

Для аннотаций определяется система типов, обеспечивающая проверку совместимости аннотаций между различными компонентами. В случае несовпадения систем типов, отображение между ними может быть произведено путем реализации соответствующего компонента.

Система UIMA доступна для свободного использования и может быть загружена с сайта системы.

OpenPipeline

Система OpenPipeline (http://www.openpipeline.org/) предоставляет возможности для автоматизированной обработки документов в серверном приложении. Для системы задается расписание выполнения работ, каждая из которых состоит из получения данных из какого-либо источника и последовательного применения определенных этапов преобразования.

Система реализована как серверное J2EE-приложение.

TESLA

Система TESLA (http://tesla.spinfo.uni-koeln.de/index.html) предоставляет удобный графический интерфейс на базе среды разработки Eclipse для построения приложений естественно-языковой обработки. Компоненты в системе связываются каналами в ориентированный граф.

Система имеет клиент-серверную архитектуру - графический интерфейс выступает в роли клиента и сам не выполняет задач по обработке текстов, а передает их серверу.

Список литературы

[1] Enrique Alfonseca, Antonio Moreno-s, JosДе MarДа Guirao, и Maria Ruiz-casado. The wraetlic NLP suite. 2006.

[2] Pranav Anand, David Anderson, John Burger, John Griffith, Marc Light, Scott Mardis, и Alex Morgan. Qanda and the Catalyst Architecture. 2002.

[3] F. Bilhaut и A. Widl\Jocher. LinguaStream: an integrated environment for computational linguistics experimentation. В Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, страница 95–98, 2006.

[4] Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, и Mark Liberman. ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation. 2000.

[5] Steven Bird и Mark Liberman. A Formal Framework for Linguistic Annotation. Speech Communication, 33:23—60, 2000.

[6] Kalina Bontcheva, Diana Maynard, Valentin Tablan, и Hamish Cunningham. GATE: A Unicode-based infrastructure supporting multilingual information extraction. In Proceedings Of Workshop On Information Extraction For Slavonic And Other Central And Eastern European Languages (Iesl’03), Borovets, 2003.

[7] U. Callmeier, A. Eisele, U. Sch\Jafer, и M. Siegel. The DeepThought core architecture

framework. В Proceedings of LREC, том 4, страница 1205–1208, 2004.
[8] A. Copestake. Robust minimal recursion semantics. unpublished draft, 2004.
[9] Ann Copestake, Dan Flickinger, Rob Malouf, Susanne Riehemann, и Ivan Sag. Translation using Minimal Recursion Semantics. In Proceedings Of The Sixth International Conference On Theoretical And Methodological Issues In Machine Translation, 1995.
[10] Berthold Crysmann, Anette Frank, Bernd Kiefer, Hans-Ulrich Krieger, Stefan MJuller, GJunter Neumann, Jakub Piskorski, Ulrich SchJafer, Melanie Siegel, Hans Uszkoreit, и Feiyu Xu. An Integrated Architecture for Shallow and Deep Processing. University Of Pennsylvania, страницы 441—448, 2002.
[11] Hamish Cunningham, Hamish Cunningham, Diana Maynard, Diana Maynard, Valentin Tablan, и Valentin Tablan. JAPE: a Java Annotation Patterns Engine. 1999.
[12] Hamish Cunningham, Kevin Humphreys, Robert Gaizauskas, и Yorick Wilks. Software Infrastructure for Natural Language Processing. 1997.
[13] Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich SchJafer, и Feiyu Xu. Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. KJunstliche Intelligenz, 1:17–23, 2004.
[14] Bernd Fischer, Ag Softwaretechnologie, и Tu Braunschweig. Resolution for Feature Logic. In Proceedings Of The, страницы 23—34, 1993.
[15] R. Grishman. TIPSTER text phase II architecture design. В Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996, страница 249–305, 1996.
[16] Jerry R Hobbs, John Bear, David Israel, и Mabry Tyson. FASTUS: A finite-state processor for information extraction from real-world text. страницы 1172—1178, 1993.
[17] Kristy Hollingshead и Brian Roark. Pipeline Iteration. В Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, страница 952–959, Prague, Czech Republic, июнь 2007. Association for Computational Linguistics.
[18] Jochen L Leidner. Current Issues in Software Engineering for Natural Language Processing. Proc. Of The Workshop On Software Engineering And Architecture Of Language Technology Systems (Sealts), The Joint Conf. For Human Language Technology And The Annual Meeting Of The Noth American Chapter Of The Association For Computational Linguistics (Hlt, 8:45—50, 2003.
[19] Tom Mahieu, Stefan Raeymaekers, и Stefan Raeymaekers Et Al. Base Architectures for NLP.
[20] Diana Maynard, Hamish Cunningham, Kalina Bontcheva, Roberta Catizone, George Demetriou, Robert Gaizauskas, Oana Hamza, Mark Hepple, и Patrick Herring. A Survey of Uses of GATE.
[21] David McKelvie, Chris Brew, и Henry Thompson. Using SGML as a Basis for Data-Intensive NLP. In Proceedings Of The Fifth Conference On Applied Natural Language Processing (ANLP-97, 1997.
[22] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Ion Androutsopoulos, и Constantine D Spyropoulos. Ellogon: A New Text Engineering Platform. In Proceedings Of The Third International Conference On Language Resources And Evaluation (Lrec 2002), Las Palmas, Canary Islands, 2002:72—78, 2002.
[23] N. Rizzolo и D. Roth. Learning Based Java for Rapid Development of NLP Systems. В Proceedings of the International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, 2010.
[24] Ulrich SchJafer. Middleware for Creating and Combining Multi-dimensional NLP Markup. IN PROCEEDINGS OF THE WORKSHOP ON MULTI-DIMENSIONAL

MARKUP IN NLP, 2006.
[25] Ulrich SchJafer. Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures. Кандидатская диссертация, Faculty of Mathematics and Computer Science, Saarland University, SaarbrJucken, Germany, 2007. Doctoral Dissertation; also available as Vol. 22 of the SaarbrJucken Dissertations in Computational Linguistics and Language Technology series (http://www.dfki.de/lt/diss), ISBN 978-3-933218-21-6.
[26] I. Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. В In Proc. of MLMTA-2003, Las Vegas, 2003.
[27] Stuart C Shapiro и Shane Axtell. Natural Language Tools for Information Extraction for Soft Target Exploitation and Fusion. 2007.
[28] Stuart M Shieber. An Introduction to Unification-Based Approaches to Grammar. 1986.
[29] Scott C Stoness. Continuous Understanding: A First Look at CAFE. 2001.
[30] Remi Zajac, Mark Casper, и Nigel Sharples. An Open Distributed Architecture for Reuse and Integration of Heterogeneous NLP Components. In Proceedings Of The 5th Conference On Applied Natural Language Processing (ANLP-97, 1997.
[31] Копотев М. В.. Между Сциллой языкознания и Харибдой языка: о русскоязычных корпусах текстов. Труды международной конференции Диалог-2005, страницы 282–285, 2005.
[32] Сокирко А. В.. Морфологические модули на сайте www.aot.ru. Труды международной конференции ЂДиалог-2004. Компьютерная лингвистика и интеллектуальные технологииЛ, страница 559, 2004.
[33] Хорошевский В.Ф.. Управление знаниями и обработка ЕЯ-текстов. В Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. В 3-х т., т. 2, страницы 565–572. М.: Физматлит, 2004.
[34] Большакова Е.И. и Носков А.А.. Программные средства анализа текста на основе лексико-синтаксических шаблонов языка LSPL. В Программные системы и инструменты: Тематический сборник, № 11 / Под ред. Королева Л.Н., страницы 71–73. М.: Изд. отдел факультета ВМиК МГУ; МАКС Пресс, 2010.
[35] Резникова Т. И. и Копотев М. В.. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов). Национальный корпус русского языка: 2003—2005, страницы 31–61, 2005.

ЧАСТЬ V. АЛГОРИТМЫ КЛАССИФИКАЦИИ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ *(ПЕСКОВА О.В.)*

Глава 1. Алгоритмы классификации с учителем

Алгоритмы *классификации с учителем* сортируют полнотекстовые документы по заранее известным категориям (классам). В роли учителя выступает выборка документов, для которых заранее известна принадлежность той или иной категории, называемая обучающим множеством. Множество категорий $\boldsymbol{\mathcal{C}} = \{c_j\}, j = \overline{1, |\boldsymbol{\mathcal{C}}|}$ и обучающее множество документов $\Omega \subset \boldsymbol{\mathcal{D}}$, где $\boldsymbol{\mathcal{D}} = \{d_i\},\ i = \overline{1, |\boldsymbol{\mathcal{D}}|}$ – полное множество документов, формируют эксперты. Алгоритм классификации с учителем – *алгоритм категоризации* – использует обучающее множество Ω, чтобы построить классификатор $\Phi: \boldsymbol{\mathcal{D}} \times \boldsymbol{\mathcal{C}} \to \{\text{истина, ложь}\}$, обеспечивающий высокую точность на всем множестве документов $\boldsymbol{\mathcal{D}}$, используя предположение, что обучающие и новые данные похожи. Обычно множество документов Ω делят на две части: одна часть – данные для обучения алгоритма, вторая – тестовые данные для оценки качества полученного классификатора.

Алгоритмы классификации документов называют по имени метода обучения, положенного в его основу. Далее рассмотрим наиболее известные из них, обсудив сначала представление полнотекстовых документов, которым оперируют алгоритмы. Описание каждого алгоритма сопроводим примером, в котором используется следующая коллекция документов:

docId	Слова в документе	c = «Китай»
1	китайский пекин китайский	*c*
2	китайский китайский шанхай	*c*
3	китайский макао	*c*
4	токио япония китайский	$\overline{c}$
5	китайский китайский китайский токио япония	*?*

Изначально идея такого примера была заимствована из [1], где на указанной коллекции из 5 документов было продемонстрировано функционирование трех алгоритмов – наивного байесовского классификатора, алгоритма Роккио и алгоритма k-ближайших соседей. Затем этот пример вырос в сквозной пример для всех рассматриваемых здесь алгоритмов.

§ 1.1. Представление данных в задачах классификации текстов

Образы полнотекстовых документов. Входными данными алгоритма классификации является не сама коллекция документов $\mathcal{D} = \{d_i\}, i = \overline{1, |\mathcal{D}|}$, а множество *образов* каждого документа $\vec{\mathcal{D}} = \{\vec{d}_i\}, i = \overline{1, |\mathcal{D}|}$, где $\vec{d}_i \in \vec{\mathcal{D}}$ – образ документа $d_i \in \mathcal{D}$. Существует несколько подходов к формированию образов, применяют тот, который соответствует модели, положенной в основу конкретного алгоритма классификации. Образы документов в тех алгоритмах, которые мы будем рассматривать, представлены в следующем виде:

а) мультимножеств терминов документов (например, наивный байесовский классификатор);

б) векторов в пространстве терминов (например, алгоритм Роккио, алгоритмы классификации без учителя).

Под *терминами* документов будем понимать все одиночные слова, встреченные в тексте хотя бы одного документа коллекции, за исключением стоп-слов, то есть распространённых слов, не характеризующих документы по смыслу, например, предлогов, союзов и т. п. Вдобавок, каждой встреченной форме слова, например, в разных падежах и числах, будет соответствовать один и тот же термин, например, данное слово в начальной форме. В результате получаем множество всех терминов коллекции $\mathcal{T} = \{t_k\}, k = \overline{1, |\mathcal{T}|}$.

Образом документа как вектора в пространстве терминов является вектор действительных чисел $\overrightarrow{d_i} = \left(d_{i1}, \dots, d_{i|\mathcal{T}|}\right)^T$, где каждое действительное число является координатой вектора, соответствующей конкретному термину, и равняется *весу термина* в данном документе. Наиболее часто используют следующий подход к вычислению веса термина:

$$d_{ij} = \frac{w_{ij}}{\|\overrightarrow{w_i}\|}, w_{ij} = tf_{ij} \times log\frac{|\mathcal{D}|}{df_j}, \quad (1)$$

где tf_{ij} – частота термина в документе, то есть количество раз, которое j-ый термин встретился в i-ом документе; df_j – документная частота, то есть количество документов, в которых встретился j-ый термин; $\|\overrightarrow{w_i}\|$ – евклидова норма $\overrightarrow{w_i}$. Такие веса d_{ij} называют нормированными весами по формуле «TF-IDF» («частота термина – обратная документная частота»), $0 \le d_{ij} \le 1$. Они обладают следующими свойствами: (а) имеют высокие значения, если термин часто встречается в небольшом числе документов, тем самым усиливая отличие этих документов от других, (б) имеют низкие значения, если термин редко встречается в каком-то документе или встречается во многих документов, тем самым снижая различие между документами.

Процесс классификации документов как векторов основан на гипотезе о том, что тематически близкие документы окажутся в пространстве терминов геометрически близко расположенными. Поэтому в основе алгоритмов классификации лежит понятие сходства или расстояния между документами в пространстве терминов.

Меры сходства и различий между образами документов. В данном случае понятия расстояния и сходства являются взаимнообратными, расстояние можно было бы называть различием. Выбор способа вычисления расстояния влияет на результат классификации. Часто применяют следующие варианты:

$$dist\left(\vec{d}_i, \vec{d}_j\right) = \left(\sum_{k=1}^{|\mathcal{T}|} \left|d_{ik} - d_{jk}\right|^r\right)^{\frac{1}{r}}, \quad (2)$$

где r – это параметр, заданный пользователем, $r \in \mathbb{R}, r > 0$. Распространённые примеры:

а) при r = 1: *манхэттенское расстояние*, или расстояние городских кварталов;

б) при r = 2: *евклидово расстояние*;

в) при r → ∞ получим *расстояние Чебышева*, которое вычисляется как максимум модуля разности компонент этих векторов $dist\left(\vec{d}_i, \vec{d}_j\right) = \max_{k=1,..,|\mathcal{T}|}\left|d_{ik} - d_{jk}\right|$.

Другой часто используемой на практике мерой сходства является *косинусная мера*:

$$sim(\vec{d}_i, \vec{d}_j) = \cos\left(\angle(\vec{d}_i, \vec{d}_j)\right) = \frac{\sum_{k=1}^{|\mathcal{T}|} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{|\mathcal{T}|} d_{ik}^2} \times \sqrt{\sum_{k=1}^{|\mathcal{T}|} d_{jk}^2}}. \quad (3)$$

Если вектора весов документов нормированы как в (1), то косинусная мера есть скалярное произведение векторов. Если векторы ортогональны, то мера близости равна 0, если совпадают, то 1.

Заметим, что в случае, когда вектора весов терминов нормированы, значения евклидова расстояния и косинусной меры соответствуют друг другу.

§ 1.2. Отбор терминов для классификации

Большое количество терминов (признаков) документов в задаче классификации приводит к ряду проблем, среди которых: (1) высокие вычислительные затраты, связанные, например, с получение значений меры близости между документами и др., (2) низкое качество классификации, вызванное наличием большого числа признаков со слабой классификационной способностью. Такие признаки часто называют шумовыми, при их добавлении к представлению документа ошибка классификации на новых данных возрастает.

В частности, (2) в классификации с учителем может привести к переобучению классификатора, то есть эффекту, возникающему, когда классификатор настраивался в большей степени на случайных (шумовых) характеристиках документов, а не на существенных для их тематик (категорий). В такой ситуации алгоритм хорошо работает на тех данных, на которых он был обучен, и значительно хуже на новых.

Таким образом, стремятся сократить число термином из множества T так, чтобы новое (редуцированное) множество терминов $\mathcal{T}'$ ($|\mathcal{T}'| \ll |\mathcal{T}|$) содержало наиболее информативные в некотором смысле термины.

Техники сокращения размерности пространства терминов (редукции) применяют двумя способами: локально (сокращают множество терминов для каждой категории в отдельности) и глобально (работают с общим множеством термином для всех документов). Первый случай применим для классификации с учителем, второй – как для классификации с учителем, так и без него. Мы рассмотрим такие техники, которые пригодны для обоих подходов – локального и глобального.

Другое существенное различие между техниками редукции заключается в природе итоговых терминов. Одни техники достигают сокращения числа терминов путём отсева некоторого числа исходных терминов, руководствуясь заданным критерием отсева, - *техники отбора признаков*. Тогда $\mathcal{T}' \subset \mathcal{T}$. Другие техники формируют новые термины путём комбинации или преобразования исходных терминов, другими словами извлекают из итоговые признаки из исходных данных – *техники извлечения признаков*. Тогда элементы множества T' имеют тип, отличный от элементов множества T. Отбор и извлечение терминов реализуются различными техниками. Мы рассмотрим только техники отбора признаков (терминов).

Отбор признаков документов. Итак, необходимо отобрать такие термины, которые повысят качество классификатора. Для этого поместим в $\mathcal{T}'$ все термины из $\mathcal{T}$, которые имеют высокое значение «важности для разбиения по категориям/классам». Для определения «важности» термина и способа её вычисления используют разные подходы.

Документная частота (DF). Самая простая и вполне эффективная техника оценки «важности терминов для классификации» основана на наблюдении того, что значительное число терминов коллекции встречаются в малом числе документов, а наибольшую информативность имеют термины со средней или даже высокой частотой, если предварительно были удалены стоп-слова. На практике часто используют пороговое значение τ, равное 1-5 документам. Таким образом, $\mathcal{T} = \{t_k \in \mathcal{T}: DF(t_k) > \tau\}$, где $DF(t_k)$ – это количество документов, в которых встречается термин t_k. Данная техника может применяться как единственная, так и предшествовать другой технике отбора признаков.

Следующие техники берут своё начало из теории информации: взаимная информация, информационная выгода и критерий хи-квадрат. Мы рассмотрим их локальные значения $f(t_k, c_j)$, чтобы получить значение глобально (вне зависимости от конкретной категории), следует вычислить либо простую сумму $\sum_{j=1}^{|C|} f(t_k, c_j)$, либо взвешенную сумму $\sum_{j=1}^{|C|} P(c_j) f(t_k, c_j)$, либо найти максимум $\max_{j=\overline{1,|C|}} f(t_k, c_j)$.

Взаимная информация (MI). Величина взаимной информации термина *t* и категории *c*:

$$MI(t_k, c_j) = \log_2 \frac{P(t_k, c_j)}{P(t_k) \times P(c_j)}. \quad (4)$$

Пусть A – количество документов, принадлежащих категории c и содержащих термин *t*;

B – количество документов, не принадлежащих категории c и содержащих термин *t*;

C – количество документов, принадлежащих категории c и не содержащих термин t.

Тогда выражение (4) можно записать следующим образом:

$$MI(t_k, c_j) = \log_2 \frac{A \times |\Omega|}{(A + C) \times (A + B)}, \quad (5)$$

где Ω – обучающее множество документов.

$MI(t_k, c_j)$ принимает значение 0, если термин *t* и категория *c* независимы.

Недостаток взаимной информации заключается в том, что её значение сильно подвержено влиянию безусловной вероятности терминов, так как $MI(t_k, c_j) = \log_2 P(t_k|c_j) - \log_2 P(t_k)$ (это следует из (4)). Если два термина имеют одинаковую условную вероятность, более высокое значение MI будет у более редкого. Следовательно, значения взаимной информации несравнимы для терминов с существенно различающейся частотой встречаемости в документах.

*Информационная выгода (IG).*Информационную выгоду часто называют *ожидаемой взаимной информацией* (EMI). Этот показатель измеряет количество информации о принадлежности категории c, которое несёт наличие/отсутствие термина *t*.

$$IG(t_k, c_j) = \sum_{c \in \{c_j, \overline{c_j}\}} \sum_{t \in \{t_k, \overline{t_k}\}} P(t, c) \times \log_2 \frac{P(t, c)}{P(t) \times P(c)}, \quad (6)$$

где $\overline{c_j}$ – все категории, кроме c_j; $t_k, \overline{t_k}$ – признаки наличия и отсутствия термина t_k соответственно.

На практике формула (6) эквивалентна следующей:

$$IG(t_k, c_j) = \frac{A}{|\Omega|} \times \log_2 \frac{|\Omega| \times A}{(A+B) \times (A+C)} + \frac{C}{|\Omega|} \times \log_2 \frac{|\Omega| \times C}{(C+D) \times (A+C)} + \frac{B}{|\Omega|} \times \log_2 \frac{|\Omega| \times B}{(A+B) \times (B+D)} + \frac{D}{|\Omega|} \times \log_2 \frac{|\Omega| \times D}{(C+D) \times (B+D)}, \quad (7)$$

где D – количество документов, не принадлежащих категории c и не содержащих термин t.

Мера информационной выгоды достигает своего максимум, когда термин является идеальных индикатором категории, то есть присутствует в документе тогда и только тогда, когда документ принадлежит классу. Если распределение термина в категории соответствует распределению термина в коллекции, то информационная выгода равна 0.

При заданном обучающем множестве для каждого термина вычисляют значение IG и удаляют из $\boldsymbol{\mathcal{T}}$ такие термины, значение информационной выгоды которых ниже некоторого заранее выбранного порогового значения.

Критерий хи-квадрат (CHI). Критерий χ^2 используется для проверки независимости двух случайных событий: появление термина X и появление класса Y. Если X и Y независимы, то P(XY)=P(X)P(Y). Критерий χ^2 вычисляется по формуле:

$$CHI(t_k, c_j) = \sum_{c \in \{c_j, \overline{c_j}\}} \sum_{t \in \{t_k, \overline{t_k}\}} \frac{\left(P(t,c) - P_{exp}(t,c)\right)^2}{P_{exp}(t,c)}, \quad (8)$$

где $P(t,c)$ – наблюдаемая на обучающем множестве, $P_{exp}(t,c)$ – ожидаемая при условии, что термин и класс являются независимыми. Величина критерия χ^2 позволяет судить о том, насколько ожидаемая и наблюдаемая вероятности отклоняются друг от друга, и принимает значение 0, если термин и категория независимы. Критерий вычисляют локально (для каждой категории), затем получают его глобальное значение, по которому ранжируют признаки коллекции документов.

На практике формула (8) эквивалентна следующей:

$$CHI(t_k, c_j) = \frac{|\Omega| \times (A \times D - C \times B)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}, \quad (9)$$

В отличие от взаимной информации критерий χ^2нормализован, что позволяет сравнивать между собой его значения для разных термов одной категории, исключением являются только редкие термы.

Пример.

MI(китайский, c) = $\log_2 \frac{3\times4}{(3+0)\times(3+1)} = 0$.

MI(китайский, $\overline{c}$) = $\log_2 \frac{1\times4}{(1+0)\times(1+3)} = 0$.

MI(пекин, c) = $\log_2 \frac{1\times4}{(1+2)\times(1+0)} \approx 0{,}41$.

MI(пекин, $\overline{c}$) = $\log_2 \frac{0\times4}{(0+2)\times(0+1)} =$ не опр.

IG(китайский, c) = $\frac{3}{4}\log_2 \frac{4\times3}{(3+1)\times(3+0)} + \frac{0}{4}\log_2 \frac{4\times0}{(0+0)\times(3+0)} + \frac{1}{4}\log_2 \frac{4\times1}{(3+1)\times(1+0)} + \frac{0}{4}\log_2 \frac{4\times0}{(0+0)\times(1+0)} = 0$.

IG(китайский, $\overline{c}$) = 0.

IG(пекин, c) = $\frac{1}{4}\log_2\frac{4\times1}{(1+0)\times(1+2)} + \frac{2}{4}\log_2\frac{4\times2}{(2+1)\times(1+2)} + 0 + \frac{1}{4}\log_2\frac{4\times1}{(2+1)\times(0+1)} = 0{,}21.$

IG(пекин, $\overline{c}$) = $0 + \frac{1}{4}\log_2\frac{4\times1}{(1+2)\times(0+1)} + \frac{1}{4}\log_2\frac{4\times1}{(0+1)\times(1+2)} + \frac{2}{4}\log_2\frac{4\times2}{(1+2)\times(1+2)} = 0{,}12.$

Принято читать, что при $p = 0$ выражение $p\log p$ равно нулю.

CHI(китайский, c) = $\frac{4\times(3\times0-0\times1)^2}{(3+0)\times(1+0)\times(3+1)\times(0+0)} =$ не опр.

CHI(китайский, $\overline{c}$) = $\frac{4\times(1\times0-0\times3)^2}{(1+0)\times(3+0)\times(1+3)\times(0+0)} =$ не опр.

CHI(пекин, c) = $\frac{4\times(1\times1-2\times0)^2}{(1+2)\times(0+1)\times(1+0)\times(2+1)} \approx 0{,}44.$

CHI(пекин, $\overline{c}$) = $\frac{4\times(0\times2-1\times1)^2}{(0+1)\times(1+2)\times(0+1)\times(1+2)} \approx 0{,}44.$

Видим, что по всем критериям термин «китайский» − шумовой для наших категорий и обучающего множества.

Извлечение признаков документов. Необходимо синтезировать новые (искусственные) признаки документов так, чтобы повысить качество классификации, например, путём разрешения неоднозначностей естественного языка, например, синонимии, омонимии, полисемии. Затем следует отобразить документы коллекции в новое признаковое пространство, которое лишено старых проблем и лучше, чем исходное, представляет содержание документов. Примерами техник извлечения признаков документов являются латентно-семантическое индексирование и кластеризация терминов документов.

Техники данной группы мы рассматривать не будем.

§ 1.3. Алгоритм "наивной" байесовской классификации

Алгоритм наивной байесовской классификации использует формулу Байеса для оценки вероятности принадлежности документа классу на обучающем множестве. Образ документа рассматривается как мультимножество терминов. Например, имеем документ с текстом «Китай лидирует по темпам роста ВВП среди развитых стран», тогда образом документа является {китай, лидировать, темп, рост, ввп, развитый, страна}.

Целью классификации является поиск наилучшего класса для документа, то есть имеющего наибольшую апостериорную вероятность $P(c_i|d_j)$:

$$c^* = \arg_{c_j\in\boldsymbol{C}}\max P(c_j|d_i), \tag{10}$$

где $d_i \in \boldsymbol{\Omega}, c_j \in \boldsymbol{C}$.

По формуле Байеса:

$$P(c_j|d_i) = \frac{P(c_j)P(d_i|c_j)}{P(d_i)} \approx P(c_j)P(d_i|c_j), \tag{11}$$

где $P(c_j)$ – априорная вероятность, что документ принадлежит c_j;

$P(d_i|c_j)$ – вероятность встретить документ типа d_i среди документов, класса c_j.

Поскольку $P(d_i)$ не влияет на выбор класса, итоговое ранжирование классов по априорной вероятности можно провести без учёта знаменателя в формуле (11).

Наивным данный алгоритм называют потому, что он использует *наивное допущение*, что слова, входящие в текст документа, не зависят друг от друга.

Следовательно, $P(d_i|c_j)$ можно вычислить как произведение вероятностей встретить термин t_k документах класса c_j:

$$P(d_i|c_j) = \prod_{k=1}^{|\mathcal{T}_{d_i}|} P(t_k|c_j), \tag{12}$$

где $\mathcal{T}_{d_i}$ – множество терминов документа d_i; $P(t_k|c_j)$ – оценка термина t_k вклада в то, что $d_i \in c_j$.

Тогда решающее правило (10) принимает окончательный вид:

$$c^* = \arg_{c_j \in \boldsymbol{C}} \max P(c_j) \prod_{k=1}^{|\mathcal{T}_{d_i}|} P(t_k|c_j), \tag{13}$$

На практике может наблюдаться потеря значащих разрядов при умножении $|\mathcal{T}_{d_i}|$ условных вероятностей. Тогда в выражении (13) вместо самих оценок вероятностей используют логарифм этих вероятностей. Поскольку логарифм – монотонно возрастающая функция, то класс c_j с наибольшим значением логарифма вероятности останется наиболее вероятным. Тогда

$$c^* = \arg_{c_j \in \boldsymbol{C}} \max \left[\log P(c_j) + \sum_{k=1}^{|\mathcal{T}_{d_i}|} \log P(t_k|c_j) \right], \tag{14}$$

Оценки вероятностей на обучающем множестве:

$$P(c_j) = \frac{\left|\boldsymbol{\mathcal{D}}_{c_j}\right|}{|\boldsymbol{\mathcal{D}}|}, \tag{15}$$

$$P(t_k|c_j) = \frac{tf(t_k, c_j)}{\sum_{i=1}^{|\boldsymbol{\mathcal{T}}|} tf(t_i, c_j)},$$

где $\boldsymbol{\mathcal{D}}_{c_j}$– множество документов в классе c_j; $\boldsymbol{\mathcal{D}}$ – количество всех документов ($\boldsymbol{\mathcal{D}} = \boldsymbol{\Omega}$);

$tf(t_k, c_j)$ – количество вхождений термина t_k в документе класса c_j;

$\boldsymbol{\mathcal{T}}$ – словарь всей коллекции документов.

Поскольку обучающее множество не может быть достаточно большим, чтобы содержать все термины, которые могут встретиться в новых документах, тогда если новый документ, содержит новый (редкий) термин, то вероятность принадлежности своему классу будет равна нулю (следует из (15) и (12)). Для решения этой проблемы на практике применяют сглаживание, например, следующего вида:

$$P(t_k|c_j) = \frac{tf(t_k, c_j) + 1}{\sum_{i=1}^{|\boldsymbol{\mathcal{T}}|}(tf(t_i, c_j) + 1)} = \frac{tf(t_k, c_j) + 1}{\sum_{i=1}^{|\boldsymbol{\mathcal{T}}|} tf(t_i, c_j) + |\boldsymbol{\mathcal{T}}|}, \tag{16}$$

Добавление единицы к каждой частоте встречаемости термина можно интерпретировать как априорное равномерное распределение (каждый термин встречается в каждом классе по одному разу), которое затем на обучающем множестве уточняется.

Алгоритм в общем виде.

Обучение.

Вход: $\boldsymbol{C}$ и $\boldsymbol{\mathcal{D}} = \Omega$.

Шаг 1. Составить словарь $\boldsymbol{\mathcal{T}}$ из $\boldsymbol{\mathcal{D}}$.

Шаг 2. Для каждого $c_j \in \boldsymbol{\mathcal{C}}$:

Шаг 3. *prior*[j] := $\left|\boldsymbol{\mathcal{D}}_{c_j}\right| / |\boldsymbol{\mathcal{D}}|$;

Шаг 4. *text*[j] := <склеить тексты всех документов $d_i \in \boldsymbol{\mathcal{D}}$: $d_i \in c_j$>;

Шаг 5. Для каждого $t_k \in \boldsymbol{\mathcal{T}}$:

Шаг 6. tf[k][j] := <вычислить количество вхождений термина t_k в text[j]>;

Шаг 7. Для каждого $t_k \in \boldsymbol{\mathcal{T}}$:

Шаг 8. cp[k][j] := $(tf[\,k\,][\,j\,] + 1)/\left(\sum_{i=1}^{|\boldsymbol{\mathcal{T}}|} tf[\,i\,][\,j\,] + |\boldsymbol{\mathcal{T}}|\right)$;

Выход: $\boldsymbol{\mathcal{T}}$, *prior*, *cp*.

Тестирование (применение).

Вход: $\boldsymbol{\mathcal{C}}$; $\boldsymbol{\mathcal{T}}$; *prior; cp;* $d \in \boldsymbol{\mathcal{D}}$.

Шаг 1. *terms* := <извлечь термины из *d* с учётом $\boldsymbol{\mathcal{T}}$ и дополнениями cp>;

Шаг 2. Для каждого $c_j \in \boldsymbol{\mathcal{C}}$:

Шаг 3. score[j] := log prior[j];

Шаг 4. Для каждого $t_k \in terms$:

Шаг 5. score[j] := score[j] + log cp[k][j];

Шаг 6. c^* := $\arg_j \max$ score[j];

Выход: c^*.

Пример. Определим класс документа d5 для коллекции документов, заданной в разделе 1.

Обучение:

1) $P(c) = \frac{3}{4}; P(\overline{c}) = \frac{1}{4}$;
2) $P(\text{китайский}|c) = \frac{5+1}{8+6} = \frac{3}{7};\ P(\text{токио}|c) = P(\text{япония}|c) = \frac{0+1}{8+6} = \frac{1}{14}$;
3) $P(\text{китайский}|\overline{c}) = P(\text{токио}|\overline{c}) = P(\text{япония}|\overline{c}) = \frac{1+1}{3+6} = \frac{2}{9}$.

Применение:

$$P(d_5|c) = \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \approx 0{,}0003;$$

$$P(d_5|\overline{c}) = \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \frac{2}{9} \times \frac{2}{9} \approx 0{,}0001.$$

Следовательно, $c^* = c$, то есть «Китай».

Вычислительная сложность.

Обучение: линейная сложность относительно размера коллекции документов $O(|\Omega|)$.

Тестирование: линейная сложность относительно числа категорий $O(|\mathcal{C}|)$.

§ 1.4. Алгоритм Роккио

Алгоритм Роккио рассматривает документы в векторном пространстве терминов и ищет границы между классами как множества точек, равноудалённых от центроидов этих классов. *Центроидом* класса называется усреднённый вектор, или центр масс членов класса:

$$\vec{\mu}_{c_j} = \frac{1}{|\boldsymbol{\mathcal{D}}_{c_j}|} \sum_{i:\, d_i \in c_j} \overrightarrow{d_i}, \tag{17}$$

где $\boldsymbol{\mathcal{D}}_{c_j}$ – множество документов в классе c_j.

Граница между двумя классами в многомерном пространстве терминов имеет вид гиперплоскости:

$$\vec{w}^T \vec{x} = b, \tag{18}$$
$$\vec{w} = \vec{\mu}_1 - \vec{\mu}_2,$$
$$b = \frac{1}{2}(\|\vec{\mu}_1\|^2 - \|\vec{\mu}_2\|^2),$$

где $\vec{w}$ – вектор нормали; b – константа.

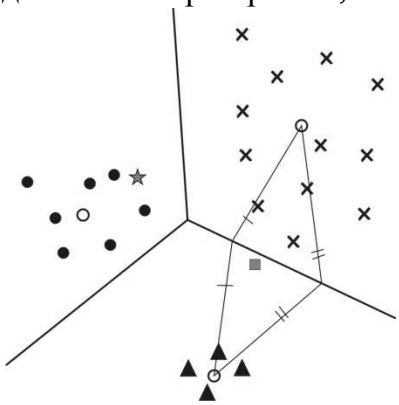


Рис. 1. Иллюстрация работы алгоритма Роккио

Правило классификации заключается в определении области, в которую попадает новый документ, то есть в поиске центроида, к которому образ нового документа ближе, чем к остальным центроидам. На рис. 1 к документу «звёздочка» ближе всех центроид класса «кружков».

Алгоритм Роккио предполагает, что классы имеют форму сфер с примерно одинаковыми радиусами. Если это предположение не выполняется, то алгоритм может привести к неудовлетворительным результатам. Например, на рис. 1 документ «квадрат» больше подходит классу «крестиков», а алгоритм отнесёт его к классу «треугольников».

Алгоритм в общем виде.

Обучение.

Вход: $\boldsymbol{\mathcal{C}}$ и $\boldsymbol{\mathcal{D}} = \Omega$.

Шаг 1. Для каждого $c_j \in \boldsymbol{\mathcal{C}}$:

Шаг 2. $\vec{\mu}_j = \frac{1}{|\boldsymbol{\mathcal{D}}_{c_j}|} \sum_{i:\, d_i \in c_j} \overrightarrow{d_i}$;

Выход: $\{\vec{\mu}_1, \dots, \vec{\mu}_{|\boldsymbol{\mathcal{C}}|}\}$.

Тестирование (применение).

Вход: $\{\vec{\mu}_1, \dots, \vec{\mu}_{|\boldsymbol{\mathcal{C}}|}\}; d \in \boldsymbol{\mathcal{D}}$.

Шаг 1. $\vec{d} := \left(d_1, \dots, d_{|\mathcal{T}|}\right)^T$ по формуле (1);

Шаг 2. $c^* := \arg_j \min \left\|\vec{\mu}_j - \vec{d}\right\|$;

Выход: c^* .

Пример. Рассчитаем веса терминов для нашей коллекции из 5 документов: 4 документа – обучающее множество Ω, и 1 документ из тестового множества.

	t1	***t2***	***t3***	***t4***	***t5***	***t6***
	китайский	*пекин*	*шанхай*	*макао*	*япония*	*токио*
df на Ω	4	1	1	1	1	1
d1	tf=2 | w=0| **0**	1 | 0,6 | **1**	**0**	**0**	**0**	**0**
d2	2 | 0 | **0**	**0**	1 | 0,6 | **1**	**0**	**0**	**0**
d3	1 | 0 | **0**	**0**	**0**	1 | 0,6 | **1**	**0**	**0**
d4	1 | 0 | **0**	**0**	**0**	**0**	1 | 0,6 | **0,7**	1 | 0,6 | **0,7**
d5	3 | 0 |**0**	**0**	**0**	**0**	1 | 0,6 | **0,7**	1 | 0,6 | **0,7**

Обучение. Вычислим центроиды классов c и $\overline{c}$:

	t1	***t2***	***t3***	***t4***	***t5***	***t6***
	китайский	*пекин*	*шанхай*	*макао*	*япония*	*токио*
μ_c	**0**	**0,33**	**0,33**	**0,33**	**0**	**0**
$\mu_{\overline{c}}$	**0**	**0**	**0**	**0**	**0,7**	**0,7**

Тестирование.

$$\left\|\overrightarrow{d_5} - \overrightarrow{\mu_c}\right\| = \sqrt{0 + 0{,}33^2 + 0{,}33^2 + 0{,}33^2 + 0{,}7^2 + 0{,}7^2} \approx 1{,}14,$$

$$\left\|\overrightarrow{d_5} - \overrightarrow{\mu_{\overline{c}}}\right\| = \sqrt{0 + 0 + 0 + 0 + 0 + 0} = 0{,}00.$$

Следовательно, c* = $\overline{c}$, то есть «не Китай».

Разделяющая гиперплоскость имеет следующий вид:

$$\vec{w} \approx (0; 0{,}33; 0{,}33; 0{,}33; -0{,}7; -0{,}7)^T;$$

$$b \approx \frac{1}{2} \times (0{,}33 - 0{,}98) \approx -0{,}33$$

Документы класса c и $\overline{c}$ лежат по разные стороны от гиперплоскости:

$$\vec{w}^T \times \overrightarrow{d_1} = \vec{w}^T \times \overrightarrow{d_2} = \vec{w}^T \times \overrightarrow{d_3} \approx 0{,}33 > b;$$

$$\vec{w}^T \times \overrightarrow{d_4} = \vec{w}^T \times \overrightarrow{d_5} \approx -0{,}98 < b.$$

Вычислительная сложность.

Обучение: линейная сложность относительно размера коллекции документов $O(|\Omega|)$.

Тестирование: линейная сложность относительно числа категорий $O(|\mathcal{C}|)$.

§ 1.5. Алгоритм k-ближайших соседей

Алгоритм k-ближайших соседей использует гипотезу компактности векторного пространства, которая заключается в том, что документы одного класса образуют в пространстве терминов компактную область, причём области разных классов не пересекаются. Тогда можно ожидать, что тестовый документ будет иметь такую же метку класса, как и окружающие его документы из обучающего множества. Алгоритм k-ближайшего соседа относит тестовый документ к преобладающему классу его k соседей. При k = 1 алгоритм относит документ к классу, самого ближайшего ему документа.

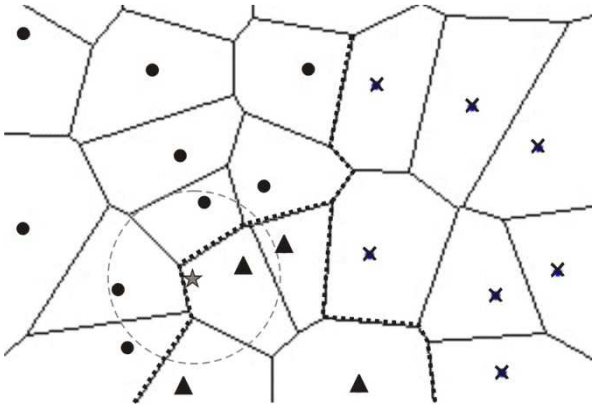


Рис. 2. Иллюстрация работы алгоритма k-ближайших соседей при k = 1

Данный алгоритм лучше справляется с несферическими или несвязными классами, чем алгоритм Роккио, поскольку определяет границы между классами локально. Для всех документов обучающего множества пространство терминов представляется разделенным на ячейки (выпуклые многогранники), состоящие из точек, которые ближе к данному объекту, чем к другим. Это в случае k = 1. В случае k > 1 внутри ячеек также множество k-ближайших соседей остаётся инвариантным.

На рис. 2 видно, что новый документ «звёздочка» попадает в ячейку объекта класса «треугольников», и при k = 1 будет отнесён к этому же классу. Однако при k = 3 «звёздочка» будет отнесён к классу «кружочков».

При k = 1 алгоритм неустойчив, так как классификация зависит всего от одного обучающего документа, а он может быть нетипичным или иметь неверную метку класса. На практике значение k выбирают на основе опыта эксперта и имеющихся знаний о решаемой задаче. Кроме того, число соседей можно подобрать на обучающем множестве так, чтобы максимизировать качество классификации.

На практике для повышения точности выбора класса могут учитываться веса «голосов» соседей, которые используются для ранжирования классов:

$$\text{ранг}(c_j, d) = \sum_{d' \in S_k} I_{c_j}(d') sim\left(\vec{d'}, \vec{d}\right), \quad (19)$$

где $I_{c_j}(d') = 1$, если $d' \in c_j$, иначе $I_{c_j}(d') = 0$;

$sim\left(\vec{d'}, \vec{d}\right)$ – косинусная мера близости (см. (3)).

Документ приписывается классу с наибольшим рангом. Так, например, если одинаковое число соседей принадлежит двум разным классам, то выбирается класс с более близкими соседями.

Алгоритм в общем виде.

Обучение.

Вход: $\boldsymbol{\mathcal{C}}$ и $\boldsymbol{\mathcal{D}} = \Omega$.

Шаг 1. k := <подходящее значение числа соседей>;

Выход: k .

Тестирование (применение).

Вход: $\boldsymbol{\mathcal{C}}$; Ω; $d \in \boldsymbol{\mathcal{D}}$, k.

Шаг 1. S_k := <множество k ближайших соседей для d>;

Шаг 2. Для каждого $c_j \in \boldsymbol{\mathcal{C}}$:

Шаг 3. $p[\,j\,] := \left|S_k \cap \boldsymbol{\mathcal{D}}_{c_j}\right|/k$; {где $p[\,j\,]$ – оценка вероятности того, что $d \in c_j$}

Шаг 4. $c^* := \arg_j \max p[\,j\,]$;

Выход: c^* .

Пример. Пусть k = 1. Тогда

$$\left\|\vec{d}_5 - \vec{d}_1\right\| = \left\|\vec{d}_5 - \vec{d}_2\right\| = \left\|\vec{d}_5 - \vec{d}_3\right\| \approx 1{,}92;$$
$$\left\|\vec{d}_5 - \vec{d}_4\right\| = 0{,}00.$$

Следовательно, c* = $\overline{c}$, то есть «не Китай».

Вычислительная сложность.

Обучение: сложность $O(1)$ в случае, если не используются никакие дополнительные техники подбора значения k.

Тестирование: линейная сложность относительно числа документов $O(|\boldsymbol{\Omega}|)$.

§ 1.6. Алгоритм опорных векторов

Алгоритм опорных векторов (SVM, Support Vector Machines), разработанный В. Н. Вапником в 1990-е годы, ищет в векторном пространстве документов разделяющую поверхность между двумя классами, максимально удалённую от всех точек обучающего множества. Расстояние между найденной поверхностью и ближайшей точкой данных называется *зазором классификации*. В алгоритме опорных векторов решающая поверхность полностью определяется небольшим подмножеством документов. Элементы данного подмножества называются *опорными векторами.*

Итак, пусть алгоритм ищет разделяющую поверхность заданную уравнением:

$$\vec{w}^T\vec{x} = -b, \qquad (20)$$

где $\vec{w}$ – вектор нормали к разделяющей поверхности, или *вектор весов*; b – параметр сдвига.

Обучающее множество представляет собой множество пар $\Omega = \{(\vec{x}_i, y_i)\}$, где $\vec{x}_i$ – это документы обучающего множества, а y_i – это соответствующая метка класса, причём в алгоритме опорных векторов метка принимает одно из двух значений +1 и -1.

Тогда линейный классификатор описывается следующей формулой:

$$f(\vec{x}) = sign\,(\vec{w}^T\vec{x} + b), \qquad (21)$$

где значение классификатора -1 соответствует одному классу, а +1 – другому; $\vec{x}$ – тестовый документ.

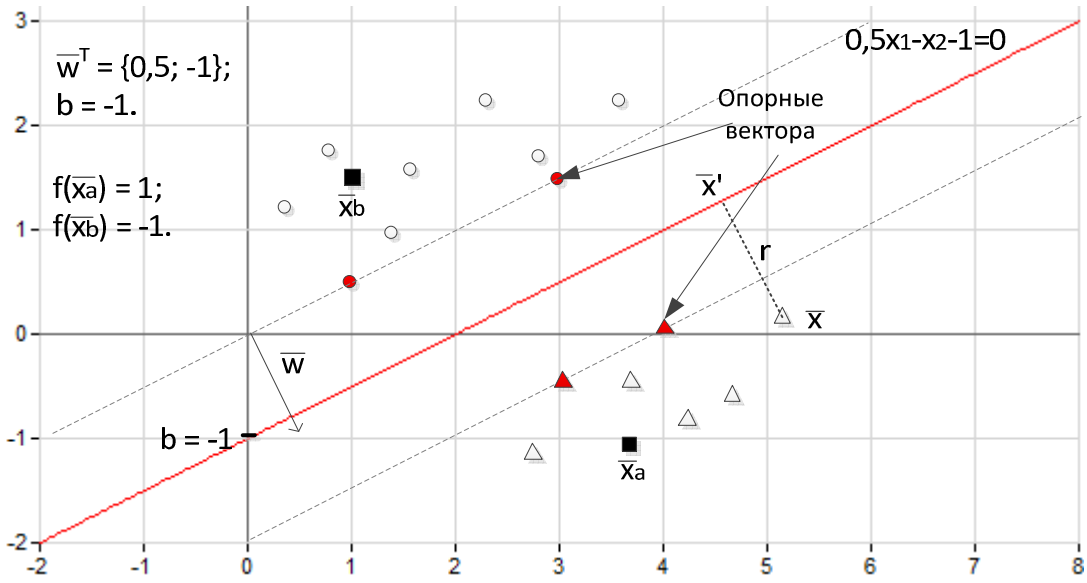


Рис. 3. Опорные вектора и разделяющая поверхность

Разделяющая поверхность ищется таким образом, чтобы максимизировать зазор классификации. Геометрически зазор – это максимальная ширина полосы, которую можно провести между опорными векторами двух классов, то есть значение, вдвое превышающее минимальное значение r (рис. 3). Значение r вычисляется так: пусть $\vec{x}'$ – документ, лежащий на разделяющей гиперплоскости и ближайший к точке $\vec{x}$; перпендикуляр из $\vec{x}'$ в $\vec{x}$ параллелен вектору $\vec{w}$; единичный вектор в этом направлении имеет вид $\vec{w}/\|\vec{w}\|$; тогда $\overrightarrow{\vec{x}' = \vec{x} - yr(w/\|\vec{w}\|)}$, где умножение на y просто меняет знак. Точка $\vec{x}'$ лежит на поверхности, следовательно:

$$\vec{w}^T\left(\vec{x} - yr\frac{\vec{w}}{\|\vec{w}\|}\right) + b = 0, \tag{22}$$

$$r = y\frac{\vec{w}^T\vec{x} + b}{\|\vec{w}\|}. \tag{23}$$

Потребуем, чтобы для всех точек $(\vec{x}_i, y_i) \in$? выполнялось следующее неравенство:

$$y_i(\vec{w}^T\vec{x}_i + b) \geq 1. \tag{24}$$

На опорных векторах неравенство (24) превращается в равенство. Из (24) и (23) следует, что геометрический зазор равен $\rho = \frac{2}{\|\vec{w}\|}$.

Таким образом, задача алгоритма опорных векторов заключается в поиске параметров $\vec{w}$ и b, удовлетворяющих следующим условия:

а) величина $\frac{\|\vec{w}\|}{2} = \frac{1}{2}\vec{w}^T\vec{w}$ достигает минимума;

б) при всех $(\vec{x}_i, y_i) \in \Omega$ выполняется неравенство (24).

Имеем задачу минимизации квадратичной функции при линейных ограничениях. Для решения задачи квадратичной оптимизации разработано множество алгоритмов, рассмотрение, которых выходит за рамки наших лекций. Однако для понимания алгоритма опорных векторов необходимо привести следующую информацию о её решении. Для решения данной задачи формулируют двойственную задачу, в которой с каждым ограничением вида (24) прямой задачи связан соответствующий множитель Лагранжа α_i, и задача заключается в поиске

значений $\alpha_1, \dots, \alpha_{|\Omega|}$ при которых величина $\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j$ достигает максимума, $\sum_i \alpha_i y_i = 0$, $\alpha_i \geq 0$ для всех $1 \leq i \leq |\Omega|$.

Решение задачи имеет следующий вид:

$$\vec{w} = \sum_{i=1}^{|\Omega|} \alpha_i y_i \vec{x}_i, \tag{25}$$

$$b = y_k - \vec{w}^T \vec{x}_k, \text{для } \forall \vec{x}_k, \text{таких что } \alpha_k \neq 0,$$

$$f(\vec{x}) = sign\left(\sum_{i=1}^{|\Omega|} \alpha_i y_i \vec{x}_i^T \vec{x} + b\right), \tag{26}$$

Большинство параметров α_i равны нулю, ненулевое значение означает, что соответствующий вектор $\vec{x}_i$ является опорным.

Пример.

Обучение. С помощью статистического программного пакета получим значения параметров $\alpha_1, \dots, \alpha_{|\Omega|}$:

$\alpha_1 \approx 0{,}31$; $\alpha_2 \approx 0{,}23$; $\alpha_3 \approx 0{,}23$; $\alpha_4 \approx 0{,}78$. Все обучающие документы стали опорными.

$y_1 = -1$; $y_2 = -1$; $y_3 = -1$; $y_4 = 1$.

Из (25):

$$\vec{w} = 0{,}31 \times (-1) \times \begin{pmatrix}0\\1\\0\\0\\0\end{pmatrix} + 0{,}23 \times (-1) \times \begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix} + 0{,}23 \times (-1) \times \begin{pmatrix}0\\0\\0\\1\\0\end{pmatrix} + 0{,}78 \times (+1) \times \begin{pmatrix}0\\0\\0\\0\\0{,}7\\0{,}7\end{pmatrix} = \begin{pmatrix}0\\-0{,}31\\-0{,}23\\-0{,}23\\0{,}55\\0{,}55\end{pmatrix}.$$

$$b = \frac{1}{4}\left(\left[(-1) - (0 \quad -0{,}31 \quad -0{,}23 \quad -0{,}23 \quad 0{,}55 \quad 0{,}55) \times \begin{pmatrix}0\\1\\0\\0\\0\\0\end{pmatrix}\right]\right.$$

$$+ \left[(-1) - (0 \quad -0{,}31 \quad -0{,}23 \quad -0{,}23 \quad 0{,}55 \quad 0{,}55) \times \begin{pmatrix}0\\0\\1\\0\\0\\0\end{pmatrix}\right]$$

$$+ \left[(-1) - (0 \quad -0{,}31 \quad -0{,}23 \quad -0{,}23 \quad 0{,}55 \quad 0{,}55) \times \begin{pmatrix}0\\0\\0\\1\\0\\0\end{pmatrix}\right]$$

$$\left. + \left[(1) - (0 \quad -0{,}31 \quad -0{,}23 \quad -0{,}23 \quad 0{,}55 \quad 0{,}55) \times \begin{pmatrix}0\\0\\0\\0\\0{,}7\\0{,}7\end{pmatrix}\right]\right)$$

$$= \frac{1}{4}(-0{,}69 - 0{,}77 - 0{,}77 + 0{,}23) = -0{,}5.$$

Тестирование. Из (26):

$$f(\overrightarrow{d_5}) = sign\left(\begin{pmatrix} 0 & -0{,}31 & -0{,}23 & -0{,}23 & 0{,}55 & 0{,}55 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0{,}7 \\ 0{,}7 \end{pmatrix} - 0{,}5 \right) = 1.$$

Следовательно, $c^* = \overline{c}$, то есть «не Китай».

Вычислительная сложность.

Обучение: $O(|\mathcal{C}||\Omega|^2)$.

Тестирование: $O(|\mathcal{C}|)$.

§ 1.7. Алгоритм деревьев принятия решений

Алгоритм деревьев принятия решений наглядно демонстрируют человеку процесс и результат классификации. На основе обучающего множества строится дерево, узлами которого являются термины документов, листьями – метки классов, а ребра помечены весами терминов.

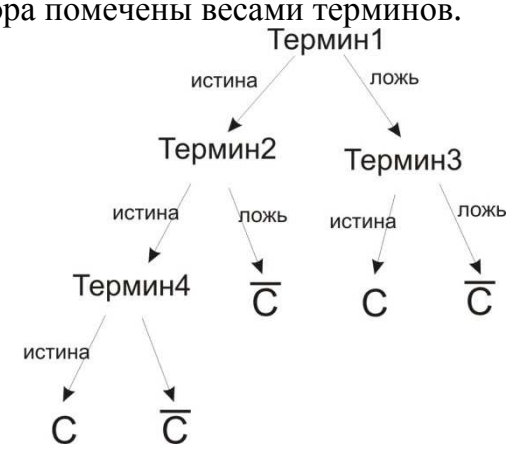


Рис. 4. Пример дерева принятия решений с бинарными весами терминов

На рис. 4 представлен пример дерева принятия решений, в котором используются бинарные веса терминов. Тестовый документ прогоняется по дереву, выбираются ветви, соответствующие терминам документа. В результате документу присваивается класс, соответствующий достигнутому листу.

При обучении используют следующую стратегию: рассматривают множество документов, проверяют, все ли документы данного множества имеют одинаковую метку класса (категорию); если нет, то ищут термин, обладающий наибольшей различительной способностью для разделения этих документов на классы; получают два подмножества документов и строят их поддеревья, повторяя всё сначала, пока не получат подмножество документов одного класса, тогда добавляют в соответствующее поддерево лист с меткой этого класса.

Для выбора очередного разделяющего термина используется понятие *информационной энтропии* – меры неопределенности. Предположим, имеем множество A из n элементов, обладающих атрибутом Q, который может принимать одно из m значений. Тогда мера неопределенности множества A по отношению к атрибуту Q вычисляется следующим образом:

$$E(A,Q) = -\sum_{i=1}^{m} \frac{m_i}{n} \times \log_2 \frac{m_i}{n}, \tag{27}$$

где m_i – число случаев, когда реализуется i-ое значение.

Иначе выражение (27) можно записать так:

$$E(A,Q) = E(p_1, \dots, p_m) = -\sum_{i=1}^{m} p_i \times \log_2 p_i. \tag{28}$$

Максимальное значение энтропия достигает, когда m значений атрибута Q равновероятны, $E(A,Q) = \log_2 m$. Если значения атрибута Q не равновероятны, то энтропия понижается, а информационная выгода от описания элементов множества A с помощью атрибута Q возрастает.

Теперь представим, что имеем множество A из n элементов, характеризующихся свойством S и обладающих атрибутом Q, который может принимать одно из m значений. Тогда информационная выгода (прирост информации) от классификации (по свойству S) посредством атрибута Q имеет следующее значение:

$$I(A,Q) = E(A,S) - \sum_{i=1}^{m} \frac{|A_i|}{|A|} \times E(A_i,S), \tag{29}$$

где A_i – множество элементов A, на которых атрибут Q имеет i-ое значение.

Применительно к задаче классификации коллекции документов по двум классам исходная энтропия вычисляется следующим образом:

$$E(A,S) = E(\mathcal{D},c) = -p(c)\log_2 p(c) - p(\overline{c})\log_2 p(\overline{c}), \tag{30}$$

Выражение (29) принимает вид:

$$I(A,Q) = E(\mathcal{D},c) - \left(p(t_k) \times E(t_k,c) + p(\overline{t_k}) \times E(\overline{t_k},c)\right) = \tag{31}$$

$$= E(\mathcal{D},c) - \left(p(t_k) \times [-p(t_k|c) \times \log_2 p(t_k|c) - p(t_k|\overline{c}) \times \log_2 p(t_k|\overline{c})] + p(\overline{t_k})\right.$$

$$\times\times \left.[-p(\overline{t_k}|c) \times \log_2 p(\overline{t_k}|c) - p(\overline{t_k}|\overline{c}) \times \log_2 p(\overline{t_k}|\overline{c})]\right),$$

Текущим разделяющим атрибутом становится тот, при котором прирост информации наибольший (а энтропия наименьшая).

Алгоритм в общем виде.

Обучение.

Вход: $\boldsymbol{\mathcal{C}}$ и $\boldsymbol{\mathcal{D}} = \Omega$.

Шаг 1. Дерево <G,E>, где G := Ø; E := Ø; {G – множество вершин, E – множество рёбер}

Шаг 2. G := G + {x}; {создать «безымянную» вершину (корень дерева)}

Шаг 3. Вызвать *ПостроитьУровень*($\boldsymbol{\mathcal{D}}$, x, G, E, $\boldsymbol{\mathcal{T}}$);

Выход: Дерево <G,E>.

ПостроитьУровень(A, x, G, E, T):

Вход: A, x, G, E, T.

Шаг 1. Если для $\forall i \neq j, d_i \in A$ и $d_j \in A,\ d_i \in c_k$ и $d_j \in c_p$: $k = p$, то

Шаг 2. x := c_k; {если все документы имеют одинаковую метку класса,

то поместить её в вершину}

Шаг 3. Выход;

Шаг 4. Иначе

Шаг 5. для каждого $t_k \in T_A$: {T_A – множество терминов документов из множества A}

Шаг 6. вычислить $I(t_k)$ по формуле (31);

Шаг 7. $t_k^* := \arg\max I(t_k)$;

Шаг 8. x := t_k^*; {поместить разделяющий термин в «безымянную» вершину}

Шаг 9. A1 :={d_i: $d_i \in A,\ t_k^* \in d_i$};

Шаг 10. G := G + {y}; E := E + <x=t_k^*, y, true>; {y – новая «безымянная» вершина}

Шаг 11. Вызвать *ПостроитьУровень*(A1, y, G, E, T – $\{t_k^*\}$);

Шаг 12. A2 :=$\{d_i: d_i \in A,\ t_k^* \notin d_i\}$;

Шаг 13. G := G + {z}; E := E + <x=t_k^*, z, false>; {z – новая «безымянная» вершина}

Шаг 14. Вызвать *ПостроитьУровень*(A2, z, G, E, T – $\{t_k^*\}$);

Выход: Дерево <G,E>.

Пример.

Обучение.

1) Исходная энтропия $E(\mathcal{D} = \Omega, c) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0{,}81.$
2) $I(\text{китайский}) = 0{,}81 - \left[\frac{4}{4} \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{0}{4} \times (\dots)\right] = 0{,}81 - 0{,}81 = 0.$
3) $I(\text{пекин}) = I(\text{шанхай}) = I(\text{макао}) = 0{,}81 - \left[\frac{1}{4} \times \left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) + \frac{3}{4} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)\right] = 0{,}81 - 0{,}69 = 0{,}12.$
4) $I(\text{токио}) = I(\text{япония}) = 0{,}81 - \left[\frac{1}{4} \times \left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{3}{4} \times \left(-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right)\right] = 0{,}81 - 0 = 0{,}81.$

Следовательно, разделять будем по термину «токио». Дальнейшего разделения не требуется, так как все документы обоих выделенных подмножеств имеют одинаковые метки класса (внутри подмножеств). Полученное дерево принятия решений представлено на рис.5.

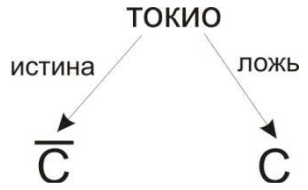


Рис.5. Дерево принятия решений для коллекции из пяти документов

Тестирование. Тестовый документ d5 содержит термин «токио».

Следовательно, $c^* = \overline{c}$, то есть «не Китай».

Вычислительная сложность.

Обучение: $O(|\Omega| \log|\Omega|)$.

§ 1.8. Алгоритм наименьших квадратов

Алгоритм наименьших квадратов (LLSF, Linear Least Squares Fit), относящийся к отряду алгоритмов регрессионного анализа, ищет линейную функциональную зависимость между средним значением наблюдаемой случайной величины (зависимой) и другими наблюдаемыми случайными величинами (независимыми). Зависимой случайной величиной является класс документа, а независимыми случайными величинами – термины документов обучающего множества.

Пусть A – матрица $|\boldsymbol{\Omega}|x|\boldsymbol{\mathcal{T}}|$, строки которой являются документами в пространстве терминов; B – матрица $|\boldsymbol{\Omega}|x|\boldsymbol{\mathcal{C}}|$, строки которой являются документами в пространстве меток классов (категорий). Тогда метод наименьших квадратов ищет

способ преобразования исходного пространства (терминов) в целевое пространство (классов). Для этого вычисляется матрица преобразования F_{LS} ($|\boldsymbol{\mathcal{C}}|x|\boldsymbol{\mathcal{T}}|$) так, чтобы минимизировать регрессионные остатки, то есть разность между фактическим значением зависимой величины и восстановленным:

$$\sum_{i=1}^{|\Omega|}\|\vec{e}_i\|^2 = \sum_{i=1}^{|\Omega|}\left\|F\vec{a}_i^T - \vec{b}_i^T\right\|^2 = \|FA^T - B^T\|_{Fr}^2, \quad (32)$$

$$F_{LS} = \arg_{\mathrm{F}}\min\|FA^T - B^T\|_{Fr}^2,$$

где $\vec{a}_i$ и $\vec{b}_i$ – i-ая пара в обучающем множестве; $\vec{e}_i$ – ошибка отображения $\vec{a}_i$ в $\vec{b}_i$ посредством *F*;

$\|M\|_{Fr} = \sqrt{\sum_{i=1}^{|\Omega|}\sum_{j=1}^{|C|}M_{ij}^2}$ – фробениусова норма матрицы M (|*C*|x|Ω|).

Матрица преобразования F_{LS} показывает степени ассоциации между терминами и классами; $f_{ij} \in \mathbb{R}$ – это оценка (вес) связи термина и класса. Метод наименьших квадратов взвешивает ассоциации так, чтобы минимизировать ошибки преобразования на всём обучающем множестве. Из анализа значений элементов этой матрицы можно получить информацию о важных/неважных терминах для всей коллекции документов. Более информативные термины имеют веса, «смещённые» к конкретным классам; менее информативные термины имеют относительно одинаковые веса ассоциаций для всех классов.

Итоговая матрица преобразований вычисляется следующим образом:

$$F_{LS} = B^T(A^+)^T, \quad (33)$$

где A^+ – матрица, псевдообратная матрице A.

Таким образом, классификационным правилом алгоритма является вычисление проекции $\vec{c}$ исходного образа тестового документа $\vec{d}$ в целевое пространство классов:

$$\vec{c} = \left(F_{LS}\vec{d}^T\right)^T, \quad (V34)$$

Пример. Для нашей коллекции из 5 документов:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}7 & 0{,}7 \end{bmatrix};\ B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

На обучающем множестве веса классов бинарные, для тестового документа – вещественные.

Обучение. Из (33):

$$F_{LS} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}72 & 0{,}72 \end{bmatrix}$$

Тестирование.

$$\overrightarrow{c_5} = \left(F_{LS}\overrightarrow{d_5}^T\right)^T = [0 \quad 1].$$

Следовательно, $c^* = \overline{c}$, то есть «не Китай».

Вычислительная сложность.

Обучение: вычислительная сложность алгоритма наименьших квадратов зависит от реализации вычисления псевдообратной матрицы и может быть кубической $O(|\boldsymbol{\Omega}|^3)$ или квадратичной $O(|\boldsymbol{\Omega}|^2)$.

Тестирование: $O(|\boldsymbol{\Omega}|\log|\boldsymbol{\Omega}|)$.

§ 1.9. Экспериментальная оценка результата классификации с учителем

Качество построенного классификатора оценивается по его ошибке на тестовом подмножестве обучающего множества документов. Ошибка – это доля неправильных решений классификатора. Решения классификатора сравнивают с решениями экспертов, формирующих обучающее множество.

Для вычисления ошибки и других классических мер качества в задачах информационного поиска – полноты, точности и F1-меры – необходимо составить следующую таблицу категорий принятых решений, для каждого $d_i \in \Omega_{\text{test}}, \Omega_{\text{test}} \subset \Omega$ и $c_j \subset \boldsymbol{C}$:

эксперт решил / классификатор решил	$d_i \in c_j$	$d_i \notin c_j$
$d_i \in c_j$	a	b
$d_i \notin c_j$	c	d

Тогда меры качества вычисляются следующим образом:

$$P = \frac{a}{a+b}, \tag{35}$$

$$R = \frac{a}{a+c}, \tag{36}$$

$$E = \frac{b+c}{a+b+d+c}, \tag{37}$$

$$A = \frac{a+d}{a+b+d+c}, A = 1 - E, \tag{38}$$

где P – *точность*, то есть доля истинно принадлежащих классу документов из всех, что классификатор записал в данный класс;

R – *полнота*, то есть доля истинно принадлежащих классу документов и записанных в этот класс классификатором среди всех документов, которые истинно ему принадлежат;

E – *ошибка* классификатора;

A – п*равильность* (аккуратность) классификатора.

Заметим, что правильность (ошибка) не пригодны для оценки результата, если есть небольшие классы, то есть классы, доля документов которых меньше 10%, поскольку в этом случае высокой правильности можно достичь, всегда отвечая «не принадлежит». Например, если относительная частота класса коллекции составляет 1%, то классификатор по принципу «всегда не принадлежит» даст правильность 99%. Надежнее использовать меры полноты и точности.

Полнота и точность – меры, противоречащие друг другу в том смысле, что 100%-ую полноту легко достичь, просто поместив все документы в класс c_j (точность будет мала), и наоборот 100%-ую точность можно достичь, строго отбрасывая документы, помещая в класс c_j малое число документов (полнота будет мала). Показатель, позволяющий найти баланс между полнотой и точность называют F_β-мерой, которая вычисляется как взвешенное среднее гармоническое:

$$F_\beta = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \beta^2 = \frac{1-\alpha}{\alpha}, \tag{39}$$

где $\alpha \in [0;1], \beta^2 \in [0;\infty]$, при $\beta > 1$ предпочтение отдаётся полноте, при $\beta < 1$ – точности.

На практике чаще всего применяют сбалансированный вариант F_β-меры – F_1-меру, то есть $\beta = 1$, или $\alpha = 0{,}5$:

$$F_1 = \frac{2PR}{P+R}, F_1 \in [0;1]. \quad (40)$$

Для обобщения мер качества для всех Ω_{test} и $\boldsymbol{C}$ применяют следующие подходы к усреднению:

а) *макроусреднение* – обобщение на уровне классов;

б) *микроусреднение* – обобщение на уровне документов.

Макроусреднение выполняется путём составления отдельных таблиц принятия решений для каждого класса, вычисления мер по каждой таблице и затем обычного усреднения значений мер по всем классам. Микроусреднение выполняется путём составления единой таблицы для всех классов, в которую сразу записываются решения по всем документам, затем по этой таблице вычисляют меры качества.

Макроусреднение приписывает равные веса решениям классификатора для каждого класса, а микроусреднение – равные веса решениям классификатора для каждого документа. Классы с бОльшим числом документов (и решений по ним) вносят бОльший вклад в микроусреднение. Таким бразом, результат микроусреднения в большей степени оценивают качество классификатора для крупных классов коллекции документов. Чтобы оценить его на малых классах следует применять макроусреднение.

§ 1.10. Выбор метода классификации с учителем

Обзор экспериментальных исследований. Исторически основным тестом для оценки систем классификации с учителем является англоязычная коллекция Reuters-21578 и её модификации. Reuters-21578 состоит из 21578 новостных сообщений, 118 классов (категорий), документы могут принадлежать 0, 1 и более классам. Эксперименты на данной коллекции [2] показали, что:

а) SVM > kNN >> {LLSF,NNet} >> NB, когда количество позитивных примеров для каждой категории мало (менее 10);

б) {SVM, kNN, LLSF} >> {NB, NNet} при обычном наполнении категорий (более 300 экземпляров).

Здесь «>>» и «>» означает «значительно эффективнее» и «эффективнее» соотвественно.

NB – наивный байесовский классификатор; NNet – простая нейронная сеть типа персептрона; kNN – алгоритм k-ближайшего соседа.

Лабораторные эксперименты показали, что наивный байесовский классификатор не может конкурировать, например, с алгоритмом опорных векторов, если обучающие и тестовые данные являются независимыми и одинаково распределёнными. Однако в реальном мире (а) обучающие выборки извлекаются из множества данных, на которых потом будут применяться классификаторы; (б) природа данных изменяется во времени; (в) данные содержат ошибки. В итоге разница становится не такой существенной и даже может служить свидетельством в пользу более простого подхода, такого как NB.

Сравнение и ранжирование алгоритмов классификации зависит (а) от коллекции документов, (б) от рассматриваемого класса, (в) от условий эксперимента (выбора терминов/признаков, разбиения коллекции на подмножества, знания о тестовом

подмножестве и др.). В результате сам по себе алгоритм редко является решающим фактором.

Компромисс между смещением и дисперсией. Понять, почему не существует одного оптимального алгоритма обучения, помогает концепция компромисса смещения и дисперсии. Цель классификации текстов заключается в поиске такого оптимального классификатора, который после усреднения по всем документам коллекции гарантировал бы оценку принадлежности документов классам как можно более близкую к истинной принадлежности документов классам. Более того, оптимальным алгоритмом обучения можно было бы назвать такой алгоритм обучения, который гарантировал бы построение этого оптимального классификатора в среднем по всем обучающим множествам, другими словами требуется алгоритм обучения с минимальной ошибкой обучения.

Концепция компромисса между смещением и дисперсией говорит, что ошибка обучения складывается из двух компонент − смещения и дисперсии −, которые невозможно минимизировать одновременно.

Смещение показывает, как в среднем по различным обучающим множествам прогноз классификатора отличается от истинной классификации данных. Смещение велико, если алгоритм обучения порождает плохие классификаторы. Смещение мало, если (а) порождает хорошие классификаторы, (б) разные обучающие множества порождают разные ошибки на разных документах, (в) разные обучающие множества порождают положительные и отрицательные ошибки на одних и тех же документах. Смещение можно интерпретировать как результат знаний о предметной области (или их недостаток), встроенных в классификатор.

Дисперсия показывает, как сильно зависит классификатор от того, на каком обучающем множестве он строился; насколько противоречивыми могут быть решения классификаторов независимо от того, правильные они или нет. Дисперсия велика, если разные обучающие множества порождают совершенно разные классификаторы. Дисперсия мала, если обучающее множество мало влияет на классификатор. Дисперсию можно интерпретировать как емкость запоминания алгоритма обучения. Емкость соответствует количеству независимых параметров, которые подгоняются на обучающем множестве. Например, для алгоритма Роккио – это $|\boldsymbol{C}|$ центроидов по $|\boldsymbol{\mathcal{T}}|$ параметров, соответствующих каждой размерности пространства терминов; такой алгоритм обучения не зависит от размера обучающей выборки и не помнит тонкие детали распределения документов. Другой пример – это алгоритм k-ближайших соседей, параметрами являются оценки $P(c|S_k)$, где S_k − множество соседей; каждое множество соседей порождает отдельное независимое решение о классификации документа; емкость такого алгоритма ограничена лишь размером обучающего множества; kNN «запоминает» обучающее множество. Для таких рассмотренных алгоритмов как наивный байесовский классификатор, алгоритм Роккио, алгоритм k-ближайших соседей можно обобщить оценки смещения и дисперсии следующим образом:

	смещение	*дисперсия*
линейные алгоритмы	(а) **мало**, если данные линейно разделимы; (б) **велико**, если данные не разделимы линейно	**мала** (большинство случайно выбранных обучающих множеств порождает близкие гиперплоскости)
нелинейные алгоритмы	**мало**	**велико** (например, результат kNN зависит от наличия шумовых документов в окрестности тестового)

При сравнении двух алгоритмов обучения часто оказывается, что у одного из них больше смещение и меньше дисперсия, у другого – меньше смещение и больше дисперсия. Следовательно, необходимо взвесить относительные преимущества смещения и дисперсии для конкретной прикладной задачи и на основании этой информации выбрать алгоритм.

Глава 2. Алгоритмы классификации без учителя

Алгоритмы *классификации без учителя* анализируют коллекцию полнотекстовых документов с целью разбиения их на группы так, чтобы внутри одной группы оказывались документы наиболее родственные в некотором смысле, а различные документы попадали в различные группы. При этом отсутствует «учитель» - обучающее подмножество документов и заранее известное множество категорий (рубрик). В общем случае алгоритм классификации без учителя – алгоритм *кластеризации* – должен самостоятельно принимать решения о количестве и составе *кластеров*, то есть групп родственных документов. Для этого используются понятия расстояний между документами.

Кластеризация текстов основывается на кластерной гипотезе [7], говорящей, что тесно связанные по смыслу документы стремятся быть релевантными одним и тем же запросам, т. е. документы, релевантные запросу, отделимы от тех, которые не релевантны этому запросу.

Итак, дана коллекция документов $\mathcal{D} = \{d_i\}, i = \overline{1, |\mathcal{D}|}$, существует множество тематических классов, которым принадлежат документы коллекции. Предполагается, что можно автоматически разбить множество документов на кластеры $\mathcal{C} = \{\mathrm{c}_j\}, \overline{1, |\mathcal{C}|}$, и полученные кластеры будут соответствовать внутренним тематическим классам. Тогда задача автоматической кластеризации коллекции полнотекстовых документов сводится к поиску неизвестного множества $\mathcal{C}$ таким образом, чтобы итоговое множество $\mathcal{C}$ являлось оптимальным в соответствии с некоторым критерием качества разбиения документов.

Далее мы рассмотрим несколько конкретных алгоритмов кластеризации, использующих различные подходы к формированию кластеров, и продолжим наш пример, начатый в разделе Глава 1, немного изменив исходные данные – добавим один новый документ *d6*:

docId	Слова в документе	c = «Китай»
1	китайский пекин китайский	c
2	китайский китайский шанхай	c
3	китайский макао	c
4	токио япония китайский	не c
5	китайский китайский китайский токио япония	?
6	токио пекин	?

Рассчитаем веса терминов для коллекции из 6 документов по формуле (1).

	t1	***t2***	***t3***	***t4***	***t5***	***t6***
	китайский	*пекин*	*шанхай*	*макао*	*япония*	*токио*
df	5	2	1	1	2	3
d1	tf=2 | w=0.16| **0,31**	1 | 0,48 | **0,94**	**0**	**0**	**0**	**0**
d2	2 | 0,16 | **0,20**	**0**	1 | 0,78 | **0,98**	**0**	**0**	**0**
d3	1 | 0,08 | **0,10**	**0**	**0**	1 | 0,78 | **0,98**	**0**	**0**
d4	1 | 0,08 | **0,25**	**0**	**0**	**0**	1 | 0,48 | **0,93**	1 | 0,30 | **0,32**
d5	3 | 0,24 | **0,39**	**0**	**0**	**0**	1 | 0,48 | **0,78**	1 | 0,30 | **0,49**
d6	**0**	1 | 0,48 | **0,84**	**0**	**0**	**0**	1 | 0,30 | **0,57**

§ 2.1. Иерархические алгоритмы

Иерархические алгоритмы в отличие от плоских алгоритмов (например см. § 2.2 и др.) создают структурированное множество кластеров – иерархию, которое может оказаться весьма информативным для некоторых приложений. Иерархические алгоритмы разделяются на два вида: *агломеративные* (восходящие) и *дивизимные* (нисходящие). Первые строят кластеры снизу вверх, начиная с множества кластеров, содержащих по одному одиночному документу коллекции, затем последовательно объединяют пары кластеров, пока не получат один кластер, содержащий все документы коллекции. Вторые разбивают кластеры сверху вниз, начиная с одного кластера, которому принадлежат все документы коллекции, затем этот кластер делится на два и так рекурсивно до тех пор, пока каждый документ не окажется в своём отдельном кластере. Нисходящие алгоритмы концептуально более сложные, так как необходимо на каждом шаге применять дополнительный алгоритм плоской кластеризации. Нисходящая иерархическая кластеризация может оказать весьма эффективной, если, например, нет необходимости генерировать полное дерево вплоть до отдельных документов, а ограничиться только верхними уровнями. Если использовать эффективный линейный плоский алгоритм, например, k-средних, то алгоритм нисходящей кластеризации окажется линейной сложности по размеру коллекции. Вдобавок, восходящий алгоритм начинает разбиение, основываясь только на информации о локальных документах, а нисходящий начинает с анализа полной информации о глобальном распределении документов.

Рассмотрим подробнее агломеративные алгоритмы. Основное их различие заключается в выборе критерия, используемого для принятия решения о том, какие кластеры следует объединить на текущем шаге алгоритма. Большое распространение получили три следующих критерия:

а) *одиночная связь* (минимальное расстояние, или максимально сходство): сходство двух кластеров есть сходство между их *наиболее похожими* документами;

б) *полная связь* (максимальное расстояние, или минимальное сходство): сходство двух кластеров есть сходство между их *наиболее непохожими* документами;

в) *групповое усреднение* (усреднение всех показателей сходства): сходство двух кластеров есть среднее сходство всех пар документов, включая пары документов из одного кластера, исключая близость документа самому себе, см. формулу (41).

$$sim^{GA}(C_i, C_j) = \frac{1}{(|C_i| + |C_j|)(|C_i| + |C_j| - 1)} \sum_{d_k \in C_i \cup C_j} \sum_{d_k \in C_i \cup C_j : d_k \neq d_p} sim(\vec{d}_k, \vec{d}_p), \quad (41)$$

где C_i и C_j – i-ый и j-ый кластер; sim – мера сходства, например, косинусная мера близости (3).

Кластеризация с одиночной связью создаёт протяженные («цепочные») кластеры, «сцепленные вместе» элементами, возможно, случайно оказавшимися ближе остальных друг к другу (рис. 6). Этот критерий носит локальный характер, так как не учитывает всю структуру кластера, например, его другие более удалённые части. Кластеризация с полной связью создаёт компактные кластеры (рис. 7) и носит глобальный характер, так как на решение об объединении кластеров влияет вся структура кластера, однако это одновременно повышает чувствительность к выбросам. Кластеризация с групповым усреднением позволяет избежать недостатков, свойственных критериям одиночной и полной связи.

Иерархические алгоритмы в общем случае не требуют исходного задания количества кластеров, однако на практике часто бывает нужно разбиение на непересекающиеся кластеры, как при плоской кластеризации. Тогда следует иерархию кластеров отсечь на некотором уровне. Выбор уровня может выполняться, например, так:

а) указать точное число кластеров k, на рис. 6 и рис. 7 разбиение при k = 2 показано пунктирной линией: $\boldsymbol{\mathcal{C}}1$ = {d1, d2, d4, d5, d6}, $\boldsymbol{\mathcal{C}}2$ = {d3} и $\boldsymbol{\mathcal{C}}1$ = {d1, d4, d5, d6}, $\boldsymbol{\mathcal{C}}2$ = {d2, d3} соответственно;

б) указать минимальное значение близости между кластерами и провести сечение на этом уровне;

в) отсечь на уровне максимальной разницы между двумя последовательными мерами сходства (различия) между кластерами, на рис. 6 и рис. 7 эта, так называемая «естественная кластеризация», показана штриховой линией: $\boldsymbol{\mathcal{C}}1$ = {d4, d5}, $\boldsymbol{\mathcal{C}}2$ = {d1, d6}, $\boldsymbol{\mathcal{C}}3$ = {d2}, $\boldsymbol{\mathcal{C}}4$ = {d3} и $\boldsymbol{\mathcal{C}}1$ = {d4, d5}, $\boldsymbol{\mathcal{C}}2$ = {d1, d6}, $\boldsymbol{\mathcal{C}}3$ = {d2}, $\boldsymbol{\mathcal{C}}4$ = {d3} соответственно.

Алгоритм в общем виде (алгомеративная иерархическая кластеризация).

Вход: множество проиндексированных документов $\boldsymbol{\mathcal{D}}$.

Шаг 1. Для каждого $d_i \in \boldsymbol{\mathcal{D}}$:

Шаг 2. Для каждого $d_j \in \boldsymbol{\mathcal{D}}$:

Шаг 3. $s_{ij} \coloneqq sim(\vec{d}_i, \vec{d}_j)$, где S = $\{s_{ij}\}$ – матрица сходства между кластерами;

Шаг 4. $F_i \coloneqq 1$, где F – массив, сообщающий о кластерах, доступных для объединения.

Шаг 5. A := ∅, где A – последовательность объединений кластеров;

Шаг 6. Для всех k от 1 до N – 1:

Шаг 7. $\langle i, m \rangle \coloneqq \arg\max_{\langle i,m \rangle : i \neq m; F_i = 1;\, F_m = 1} s_{im}$;

Шаг 8. A := A + $\langle i, m \rangle$;

Шаг 9. Для всех j от 1 до $|\mathcal{D}|$:

Шаг 10. $s_{ij} \coloneqq sim(c_{<i,m>}, c_j)$;

Шаг 11. $s_{ji} \coloneqq sim(c_{<i,m>}, c_j)$,

где $c_{<i,m>}$ – кластер, полученный путём объединения кластеров i и m,

c_j – j-ый кластер;

Шаг 12. $F_m \coloneqq 0$;

Выход: список объединений A.

Пример. Вычислим матрицу расстояний для всех документов из нашего примера.

Sim	d1	d2	d3	d4	d5	d6
d1	0					
d2	1,36	0				
d3	1,37	1,39	0			
d4	1,36	1,39	1,40	0		
d5	1,32	1,36	1,38	0,27	0	
d6	0,66	1,43	1,41	1,30	1,21	0

Результат иерархической агломеративной кластеризации по правилу одиночной связи представлен на рис. 6.

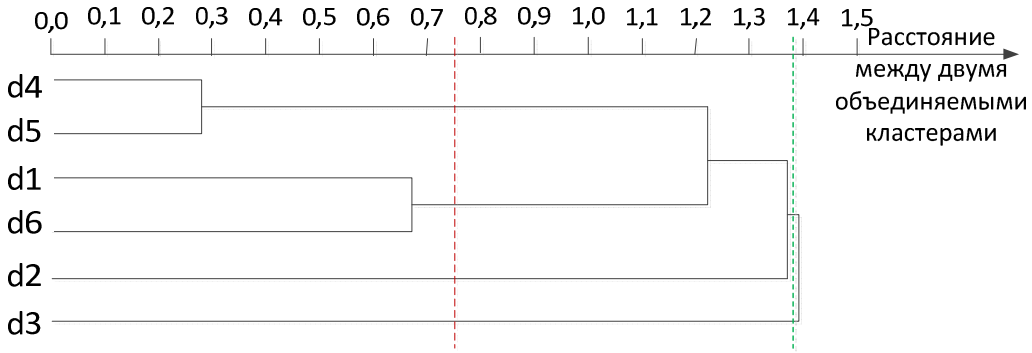


Рис. 6. Дендрограмма кластеризации с одиночной связью для примера из 6 документов

Результат иерархической агломеративной кластеризации по правилу полной связи представлен на рис. 7.

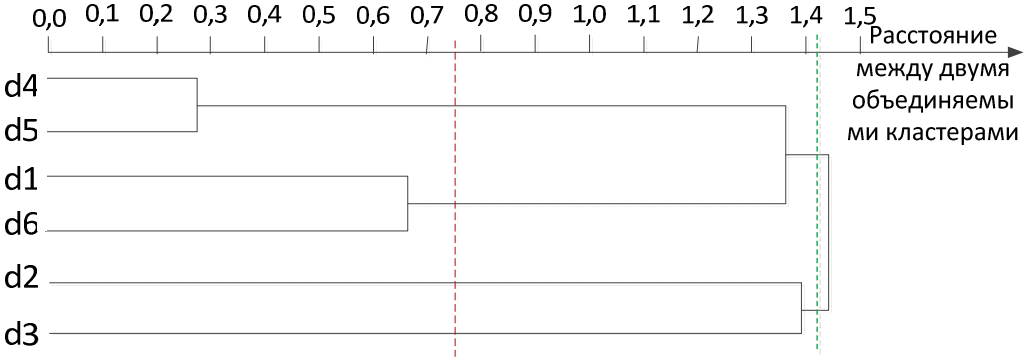


Рис. 7. Дендрограмма кластеризации с полной связью для примера из 6 документов

Вычислительная сложность. Сложность агломеративного иерархического алгоритма зависит от выбранной меры близости (различия) между кластерами и способа реализации идеи алгоритма. Известны реализации со следующими оценками вычислительной сложности: одиночная связь – $O(|\mathcal{D}|^2)$; полная связь –

$O(|\mathcal{D}|^2 \log|\mathcal{D}|)$; усреднение по группе – $O(|\mathcal{D}|^2)$. Сложность дивизимного алгоритма зависит от выбранного дополнительного алгоритма плоской кластеризации, если выбран алгоритм k-средних, то – $O(|\mathcal{D}|)$.

§ 2.2. Алгоритм k-средних

Первые применения алгоритма k-средних были описаны в работе Джеймса МакКуина в 1967 году. При заранее известном числе кластеров k алгоритм k-средних начинает с некоторого начального разбиения документов и уточняет его, оптимизируя целевую функцию – среднеквадратичную ошибку кластеризации как среднеквадратичное расстояние между документами и центрами их кластеров:

$$e(\mathcal{D}, \mathcal{C}) = \sum_{j=1}^{k} \sum_{i: d_i \in C_j} \left\| \vec{d}_i - \vec{\mu}_j \right\|^2, \quad (42)$$

где μ_j – центр, или центроид, кластера $C_j, j = \overline{1, |\mathcal{C}|}$, $|\mathcal{C}| = k$, вычисляющийся по формуле:

$$\vec{\mu}_j = \frac{1}{|C_j|} \sum_{i:\, d_i \in C_j} \overrightarrow{d_i}, \quad (43)$$

где $|C_j|$ – количество документов в C_j.

Идеальным кластером алгоритм k-средних считает сферу с центроидом в центре сферы.

Действие алгоритма начинается с выбора k начальных центров кластеров. Обычно исходные центры кластеров выбираются случайным образом. Затем каждый документ присваивается тому кластеру, чей центр является наиболее близким документу, и выполняется повторное вычисление центра каждого кластера как центроида, или среднего своих членов. Такое перемещение документов и повторное вычисление центроидов кластеров продолжается до тех пор, пока не будет достигнуто условие остановки. Условием остановки может служить следующее: (а) достигнуто пороговое число итераций, (б) центроиды кластеров больше не изменяются и (в) достигнуто пороговое значение ошибки кластеризации. На практике используют комбинацию критериев остановки, чтобы одновременно ограничить время работы алгоритма и получить приемлемое качество.

В общем случае алгоритм k-средних достигает локального минимума целевой функции, что приводит к субоптимальному разбиению документов. Поэтому важен способ выбора начальных значений центроидов. Для этого известны различные эвристические правила, например, получить начальные центры с помощью другого алгоритма – детерминированного, например, иерархического агломеративного.

Алгоритм в общем виде.

Вход: множество проиндексированных документов $\boldsymbol{\mathcal{D}}$, количество кластеров k.

Шаг 1. Инициализация центров кластеров $\{\vec{\mu}_j\}, j = \overline{1, k}$, например, случайными числами.

Шаг 2. $C_j \coloneqq \{\}, j = \overline{1, k}$.

Шаг 3. Для каждого $d_i \in \boldsymbol{\mathcal{D}}$:

Шаг 4. $j^* \coloneqq arg\ \min_j \left\| \vec{\mu}_j - \vec{d}_i \right\|, j = \overline{1, k}$;

Шаг 5. $C_{j^*} \coloneqq C_{j^*} + \{d_i\}$;

Шаг 6. Для каждого $C_j \in \boldsymbol{\mathcal{C}}$:

Шаг 7. $\vec{\mu}_j := \frac{1}{|C_j|}\sum_{i:\, d_i \in C_j} \overrightarrow{d_i}$,

Шаг 8. Если не достигнуто условие остановки, то повторить с *шага 2*.

Выход: множество центров кластеров $\{\vec{\mu}_j\}$ и множество самих кластеров $\boldsymbol{\mathcal{C}}$.

Пример. Пусть k=2.

Итерация 1. Начальные $\{\vec{\mu}_j\}, j = \overline{1,k}$ инициализированы случайным образом:

$$\mu_1 = [0.96\ \ 0.80\ \ 0.42\ \ 0.79\ \ 0.66\ \ 0.85];$$
$$\mu_2 = [0.49\ \ 0.14\ \ 0.91\ \ 0.96\ \ 0.04\ \ 0.93]$$

dist	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**
$\boldsymbol{\mu_1}$	1.55	1.81	1.66	1.51	1.38	0.85
$\boldsymbol{\mu_2}$	1.82	1.38	1.37	1.74	1.59	0,93

⇨ C1 := {d1, d4, d5, d6}; C2 := {d2, d3}.

Итерация 2.

$$\mu_1 = [0.24\ \ 0.45\ \ 0\ \ 0\ \ 0.43\ \ 0.35];$$
$$\mu_2 = [0.159\ \ 0\ \ 0.49\ \ 0.49\ \ 0\ \ 0]$$

dist	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**
$\boldsymbol{\mu_1}$	0.74	1.21	1.22	0.68	0.61	0.67
$\boldsymbol{\mu_2}$	1.18	0.69	0.69	1.21	1.18	1.24

⇨ C1 := {d1, d4, d5, d6}; C2 := {d2, d3}.

Разбиение на кластеры не изменилось – условие остановки выполнено, мы получили итоговые два кластера.

Вычислительная сложность. Алгоритм k-средних линейно зависит от всех своих факторов: от количества документов, количества кластеров, количества терминов и количества итераций. Для сохранения линейной сложности ($O(|\mathcal{D}|)$) при комбинировании с иерархическим с целью эффективного задания начальных центроидов кластеров, предлагается квадратичный иерархический алгоритм применить к выборке документов размером $\sqrt{|\mathcal{D}|}$. Такой подход получил название алгоритм картечи.

§ 2.3. Плотностный алгоритм DBSCAN

Алгоритм DBSCAN (Density Based Spatial Clustering of Applications with Noise), плотностный алгоритм для кластеризации пространственных данных с присутствием шума), был предложен Мартином Эстер, Гансом-Питером Кригель и коллегами в 1996 году как решение проблемы разбиения (изначально пространственных) данных на кластеры произвольной формы [4]. Большинство алгоритмов, производящих плоское разбиение, создают кластеры по форме близкие к сферическим, так как минимизируют расстояние документов до центра кластера.

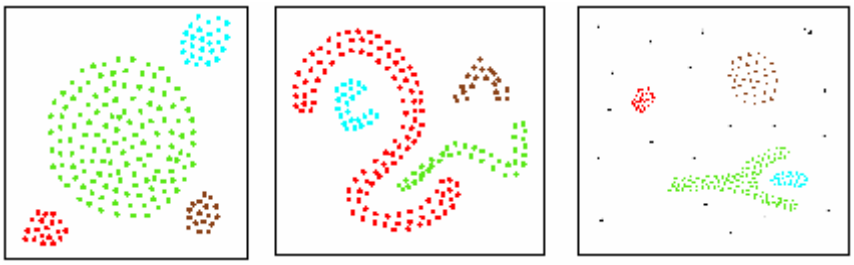


Рис. 8. Примеры кластеров произвольной формы

Авторы DBSCAN экспериментально показали, что их алгоритм способен распознать кластеры различной формы, например, как на рис. 8.

Идея, положенная в основу алгоритма, заключается в том, что внутри каждого кластера наблюдается типичная плотность точек (объектов), которая заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров. Ещё точнее, что для каждой точки кластера её соседство заданного радиуса должно содержать не менее некоторого числа точек, это число точек задаётся пороговым значением. Перед изложением алгоритма дадим необходимые определения.

Определение 1. Eps-соседство точки p, обозначаемое как $N_{Eps}(p)$, определяется как множество документов, находящихся от точки p на расстояния не более *Eps*: $N_{Eps}(p) = \{q \in \mathcal{D} | \, dist(p,q) \leq Eps\}$. Поиска точек, чьё $N_{Eps}(p)$содержит хотя бы минимальное число точек (*MinPt*) не достаточно, так как точки бывают двух видов: ядровые и граничные.

Определение 2. Точка p *непосредственно плотно-достижима* из точки q (при заданных *Eps* и *MinPt*), если $p \in N_{Eps}(p)$ и $|N_{Eps}(p)| \geq MinPt$ (рис. 9).

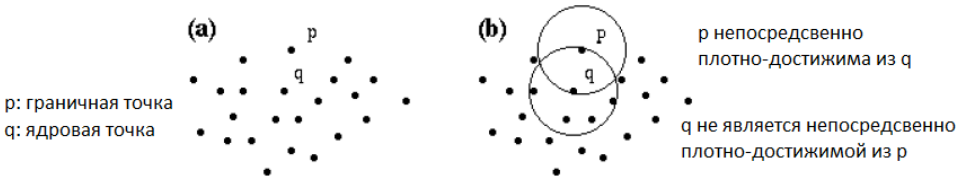


Рис. 9. Пример точек, находящихся в отношении *непосредственно плотной достижимости*

Определение 3. Точка p *плотно-достижима* из точки q (при заданных *Eps* и *MinPt*), если существует последовательность точек $q = p_1, p_2 \dots, p_n = p: p_{i+1}$ непосредственно плотно-достижимы из p_i. Это отношение транзитивно, но не симметрично в общем случае, однако симметрично для двух ядровых точек.

Определение 4. Точка p *плотно-связана* с точкой q (при заданных *Eps* и *MinPt*), если существует точка o: p и q плотно-достижимы из o (при заданных *Eps* и *MinPt*), см. рис. 10.

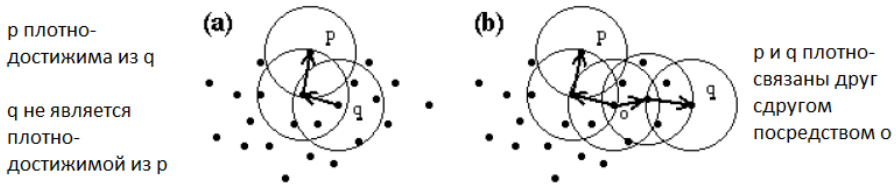


Рис. 10. Пример точек, находящихся в отношении *плотной связанности*

Теперь мы готовы дать определения кластеру и шуму.

Определение 5. Кластер C_j(при заданных *Eps* и *MinPt*) – это не пустое подмножество документов, удовлетворяющее следующих условиям:

1) $\forall p, q$: если $p \in C_j$ и q плотно-достижима из p (при заданных *Eps* и *MinPt*), то $q \in C_j$.
2) $\forall p, q \in C_j$: p плотно-связана с q (при заданных *Eps* и *MinPt*).

Итак, кластер – это множество плотно-связанных точек. В каждом кластере содержится хотя бы *MinPt* документов.

Шум – это подмножество документов, которые не принадлежат ни одному кластеру: $\{p \in \mathcal{D} | \forall j: p \notin C_j, j = \overline{1, |\mathcal{C}|}\}$.

Алгоритм DBSCAN для заданных значений параметров *Eps* и *MinPt* исследует кластер следующим образом: сначала выбирает случайную точку, являющуюся ядровой, в качестве затравки, затем помещает в кластер саму затравку и все точки, плотно-достижимые из неё.

Алгоритм в общем виде.

DBSCAN

Вход: множество индексированных документов $\boldsymbol{\mathcal{D}}$, *Eps* и *MinPt*.

Шаг 1. Установить всем элементам множества $\boldsymbol{\mathcal{D}}$ флаг «не посещён». Присвоить текущему кластеру C_j нулевой номер, j := 0. Множество шумовых документов Noise := $\emptyset$.

Шаг 2. Для каждого $d_i \in \boldsymbol{\mathcal{D}}$ такого, что флаг(d_i) = «не посещён», выполнить:

Шаг 3. флаг(d_i) := «посещён»;

Шаг 4. $N_i := N_{Eps}(d_i) = \{q \in \mathcal{D} | \, dist(d_i, q) \leq Eps\}$

Шаг 5. Если $|N_i| <$ *MinPt*, то

Noise := Noise + $\{d_i\}$

иначе

номер следующего кластера j := j + 1;

EXPANDCLUSTER($d_i, N_i, C_j, Eps, MinPt$);

Выход: множество кластеров $\boldsymbol{\mathcal{C}}$=$\{C_j\}$.

EXPANDCLUSTER

Вход: текущий документ d_i, его eps-соседство N_i, текущий кластер C_j и $Eps, MinPt$.

Шаг 1. $C_j := C_j + \{d_i\}$;

Шаг 2. Для всех документов $d_k \in N_i$:

Шаг 3. Если флаг(d_k) = «не посещён», то

Шаг 4 флаг(d_k) := «посещён»;

Шаг 5. $N_{ik} := N_{Eps}(d_k)$;

Шаг 6. Если $|N_{ik}| >=$ *MinPt*, то $N_i := N_i + N_{ik}$;

Шаг 7. Если $\nexists p$: $d_k \in C_p$, $p = \overline{1, |\mathcal{C}|}$, то $C_j := C_j + \{d_k\}$;

Выход: кластер C_j.

Эвристический подход к заданию начальных параметров. Для некоторого значения k, построить график, на котором каждому документу коллекции поставить в соответствие значение k-dist – расстояние до его k-ого соседа, при этом документы должны быть отсортированы по убыванию значения k-dist. Тогда полученный график может породить догадки о распределении плотности в массиве документов. Ищем пороговую точку p на графике с наибольшим значением k-dist для «самого разреженного» кластера: предполагаем, то это первая точка первой «равнины».

Точки, расположенные левее, предположительно считаются шумовыми, а правее – принадлежащими одному из кластеров. Тогда значения параметров *Eps = k-dist(p)* и *MinPt=k*.

Пример. Воспользуемся эвристикой авторов DBSCAN для настройки начальных параметров, построим 3-dist: (d2;1,43), (d3;1,39), (d1;1,36), (d4;1,36), (d5;1,32), (d6;1,30).

Рис. 11. Поиск значений входных параметров алгоритма DBSCAN

Итак, *Eps* = 1,36 и *MinPt* = *3*.

j := 0; Noise := ∅; Флаг := [0 0 0 0 0 0], где 0 соответствует «не посещён».

1) d_i := **d1**. Флаг = [1 0 0 0 0 0]. $N_i = \{d1, d2, d4, d5, d6\}$. $|N_i| = 5 > MinPt$; j = 1;

2) C_j = {d1}.

3) d_k := **d2**; Флаг = [1 1 0 0 0 0] N_{ik} = {d1, d2, d5}; $|N_{ik}| = 3 >= MinPt$; $N_i = \{d1, d2, d4, d5, d6\}$. C_j = {d1, d2}.

4) d_k := **d4**; Флаг = [1 1 0 1 0 0] N_{ik} = {d1, d4,d5,d6}; $|N_{ik}| = 4 >= MinPt$; $N_i = \{d1, d2, d4, d5, d6\}$. C_j = {d1, d2, d4}.

5) d_k := **d5**; Флаг = [1 1 0 1 1 0] N_{ik} = {d1, d2, d4, d5,d6}; $|N_{ik}| = 5 >= MinPt$; $N_i = \{d1, d2, d4, d5, d6\}$. C_j = {d1, d2, d4, d5}.

6) d_k := **d6**; Флаг = [1 1 0 1 1 1] N_{ik} = {d1, d4, d5,d6}; $|N_{ik}| = 4 >= MinPt$; $N_i = \{d1, d2, d4, d5, d6\}$. C_j = {d1, d2, d4, d5, d6}.

7) d_i := **d3**. Флаг = [1 1 1 1 1 1]. $N_i = \{d3\}$. $|N_i| = 1 < MinPt$; Noise := {d3}.

Получаем один кластер и шумовые документы:

$\mathcal{C}1$ = {d1, d2, d4, d5, d6} и Noise = {d3}. Разницу между ожиданиями, связанными с графиком 3-dist (2 шумовых документа), можно объяснить особенностью построения графика 3-dist для данного частного примера: при вычислении расстояния до k-ого соседа k-соседом становился тот, расстояние до которого отлично от расстояния до (k-1)-соседа.

Вычислительная сложность. В общем случае алгоритм DBSCAN имеет квадратичную вычислительную сложность из-за поиска *Eps*-соседства. Однако авторы алгоритма использовали для этой цели специальную структуру данных – R*-деревья, в результате поиск *Eps*-соседства для одной точки – O(log n). Общая вычислительная сложность DBSCAN – O(n * log n).

§ 2.4. Нечёткий алгоритм c-средних

Нечёткий алгоритм c-средних был предложен Джоном С. Данном в 1973 году (позднее усовершенствован Дж. Бездеком в 1981 году) как решение проблемы мягкой кластеризации, то есть присвоения каждого документа более чем одному кластеру. Как и его чёткий вариант – алгоритм k-средних – данный алгоритм, начиная с некоторого начального разбиения данных, итеративно минимизирует целевую функцию, которой является следующее выражение:

$$e_m(\mathcal{D}, \mathcal{C}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{C}|} u_{ij}^m \left\| \vec{d}_i - \vec{\mu}_j \right\|^2, \quad (44)$$

где m – спепень нечёткости, $1 < m < \infty$;

u_{ij} – степень принадлежности i-ого документа j-ому кластеру,

$u_{ij} \in [0;1]$, $\sum_{j=1}^{|\mathcal{C}|} u_{ij} = 1$ для любого $d_i \in \boldsymbol{\mathcal{D}}$;

$$u_{ij} = \frac{1}{\sum_{k=1}^{|\mathcal{C}|} \left(\frac{\left\| \vec{d}_i - \vec{c}_j \right\|}{\left\| \vec{d}_i - \vec{c}_k \right\|} \right)^{\frac{2}{m-1}}} \quad (45)$$

μ_j – центроид, кластера $C_j, j = \overline{1, |\mathcal{C}|}$, вычисляющийся по формуле:

$$\overrightarrow{\mu_j} = \frac{\sum_{i=1}^{|\boldsymbol{\mathcal{D}}|} u_{ij}^m \times \vec{d}_i}{\sum_{i=1}^{|\boldsymbol{\mathcal{D}}|} u_{ij}^m}. \quad (46)$$

Алгоритм в общем виде.

Вход: множество проиндексированных документов $\boldsymbol{\mathcal{D}}$, количество кластеров $|\boldsymbol{\mathcal{C}}|$=k.

Шаг 1. Инициализация матрицы $U^0 = (u_{ij}), i = \overline{1, |\mathcal{D}|}\ j = \overline{1, k}$, например, случайными числами.

t = 0, где t – номер итерации.

Шаг 2. t := t +1.

Шаг 3. Вычислить текущие центроиды кластеров $\{\vec{\mu}_j\}^t, j = \overline{1, k}$ по формуле (46).

Шаг 4. Обновить матрицу нечёткого разбиения, то есть вычислить $U^t = (u_{ij})$ по формуле (45).

Шаг 5. Если не достигнуто условие остановки, например, $\|U^t - U^{t-1}\| < \varepsilon$, где $0 < \varepsilon < 1$, то повторить с *шага 2*.

Выход: матрица степеней принадлежности документов кластерам $U^t = (u_{ij})$.

Пример. Пусть k=2; m=2; ε = 0,01.

Итерация 0. Инициализация матрицы $U^0 = (u_{ij})$ случайными числами:

0.9572	0.0428
0.4854	0.5146
0.8003	0.1997
0.1419	0.8581
0.4218	0.5782
0.9157	0.0843

Итерация 1.

$U^1 = (u_{ij}) =$

0.8442	0.1558
0.4953	0.5047
0.6191	0.3809
0.0788	0.9212
0.0889	0.9111
0.8160	0.1840

$\{\vec{\mu}_j\}^1 =$

0.1660	0.5535	0.0816	0.2219	0.0557	0.2021
0.2687	0.0056	0.1875	0.0282	0.6831	0.2915

И т. д.

Итерация 4.

$U^4 = (u_{ij})$ =

0.8801	0.1199
0.5411	0.4589
0.5537	0.4463
0.0560	0.9440
0.0493	0.9507
0.8369	0.1631

$\{\vec{\mu}_j\}^4 =$

0.1589	0.6297	0.1369	0.1489	0.0024	0.1920
0.2860	0.0168	0.0933	0.0843	0.6832	0.3314

Выполнено условие остановки: $\|U^4 - U^3\| < \varepsilon$.

Если привести результат к чёткой кластеризации, то получим следующие два кластера:

$\boldsymbol{\mathcal{C}}1$ = {d2, d3, d4}; $\boldsymbol{\mathcal{C}}2$ = {d1, d5, d6}.

Вычислительная сложность. Алгоритм имеет линейную сложность – $O(|\mathcal{D}|)$.

§ 2.5. Инкрементный алгоритм C^{2}ICM

Алгоритм C^{2}ICM (Cover-Coefficient-based Incremental Clustering Methodology), алгоритм инкрементной кластеризации на основе анализа коэффициентов покрытия, был предложен Фазли Кэном и коллегами в 1991 как решение проблемы поддержки актуальности кластерной структуры в пополняемых базах документов. Все рассмотренные нами ранее алгоритмы кластеризации не учитывают динамическую природу массивов документов, будем называть их статическими алгоритмами. Если в массив будут добавлены новые документы или удалены некоторые старые, то с помощью статических алгоритмов возможно только класстеризовать обновленный массив с самого начала. Ф. Кэн предлагает инкрементный алгоритм C^{2}ICM, позволяющий время от времени модифицировать кластерную структуру изменившегося массива документов.

В основе инкрементного алгоритма C^{2}ICM лежит статический алгоритм C^{3}M (Cover-Coefficient-based Clustering Methodology), алгоритм кластеризации на основе анализа коэффициентов покрытия, предложенный в Ф. Кэном и коллегами 1990 г. Первичное разбиение массива документов выполняется с помощью C^{3}M. Анализируется сходство между документами, автоматически определяется количество кластеров, каждый документ коллекции оценивается на возможность стать затравкой для кластера (оценивается затравочная сила каждого документа). Затем все остальные документы присваиваются тем кластерам, чей документ-затравка ближе других. Алгоритм C^{2}ICM используется, когда массив документов обновлён. Снова анализируется сходство между документами, автоматически определяется новое количество кластеров, каждый документ коллекции оценивается на возможность стать затравкой для кластера. В итоге формируется подмножество документов $\boldsymbol{\mathcal{D}}_г$, которые следует кластеризовать. В него попадают новые документы и документы из тех кластеров, которые теперь признаны недействительными. Кластер признаётся недействительным, если: (а) его документ-затравка больше не является затравкой (это действует и в случае, когда затравка удаляется из массива) и (б) один или более документов, не являющихся ранее затравками, после обновления массива стали затравочными. Затем полученное подмножество снова разбивается алгоритмом C^{3}M.

Таким образом, не происходит повторное разбиение всей коллекции, а только её части. Эксперименты, проведённые Ф. Кэном, показали, что после обновления массива документов C^{2}ICM производит очень близкие кластеры к тем, что и C^{3}M путём полной перекластеризации.

Идея алгоритма C^{3}M заключается в количественной оценке взаимоотношения каждой пары документов: насколько первый документ «покрывает» второй и наоборот. Эта оценка по сути является ассиметричной оценка сходства двух документов, учитывающей сколько общих терминов у документов, сколько всего терминов в первом документе и сколько раз каждый общий термин встречается в коллекции. Матрица *коэффициентов покрытия* C ($|\boldsymbol{\mathcal{D}}|x|\boldsymbol{\mathcal{D}}|$) вычисляется следующим образом:

$$c_{ij} = \alpha_i \times \sum_{k=1}^{|\mathcal{D}|} d_{ik} \times \beta_k \times d_{jk}, \tag{47}$$

где$i, j = \overline{1, |\mathcal{D}|}$, $\alpha_i = 1/\sum_{p=1}^{|\mathcal{T}|} d_{ip}$, $\beta_k = 1/\sum_{p=1}^{|\mathcal{D}|} d_{pk}$.

c_{ij} показывает, с какой степенью документ d_iпокрывается документом d_j.

Коэффициенты c_{ij} обладают следующими свойствами:

1) $0 < c_{ij} < 1$, $0 < c_{ii} < 1$;
2) $c_{ii} \geq c_{ij}, \min(c_{ii}) = 1/|\mathcal{D}|$ для бинарных весов терминов в документах;
3) $\left(c_{i1} + c_{i2} + \cdots + c_{i|\mathcal{D}|}\right) = 1$;
4) $c_{ij}=0 \leftrightarrow c_{ji}=0$; $c_{ij} > 0 \leftrightarrow c_{ji} > 0$; в общем случае $c_{ij} \neq c_{ji}$;
5) $c_{ij} = c_{ji} = c_{ii} = c_{jj} \leftrightarrow d_i$ и d_j – идентичны;
6) d_i является отличным ото всех $\leftrightarrow c_{ii} = 1$.

Идентичные документы с равной степенью покрываются всеми другими документами. Если у двух документов нет общих терминов, то их коэффициенты покрытия cij и cji будут равны 0. Если d_i и d_j имеют общие редкие термины, то cij возрастает.

Поскольку в общем случае, если документ имеет общие термины с малым числом других документов, то значение c_{ii} будет близко в 1, иначе к 0, то $\delta_i = c_{ii}$ называют *коэффициентом разделимости*, а $\varphi_i = 1 - \delta_i$ – *коэффициентов объединяемости*. К коллекции, состоящей из близких документов, коэффициент разделяемости будет низким, а в коллекции из различных документов – высоким, поэтому количество кластеров предлагается вычислять следующим образом:

$$nc = \sum_{i=1}^{|\mathcal{D}|} \delta_i, \text{ где } 1 \leq nc \leq \min(|\mathcal{D}|, |\boldsymbol{\mathcal{T}}|). \tag{48}$$

Затем для каждого документа $d_i \in \boldsymbol{\mathcal{D}}$ вычисляется затравочная сила p_i:

$$p_i = \delta_i \times \varphi_i \times \sum_{j=1}^{|\mathcal{T}|} d_{ij}, \text{ если веса терминов в документах бинарные,} \tag{49}$$

$$\text{иначе } p_i = \delta_i \times \varphi_i \times \sum_{j=1}^{|\mathcal{T}|} (d_{ij} \times \delta'_j \times \varphi'_j), \text{ где } \delta'_j = c'_{jj},\ \varphi'_j = 1 - \delta'_{jj}, \tag{50}$$

c'_{jj} – элемент матрицы коэффициентов покрытия для каждой пары терминов, вычисляется аналогично по формуле (47) только не для документов, а для терминов $j = \overline{1, |\mathcal{T}|}$.

Определим документы-затравки как *nc* документов с наибольшей затравочной силой. Если в массиве документов есть идентичные документы, то выбираем только один из них (любой). Для остальных документов коллекции определяем документы-затравки, которые максимально покрывают их, и помещаем эти документы в соответствующие кластеры.

Алгоритм в общем виде.

Начальная итерация (t = 0) – новая коллекция документов, тогда разбить посредством C^3M.

C^3M:

Вход: множество проиндексированных документов $\boldsymbol{\mathcal{D}}$.

Шаг 1. вычислить матрицу коэффициентов покрытия C = $\{c_{ij}\}$ по формуле (47);

Шаг 2. вычислить количество кластеров nc по формуле (48);

Шаг 3. для каждого документа $d_i \in \boldsymbol{\mathcal{D}}$**:**

Шаг 4. вычислить затравочную силу p_i по формуле (49);

Шаг 5. упорядочить $d_i \in \boldsymbol{\mathcal{D}}$ по убыванию p_i;

Шаг 6. τ := <пороговое значение>; $\{0 < \tau < 1$; например, $\tau := 0{,}001$ }

Шаг 7. выбрать первые nc документов в качестве затравок:

$$s_j := d_k, j = \overline{1, nc}, \text{так, чтобы } p_{s_j} - p_{s_{j+1}} > \tau;$$

Шаг 8. $C_j := \emptyset, j = \overline{1, nc}$;

Шаг 9. для каждого $d_i \in \boldsymbol{\mathcal{D}}$**:**

Шаг 10. для каждого $s_j = d_k \in \boldsymbol{\mathcal{D}}$**:**

Шаг 11. $j^* = \arg\max_{j:s_j=d_k} c_{ki}$;

Шаг 12. $C_{j^*} := C_{j^*} + \{d_i\}$;

Выход: множество кластеров $\boldsymbol{\mathcal{C}}$=$\{C_1, \dots, C_{nc}\}$; множество затравок $\{s_j\}$, j от 1 до nc.

(t > 0) – изменения в коллекции документов, требуется модифицировать начальное разбиение.

C^2ICM:

Вход: множество проиндексированных документов $\mathcal{D}^{t-1}$; $\mathcal{C}^{t-1}$=$\{C_j | s_j \in \boldsymbol{\mathcal{D}} -$ затравка$\}$;, $\mathcal{D}^{t-1}_{ragbag}$;

$\boldsymbol{\mathcal{D}}'$ – множество новых документов; $\boldsymbol{\mathcal{D}}''$ – множество удаляемых из $\boldsymbol{\mathcal{D}}$ документов.

Шаг 1. обновить массив документов $\mathcal{D}^t := \mathcal{D}^{t-1} + \boldsymbol{\mathcal{D}}' - \boldsymbol{\mathcal{D}}''$;

Шаг 2. обновить словарь $\mathcal{T}^t$: удалить термины, которые не принадлежат ни одному документу; добавить термины из новых документов, если их раньше не было в словаре.

Шаг 3. *C^3M:* обновить матрицу C = $\{c_{ij}\}$, массив p_i, вычислить nc и $\{s_k\}$;

Шаг 4. Сформировать множество документов, подлежащих кластеризации $\mathcal{D}^t_r := \mathcal{D}' + \mathcal{D}^t_{fs} + \mathcal{D}^{t-1}_{ragbag}$. где $\mathcal{D}^{t-1}_{ragbag}$– множество документов, не покрытых ни одним кластером на шаге (t-1);

$\mathcal{D}^t_{fs} = \{d_i \in \mathcal{D}^{t-1} | \, d_i \in C_j, C_j -$ недействительный кластер на шаге $t\}$;

Шаг 5. Кластеризовать $\mathcal{D}^t_r$ с помощью *C^3M: шаги 5-12;*

Шаг 6. Если есть документы из $\mathcal{D}^t_r$, которые не попали ни в один кластер, то поместить их в $\mathcal{D}^t_{ragbag}$;

Выход: множество кластеров $\mathcal{C}^t$=$\{C_1, \dots, C_{nc}\}$; множество затравок $\{s_j\}^t$, j от 1 до nc, $\mathcal{D}^t_{ragbag}$.

Пример. Для простоты вычислений будем использовать бинарные веса терминов в наших документах.

	t1	***t2***	***t3***	***t4***	***t5***	***t6***
	китайский	*пекин*	*шанхай*	*макао*	*япония*	*токио*
d1	1	1	0	0	0	0
d2	1	0	1	0	0	0
d3	1	0	0	1	0	0
d4	1	0	0	0	1	1
d5	1	0	0	0	1	1
d6	0	1	0	0	0	1

t=0. Целиком новый массив документов. Алгоритм C^3M.

Матрица коэффициентов покрытия C: Затравочная сила документов p_i:

0.3500	0.1000	0.1000	0.1000	0.1000	0.2500	0.4550
0.1000	0.6000	0.1000	0.1000	0.1000	0	0.4800
0.1000	0.1000	0.6000	0.1000	0.1000	0	0.4800
0.0667	0.0667	0.0667	0.3444	0.3444	0.1111	0.6774
0.0667	0.0667	0.0667	0.3444	0.3444	0.1111	0.6774
0.2500	0	0	0.1667	0.1667	0.4167	0.4861

Число кластеров nc = 3.

Выбираем 3 документа-затравки с наибольшей затравочной силой. Наибольшая сила 0.6774 у d4 и d5, проверим, являются ли они «идентичными» документами (см. определение матрицы коэффициентов покрытия): c44 = c55 = c45 = c54 = 0.3444. Только один из «идентичных» документов может быть выбран. выбираем d5. Итак, затравочные документы: d5, d6 и d2, обладающие затравочной силой 0.6774, 0.4861 и 0.4800 соответственно.

Получаем три кластера:

$\boldsymbol{C}1$ = {d5,d4}; $\boldsymbol{C}2$ = {d6,d1}; $\boldsymbol{C}3$ = {d2,d3}.

t=1. Обновление массива документов. Алгоритм C^2ICM.

Добавим новый документ d7 = [0 1 0 1 0 0] и удалим d2 = [1 0 1 0 0 0]:

	t1	***t2***	***t4***	***t5***	***t6***
	китайский	*пекин*	*макао*	*япония*	*токио*
d1	1	1	0	0	0
d3	1	0	1	0	0
d4	1	0	0	1	1
d5	1	0	0	1	1
d6	0	1	0	0	1
d7	0	1	1	0	0

Заметим, что признак *t3* больше не принадлежит ни одному документу.

Матрица коэффициентов покрытия C: Затравочная сила документов p_i:

0.2917	0.1250	0.1250	0.1250	0.1667	0.1667	0.4132
0.1250	0.3750	0.1250	0.1250	0	0.2500	0.4688
0.0833	0.0833	0.3611	0.3611	0.1111	0	0.6921
0.0833	0.0833	0.3611	0.3611	0.1111	0	0.6921
0.1667	0	0.1667	0.1667	0.3333	0.1667	0.4444
0.1667	0.2500	0	0	0.1667	0.4167	0.4861

Число кластеров nc = 2.

Выбираем 2 документа-затравки с наибольшей затравочной силой: d5 (старая затравка) и d7 (новая затравка), обладающие затравочной силой 0.6921 и 0.4861 соответственно.

Следовательно, $\boldsymbol{\mathcal{D}}_{\text{г}}$ = {d6, d1, d3};

Ищем документы-затравки, которые максимально покрывают элементы множества $\boldsymbol{\mathcal{D}}_{\text{г}}$.

Получаем два кластера:

$\boldsymbol{\mathcal{C}}1$ = {d5,d4,d6}; $\boldsymbol{\mathcal{C}}2$ = {d7,d1,d3}.

Вычислительная сложность. Относительно размера коллекции документов алгоритмы C^3M и C^2ICM имеют линейную вычислительную сложность.

§ 2.6. Нейросетевой алгоритм SOM

Алгоритм самоорганизующихся карт (SOM, Self Organizing Maps) был предложен Тойво Кохоненом в 1982 году как решение проблемы визуализации и кластеризации данных. Визуализация данных осуществляется путём проецирования многомерного пространства данных в двумерное пространство – карту данных. Такая карта, построенная для массива полнотекстовых документов, может служить как поисковый механизм, альтернативный поиску по запросу, предлагающий пользователю обзор/навигацию по коллекции документов. Документы близких тематик оказываются на карте рядом.

Идея алгоритма заключается в том, чтобы обучить нейронную сеть без учителя. Сеть состоит из некоторого числа нейронов, упорядоченных по узлам двумерной сетки. Каждый нейрон имеет координаты в исходном $|\boldsymbol{\mathcal{T}}|$-мерном пространстве документов и в двумерном пространстве карты. В процессе обучения нейроны упорядочиваются в пространстве документов так, чтобы наилучшим образом описать входной массив документов. Этот процесс является итерационным, на каждой итерации *t*:

а) случайным образом выбирают из входного массива $d_i \in \boldsymbol{\mathcal{D}}$;

б) находят нейрон-победитель $m_c \in \boldsymbol{\mathcal{M}}$, то есть ближайший к документу d_i:

$$\text{c} = \arg\min_j \left\| d_i - m_j \right\|, \text{для } \forall m_j \in \boldsymbol{\mathcal{M}}, j = \overline{1, |\mathcal{M}|}; \quad (51)$$

$\|\quad\|$ – евклидово расстояние между векторами в пространстве терминов;

в) корректируют веса (координаты в пространстве терминов) нейрона-победителя и его соседей:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[d_i - m_i(t)], \quad (52)$$

где $h_{ci}(t)$ – это функция соседства, которая определяет, у какого количества нейронов (узлов сетки), окружающих нейрон-победитель, изменятся веса и насколько сильно они изменятся. Часто функция соседства имеет следующий вид:

$$h_{ci}(t) = \alpha(t) \times e^{\left(-\frac{\|r_i - r_c\|^2}{2\sigma(t)^2}\right)}, \quad (53)$$

где $\alpha(t)$ – коэффициент обучения, монотонно убывающий с ростом номера итерации t, $0 < \alpha(t) < 1$; на начальных шагах работы сети происходит заметное упорядочивание векторов нейронов, а на остальных –уточняющая подстройка карты; часто $\alpha(t)$ задают как линейную, экспоненциальную или обратно пропорциональную функцию $\alpha(t) = A/(t+B)$, где A и B – константы;

$r_i, r_c \in \mathbb{R}^2$ – координаты нейронов как узлов сетки;

$\sigma(t)$ – ширина соседства, монотонно убывающая с ростом номера итерации *t*.

Процесс обучения сети завершается, когда ошибка *E* как среднее расстояние между документами и их ближайшими нейронами становится меньше требуемого порогового значения:

$$E = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\|x_i - m_c\|. \quad (54)$$

Координатами документов на карте являются узлы, соответствующие ближайшим им нейронам (нейронам-победителям).

Для визуализации карты вычисляют матрицу расстояний между нейронами в пространстве терминов. Область карты, соответствующую очередному нейрону окрашивают цветом, определённым пропорционально среднему расстоянию данного нейрона до всех его ближайших соседей (узлов на карте). Если для окраски используются градации серого цвета, то чем ближе расположились в результате обучения соседние нейроны в пространстве терминов, тем светлее будут соответствующие им ячейки карты, и наоборот, чем дальше, тем темнее. По характеру окраски карты можно делать выводы о количестве и составе кластеров документов: темные области карты соответствуют границам между кластерами. Другим способом определения кластеров является кластеризация итоговых нейронов любым алгоритмом, например, алгоритмом k-средних.

Алгоритм в общем виде. Построение самоорганизующейся карты.

Вход: множество проиндексированных документов $\boldsymbol{\mathcal{D}}$, размеры сетки (*len* x *len*).

Шаг 1. Инициализация карты: распределение нейронов $m_i \in \boldsymbol{\mathcal{M}}, i = \overline{1, |\boldsymbol{\mathcal{M}}|}, |\boldsymbol{\mathcal{M}}| = len \times len,$ по узлам карты и присвоение случайных значений весам нейронов (координатам в пространстве терминов).

t := 0; задать пороговое значение допустимой ошибки обучения τ.

Шаг 2. $\boldsymbol{\mathcal{D}}_{tr} := \boldsymbol{\mathcal{D}}$.

Шаг 3. Извлечь случайный документ $d_i \in \boldsymbol{\mathcal{D}}_{tr}$.

Шаг 4. Вычислить нейрон-победитель $m_c \in \boldsymbol{\mathcal{M}}$ по формуле (51).

Шаг 5. Скорректировать веса нейрона-победителя и его соседей по формуле (52).

Шаг 6. *t* := *t* + 1. Если $\mathcal{D}_{\text{tr}} \neq \emptyset$, то продолжить с *шага 3*, иначе перейти к *шагу 7*.

Шаг 7. Вычислить текущую ошибку обучения *E* по формуле (54). Если $E > \tau$, то повторить с *шага 2*.

Выход: множество нейронов $\boldsymbol{\mathcal{M}}$, представленных как в пространстве терминов, так и в двумерном пространстве карты (сетки).

Пример. Продолжим наш пример с шестью документами, построим для них карту в виде прямоугольной сетки, содержащей 10 на 10 узлов (рис. 12).

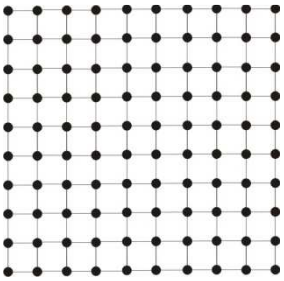


Рис. 12. Исходный порядок нейронов на карте

Инициализацию весов нейроной выполним случайным образом. Коэффициент обучения зададим следующим образом:

$\alpha(t) = 0{,}1 \times e^{\left(-\frac{t}{1000}\right)}$.

Ширину соседства будем вычислять по формуле:

$\sigma(t) = len \times e^{\left(-\frac{t}{C}\right)}$,

где len – это число узлов на каждой стороне сетки, C – константа, $C = 1000/\ln(len)$.

Пороговому значению ошибки E присвоим значение 0,0000001.

Запустим алгоритм SOM.

В результате было выполнено 1384 итерации, итоговые координаты документов в двумерном пространстве карты следующие: d1 (0;0), d2 (5;9), d3 (5;0), d4 (9;1), d5 (9;5), d6 (0;5). Левый верхний узел карты – координаты (0;0). Сама карта, построенная с помощью вычисления матрицы расстояний между нейронами, представлена на рис. 13.

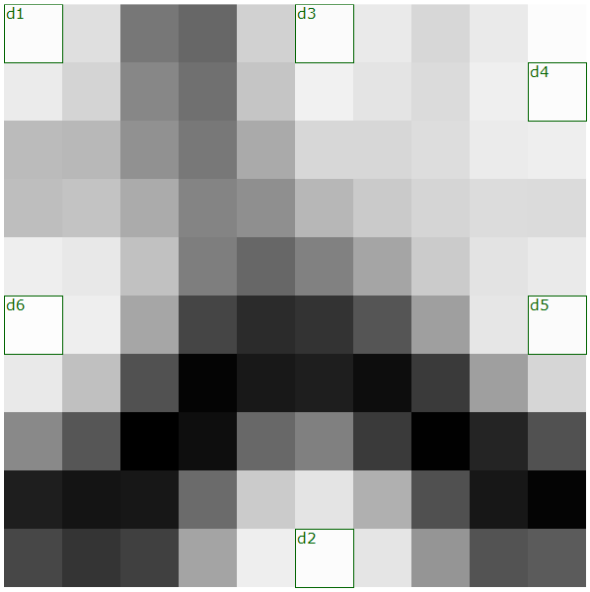


Рис. 13. Самоорганизующаяся карта для примера из шести документов

По цвету ячеек на карте видим, что образовалось три кластера:

$\boldsymbol{C}1$ = {d1,d6}; $\boldsymbol{C}2$ = {d3,d4,d5}; $\boldsymbol{C}3$ = {d2}.

Вычислительная сложность. Алгоритм имеет квадратичную сложность по числу документов – $O(|\mathcal{T}| \times |\mathcal{D}|^2)$.

§ 2.7. Экспериментальная оценка результата классификации без учителя

Разбиение документов на кластеры оценивают путём вычисления мер качества, которые бывают двух видов:

а) *внешние меры*, сравнивают созданное системой разбиение документов с «эталонным» разбиением;

б) *внутренние меры*, автоматически анализируют внутренние свойства, присущие конкретному массиву документов.

Внешние меры. Ключевым понятием в сравнении «эталонного» и автоматически полученного разбиения является анализ сходства предсказаний экспертов и предсказаний системы относительно принадлежности каждой пары документов одному или разным кластерам. Для каждой пары документов d_i, d_j, где $d_i, d_j \in \boldsymbol{\mathcal{D}}$, на основе знания о двух разбиениях $\boldsymbol{\mathcal{C}}^*$ и $\boldsymbol{\mathcal{C}}$, ($\boldsymbol{\mathcal{C}}^*$ получено от экспертов, $\boldsymbol{\mathcal{C}}$ – алгоритмом кластеризации) необходимо составить таблицу следующего вида:

	d_i, d_j принадлежат одному кластеру в $\boldsymbol{\mathcal{C}}^*$	d_i, d_j принадлежат разным кластерам в $\boldsymbol{\mathcal{C}}^*$
d_i, d_j принадлежат одному кластеру в $\boldsymbol{\mathcal{C}}$	a	c
d_i, d_j принадлежат разным кластерам в $\boldsymbol{\mathcal{C}}$	b	d

Дальнейший анализ полученной таблицы аналогичен описанному в § 1.9: вычисляют меры качества, заданные формулами (35)-(40), применяют микро-, макро-усреднение.

Внутренние меры. Внутренние меры предназначены как для сравнения разбиения на кластеры разными алгоритмами, так и одним и тем же алгоритмом, но с разными значениями входных параметров. Примером второго случая является попытка автоматически установить, какое количество кластеров приведёт к оптимальному в определённом смысле разбиению. Внутренние меры строятся на основе предположения, что оптимальное разбиение обладает свойствами компактности и отделимости. Компактность означает, что члены одного кластера должны быть настолько близкими друг другу, насколько это возможно. Отделимость – что сами кластеры должны достаточно далеко отстоять друг от друга.

Приведём примеры трёх внутренних мер: для чёткого плоского, нечёткого плоского и иерархического разбиения. Подробнее об этих и других мерах можно узнать, например, из работы [8].

Внутренняя мера чёткого плоского разбиения. Индекс Дана *DI*:

$$DI(C) = \frac{\min\limits_{i \neq j} \delta(c_i, c_j)}{\max\limits_{1 \le l \le N_c} \Delta(c_l)}, \tag{55}$$

где $C = \{c_1, \ldots, c_{N_C}\}$ – множество кластеров; $N_C = |\mathcal{C}|$;

$\delta(c_i, c_j) = \frac{1}{|c_i| \times |c_j|} \times \sum_{d_k \in c_i, d_p \in c_j} dist(\vec{d}_k, \vec{d}_p)$ – мера расстояния между кластерами;

$\Delta(c_l) = 2 \times \left(\frac{\sum_{d_j \in c_l} dist(\vec{d}_j, \vec{c}_l)}{|c_l|} \right)$ – мера диаметра кластера.

$\vec{c}_i$ – вектор центроида кластера c_i.

Оптимальному разбиению данных соответствует максимальное значение индекса Данна.

Внутренняя мера нечёткого разбиения. Модифицированный коэффициент разбиения *MPC*:

$$MPC(|\boldsymbol{\mathcal{C}}|) = 1 - \frac{|\mathcal{C}|}{}. \tag{56}$$

$$MPC(|\mathcal{C}|) = 1 - \frac{|\mathcal{C}|}{|\mathcal{C}| - 1}\left(1 - PC(|\mathcal{C}|)\right), \tag{57}$$

$$PC(|\mathcal{C}|) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{D}|} u_{ij}^2,$$

где $0 \le MPC(|\mathcal{C}|) \le 1$,

$PC(|\mathcal{C}|)$ – коэффициент разбиения, $\frac{1}{|\mathcal{C}|} \le PC(|\mathcal{C}|) \le 1$;

u_{ij} – элемент матрицы нечёткого разбиения.

Оптимальному разбиению данных соответствует максимальное значение модифицированного коэффициента разбиения.

Внутренняя мера иерархического разбиения. Кофенетический коэффициент корреляции (*CPCC*), для вычисления которого необходимо сформировать кофенетическую матрицу S_C. Каждым элементом данной матрицы является номер уровня в иерархии кластеров, на котором документы d_i и d_j впервые встретились в одном кластере. Мера *CPCC* оценивает степень сходства кофенетической матрицы S_C и действительной матрицы близости документов коллекции *S*.

$$CPCC(C) = \frac{(1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{|N_D|}\left(s_{ij}sc_{ij} - \mu_S\mu_{S_C}\right)}{\sqrt{\left((1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{N_D}\left(s_{ij}^2 - \mu_S\right)\right)\left((1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{N_D}\left(sc_{ij}^2 - \mu_{S_C}\right)\right)}}, \tag{58}$$

где $-1 \le CPCC \le 1;\ M = \frac{N_D(N_D - 1)}{2};\ N_D = |\mathcal{D}|;$

s_{ij} и $sc_{ij} - (i, j)$-ые значения матриц S и S_C соответственно;

$\mu_S = \frac{1}{M}\sum_{i=1}^{N_D}\sum_{j=i+1}^{N_D} s_{ij}$, $\mu_{S_C} = \frac{1}{M}\sum_{i=1}^{N_D}\sum_{j=i+1}^{N_D} sc_{ij}$ – здесь средние значения матриц S и S_C соответственно.

Чем ближе к нулю значение *CPCC*, тем ниже сходство между матрицами.

§ 2.8. Выбор метода классификации без учителя

Обзор экспериментальных исследований. В области классификации без учителя сложилась не такая благоприятная среда для сравнительного анализа экспериментов, как в области классификации с учителем. В первую очередь это связано с высокой трудоёмкостью формирования тестовых данных. В соответствии с определением внешних критериев необходимо, чтобы эксперт заранее оценил $(((|\mathcal{D}|)*(|\mathcal{D}| - 1))/2)$ пар документов, что является непосильной задачей для реальных коллекций документов, содержащих десятки тысяч документов и более. Неизвестны готовые наборы данных для экспериментов, и большинство коллективов проводят эксперименты на собственных данных и применяют собственные методики для оценки, что затрудняет сравнительный анализ разных алгоритмов по результатам из различных публикаций.

Таким образом, главным основанием для выбора алгоритма кластеризации является знание о его теоретических характеристиках и оценка пригодности для решения частной задачи разбиения текстов.

Подход к поиску разбиения. Алгоритмы, применяющие теоретико-графовый подход, к ним относятся рассмотренные иерархические агломеративные алгоритмы, имеют как минимум квадратичную сложность вычислений, что делает их малопригодными для приложений, где на первом месте стоит производительность системы. С другой стороны, этот вид алгоритмов может приносить выигрыш в эффективности, то есть в качестве классификации, поскольку имеет глобальную сходимость, детерминирован и частично лишен таких недостатков многих плоских алгоритмов как, например, необходимость заранее знать число кластеров.

Итеративные алгоритмы, пытающиеся улучшить изначальное разбиение массива документов путем оптимизации некоторой целевой функции, к ним относятся алгоритм k-средних и его модификации, имеют линейную вычислительную сложность, что делает их привлекательными для реализации в составе программной системы. И не смотря на их теоретическую локальную сходимость, нередко показывают приемлемый уровень эффективности системы. Однако эти алгоритмы чувствительны к шуму. Если известно, что массив документов содержит большой процент шума, то следует применять либо модификации k-средних, либо иерархические алгоритмы, либо алгоритмы, специально спроектированные для борьбы с шумом, например, плотностный алгоритм DBSCAN. У последнего имеется ещё одно преимущество относительно k-средних – это распознавание кластеров произвольной формы. Но достичь этого возможно только при удачном подборе параметров плотности, что на практике не всегда удаётся. В вопросе количества настроечных параметров и важности способа их инициализации самыми неприхотливыми являются иерархические агломеративные алгоритмы, которые в общем случае не используют никаких дополнительных параметров.

Нечёткие алгоритмы, относящие документы к нескольким кластерам одновременно, имеют преимущество над чёткими в тех приложениях, где природа данных подразумевает такую нечёткость. Они имеют те же недостатки, что и их чёткие предшественники, например, сказанное в этом подразделе про алгоритм k-средних верно и для алгоритма нечётких с-средних.

И, наконец, если приложение требует визуализировать полученные кластеры, то имеет смысл обратить внимание алгоритмы, специально на это нацеленные, к ним в первую очередь относится алгоритм самоорганизующихся карт, и в некотором смысле можно отнести иерархические алгоритмы. Однако не следует ожидать, что такие карты будут востребованы широкой аудиторией пользователей.

Список используемой литературы

1. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск.: Пер. с англ. – М.: ООО «Вильямс», 2011. – 528 с.: ил.
2. Yang Y., Liu X. A re-examination of text categorization methods, School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213-3702, USA, 1999 – p. 8.
3. Yang Y., Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization // The Fourteenth International Conference on Machine Learning: Proceedings of ICML'97. – San Francisco, 1997. – P. 412-420.
4. Ester M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P .Kriegel, J. Sander, X. Xu // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). – Portland, 1996. – P. 226-231.
5. Can F. Experiments on Incremental Clustering. – Miami University,1991.– Access mode: http://sc.lib.muohio.edu/bitstream/handle/2374.MIA/187/fulltext.pdf?sequence=1
6. Kohonen T. Self organization of a massive document collection / T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela // IEEE Transactions on neural networks. – 2000. – Vol. 11, No. 3. – P. 574 - 585.
7. van Rijsbergen C. J. Information retrieval [Electronic resource]. – Electronic text and graphic data. – 1979. – Access mode: http://www.dcs.gla.ac.uk/Keith/Preface.html
8. Halkidi M. On Clustering Validation Techniques / M. Halkidi, V. Batistakis, M. Vazirgiannis // Journal of Intelligent Information Systems, Kluwer Academic Publishers. Manufactured in The Netherlands. – 2001. – 17:2/3. – P. 107-145.
9. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, No. 1. – 47 p.

ЧАСТЬ VI. ИНФОРМАЦИОННЫЕ ПОТОКИ И СЛОЖНЫЕ СЕТИ *(Д.В. ЛАНДЭ)*

Глава 1. Основы анализа информационного пространства и информационных потоков

§ 1.1. Понятие информационного пространства

Под информационным пространством принято понимать совокупность информационных ресурсов, технологий их сопровождения и использования, информационных и телекоммуникационных систем, образующих некую информационную инфраструктуру. Элементами информационного пространства могут быть, в частности, документы, обобщающие самые различные виды информации – файлы, электронные письма, веб-страницы не зависимо от форматов их представления.

Естественно, приведенное определение информационного пространства является качественным. Конечно же, термин «пространство» в данном случае, не совпадает с понятием «пространство» в математике или физике. В качестве примеров удачных моделей информационного пространства можно приветсти «векторно-пространственную» модель Г. Солтона [1] или модель старения информации Бартона-Кеблера [2]. Модель такого информационного пространства, как сеть WWW была построена А. Брёдером и его соавторами из компаний IBM и Altavista [3].

Во многих моделях информационного пространства изучаются структурные связи между тематическими множествами его элементов – документами.

Информационное пространство можно рассматривать и как множество связанных по смыслу элементов (документов), образующих информационные системы – кластеры близких по тематике документов. При этом оно за все время существования сохраняет свои устойчивые закономерности. Многочисленными исследованиями показано, что параметры частотного и рангового распределений документов во многих информационных системах остаются одинаковыми, и определяются параметрами, зависящими от содержания, тематики информации. В связи с этим С.А. Иванов [4] заметил, что «информационное пространство – это документальная среда, в которой формируются кластерные структуры научных публикаций в периодических изданиях, являющиеся фракталами». Информационные системы отражают в информационном пространстве коммуникационные процессы в своей тематической области, появление новых тематик сопровождается возникновением новых фрактальных массивов в информационном пространстве.

Как и многие другие сложные системы, информационное пространство можно представить как коммуникационную среду – в виде системы с комплексом связей информационных источников и преобразователей между собой, влияющих друг на друга в зависимости от уровня восприятия генерируемых и преобразуемых ими отдельных информационных сообщений.

При этом для моделирования источников и преобразователей информации, с одной стороны, вполне подходит классическая теория информации как математическая теория связи, разработанная Шенноном в 40-х годах XX столетия и

существенно дополненная и расширенная в последующие годы работами Н. Винера, В. А. Котельникова и А. Н. Колмогорова. Однако классическая теория информации не учитывает взаимодействия между источниками и преобразователями информации, что, с другой стороны, вполне укладывается в идеологию современной теории сложных систем

§ 1.2. Информационный поток как объект исследования

Сетевые структуры в информационном пространстве состоят из отдельных элементов, образующих в динамике своей эволюции (появление, развитие, модификация, уничтожение) информационные потоки. Следовательно, живучесть информационных систем напрямую зависит от свойств информационных потоков.

Для исследования современных информационных потоков в Интернет, то есть потоков сообщений, которые публикуются на страницах веб-сайтов, в социальных сетях, блогах, и тому подобное, должен применяться принципиально новый инструментарий, потому что классические методы обобщения информационных массивов (классификации, фазового укрупнения, кластерного анализа и тому подобное) не всегда способны адекватно отражать состояние динамической составляющей информационного пространства. В этом случае речь идет не столько об анализе документальных массивов фиксированных размеров, пусть даже очень больших, сколько об обобщении динамического потока гипертекстовых данных.

Конечно, большая часть информации, которая представлена в Интернет, находит своего потребителя. Однако если рассматривать всю совокупность сетевых публикаций как какую-то общность по отношению к конкретному пользователю (или группы пользователей), то можно увидеть ряд проблем, связанных с полнотой, релевантностью и оперативностью получения данных. Поиск, фильтрация, сбор информации в Интернет требуют достаточной квалификации персонала и, к сожалению, при этом не могут учитываться все особенности информационной структуры сети и представления в ней данных. Это, в свою очередь, ведет к тому, что единичные выборки информации из веб-пространства не могут считаться репрезентативными.

При этом информационный поток, который «потребляется» конкретным пользователем носит, как правило, выраженную предметную направленность, которая характеризуется областью его интересов. Поиск и обработка информации в ручном режиме – достаточно трудоемкий, а главное, длительный процесс, который чаще всего не дает желаемого результата. Решение проблемы на практике возможно путем создания автоматизированных систем сбора, фильтрации и анализа информации, так называемых «интеллектуальных посредников» между пользователем или корпоративной информационной системой и сетью Интернет. Подобная система должна осуществлять сбор и селекцию информации из Интернет и создавать документальную базу данных, специфицированную предметной областью пользователя, то есть выполнять функции интеграции информационных потоков. Загрузка информации в базу данных должна сопровождаться ее классификацией и структуризацией. Для последующей информационно-аналитической работы пользователю должны предоставляться эффективные средства навигации, поиска и обобщения информации, которая сохраняется в соответствующей динамической документальной базе данных.

Современный уровень развития информационного пространства обуславливает интерес к подходам, основанным на понимании информации как меры упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки. Для организации эффективной коммуникации в сетях сегодня приходится постоянно возвращаться к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др., широкие перспективы применения мощного аппарата математики и физики в решении теоретико-информационных задач [5].

Для формального описания информационных потоков введем некоторые общие для всего последующего изложения предположения. Дадим определение информационного потока, какое корреспондируется с классическим определением из теории информации.

Рассмотрим отрезок (a,τ) действительной оси (оси времени), где $\tau > a$. Допустим, что на этом отрезке времени в соответствии с некоторыми закономерностями в сети публикуется некоторое количество информационных документов – k. На оси времени моменты публикации отдельных документов обозначим как τ_1, τ_2 ,..., τ_k $(a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$. Информационным потоком будем называть процесс $N_\alpha(\tau)$, реализация которого характеризуется количеством точек (документов), появившихся в интервале (a,τ), как функцию правого конца отрезка τ. В соответствии с этим определением реализация информационного потока является неубывающей ступенчатой всегда целочисленной функцией $N_\alpha(\tau)$.

Приведено определение на локальных временных областях соответствует действительности, но не учитывает такой эффект, как старение информации, какое противоречит «накопительной» способности информационного потока $N_\alpha(\tau)$ на больших промежутках времени.

Так определенный информационный поток учитывает лишь количество информационных сообщений, вне зависимости от их содержания. В общем случае, определение содержания, тематики отдельных документов является достаточно субъективным процессом. Для строгого моделирования тематических информационных потоков используют модели, которые различают документы по отдельным словам или словосочетаниям (обычно их называют термами, от англ. *Terms*).

Задачи мониторинга информационных потоков большого объема в компьютерных сетях, их адаптивного агрегирования и обобщения осложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов. В настоящее время исследования по проблемам анализа информационных потоков большого объема в компьютерных сетях носят чаще всего узко специализированный характер. Вместе с тем, опыт создания и внедрения корпоративных информационных систем свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения разнообразных аналитических сведений, навигации в документальных информационных потоках больших объемов.

При моделировании этих процессов используются методы нелинейной динамики, теории клеточных автоматов и самоорганизованной критичности. При моделировании информационных потоков изучаются структурные связи между входящими в них массивами документов. Сегодня при этом все чаще применяется фрактальный анализ, подход, базирующийся на свойствах сохранения внутренней

структуры массивов документов при изменениях их размеров или масштабов рассмотрения. Теория информации, которая ранее находила свое основное применение в области передачи данных, становится полезной и для анализа текстовых массивов, динамически порождаемых в сетях.

Предусматривается, что новостные сообщения обладают свойством старения, т.е. теряют свою актуальность со временем. Все информационное пространство можно с достаточной мерой условности разделить на две составляющие – стабильную и динамическую, которые имеют очень разные характеристики своего развития. В частности, процесс старения информации в известной модели Бартона-Кеблера описывается уравнением, которое состоит из двух компонент:

$$m(t) = 1 - ae^{-T} - b\,\mathrm{e}^{-2T},$$

где $m(t)$ – часть полезной информации в общем потоке через время T, первое вычитаемое соответствует стабильным ресурсам, а второе – динамическим, новостным. Это уравнение также в полной мере соответствует объемам информации, которые формируются в информационном пространстве по определенными тематиками, которые время от времени возникают и исчезают. Стабильная составляющая информационного пространства содержит информацию «долгосрочного» плана, в то время, как динамическая составляющая содержит ресурсы, которые постоянно обновляются. Некоторая часть последней составляющей впоследствии вливается в стабильную, однако большая часть «исчезает» из информационного пространства или попадает в сегмент так называемой его «скрытой» части, не доступной пользователям с помощью обычных информационно-поисковых систем (ИПС).

§ 1.3. Тематические информационные потоки

Под тематическим информационным потоком будем понимать последовательность сообщений, соответствующих определенной тематике. Таким образом информационные системы в нашем понимании также являются тематическими информационными потоками, но в отличие от следующих друг за другом сообщений в простых информационных потоках, информационные системы – это сетевые структуры, учитывающие многочисленные информационные связи.

В узком смысле под тематическим информационным потоком будем понимать количество документов, которые в некотором смысле соответствуют заданной теме. Рассмотрим общую картину динамики тематических информационных потоков, ограничившись механизмами, типичными для динамического сегмента веб-пространства.

Многочисленные факты свидетельствуют о том, что в действительности динамика тематических информационных потоков определяется комплексом внутренних нелинейных механизмов, которые лишь частично коррелируют с объективным окружением. Очевидно, что эта динамика в принципе не может быть объяснена некоторым одним фактором, который полностью отвечает за все разнообразие наблюдаемых эффектов. Именно это обстоятельство и объясняет большую актуальность проблемы моделирования динамики тематических информационных потоков.

Информационный поток, измеряемый количеством сообщений, является величиной относительно стабильной. Изменяются во времени лишь объемы массивов сообщений, соответствующие той или иной тематике, той или иной информационной

системе. Другими словами, рост количества публикаций по одной теме при ограниченной способности их генерации (что вполне соответствует действительности) сопровождается уменьшением публикаций на другие темы, так что для каждого промежутка времени T имеем:

$$\int_0^T \sum_{i=1}^{M} n_i(t)dt = NT ,$$

где $n_i(t)$ – количество публикаций в единицу времени по теме i, а M – общее количество всех возможных тем. То есть для локальных временных промежутков можно наблюдать так называемый «тематический баланс».

Основной интерес в такой формулировке представляет изучение динамики отдельного тематического потока, который описывается плотностью $n_i(t)$.

Теоретически можно допустить, что множества публикаций, ассоциируемых с определенным набором тематик, пересекаются, то есть существуют публикации, которые могут быть отнесены одновременно к нескольким различным тематикам. В реальности такая политематичность действительно наблюдается, она является эффектом, который необходимо учитывать, но в первом приближении будем считать, что его вклад не искажает общей картины.

Каждая тематика также имеет ряд характерных свойств, которые допускают некоторую классификацию, например, на основе особенностей ее образования и воспроизведения во времени:

- публикации на «разовую» тему, временная зависимость количества которых резко растет, выходит на насыщение, а затем убывает и далее асимптотически стремиться к нулю;
- публикации по темам, которые периодически появляются в общем информационном потоке, а затем через некоторое время практически исчезают из него;
- публикации по теме, временная зависимость количества которых колеблется вокруг некоторого значения и никогда не исчезает полностью.

Таким образом сообщения могут подразделяться на аналогичные категории, причем каждая из них имеет собственную специфику развития во времени.

Еще сложнее выглядит синхронное изменение количества сообщений из нескольких тематических информационных потоков. Их поведение четко напоминает процессы взаимодействия популяций в биоценозе. Так, например, в ряде случаев увеличение числа публикаций по одной теме сопровождается сокращением числа публикаций по другим темам. Общая динамика в этом случае может описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку. Подчеркнем, что общие политематические потоки являются стационарными по количеству публикаций, динамика же в основном определяется «конкурентной борьбой» отдельных тематик.

Вместе с тем в практическом плане часто оказывается полностью удовлетворительным упрощеное понимание информационного потока как некоторой зависимой от времени величины $n(t)$, которая описывается уравнением:

$$\frac{dn(t)}{dt} = F\left(n(t), t\right).$$

В мнорготисленной литературе описаны много разновидностей систем «конкурентной борьбы» для разных модификаций модели в зависимости от целого

ряда предположений о реальных условиях протекания процессов. В самом простом виде такие уравнения могут иметь следующий вид:

$$\frac{dm_i(t)}{dt} = p_i \cdot m_i(t) - \sum_{j=1}^{N_m} r_{ij} \cdot m_i(t) \cdot m_j(t),$$

где N_m – количество тематик.

Приведенная система уравнений описывает перераспределение публикаций между тематиками, образующими фиксированный набор. Но в реальной жизни тематики (сюжеты) появляются и со временем исчезают, потому необходимо ввести в эти уравнения соответствующие коррективы. Это можно сделать по-разному, например, определив коэффициенты p_i и r_{ij} зависящими от времени так, чтобы каждый сюжет имел собственный максимум активности на определенном промежутке времени.

§ 1.4. Моделирование информационных потоков

Анализ динамики тематических информационных потоков, которые генерируются в веб-пространстве становится сегодня одним из наиболее информативных методов исследования актуальности тех или других тематических направлений [5]. Эта динамика обусловлена факторами, много из которых не поддаются точному анализу. Однако общий характер временной зависимости количества тематических публикаций в Интернете все же допускает построение математических моделей.

В поведении информационных потоков наблюдаются две характерных черты: во-первых, выразительная тенденция к постоянному росту их объемов, а во-вторых, усложнение динамической структуры. Наблюдения временных зависимостей числа сообщений в сетевых информационных потоках убедительно свидетельствуют о том, что механизмы их генерации и распространения, очевидно, связанны со сложными нелинейными процессами общей сетевой динамики.

Традиционными считаются два класса моделей информационных потоков: линейные и экспоненциальные. Оба класса имеют существенную ограниченность – монотонный характер временной зависимости. То есть они мало пригодны для изучения реальной динамики сетевых информационных потоков в течение длительных интервалов времени.

Линейная модель

В некоторых случаях динамика тематических информационных потоков, выражаемых количеством публикаций за определенный период, их интенсивностью, обусловленной, например, изменением активности тематики (ее повышением или старением), происходит линейно, то есть количество сообщений в момент времени t можно, соответственно, представить формулой:

$$y(t) = y(t_0) + v(t - t_0),$$

где t_0 – стартовое время отсчета, $y(t)$ – количество сообщений к моменту времени t, v – средняя скорость увеличения (уменьшения) интенсивности тематического информационного потока.

Важные характеристики информационного потока могут быть количественно оценены флуктуацией этого потока – изменением среднеквадратичного отклонения $\sigma(t)$, вычисляемого по формуле:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left[y(t_i) - \left(y(t_0) + v(t_i - t_0) \right) \right]^2}.$$

Если эта величина изменяется пропорционально квадратному корню от времени, то процесс изменения количества публикаций по избранной теме можно считать процессом с независимыми приращениями. При этом связями с предыдущими тематическими публикациями можно пренебречь.

В случае, когда среднеквадратичное отклонение пропроционально некоторой степени от времени: $\sigma(t) \propto t^{\mu}$ $(1/2 \le \mu \le 1)$, чем большее значение μ, тем выше корреляция между текущими и предыдущими сообщениями в информационном потоке.

Экспоненциальная модель

В некоторых случаях процесс изменения актуальности тематики (увеличения или уменьшения количества тематических сообщений в информационном потоке в единицу времени) аппроксимируется экспоненциальной зависимостью, которая выражается формулой:

$$y(t) = y(t_0)\exp\left[\lambda(t - t_0)\right],$$

где λ – среднее относительное изменение интенсивности тематического информационного потока.

В реальности актуальность тематики является дискретной величиной, измеряемой в моменты времени $t_0, ..., t_n$, которая лишь аппроксимируется приведенной выше зависимостью. В рамках данной модели справедливо:

$$y(t_i) = y(t_0)\exp\left[\lambda(t_i - t_0)\right] =$$

$$= y(t_0)\exp\left[\lambda(t_i - t_{i-1} + t_{i-1} - t_0)\right] = y(t_{i-1})\exp\left[\lambda(t_i - t_{i-1})\right].$$

Откуда:

$$\frac{y(t_i)}{y(t_{i-1})} = \exp\left[\lambda(t_i - t_{i-1})\right].$$

Введем обозначение: $\lambda(t_i)$ – относительное изменение интенсивности тематического информационного потока в момент времени t_i:

$$\lambda(t_i) = \lambda(t_i - t_{i-1})$$

и прологарифмируем приведенное выше уравнение:

$$\lambda(t_i) = \ln\frac{y(t_i)}{y(t_{i-1})}.$$

Относительное изменение интенсивности в момент времени t_i на практике также часто вычисляется как соотношение:

$$\lambda(t_i) = \ln\frac{y(t_i)}{y(t_{i-1})} \approx \frac{y(t_i) - y(t_{i-1})}{y(t_{i-1})}.$$

Изменение флуктуаций величины $\lambda(t_i)$ относительно среднего значения может оцениваться по стандартному отклонению:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\lambda(t_i) - \lambda)^2}.$$

В этом случае также, если $\sigma(t)$ изменяется пропорционально корню квадратному от времени, то можно говорить о процессе с независимыми приращениями – корреляция между отдельными сообщениями несущественна. В случае наличия значительной зависимости сообщений наблюдается соотношение: $\sigma(t) \propto t^{\mu}$, причем значение μ превышает 1/2, но ограниченно 1.

Значение μ, которое превышает 1/2, свидетельствует о наличие долгосрочной памяти в информационном потоке. Такой класс процессов получил название автомодельных, для которых предусматривается корреляция между количеством сообщений, публикуемых в разные моменты времени.

Логистическая модель

В отличие от модели Бартона-Кеблера в реальной динамике информационных потоков имеют место процессы как роста, так и спада количества документов. Поэтому для построения реалистичной картины, безусловно, необходимо применять более гибкие модели.

В первую очередь, стоит сказать, что документы в информационном потоке во многих отношениях напоминают популяции живых организмов. Они в определенном смысле «рождаются», «умирают» и дают «потомство» (документы, содержащиет информацию, ранее появившуюся в других документах). В современной научной литературе понятие популяции часто используется в широком смысле, и потому полностью обосновано введение его и при моделировании информационных потоков.

Во второй половине XX века были достигнуты значительные успехи в построении различных математических моделей динамики популяций, в частности, логистическая модель, которая оказалась применимой во многих отраслях науки и техники.

Логистическую модель можно рассматривать как обобщение экспоненциальной модели Мальтуса, предусматривающей пропорциональность скорости роста функции $y(t)$ в каждый момент времени ее значению:

$$\frac{dy(t)}{d(t)} = ky(t),$$

где k – некоторый коэффициент.

В реальной жизни, как правило, динамические системы имеют достаточно эффективные обратные связи, позволяющие корректировать характер процессов, происходящих в них, и тем самым удерживать их в определенных рамках. Информационные операции, корректируя эти обратные связи в определенные периоды эволюционного процесса, могут эффективно повлиять на характер поведения всей системы.

Наиболее простым обобщением закона Мальтуса, позволяющим уйти от неограниченного роста решения, является замена постоянного коэффициента k некоторой функцией времени $k(t)$. Естественно, эта функция должна быть выбрана таким образом, чтобы выполнялись условия:

– решение уравнения имело бы приемлемое поведение;

– структура функции имела бы определенный смысл с точки зрения исследуемого явления.

Главная идея логистической модели заключается в том, что для ограничения скорости роста на функцию $y(t)$ накладывается дополнительное условие, в соответствии с которым ее значением не должно превышать некоторую величину [6]. Для этого выберем $k(t)$ такого вида:

$$k(t) = k \cdot [N - ry(t)],$$

где N – предельное значение, которое функция $y(t)$ не может превысить, r – коэффициент, который описывает негативные для данной тенденции процессы, k – коэффициент пропорциональности. Причем предусматривается, что всегда $n_0 \le N$. Тогда вместо первого уравнения имеем:

$$\begin{cases} \dfrac{dy(t)}{d(t)} = ky(t)(N - ry(t)), \\ y(t_0) = y_0. \end{cases}$$

Модель, основанная на приведенном выше уравнении, называется логистической. Несмотря на мнимую простоту, подобное обобщение закона Мальтуса никоим образом не является примитивным. Напротив, оно позволяет явно включить в описание динамики популяций исключительно важную обратную связь. Логистическое уравнение, можно считать феноменологическим: исследователям не обязательно знать, как действуют конкретные механизмы, которые по мере роста $y(t)$ снижают скорость ее изменения.

Приведенное выше логистическое уравнение имеет два равновесных решения: $y(t) = 0$ и $y(t) = N$. С формальной точки зрения первое из них неустойчиво, однако на практике это не совсем так. Дело в том, что реальные объемы информационных потоков выражаются дискретными числами, и если в какой-то момент $y(t)$ принимает значение, меньшее единицы, то в дальнейшем расти оно уже не сможет. Поэтому в реальности решение $y(t) = 0$ также можно считать равновесным.

Второе же решение $y(t) = N$ является равновесным в любом смысле. Действительно, при $y(t) > N$ включаются механизмы спада зависимости, а при $y(t) < N$, соответственно, роста.

Рассмотрим, как логистическая модель может применяться во время анализа информационных потоков, а именно определение минимального начального количества c сообщений (которое можно, например, выделить для начала некоторой информационной операции). Пусть x – объем тематического информационного потока. На динамику этой величины осуществляется влияние других тематик, уменьшающих ее распространение, которое описывается таким образом: $\dot{x} = x - x^2 - c.$

Вычисления показывают, что поведение системы резко изменяется при некотором критическом значении c.

Очевидно, что при наличии благоприятных внешних условий (при некоторой плотности ресурса) объем информационного потока растет свободно, что способствует логистическому росту. В этом случае даже более сложные модели должны давать результаты, подобные приведенным. С другой стороны это означает, что основные параметры для конкретизации общей модели могут определяться в результате анализа упрощенной логистической модели.

Следовательно, логистическая модель успешно описывает достижение тематическим информационным потоком некоторого равновесного состояния.

Информационную динамику в общем случае можем представить как процесс, обусловленный возникновением и исчезновением отдельных тематик, которые происходят на фоне общих тенденций информационного пространства. Зафиксируем определенную тематику и допустим, что в момент времени $t=0$ существует n_0 фоновых публикаций. В силу того, что (в рамках принятой модели) актуальность тематики сохраняется в течение промежутка времени λ, можно рассматривать отдельно две временных области: $0<t\le\lambda$ с $D>0$ и $t>\lambda$ с $D=0$ (в рамка данной модели $D=const$ для каждой области – уровень актуальности темы) и, соответственно, функции $u(t)$ и $v(t)$, которые являются решениями для этих областей и «сшиваются» в точке λ:

$$y(t)=\begin{cases} u(t), & 0<t<\lambda, \\ v(t), & t>\lambda, \\ u(t)=v(t), & t=\lambda. \end{cases}$$

Первой области соответствует процесс роста количества публикаций в условиях ненулевой актуальности темы и, возможно, переход к состоянию насыщения.

Реакция медийных средств никогда не бывает мгновенной: всегда существует определенная задержка во времени. Этот аспект учитывается в модели путем введения фактора запаздывания τ.

Соответствующая динамика описывается уравнением, которое после переопределения коэффициентов и их нормировки для функции $u(t)$ можно представить в виде:

$$\frac{du(t-\tau)}{dt}=pu(t-\tau)(1-qu(t-\tau))+Du(t-\tau),$$

$$u(0)=n_0.$$

Подчеркнем, что содержательно величина p определяет нормируемую вероятность появления публикации в единицу времени независимо от актуальности темы. Этот фактор отображает фоновые механизмы генерации информации (типичным примером может быть механическая перепечатка материалов из престижных информационных источников). Величина же D характеризует непосредственное влияние актуальности данной темы. Параметр q характеризует уменьшение скорости роста количества публикаций и является величиной, обратной к асимптотическому значению зависимости $u(t)$ при $D=0$.

Для второй области, описываемой функцией $v(t)$, соответственно, имеем:

$$\frac{dv(t-\lambda)}{d(t)}=pv(t-\lambda)(1-qv(t-\lambda)).$$

При этом должно учитываться условие равенства функций $u(t)$ и $v(t)$ в момент $t=\lambda$:

$$v(\lambda)=u(\lambda).$$

Приведеные выше нелинейные дифференциальные уравнения являются вариантами записи уравнения Бернулли:

$y'=ay^2+by$,

которое линеаризуется стандартной заменой $z=1/y$:

$z'+bz+a=0$.

Общее решение этого уравнения имеет вид:

$$z = \frac{1}{\mu(x)}\left[C - a\int \mu(x)dx\right]$$

с интегрирующим множителем:

$\mu(x) = e^{bx}$.

Переменные C определяются: для первой области из начальных условий, а для второй – из условия «сшивки». Путем несложных преобразований находим решение для первой области:

$$u(t) = \frac{u_s}{1 + (u_s / n_0 - 1)\exp\left[-(p+d)(t-\tau)\right]},$$

где u_s – асимптотическое значение u, величина которого определяет область насыщения:

$$u_s = \frac{p+D}{pq}.$$

Таким образом, модель описывает зависимость, которая имеет *S*–подобную (логистическую) форму, представленную на рис. 6.1.

Заметим, что решение не зависит от значения n_0, что свидетельствует о несущественности начальных условий для информационной динамики. Каким бы не было начальное количество публикаций, насыщение будет определяться исключительно параметрами, которые характеризуют фоновую скорость роста количества публикаций, количественную меру актуальности и негативные для процесса факторы.

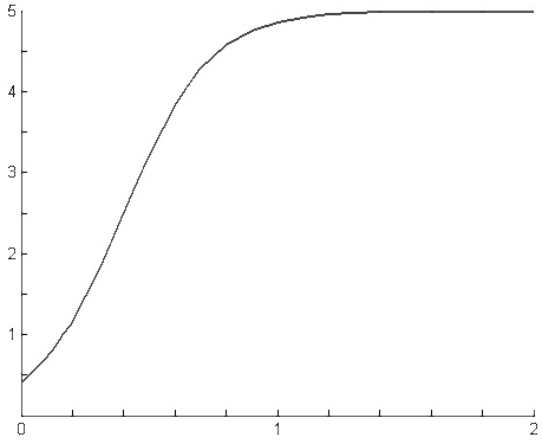


Рис. Часть VI.1. Функция роста

Кривая, представленная на рис. 6.1 имеет точку перегиба:

$$t_{\inf} = \frac{1}{p+D}\ln(u_s / n_0 - 1) + \tau.$$

Таким образом, для первой области имеем так называемую *S*-подобную зависимость, а при $t \sim t_{\inf}$ поведение $u(t)$ приближается к линейной и соответствует линейной модели.

Представим теперь выражение для $u(t)$ следующем виде:

$$u(t) = \frac{u_s}{\exp[(p+D)t] + (u_s / n_0 - 1)\exp[(p+D)\tau]},$$

откуда видно, что при условии

$$t < \frac{1}{p+D}\ln(u_s / n_0 - 1) + \tau = t_{\inf}$$

зависимость $u(t)$ имеет экспоненциальный характер, то есть для значений *t*, значительно меньших $t_{\inf}$, модель совпадает с экспоненциальной моделью.

Для второй области, соответственно, имеем (рис. 6.2):

$$v(t) = \frac{v(\lambda)}{qv(\lambda) + (1 - qv(\lambda))\exp[-p(t-\lambda)]^2},$$

учитывая условие «сшивки»:

$v(\lambda) = u(\lambda)$

Если зависимость $u(t)$ успевает достичь насыщения за промежуток времени $t < \lambda$, то приведенное выше уравнение можно упростить, представив его следующим образом:

$$v(t) = \frac{v_s(p+D)}{p + D(1 - \exp[-p(t-\lambda)])},$$

где $v_s = 1/q$ – асимптотическое значение зависимости $v(t)$.

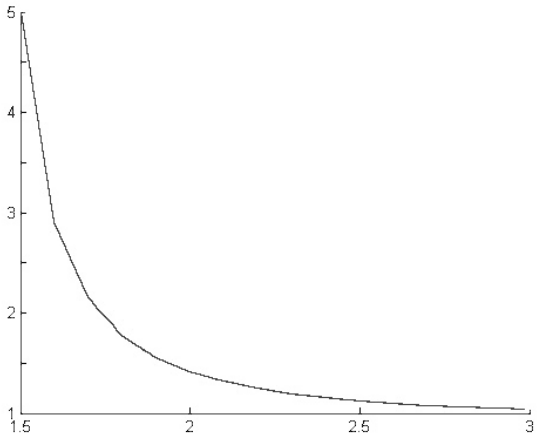


Рис. Часть VI.2. Функция спада

Как и ожидалось, величина v_s также не зависит ни от начального условия, ни от условия «сшивки» с функцией $u(t)$ на границе областей. Как видно, полученная зависимость имеет область насыщения u_s при $t \le \lambda$ и асимптотику v_s, которая описывает постепенное уменьшение числа публикаций до фонового уровня. То есть она, по крайней мере, на качественном уровне, согласовывается с общими соображениями о характере информационной динамики, полученными на основе опытных данных. Кроме того, на локальных участках она неплохо аппроксимируется линейной и экспоненциальной моделями.

В случае информационных потоков, которые ассоциируются с конкретными темами, необходимо описывать динамику каждого из таких потоков отдельно, принимая во внимание то, что рост одного из них автоматически приводит к уменьшению других и наоборот. Поэтому ограничение на количество сообщений по всем тематикам распространяется и на совокупность всех монотематических потоков.

В случае изучения общего информационного потока наблюдается явление «перетекания» публикаций из одних, теряющих актуальность тематик, в другие.

Общая динамика должна описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку. Подчеркнем, что общие политематические потоки являются стационарными по количеству публикаций, динамика же в основном определяется «конкурентной борьбой» отдельных тематик.

Приведенную выше систему уравнений «конкурентной борьбы» в рамках обобщенной логистической модели можно представить в таком виде:

$$\frac{dy_i(t)}{dt} = \left(p_i + D_i(t, \lambda_i) \right) \cdot \left(y_i(t) - \sum_j r_{ij} \cdot y_i(t) y_j(t) \right).$$

В этих соотношениях коэффициенты p_i и D_i имеют тот же смысл, что и ранее, а λ_i являются точками, в которых соответствующие D_i достигают максимальных значений.

§ 1.5. Модель диффузии информации

Обратимся к еще одному направлению в изучении процессов, связанных с информационными потоками - к диффузии информации.

Напомним, что в естественных науках под диффузией понимают взаимное проникновение друг в друга соприкасающихся веществ, вызванное, например, тепловым движением их частиц.

Для понимания сути дела необходимо, прежде всего, учитывать, что информация также в определенном смысле состоит из «частиц» - документов (сообщений). Множество процессов, близких к динамике информационных потоков, можно моделировать довольно точно, если четко параметризовать и установить их предельные параметры.

Процессы диффузии информации, как и процессы диффузии в физике, достаточно точно моделируются с помощью методов клеточных автоматов.

Концепция клеточных автоматов была впервые предложена больше полстолетия тому назад Дж. Фон Нейманом (J. Von Neumann) [7] и развита С. Вольфрамом (S. Wolfram) [8].

Клеточные автоматы являются полезными дискретными моделями для исследования динамических систем. Дискретность модели, а точнее, возможность представить модель в дискретной форме, может считаться важным преимуществом, поскольку открывает широкие возможности использования компьютерных технологий. Клеточные автоматы в этом смысле занимают особое место, поскольку их дискретность объединяется с другими преимуществами.

Главным достоинством клеточных автоматов является их абсолютная совместимость с алгоритмическими методами решения задач. Оконченный набор формальных правил, заданный на ограниченном множестве элементов (клеток), допускает точную реализацию в виде алгоритмов. Однако отсюда вытекает и главный недостаток клеточных автоматов: вычислительные трудности, которые возникают при расчетах соответствующих масштабов. Ведь на каждой итерации необходимо сканировать весь набор клеток и для каждой из них выполнять необходимые операции. Когда и клеток, и итераций действительно много, требуются значительные ресурсы, в том числе вычислительные и временные.

Поэтому продолжительное время клеточные автоматы воспринимались в основном как забавная, хотя и поучительная игра, которая не имеет практической ценности. Но в последние годы, в связи с бурным развитием компьютерных технологий, они начинают быстро входить в арсенал инструментальных средств, которые используются на практике в различных областях науки и техники.

Клеточный автомат представляет собой дискретную динамическую систему, совокупность одинаковых клеток, одинаковым образом соединенных между собой. Все клетки образуют сеть (решетку) клеточных автоматов. Состояние каждой клетки определяются состоянием клеток, входящих в ее локальную окрестность и называемых ближайшими соседями. Окрестностью конечного автомата с номером j называется множество его ближайших соседей. Состояние j-го клеточного автомата в момент времени $t+1$, таким образом, определяется следующим образом:

$$y_j(t+1) = F\left(y_j,\ O(j),\ t\right),$$

где F – некоторое правило, которое можно выразить, например, языком булевой алгебры. Во многих задачах считается, что сам элемент относится к своим ближайшим соседям, т.е. $y_j \in O(j)$, в этом случае формула упрощается: $y_j(t+1) = F\left(O(j),\ t\right)$. Клеточные автоматы в традиционном понимании удовлетворяют таким правилам:

- изменение значений всех клеток происходит одновременно (единица измерения - такт);
- сеть клеточных автоматов однородная, т.е. правила изменения состояний для всех клеток одинаковые;
- на клетку могут повлиять лишь клетки из ее локальной окрестности;
- множество состояний клетки конечное.

Теоретически клеточные автоматы могут иметь любую размерность, однако чаще всего рассматривают одномерные и двумерные системы клеточных автоматов.

Модель диффузии информации [9], которую будем рассматривать в дальнейшем, является двумерной, поэтому дальнейший формализм касается этого случая. В двумерном клеточном автомате решетка реализуется двумерным массивом. Поэтому в этом случае удобно перейти к двум индексам, что вполне корректно для двумерных конечных решеток.

В случае двумерной решетки, элементами которой являются квадраты, ближайшими соседями, входящими в окрестность элемента $y_{i,j}$, можно считать или только элементы, расположенные вверх-вниз и влево-вправо от него (так называемая окрестность фон Неймана: $y_{i-1,j}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j}$), либо добавленные к ним еще и диагональные элементы (окрестность Мура (G. Moore $y_{i-1,j-1}$, $y_{i-1,j}$, $y_{i-1,j+1}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j-1}$, $y_{i+1,j}$, $y_{i+1,j+1}$).

В модели Мура каждая клетка имеет восемь соседей. Для устранения краевых эффектов решетка топологически «сворачивается в тор», т.е. первая строка считается продолжением последней, а последняя – предшествующей первой. То же самое относится и к столбцам.

Это позволяет определять общее соотношение значения клетки на шаге $t+1$ по сравнению с шагом t:

$$y_{i,j}(t+1) = F\left(y_{i-1,j-1}(t), y_{i-1,j}(t), y_{i-1,j+1}(t), y_{i,j-1}(t), y_{i,j}(t), y_{i,j+1}(t), y_{i+1,j-1}(t), y_{i+1,j}(t), y_{i+1,j+1}(t)\right).$$

С. Вольфрам, классифицируя различные клеточные автоматы, выделил те, динамика которых существенным образом зависит от начального состояния. Подбирая различные начальные состояния, можно получать разнообразнейшие конфигурации и типы поведения. Именно к таким системам относится классический пример - игра "Жизнь", изобретенная Дж. Конвеем (J. Conway) и известная широкому кругу читателей благодаря публикации в книге М. Гарднера (M. Gardner) [10].

Клеточные автоматы с успехом применяются при моделировании процессов распространения инноваций [11]. Клеточные автоматы также используются при моделировании электоральных процессов, в этом случае предполагается, что избирательные преференции человека определяются установками его ближайшего окружения.

В одной из моделей предполагается, что индивид принимает решение голосовать в момент $t+1$ за республиканцев или демократов в соответствии с правилом простого большинства. В этой модели учитывались взгляды индивида и четырех его ближайших соседей в момент t (окрестность фон Неймана). Модель исследовалась на большом временном отрезке - до 20 000 тактов. Оказалось, что партийная борьба приводит к очень сложным конфигурациям, которые существенным образом зависят от исходного распределения.

Как упрощенную модель диффузии информации сначала рассмотрим признанную модель распространения инноваций [11].

Подобная модель функционирует по следующим правилам: каждый индивид, который способен принять инновацию, соответствует одной квадратной клетке на двумерной плоскости. Каждая клетка может находиться в двух состояниях: 1 - новинка принята; 0 - новинка не принята. Предполагается, что автомат, восприняв инновацию один раз, запоминает ее навсегда (состояние 1, которое не может быть измененным). Автомат одобряет решение относительно принятии новинки, ориентируясь на мнение восьми ближайших соседей, т.е. если в окрестности данной клетки (используется окрестность Мура) есть m приверженцев новинки, p - вероятность ее принятия (генерируется в ходе работы модели) и если $pm > R,$ (R - фиксированное предельное значение), то клетка принимает инновацию (значение 1). По мнению авторов этой модели, клеточное моделирование позволяет строить значительно более реалистические модели рынка инноваций, чем традиционные подходы.

Вместе с тем динамике распространения информации присущи некоторые дополнительные свойства, которые были учтены в представленной ниже модели. В модели диффузии информации, наряду с теми же условиями, которые относятся к клеточному пространству, окрестности Мура и вероятностному правилу принятия новости, дополнительно к у условиям диффузии инноваций предполагалось, что клетка может быть в одном из трех состояний: 1 - «свежая новость» (клетка окрашивается в черный цвет); 2 - новость, которая устарела, но сохраненная в виде сведений (серая клетка); 3 - клетка не имеет информации, переданной новостным сообщением (клетка белая, информация не дошла или уже забыта). В модели приняты такие правила распространения сообщений:

- сначала все поле состоит из белых клеток за исключением одной, черной, которая первой «приняла» новость (рис. 6.3 а);
- белая клетка может перекрашиваться только в черный цвет или оставаться белой (она может получать новость или оставаться «в неведении»);

- белая клетка перекрашивается, если выполняется условие, аналогичное модели диффузии инноваций: $pm > 1$ (это условие несколько модифицируется для $m \le 2: \ 1.5 \cdot pm > 1$);
- если клетка черная, а вокруг нее исключительно черные и серые, то она перекрашивается в серые цвета (новость устаревает, но сохраняется как сведения);
- если клетка серая, а вокруг нее исключительно серые и черные, то она перекрашивается в белый цвет (происходит старение новости при ее общеизвестности).

Описанная система клеточных автоматов вполне реалистично отражает процесс распространения сообщений среди отдельных информационных источников. Авторами были реализован приведенный выше алгоритм на поле размером 40 х 40 (размеры были выбраны исключительно с целью наглядности). Выяснилось, что состояние системы клеточных автоматов полностью стабилизируется за ограниченное количество ходов, т.е. процесс эволюции оказался сходящимся. Пример работы модели приведен на рис. 6.3.

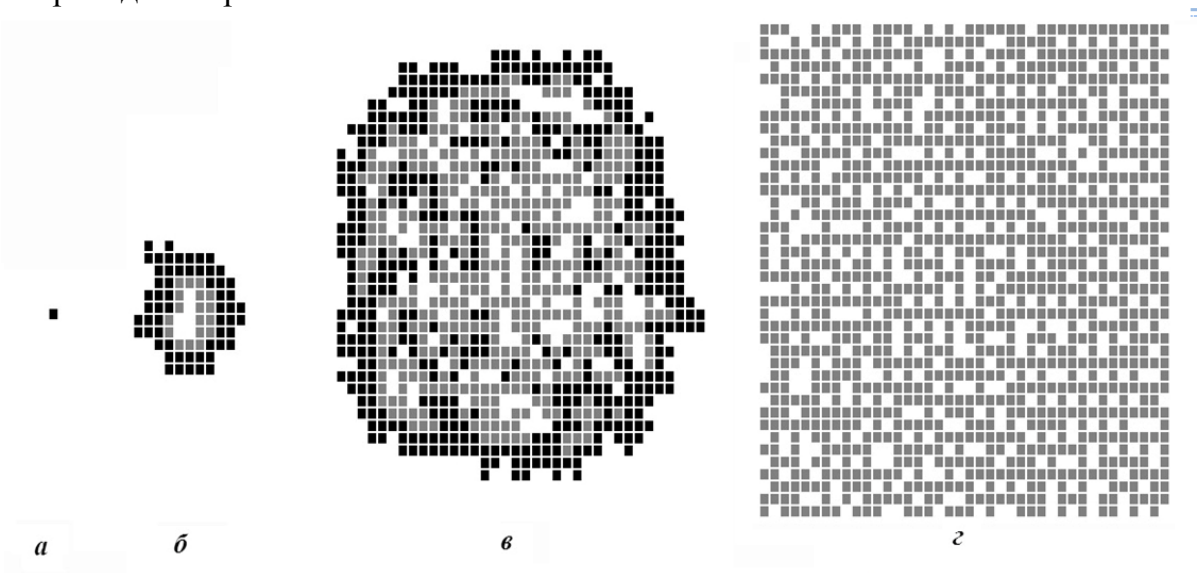


Рис. Часть VI.3. Процесс эволюции системы клеточных автоматов «диффузии новостей»: а - исходное состояние; б-в - промежуточные состояния; г - конечное состояние

Типичные зависимости количества клеток, которые находятся в разных состояниях в зависимости от шага итерации приведены на рис. 6.4. При анализе приведенных графиков следует обратить внимание на такие особенности: 1 - суммарное количество клеток, которые находятся во всех трех состояниях на каждом шагу итерации постоянно и равно количеству клеток, 2 - при стабилизации клеточных автоматов соотношения серых, белых и черных клеток приблизительно составляет: 3:1:0; существует точка пересечения всех трех кривых.

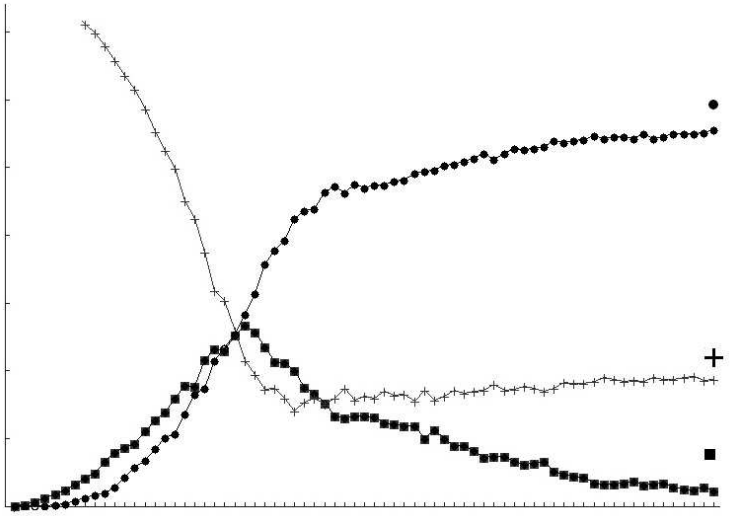


Рис. Часть VI.4. Количество клеток каждого из цветов в зависимости от шага эволюции: белые клетки - (+); серые клетки - (•); черные клетки – (■)

Детальный анализ полученных зависимостей позволил провести аналогии данной модели «диффузии информации» с некоторыми аналитическими соображениями. Результаты моделирования дают основания предположить, что эволюция серых клеток описывается некоторой непрерывной функцией:

$$x_g = f\left(t, \tau_g, \gamma_g\right),$$

где t - время (шаг эволюции), τ_g - сдвиг по времени, обеспечивающий получение необходимого фрагмента аналитической функции, γ_g - параметр крутизны данной функции. Соответственно, динамика белых клеток x_w (количество клеток в момент t) может моделироваться «перевернутой» функцией x_g с аналогичными параметрами:

$$x_w = 1 - f\left(t, \tau_w, \gamma_w\right).$$

Поскольку, как было сказано выше, всегда выполняется условие баланса, т.е. общее количество клеток в любой момент времени всегда постоянно, то условие нормирования можно записать следующим образом:

$$x_g + x_w + x_b = 1,$$

где x_w - количество черных клеток в момент времени t.

Таким образом, получаем:

$$x_b = 1 - x_g - x_w = f\left(t, \tau_w, \gamma_w\right) - f\left(t, \tau_g, \gamma_g\right).$$

Вид представленной на рис. 55 зависимости позволяет предположить, что в качестве функции $f\left(t, \tau, \gamma\right)$ может быть выбрано следующее выражение (логистическая функция):

$$f\left(t, \tau, \gamma\right) = \frac{C}{1 + e^{\gamma(t-\tau)}},$$

где C - некоторая нормирующая константа.

Следует отметить, что зависимость диффузии новостей, полученная в результате моделирования, хорошо согласуется с «жизненным» поведением тематических информационных потоков на интернет-источниках (веб-сайтах), а на локальных временных промежутках - с традиционными моделями.

Глава 2. Самоподобие в информационном пространстве

§ 2.1. Ранговые распределения в лингвистике

Поясним понятие рангового распределения из в теории вероятностей применительно к лингвистическим задачам [12].

Пусть T – множество всех слов некоторого текста, а $V=\{x\}$ – множество различных слов в этом тексте. Под $F(x)$ понимается число вхождений слова x в текст T. Таким образом каждому слову $x\in V$ соответствует подмножество $T(x)$ всех вхождений этого слова в текст T. Очевидно, что $F(x)=|T(x)|$.

Через $L=|T|$ мы обозначим длину (объем) текста, а через $N=|V|$ – объем его словаря.

Перенумеруем элементы словаря $V=\{x_1,x_2,...,x_N\}$ так, чтобы $F(x)$ была невозрастающей функцией его номера:

$$F(x_1)>F(x_2)>...>F(x_N). \qquad (6.1)$$

Ранговым распределением называется функция $\Phi(n)=F(x_n)$, которая ставит в соответствие номеру n слова $x\in V$ значение $F(x)$ этого слова. Итак,

$$\Phi:\Im\to\Re,$$

где $\Im$ – отрезок натурального ряда, а $\Re$ – множество положительных вещественных чисел.

Итак, в качестве T выступает случайным образом сформированная совокупность текстов и отрывков из текстов общей длиной в L слов, а в качестве V – список всех различных слов, обнаруженный в этой совокупности текстов. Величину $f(x)=F(x)/L$ интерпретируют как относительную частоту употребления слова x в данной выборке T, а под генеральной совокупностью понимают язык или определенный стиль данного текста.

При традиционном рассмотрении статистических языковых моделей предполагается, что каждое слово $x\in V$ имеет в языке объективную вероятность появления $f(x)$. Тогда, если сами слова упорядочить по убыванию вероятностей, то можно говорить о вероятности $f_n=f(x_n)$ появления слова ранга n. Для этих вероятностей постулируется существование теоретического закона распределения, (закона Ципфа, см. ниже):

$$f_n=\frac{c}{n^\gamma}=\frac{F_{\max}}{n^\gamma}, \qquad (6.2)$$

либо в виде закона Ципфа – Мандельброта

$$f_n=\frac{c}{(n+a)^\gamma}=\frac{F_{\max}}{n^\gamma}, \qquad (6.2')$$

где n – ранг слова x; a, γ, c – константы, удовлетворяющие условию нормировки:

$$c\sum_{n=1}^{\infty}\frac{1}{(n+a)^\gamma}=1 \qquad (6.3)$$

f_n в данном случае является теоретической оценкой значения наблюдаемой частоты n-ого по рангу слова в выборке T длиной в L слов. При этом предполагается высокая вероятность выполнения серии неравенств:

$$\left| f_n L - F_n \right| < \varepsilon,$$

при удовлетворяющих исследователя значениях ε.

В монографии Хердана [13] показано, что в различных текстах одни и те же слова имеют существенно различные ранги, а общим является только вид закономерности (6.2).

В книге [14] излагается следующая концепция текста. Предполагается, что существует вероятность появления определенного знака (буквы, слога, слова) после группы из k знаков. Тогда можно говорить, что порождение данного текста "разыгрывается" в зависимости от накопившейся предыстории. Каждый текст приобретает при этом определенную вероятность, а совокупность реализаций эргодического марковского процесса как раз и дает нам искомый однородный статистический ансамбль. Именно благодаря свойству эргодичности имеет смысл говорить и о вероятности того, что данный знак (элемент словаря V) появляется в данном месте текста.

В действительности и в этом подходе существенна только идея стохастического порождения текста, идея поиска механизма, гарантирующего статистическую однородность ансамбля текстов.

Таким образом, важен момент, который не учитываются или не объясняются в рамках традиционного подхода. В самом естественном языке устойчивость частот слов (существование ансамбля статистически однородных текстов) вызывает сомнение. Любой целостный текст обладает индивидуальностью. Попытка найти реальные статистически однородные ансамбли текстов никому еще не удавалось. Точнее говоря, не удавалось наблюдать такой набор текстов, в которых слова встречались с одинаковым спектром частот. В то же время словник любого текста, который по разумным содержательным соображениям удается считать замкнутым, можно упорядочить, так, что для частот достаточно хорошо выполняется соотношение (6.2). При традиционном подходе основным понятием является вероятность f_n появления в тексте слова с рангом n, а понятие текста вводится как вторичное, как нечто, случайным образом порождаемое из элементов, имеющих данные вероятности.

Если, наоборот, рассмотреть понятие "элемент текста" как вторичное, то существенным оказывается именно понятие рангового распределения. При изучении ранговых распределений устойчивыми являются лишь общие свойства формы распределения в целом; место же в этом распределении отдельных элементов текста окказионально и не может быть объектом прогноза (что вполне соответствует действительности). Можно предсказывать форму рангового распределения в тексте, некоторые свойства гармонических отношений между его лексическими и синтаксическими элементами, но практически невозможно сколько-нибудь достаточно предсказать частоту появления в будущем тексте каких-либо определенных слов или конструкций.

Словарь текста V можно интерпретировать как конечное разбиение X множества T, содержащее N классов эквивалентности $x_1, x_2, ..., x_N$, $T = \bigcup_{i=1}^{N} x_i$. Иной способ описаний заключается в том, что задается отображение φ множества T на множество V:

$$\varphi: T \rightarrow V$$

тогда $F(x) = |\varphi^{-1}(x)|$ (под $\varphi(x)$ имеется в виду множество всех таких y, что $\varphi(y) = x$). Очевидно, что $\sum_{x \in V} F(x) = L$.

Обратимся теперь к понятию ранга. Располагая элементы словаря, $x \in V$, по убыванию величины $F(x)$, мы, вообще говоря, не определяем на V единственной нумерации слов, а именно: элементы словаря, имеющие одинаковую частоту, могут произвольно меняться местами. Обозначим через $M(F)$ множество элементов словаря, имеющих в слове V одну и ту же частоту, а через $\mu(F) = |\, M(F) \,|$ – число таких элементов.

Далее, пусть $\mu_1(F) = \sum_{f<F} \mu(f)$ – количество элементов словаря V, имеющих частоту, меньшую F, а $\mu_2(F) = \sum_{f \geq F} \mu(f)$ – количество элементов словаря, имеющих частоту, большую или равную F.

Очевидно, что $\mu_2(F) - \mu_1(F) = \mu(F)$.

Тогда $m_F = [\mu_1(F), \mu_2(F)]$ – ранговый интервал, соответствующий множеству $|\, M(F) \,|$ элементов словаря с частотой F. На самом деле, ранговое распределение состоит не в выполнении формулы (6.2) или какой-либо другой аналогичной ей, а в том, чтобы каждому элементу $x \in V$ сопоставлялся ранговый определенный интервал.

Ниже будут оподробно бсуждаться параметры некоторых распределений, присущих многим информационным процессам, с учетом которых можно строить модели одновременно в рамках теории информационного поиска и концепции сложных сетей.

Распределение Парето

Анализируя общественные процессы, В. Парето (V. Pareto) рассмотрел социальную среду как пирамиду, на вершине которой находятся люди, представляющие элиту. Парето в 1906 году установил, что около 80 процентов земли в Италии принадлежит лишь 20 процентам ее жителей. Он пришел к заключению, что параметры полученного им распределения приблизительно одинаковы и принципиально не различаются в разных странах и в разное время. Парето также установил, что точно такая же закономерность наблюдается и в распределении доходов между людьми, которое описывается уравнением $N = A/X^p$, где X – величина дохода, N - количество людей с доходом, равным или превышающим X, A и p - параметры распределения. В математической статистике это распределение получило имя Парето, при этом предполагаются естественные ограничения на параметры: $X \geq 1,\ p > 1$. Распределению Парето присуще свойство устойчивости, т.е. сумма двух случайных переменных, которые имеют распределение Парето, также будет распределена по Парето.

Перейдем к более строгой формулировке закона Парето. Предположим, что последовательность $x_1, x_2, \ldots x_n, \ldots$ соответствует размерам доходов отдельных людей. После ранжирования этой последовательности по убыванию получается новая последовательность $x_{(1)}, x_{(2)}, \ldots, x_{(r)}, \ldots$ (элементы $x_{(r)}$ расположены в порядке убывания).

Предположим, что N - общее число людей, у которых доход составляет не менее $x_{(r)}$, т.е. $N = r$. Тогда правило Парето можно переписать в таком виде:

$$r = \frac{A}{x_{(r)}^{\,p}}.$$

Отсюда:

$$x_{(r)} = \left(\frac{A}{r}\right)^{\frac{1}{p}}$$

Рассматривается сумма первых n ($n = 1, 2, ..., N$) значений величины $x_{(r)}$, то есть общая величина дохода наиболее богатых людей - $m(n)$ составляет:

$$m(n) = \sum_{r=1}^{n} x_{(r)} = \sum_{r=1}^{n} \left(\frac{A}{r}\right)^{1/p} = \sum_{r=1}^{n} \frac{C}{r^{g}},$$

где $g = 1 - 1/p;\ C = A^{1/p}$.

Переходя от дискретных величин к непрерывным (предполагая, что $n >> 1$), имеем:

$$m(n) \approx \int_{1}^{n} \frac{C}{r^{g}} dr \approx \frac{C}{1-g} n^{1-g}.$$

В безразмерных переменных $m = m(n)/m(N)$ - и $n = n/N$ последнее равенство имеет вид (см. рис. 26):

$$m = n^{1-g}.$$

Величина m - в нашем примере - относительное количество дохода, получаемого первыми по рангу n людьми, доля которых (относительно всех людей) равна n.

Для $\nu \approx 0.2$ справедливо $\mu \approx 0.8$, т.е., действительно, 20% людей имеют 80% доходов.

Законы Ципфа

Дж. Ципф (G. Zipf) изучал использование статистических свойств языка в текстовых документах и выявил несколько эмпирических законов, которые представил как эмпирическое доказательство своего «принципа наименьшего количества усилий». Он экспериментально показал, что распределение слов естественного языка подчиняется закону, который часто называют первым законом Ципфа, относящимся к распределению частоты слов в тексте. Этот закон можно сформулировать таким образом. Если для какого-нибудь довольно большого текста составить список всех слов, которые встретились в нем, а потом ранжировать эти слова в порядке убывания частоты их появления в тексте, то для любого слова произведение его ранга и частоты появления будет величиной постоянной: $f \cdot r = c$, где f - частота встречаемости слова в тексте; r - ранг слова в списке; c - эмпирическая постоянная величина (коэффициент Ципфа). Для славянских языков, в частности, коэффициент Ципфа составляет приблизительно 0,06-0,07.

Приведенная зависимость отражает тот факт, что существует небольшой словарь, который составляет большую часть слов текста. Это главным образом служебные слова. Например, приведенный в [15] анализ романа «Том Сойер», позволил выделить 11.000 английских слов. При этом было обнаружено двенадцать слов (the, and, и др.), каждое из которых охватывает более 1 % лексем в романе. Закон Ципфа был многократно проверен на многих массивах. Ципф объяснял

приведенное выше гиперболическое распределение «принципом наименьшего количества усилий» предполагая что при создании текста меньше усилий уходит на повторение некоторых слов, чем на использование новых, т.е. на обращение к «оперативной памяти, а не к долговременной».

Ципф сформулировал еще одну закономерность, так называемый второй закон Ципфа, состоящий в том, что частота и количество слов, которые входят в текст с данной частотой, также связанны подобным соотношением, а именно:

$$N(f) = \frac{B}{f^b},$$

где $N(f)$ - количество различных слов, каждое из которых используется в тексте f раз, B - константа нормирования.

Существует простая количественная модель определения зависимости частоты от ранга. Предположим, что генерируется случайный текст обезьяной на пишущей машинке. С вероятностью p генерируется пробел, а с вероятностью $(1-p)$ - другие символы, каждый из которых имеет равную вероятность. Показано, что полученный таким образом текст будет давать результаты, близкие по форме к распределению Ципфа. Эта модель была усовершенствована в соответствии с фактическими эмпирическими данными, когда вероятности генерации отдельных символов были заданы на основе анализа большого текстового массива [16]. Полученное соответствие не доказывает закона Ципфа, но вполне его объясняет с помощью простой модели.

Более сложную модель генерации случайного текста, удовлетворяющего второму закону Ципфа, предложил Г.А. Саймон (H.A. Simon) [17].

Условия этой модели достаточно просты: если текст достиг размера в n слов, тогда то, каким будет $(n+1)$-е слово текста определяется двумя допущениями:

1. Пусть $N(f,n)$ - количество разных слов, каждое из которых использовалось f раз среди первых n слов текста. Тогда вероятность того, что $(n+1)$-ым окажется слово, которое до того использовалось f раз пропорционально $f \cdot N(f,n)$ - общему количеству появления всех слов, каждое из которых до этого использовалось f раз.
2. С вероятностью d $(n+1)$-ым словом будет новое слово.

Распределение Ципфа часто искажается на практике ввиду недостаточных объемов текстовых корпусов, что приводит к проблеме оценки параметров статистических моделей. Вместе с тем соотношение между рангом и частотой была взята Солтоном в 1975 г. как отправная точка для выбора терминов для индексирования. Далее им рассматривалась идея сортировки слов в соответствии с их частотой в текстовом массиве. Как второй шаг высокочастотные слова могут быть устранены, потому что они не являются хорошими различительными признаками для отдельных документов из текстового массива. На третьем шаге термы с низкой частотой, определяемой некоторым порогом (например, слова, которые встречаются только единожды или дважды) удаляются, потому что они встречаются так нечасто, что редко используются в запросах пользователей. Используя этот подход, можно значительно уменьшить размер индекса поисковой системы. Более принципиальный подход к подбору индексных термов – учет их весовых значений. В весовых моделях среднечастотные термы оказываются самыми весомыми, так как они являются

наиболее существенными при отборе того или иного документа (наиболее частотные слова встречаются одновременно в большом количестве документов, а низкочастотные могут не входить в документы, интересующие пользователя).

Еще один эмпирический закон, сформулированный Ципфом состоит в том, что количество значений слова коррелирует с квадратным корнем его частоты. Подразумевалось, что нечасто используемые слова более однозначны, а это подтверждает то, что высокочастотные слова не подходят для внесения в индексы информационно-поисковых систем.

Ципф также определил, что длина слова обратно пропорциональна его частоте, что может быть легко проверено путем простого анализа списка служебных слов. Последний закон действительно служит примером принципа экономии усилий: более короткие слова требуют меньше усилий при воспроизведении, и таким образом, используются более часто. Этот «закон» можно подтвердить, рассматривая приведенную выше модель генерации слов обезьяной. Легко видеть, что вероятность генерации слова уменьшается с длиной, вероятность слова из n непробельных символов равна:

$$(1-p)^n \cdot p,$$

где p - вероятность генерации пробела.

Хотя закон Ципфа дает интересные общие характеристики слов в текстовых массивах, в общем случае замечены некоторые ограничения его применимости при получении статистических характеристик документальных массивов, состоящих из множества независимых документов разных авторов.

Законам Ципфа удовлетворяют не только слова из одного текста, но многие объекты современного информационного пространства.

Закономерность Бредфорда

Закономерность С. Бредфорда (S. Bredford), известного документалиста, одного из авторов универсальной десятичной классификации – УДК, состоит в следующем: если научные журналы расположить в порядке убывания числа помещенных в них статей по конкретному предмету, то полученный список можно разбить на три зоны таким образом, чтобы количество статей в каждой зоне по заданному предмету была одинаковой. Эти три зоны представляют: ядро - профильные журналы, непосредственно посвященные рассмотренной тематике, журналы, частично посвященные заданной области и журналы, тематика которых довольно далека от рассмотренного предмета. С. Бредфорд в 1934 г. установил следующее соотношение для количества журналов в разных зонах [18]:

$$\frac{N_3}{N_2} = \frac{N_2}{N_1} = const,$$

где количество журналов в первой зоне - N_1, во второй - N_2, в третьей - N_3.

Бредфорд вначале рассматривал найденную закономерность только как специфический случай распределения Ципфа для системы периодических изданий по науке и технике. Однако в дальнейшем оказалось, что эта же закономерность справедлива и для периодических изданий из многих других предметных областей [19], а также для наборов веб-сайтов, относящихся к некоторой выбранной тематике.

Закон Хипса

В компьютерной лингвистике эмпирический закон Г.С. Хипса (H.S. Heaps) связывает объем документа с объемом словаря уникальных слов, которые входят в этот документ [20]. Казалось бы, словарь уникальных слов должен насыщаться, а его объем стабилизироваться при увеличении объемов текста.

Оказывается, это не так! Для всех известных сегодня текстов в соответствии с законом Хипса, эти значения связаны соотношением (рис. 6.5):

$$v(n) = \alpha n^{\beta},$$

где $v-$ это объем словаря уникальных слов, составленный из текста, который состоит из n уникальных слов, α и β – определенные эмпирически параметры. Для европейских языков α принимает значение от 10 до 100, а β - от 0.4 до 0.6.

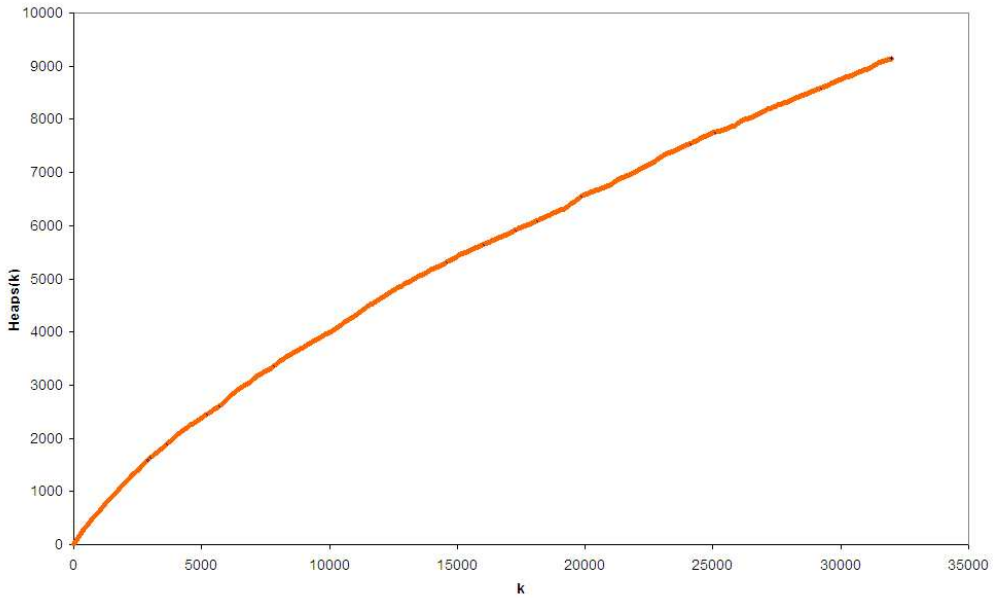


Рис. Часть VI.5. Типичный график, подтверждающий закон Хипса: по оси абсцисс – количество слов в тексте, по оси ординат – объем словаря – количество уникальных слов

Закон Хипса справедлив не только для уникальных слов, но и для многих других информационных объектов, что вполне естественно, так как уже доказано [21], что он является следствием закона Ципфа.

§ 2.2. Степенное распределение и самоподобие

Наиболее частыми (как обычно считается), универсальными законами распределения случайных величин, встречаемыми в различных естественнонаучных исследованиях, является нормальный закон – распределение Гаусса:

$$f(x) = \frac{1}{s\sqrt{2p}} e^{-\frac{x^2}{2s^2}}, \quad x > 0$$

Частая встречаемость нормального закона объясняется тем, что когда случайная величина является суммой независимых случайных величин, то ее распределение приближается к нормальному. Именно это утверждение является содержанием так называемой центральной предельной теоремы теории вероятностей. Заметим, что

часто в конкретных исследованиях гауссово распределение случайной величины принимается в силу привычки или удобства.

Б. Мандельброт был одним из первых, кто обратил пристальное внимание на то, что не менее универсальным, часто встречаемым законом распределения случайной величины является степенное (часто говорят гиперболическое) распределение с плотностью вероятности:

$$f(x)= \frac{B}{x^b},$$

или

$$P(X \text{ i } x)= \frac{A}{x^a}, \quad 0< x< \Gamma, \quad a = b - 1,$$

где $P(X \text{ i } x)$ – вероятность того, что $X \text{ i } x$, а A и a - некоторые положительные константы, параметры распределения.

Следует отметить, что приведенное выше распределение рассматривалось Б. Мандельбротом (B. Mandelbrot) как уточнение закона Ципфа и его часто называют распределением Ципфа-Мандельброта. При этом оказалось, что α - близкая к единице величина, которая может изменяться в зависимости от свойств текста и языка. Соответственно,

$$P(X \text{ i } x)= \int_x^{\Gamma} \frac{B}{x^b} dx, \qquad \frac{\P P(x)}{\P x} = - f(x).$$

Напомним, гиперболическое распределение A/x названо в честь В. Парето, а дискретный закон распределения с ранжированной переменной был назван в честь Д. Ципфа, который сформулировал его для описания частоты употребления слов.

Самоподобие или инвариантность относительно изменений масштаба или размера являет собой отличительную черту многих законов природы и бесчисленных явлений в мире, мы которого окружающих. «Самоподобие является в действительности одной из решающих симметрий, которая формирует нашу вселенную и оказывает влияние на наши попытки ее понять» [22].

Самоподобие информационного пространства выражается, в первую очередь в том, что при бурном росте этого пространства в последние десятилетия, гиперболичные частотные и ранговые распределения, получаемые в таких содержательных разрезах, как, например, источники и авторы документов, практически не изменяют свою форму. Закономерности, открытые такими учеными, как Зипф, Брэдфорд, Лотки и другие, в полной мере свидетельствуют о самоподобии информационного пространства. С другой стороны, самоподобие (скейлинг) можно рассматривать и как следствие общих структурных закономерностей информационного пространства.

Явление, которое имеет свойство самогоподобия, выглядит одинаково или одинаково себя ведет при его рассмотрении с разной степенью «увеличения» или в разном масштабе. Масштабирующей величиной может быть пространство (длина, ширина) или время. Рассматриваются, в частности, временные ряды, которые демонстрируют свойство самогоподобия.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует, например, интерфейс, представленный на веб-сайте службы News Is

Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. При этом учитывается два основных параметра отображения - ранг популярности и оперативность информации.

Четкое определения самоподобного стохастичекского процесса используется при прямом масштабировании непрерывной переменной времени.

Известно, что для последовательности сообщений тематических информационных потоков в соответствии со скейлинговым принципом, количество сообщений, резонансов на событии реального мира пропорционально некоторой степени количества источников информации (кластеров) и итерационно длится в течение определенного времени. Так же, как и в традиционных научных коммуникациях, множество сообщений в Интернет по одной тематике во времени являет собой динамическую кластерную систему, которая возникает в результате итерационных процессов. Этот процесс порождается републикациями, цитированием, разными публикациями – отражением одних и тех же событий реального мира, прямыми ссылками и тому подобное.

Если рассматривать информационные потоки как ряды публикаций в течение времени, то обнаруживается наличие таких свойств, как самоподобие (масштабная инвариантость, скейлинг), устойчивые взаимные корреляции. Анализ самоподобия информационных массивов может рассматриваться как технология, предназначенная для осуществления аналитических исследований с элементами прогнозирования, которая пригодна к экстраполяции полученных зависимостей.

Стохастический процесс $X(t)$ является стохастически самоподобным с параметром H $(0{,}5 \le H \le 1)$, если для любого действительного значения $a > 0$ процесс $a^{-H} X(at)$ имеет те же самые статистические характеристики, что и сам процесс $X(t)$. Это утверждение можно выразить тремя условиями [23]:

- среднее:
$$E\left[X(t)\right] = \frac{E\left[X(at)\right]}{a^H};$$
- дисперсия:
$$\sigma\left[X(t)\right] = \frac{\sigma\left[X(at)\right]}{a^{2H}};$$
- автокорреляция:
$$K\left[X(t), X(s)\right] = \frac{K\left[aX(t), aX(s)\right]}{a^{2H}}.$$

Параметр H, называемый параметром Херста (*Hurst parametr*) или параметром сомоподобия (*self-similarity parametr*) , представляет собой ключевую меру самоподобия. Точнее, H представляет меру устойчивости статистического явления, или меру действия долговременной зависимости статистического процесса. Значение $H = 0{,}5$ указывает на отсутствие долговременной зависимости. Чем ближе значение H к 1, тем выше степень устойчивости долговременной зависимости.

Рассмотрим для примера процесс броуновского движения $B(t)$ и докажем его самоподобие с параметром $H = 0{,}5$ в соответствии с приведенным выше определением. Рассмотрим три условия самоподобия:

– по определению, $E[B(t)]$=0. Тому $E[B(t)] = E[B(at)]/a^{0,5}$,

что удовлетворяет первому условию;

– известно, что дисперсия $\sigma[B(t)]$ равна t, поэтому $\sigma[B(at)] = at = a\sigma[B(t)]$, что удовлетворяет второму условию;

– загальновідомо, що автокореляція $K[B(t), B(s)] = \min[t, s]$. Отсюда: $K[B(at), B(as)] = \min[at, as] = a\min[t, s] = aK[B(t), B(s)]$, что удовлетворяет третьему условию.

Далее рассмотрим случай стохастического процесса, определенного в дискретных точках времени, так что стохастический процесс $X(t)$ определяется как $\{x_t,\ t = 0,1,2,...\}$. Для таких процессов определяются $m-$агрегированных временных серий $\left\{x_k^{(m)},\ k = 0,1,2,...\right\}$, получаемые в результате суммирования значений исходных серий в непересекающихся соседних блоках размером m элементов. Это может быть выражено таким образом:

$$x_k^{(m)} = \frac{1}{m}\sum_{i=km-m+1}^{km} x_i .$$

Агрегированные временные серии можно рассматривать как метод сжатия временной шкалы. При этом $x^{(1)}$ может считаться максимальным увеличением или наивысшей разрешающей способностью для этой временной серии. Процесс $x^{(5)}$, например, представляет собой тот же самый процесс, уменьшенный в пять раз. Если статистические характеристики процесса совпадают при сжатии, то можо считать, что идет речь о самоподобном процессе.

Таким образом, можно предложить функциональное определение самоподобия, а именно:

процесс x называется точно самоподобным (*exactly self-similar*) с параметром β $(0 < \beta < 1)$, если для всех $m = 1,2,...$ выполняется:

– для дисперсии:

$$\sigma\left[x^{(m)}\right] = \frac{\sigma[x]}{m^{\beta}};$$

– автокорреляция:

$$K\left[x^{(m)}, k\right] = K[x, k].$$

Можно показать, что параметр β связан с определенным ранее параметром Херста соотношением: $H = 1 - (\beta/2)$. Для стационарного эргодического процесса $\beta = 1$, а средняя дисперсия со временем стремится к нулю со скоростью $1/m$. Для самоподобного процесса средняя дисперсия времени затухает более медленно.

Вышеприведенное определение позволяет реализовать самый простой алгоритм определения того, является ли временная серия самподобной.

Если прологарифмировать вышеприведенную формулу для дисперсии, получаем:

$$\log\left(\sigma\left[x^{(m)}\right]\right) = \log\left(\sigma[x]\right) - \beta\log m .$$

Поскольку $\log\left(\sigma\left[x\right]\right)$ является монотонной константой, которая не зависит от m, то график зависимости $\log\left(\sigma\left[x^{(m)}\right]\right)$ от m в логарифмическом масштабе будет представлять собой прямую линию с наклоном, равным $-\beta$.

График можно построить (конечно, для фактических данных следует использовать выборочную дисперсию вместо теоретической), если сгенерировать процесс на разных уровнях агрегации m, а после этого вычислить дисперсию. Обычно временные ряды, которые формируются из объемов тематических информационных потоков, ложатся на прямую линию с отрицательным наклоном. В этих случаях обычно определяют значение параметра H.

Другой концепцией, связанной с самим подобием, являются медленно затухающие распределения, или распределения с "тяжелыми хвостами" (heavy - tailed distributions). Медленно затухающие распределения могут использоваться для представления плотности вероятностей, которые описывают, например, объемы данных в информационных потоках. Известно, что распределение случайной переменной X медленно затухает, если:

$$1-F(x)=\Pr\left[X>x\right]\sim\frac{1}{x^{\alpha}} \text{ при } x\to\infty,\ \ 0<\alpha.$$

В целом, случайная переменная с медленно затухающим распределением имеет бесконечную дисперсию и, возможно, бесконечное среднее. Случайная переменная с медленно затухающим распределением может принимать очень большие значения с вероятностью, которой невозможно пренебречь.

Самим простым медленно затухающим распределением является распределение Парето с параметрами k и α $(k,\ \alpha<0)$ и такими статистическими показателями:

$$f(x)=F(x)=0\ \ (x\le k);$$

$$f(x)=\frac{\alpha}{k}\left(\frac{k}{x}\right)^{\alpha+1};$$

$$F(x)=1-\left(\frac{k}{x}\right)^{\alpha}\ (x>k;\ \ \alpha>0);$$

$$E\left[x\right]=\frac{\alpha}{\alpha-1}k\ \ (\alpha>1).$$

§ 2.3. Основы фрактального анализа информационных потоков

Многочисленные эксперименты, замеры параметров информационного пространства подтверждают тот факт, что при значительном возрастании объемов информационных ресурсов статистические распределения документов, получаемые в самых разнообразных содержательных разрезах (таких, например, как источники, авторы, тематики) практически не меняют своей формы.

Применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, которые составляют основы информатики. Известно, что многие информационно-поисковые системы, включающие элементы кластерного анализа, позволяют автоматически обнаруживать

новые классы и распределяют документы по этим классам. Соответственно, показано, что тематические информационные массивы представляют собой самоподобные развивающиеся структуры, однако их самоподобие справедливо лишь на статистическом уровне (например, распределение тематических кластеров документов по размерам).

Чем же определяется природа фрактальных свойств информационного пространства, порождаемого такими кластерными структурами? С одной стороны, параметрами ранговых распределений, а с другой стороны, механизмом развития информационных кластеров. Появление новых публикаций увеличивает размеры уже существующих кластеров и является причиной образования новых.

Фрактальные свойства характерны и для кластеров информационных веб-сайтов, на которых публикуются документы, соответствующие определенным тематикам. Эти кластеры, как наборы тематических документов, представляют собой структуры, обладающие рядом уникальных свойств.

Топология и характеристики моделей веб-пространства оказываются приблизительно одинаковыми его разных подмножеств, подтверждая тем самым наблюдение о том, что «веб - это фрактал».

Как показано в роботах С. Иванова [24], для последовательности сообщений тематических информационных потоков количество сообщений, резонансов на события реального мира, пропорционально некоторой степени количества источников информации (кластеров).

Известно, что все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Ципфа, могут быть обобщенные именно в рамках теории стохастических фракталов. Точно так же, как и в традиционных научных коммуникациях, множество сообщений в Интернете по одной тематике во времени представляет собой динамическую кластерную систему, которая возникает в результате итерационных процессов. Этот процесс обуславливается републикациями, односторонним или взаимным цитированием, различными публикациями - отражениями одних и тех же событий реального мира, прямыми ссылками и т.п.

Фрактальная размерность в кластерной системе, которая соответствует тематическим информационным потокам, показывает уровень заполнения информационного пространства сообщениями на протяжении определенного времени [24]:

$$N_{publ}(\varepsilon t) = \varepsilon^{\rho} N_k(t)^{\rho},$$

где N_{publ} - размер системы (общее количество сообщений в информационном потоке); N_k – размер – число кластеров (тематик или источников); ρ – фрактальная размерность информационного массива; ε – коэффициент масштабирования. В приведенном соотношении между количеством сообщений и кластеров проявляется свойство сохранения внутренней структуры множества при изменении масштабов его внешнего рассмотрения.

Изучение характеристик временных рядов, порождаемых информационными потоками, сообщения которых отражают процессы, происходящие в реальном мире, дает возможность прогнозировать их динамику, выявлять скрытые корреляции, циклы и т.п.

В этом разделе будут описаны основные алгоритмы, применяемые при исследовании фрактальных свойств рядов измерений. В качестве иллюстраций приведены результаты реальных численных экспериментов. Как база для

исследования фрактальных свойств рядов, отражающих интенсивность публикаций тематических информационных потоков, использовалась система контент-мониторинга новостей с веб-сайтов сети Интернет InfoStream. Тематика исследуемого информационного потока определялась запросом к этой системе. Данные для исследований были получены из интерфейса режима «Динамика появления понятий».

В ходе исследований обрабатывался тематический информационный поток, содержащий сообщения онлайновых СМИ - массив из 14069 документов, опубликованных с 1 января 2006 г. по 31 декабря 2007 г., по тематике компьютерной вирусологии, удовлетворяющих запросу:

«компьютерный вирус» OR «вирусная атака» OR (антивирус AND (программа OR утилита OR Windows OR Linux)).

Ниже анализируетсятся временной ряд из количества тематических публикаций за указанный период с определенной дискретностью по времени в сутки .

Остановимся подробнее на некоторых методах анализа подобного типа временных рядов, порождаемых, в частности, информационными потоками.

Метод DFA

Один из универсальных подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) [25] – универсальном методе обработки рядов измерений. Метод DFA (Detrended fluctuation analysus) также чаще всего употребится для выявления статистического самоподобия сигналов.

Этот метод является вариантом дисперсионного анализа одномерных случайных блужданий и позволяет исследовать эффекты продолжительных корреляций в рядах, которые исследуются. В рамках алгоритма DFA анализируется среднеквадратичная ошибка линейной аппроксимации в зависимости от размера участка аппроксимации (окна наблюдения). Пусть есть ряд измерений x_t, $t \in 1,...,N$. Обозначим среднее значение этого ряда измерений: $\langle x \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. Из исходного ряда строится ряд накопления:

$$X_t = \sum_{k=1}^{t} \left(x_k - \langle x \rangle \right).$$

Потом ряд X_t разделяется на временные окна длиной L, строится линейная аппроксимация ($L_{j,L}$) по значениям $X_{k,j,L}$ с $X_{j,L}$ внутри каждого окна (в свою очередь, $X_{j,L}$ – подмножество X_t, $j = 1,...,J$, $J = N/L$ – количество окон наблюдения) и рассчитывается отклонение точек ряда накопления от линейной аппроксимации:

$$E(j,L) = \sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(X_{k,j,L} - L_{k,j,L} \right)^2} = \sqrt{\frac{1}{L}\sum_{k=1}^{L} |\Delta_{k,j,L}|^2},$$

где $L_{k,j,L}$ – значение локальной линейной аппроксимации в точке $t = (j-1)L + k$.

Здесь $|\Delta_{k,j,L}|$ – абсолютное отклонение элемента $X_{k,j,L}$ от локальной линейной аппроксимации.

Далее вычисляется среднее значение:

$$F(L) = \frac{1}{J}\sum_{j=1}^{J} E(j,L),$$

после чего, в случае $F(L) \propto L^{\alpha}$, где α некоторая константа, делаются выводы о наличии статистического самоподобия и характер поведения исследуемого ряда измерений.

Этот метод был применен к ряду значений количества публикаций, полученных за представленным выше запросом. На рис. 6.6 представленная зависимость среднеквадратичной ошибки аппроксимации от длины участков аппроксимации в двойном логарифмическом масштабе.

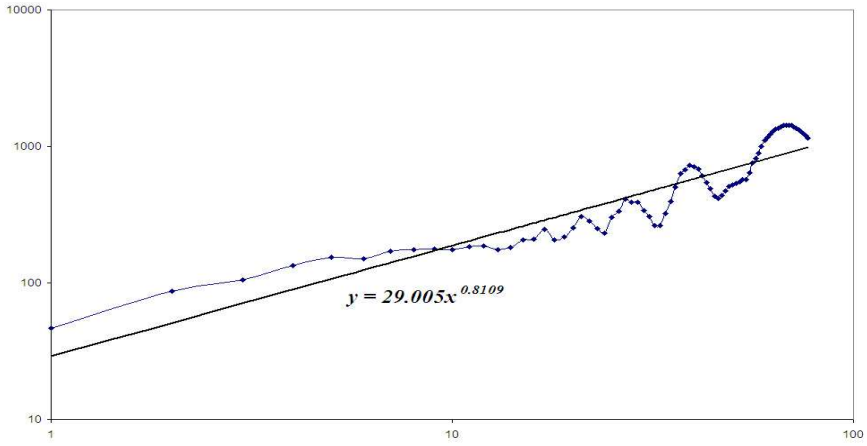


Рис. Часть VI.6. Зависимость среднеквадратичной ошибки линейной аппроксимации D от длины окна наблюдения k

Близость зависимости $D(k)$ к линейному еще раз подтверждает наличие локального скейлинга во втором полугодии 2008 года.

Визуализация на основе ΔL-анализа

С целью визуализации и анализа временных рядов, связанных с публикациями в информационном просторанстве сети Интернет разработан новый метод дисперсионного анализа, предназначенный для анализа и визуализации состояния временных рядов интенсивности публикаций по определенной тематике [26].

Задачи выявления и визуализации трендов, выявление гармонических составляющих, трендов, локальных особенностей временных рядов, фильтрации шума сегодня решаются методами фрактального, вейвлет- и Фурье-анализа.

Как и в методе DFA, рассмотрим поведения отклонения точек ряда накопления от линейной аппроксимации (но в этом случае абсолютное значение) $|\Delta_{k,j,L}|$. Построение соответствующих диаграмм значений $|\Delta_{k,j,L}|$, которые зависят фактически от двух параметров – L и $t = (j-1)L + k$ названо ΔL-методом визуализации. Такая визуализация в виде «рельефной» диаграммы представляет собой определенный интерес для изучения особенностей процессов, которые отвечают исходным рядам измерений.

ΔL-метод оказывается довольно эффективным для выявления гармонических составляющих исследуемого ряда. Применение ΔL-метода к ряду, составленному из количества публикаций, собранных системой InfoStream из Интернет без учета тематического деления, имеет явным образом выраженную гармоническую составляющую (общее количество публикаций зависит со дня недели), что можно видеть на рис. 6.7. Кроме того, на этой диаграмме заметные отклонения от общей динамики объемов публикаций в праздничные дни.

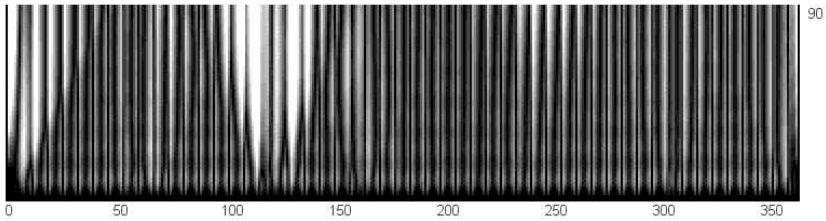


Рис. Часть VI.7. ΔL-диаграмма ряда количества публикаций, собираемых ежесуточно системой InfoStream в 2008 году

«Рельефные диаграммы», получаемые в результате ΔL-метода (более светлые тона соответствуют большим значениям $|\Delta_{k,j,L}|$), напоминают скейлограммы, полученные в результате непрерывных вейвлет-преобразований. Следует обратить внимание на то, что темные полосы в центре многих областей светлого закрашивания свидетельствуют об «стабилизации» больших значений рассмотренного ряда на высоком уровне.

ΔL-метод применяется для реальных временных рядов, например тех, которые отражают интенсивность публикаций данной тематики в Интернете. На рис. 6.8 приведена ΔL-диаграмма для рассмотренного выше временного ряда из количества публикаций сообщений через сутки по выбранной тематике в сети Интернет на протяжении года.

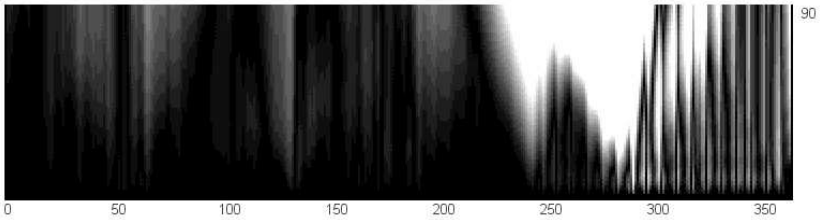


Рис. Часть VI.8. ΔL-диаграмма временного ряда интенсивности тематических публикаций (ось абсцисс – дни года, ось ординат – величина окна измерений)

На рис. 6.9 приведена ΔL-диаграмма наличного курса доллара в гривнах на протяжении 2008 года. Еще нагляднее, чем в случае применения вейвлет-анализа можно убедиться в потому, что наиболее значительные отклонения на диаграмме в этом случае наступают с некоторой временной задержкой у сравнение с диаграммой публикациями по кризисной тематике.

Предложенный метод визуализации абсолютных отклонений ΔL, как и метод вейвлет-преобразований, позволяет (и как показано на примере – не хуже) обнаруживать единичные и нерегулярные «всплески», резкие изменения значений количественных показателей в разные периоды времени.

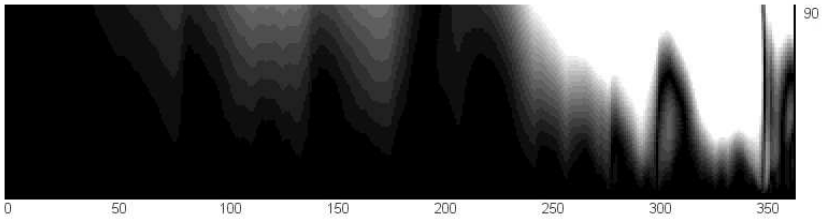


Рис. Часть VI.9. ΔL-диаграмма временного ряда значений наличных курсов доллара в гривнах (ось абсцисс – дни года, ось ординат – величина окна измерений)

Следует отметить, что метод вейвлет-преобразований может применяться с использованием разнообразных вейвлетов. В случае применения ΔL-метода не нужно решать сложную задачу выбора и обоснования применения соответствующего вейвлета; в отличие от методов фрактального анализа предложенный подход не требует значительных объемов точек ряда измерений. Этот метод довольно простой в программной реализации и базируется на такому мощной теоретической основе как DFA, оказался довольно эффективным при анализе временных рядов в таких областях, как экономика и социология.

Корреляциионный анализ

Если обозначить через X_t член ряда количества публикаций (количества электронных сообщений, поступивших, например, в день t, $t = 1, ..., N$), то функция автокорреляции для этого ряда X определяется как:

$$F(k) = \frac{1}{N-k}\sum_{t=1}^{N-k}(X_{k+t} - m)(X_t - m),$$

где m – среднее значение ряда X, которое в дальнейшем, не ограничивая общности, будем считать равным 0 (это достигается переприсвоением значению X_t значения X_t - m). Предполагается, что ряд X может содержать скрытую периодическую составляющую.

Известно, что функция автокорреляции обладает тем свойством, что если скрытая периодическая составляющая существует, то ее значение асимптотически приближается к квадрату среднего значения исходного ряда .

Если рассматриваемый ряд периодический, т.е. может быть представлен как:

$$X_t = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega t + \theta_n),$$

то его функция автокорреляции будет равна:

$$F(k) = \frac{a_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty} a_n^2 \cos n\omega k.$$

Этот результат показывает, что функция автокорреляции периодического ряда также является периодической, содержит основную частоту и гармоники, но без фазовых углов θ_n.

Рассмотрим числовой ряд X, являющийся суммой некоторой содержательной составляющей N и синусоидальной сигнала S:

$$X_t = N_t + S_t.$$

Найдем функцию автокорреляции для этого ряда (значения приведены к среднему $m = 0$):

$$F(k) = \frac{1}{N-k}\sum_{t=1}^{N-k} X_{k+t} X_t =$$

$$= \frac{1}{N-k}\sum_{t=1}^{N-k}(N_{k+t} + S_{k+t})(N_t + S_t) =$$

$$= \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t} N_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t} S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t} S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t} N_t.$$

Очевидно, первое слагаемое есть функция непериодическая, асимптотически стремящаяся к нулю. Так как взаимная корреляция между N и S отсутствует, то то третье и четвертое слагаемое также стремятся к нулю. Таким образом, самый значительный ненулевой вклад составляет второе слагаемое – автокорреляция сигнала S. Т.е. функция автокорреляции ряда X остается периодической.

Для экспериментального подтверждения рассмотренной гипотезы была сгенерирована последовательность, по своей природе напоминающая реальный информационный поток. Предполагалось, что ежедневное количество сообщений в сети растет по экспоненциальному закону (с очень небольшим значением экспоненциальной степени), и на это количество накладываются колебания, связанные с недельной цикличностью в работе информационных источников. Также принимается во внимание некоторый элемент случайности, выраженный соответствующими отклонениями.

Для получения соответствующего временного ряда были рассмотрены значения функции:

$$y = ae^{0.001x} + \sin(\pi x/7 + a),$$

которая реализует простейшую модель информационного потока – экспонента отвечает за рост количества публикаций во времени (общая тенденция), синус – за недельную периодичность, параметр a – за случайные отклонения. Количество публикаций y не может быть отрицательным числом. Исходный ряд был обработан: приведен к нулевому среднему и нормирован (каждый член разделен на среднее). После этого были рассчитаны коэффициенты корреляции, которые для рядов измерений X длиной N рассчитываются по формуле:

$$R(k) = \frac{F(k)}{\sigma^2},$$

где $F(k)$ – функция автокорреляции; σ^2 – дисперсия.

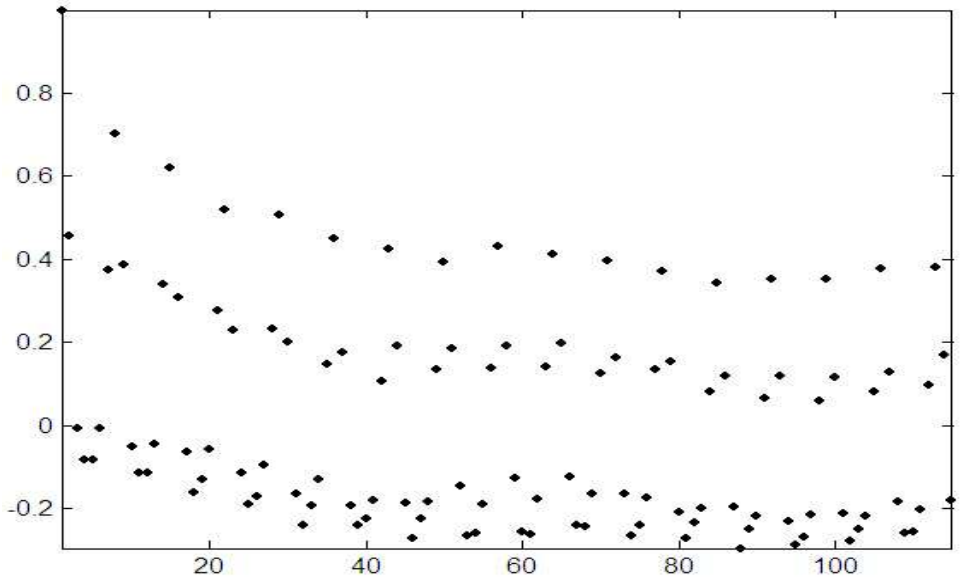


Рис. Часть VI.10. *Коэффициенты корреляции ряда наблюдений* $R(k)$ *(ось ординат) в зависимости от* k *(ось абсцисс)*

На рис. 6.10 приведены значения коэффициентов корреляций (ось абсцисс – переменная *k*, ось ординат – коэффициент корреляции *R(k)*. Коэффициенты

корреляции ряда наблюдений, усредненного по неделям, аппроксимируются гиперболической функцией, которая характеризует долгосрочную зависимость членов «урупненного» исходного ряда.

Фактор Фано

Для изучения поведения процессов принято использовать еще один показатель – индекс разброса дисперсии (IDC), так называемый фактор Фано (U. Fano). Эта величина определяется как отношение дисперсии количества событий (в нашем случае – количества публикаций) на заданном окне наблюдений k к соответствующему математическому ожиданию:

$$F(k) = \sigma^2(k)/m(k).$$

Для самоподобных процессов выполняется соотношение:

$$F(k) = 1 + Ck^{2H-1},$$

где C и H – константы.

Показатель Херста

Показатель Херста (H.E. Hurst) - H связан с коэффициентом нормированного размаха R/S, где R - вычисляемый определенным образом «размах» соответствующего временного ряда, а S - стандартное отклонение [27].

Г.Э. Херст экспериментально обнаружил, что для многих временных рядов справедливо: $R/S = (N/2)^H$. Эта закономерность связана с традиционной «клеточной» фрактальной размерностью D простым соотношением:

$$D = 2 - H.$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью в соответствии с приведенной формулой, определено Е. Федером следующим образом: «… рассматривают клетки, размеры которых малы по сравнению как с длительностью процесса, так и с диапазоном изменения функции; поэтому соотношение справедливо, когда структура кривой, описывающая фрактальную функцию, исследуется с высоким разрешением, т.е. в локальном пределе». Еще одним важным условием является самоаффинность функции. Не вдаваясь в подробности, заметим, что для информационных потоков это свойство интерпретируется как самоподобие, возникающее в результате процессов их формирования. Можно отметить, что указанными свойствами обладают не все информационные потоки, а лишь те, которые характеризуются достаточной мощностью и итеративностью при формировании. При этом временные ряды, построенные на основании мощных тематических информационных потоков, вполне удовлетворяют этому условию. Поэтому при расчете показателя Херста фактически определяется и такой показатель тематического информационного потока как фрактальная размерность.

Известно, что показатель Херста представляет собой меру персистентности - склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность — броуновское движение.

Для изучения фрактальных характеристик тематических информационных потоков за определенный период для временных рядов $F(n)$, $n=1, ..., N$, составленных из количества относящихся к ним сообщений, изучалось значение показателя Херста, которое определялось из соотношения:

$$R/S=(N/2)^H,\ \ N>>1.$$

Здесь S – стандартное отклонение:

$$S=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(F(n)-\langle F\rangle_N\right)^2},$$

$$\langle F\rangle_N=\frac{1}{N}\sum_{n=1}^{N}F(n),$$

а R - так называемый размах:

$$R(N)=\max_{1\le n\le N}X(n,N)-\min_{1\le n\le N}X(n,N),$$

где

$$X(n,N)=\sum_{i=1}^{n}\left(F(i)-\langle F\rangle_N\right).$$

Исследования фрактальных свойств рядов измерений, получаемых в результате мониторинга тематических информационных массивов из Интернет, свидетельствуют о том, что при увеличении n показатель H принимает значения 0.65 ÷ 0.75. Ввиду того, что значение H намного превышает ½, в этом ряду обнаруживается персистентность (существование долговременных корреляций, которые могут быть связаны с проявлением детерминированного хаоса). Если предположить, что ряд $F(n)$ является локально самоаффинным (этот вопрос в настоящее время открыт), то он имеет фрактальную размерность D, равную

$D = 2 – H \approx 1.35 \div 1.25$.

То есть, исследования тематических информационных потоков подтверждают предположение о самоподобии и итеративности процессов в веб-пространстве. Републикации, цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах.

В результате экспериментов было подтверждено наличие высокого уровня статистической корреляции в информационных потоках на продолжительных временных интервалах. На основе рассмотренного примера показана высокая персистентность процесса, что, в частности, свидетельствует об общей тенденции увеличения публикации по выбранной тематике.

Анализ самоподобия информационных массивов может рассматриваться как технология для осуществления прогнозирования.

Вейвлет-анализ

Основой вейвлет-анализа [28, 29] являются вейвлет-преобразование, представляющего собой особый тип линейного преобразования, базисные функции которого (вейвлеты) имеют специфические свойства.

Вейвлетом (малой волной) называется некоторая функция, сосредоточенная в небольшой окрестности некоторой точки и резко убывающая к нулю по мере удаления от нее как во временной, так и в частотной области. Существуют

разнообразные вейвлеты, имеющие разные свойства. Вместе с тем, все вейвлеты имеют вид коротких волновых пакетов с нулевым интегральным значением, локализованных на временной оси, являющихся инвариантными к сдвигу и масштабированию.

К любому вейвлету можно применить две операции:

- сдвиг, т.е. перемещение области его локализации во времени;
- масштабирование (растяжение или сжатие).

Главная идея вейвлет-преобразования заключается в том, что нестационарный временной ряд разделяется на отдельные промежутки (так называемые «окна наблюдения»), и на каждом из них выполняется вычисление скалярного произведения (величины, которая характеризует степень близости двух закономерностей) исследуемых данных с разными сдвигами некоторого вейвлета на разных масштабах. Вейвлет-преобразование генерирует набор коэффициентов, с помощью которых представляется исходный ряд. Они являются функциями двух переменных: времени и частоты, и потому образуют поверхность в трехмерном пространстве. Эти коэффициенты, показывают насколько поведение процесса в данной точке аналогично вейвлету на данном масштабе. Чем ближе вид анализируемой зависимости в окрестности данной точки к виду вейвлета, тем большую абсолютную величину имеет соответствующий коэффициент. Отрицательные коэффициенты показывают, что зависимость похожа на «зеркальное отражение» вейвлета. Использование этих операций, с учетом свойства локальности вейвлета в частотно-временной области, позволяет анализировать данные на разных масштабах и точно определять места их характерных особенностей во времени.

Технология использования вейвлетов позволяет обнаруживать единичные и нерегулярные «всплески», резкие изменения значений количественных показателей в разные периоды времени, в частности, объемов тематических публикаций в веб-пространстве. При этом могут обнаружиться моменты возникновения циклов, а также моменты, когда за периодами регулярной динамики следуют хаотические колебания.

Рассматриваемый временной ряд может аппроксимироваться кривой, которая, в свою очередь, может быть представленная в виде суммы гармонических колебаний разной частоты и амплитуды. При этом колебания, которые имеют низкую частоту, отвечают за медленные, плавные, крупномасштабные изменения значений исходного ряда, а высокочастотные – за короткие, мелкомасштабные изменения. Чем сильнее изменяется описываемая данной закономерностью величина при данном масштабе, тем большую амплитуду имеют составляющая соответствующей частоты. Таким образом, исследуемый временной ряд можно рассматривать в частотно-временной области – т.е. об исследовании закономерности, описывающей процесс в зависимости как от времени, так и от частоты.

Непрерывное вейвлет-преобразование для функции $f(t)$ строится с помощью непрерывных масштабных преобразований и переносов выбранного вейвлета $\psi(t)$ с произвольными значениями масштабного коэффициента a и параметра сдвига b:

$$W(a,b) = (f(t),\psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt .$$

Полученные коэффициенты представляются в графическом виде как карта коэффициентов преобразования, или скейлограмма. На скейлограмме по одной оси откладываются сдвиг вейвлета (ось времени), а по другой – масштабы (ось масштабов), после чего точки схемы, которая получается, раскрашиваются в

зависимости от величины соответствующих коэффициентов (чем больше коэффициент, тем ярче цвета изображения). На скейлограмме видные все характерные особенности исходного ряда: масштаб и интенсивность периодических изменений, направление и величина трендов, наличие, расположение и продолжительность локальных особенностей.

Например, известно, что комбинация нескольких разных колебаний может иметь настолько сложную форму, которая не позволяет аналитику выявить их. Периодические изменения, которые происходят для значений коэффициентов вейвлет-преобразования на некотором непрерывном множестве частот выглядят как цепочка «холмов», имеющие вершины, расположенные в точках (по оси времени), в которых эти изменения достигают наибольших значений.

Другим важным показателем является выраженная тенденция динамики временного ряда (тренд) вне зависимости от периодических колебаний. Наличие тренда может быть неочевидным при простом рассмотрении временного ряда, например, если тренд объединяется с периодическими колебаниями. Тренд отражается на скейлограмме как плавное изменение яркости вдоль оси времени одновременно на всех масштабах. Если тренд возрастающий, то яркость будет увеличиваться, если убывающий – уменьшаться.

Еще одним важным фактором, которому необходимо учитывать при анализе временных рядов, являются локальные особенности, т.е. возможные резкие, скачкообразные изменения характеристик исходного ряда. Локальные особенности представленные на скейлограмме вейвлет-преобразования как линии резкого перепада яркости, которые исходят из точки, соответствуют времени возникновения скачка. Локальные особенности могут иметь как случайный, так и систематический характер, при этом "маскировать" периодические зависимости или краткосрочный тренд. Анализ локальных особенностей позволяет восстановить информацию о динамике исходного процесса и в некоторых случаях прогнозировать подобные ситуации.

На рис. 6.11 приведенная скейлограмма – результат непрерывного вейвлет-анализа (вейвлет Гаусса) временного ряда, соответствующего рассматриваемому выше процессу.

Приведенный пример показывает, что вейвлет-анализ позволяет обнаруживать не только очевидные аномалии в исследуемом ряде, но и критические значения, которые скрыты за относительно небольшими абсолютными значениями элементов ряда. Например, на скелетоне наибольшие значения отмечены не только в 250-й день, но показаны и неявные экстремумы (25-й и 75-й дни).

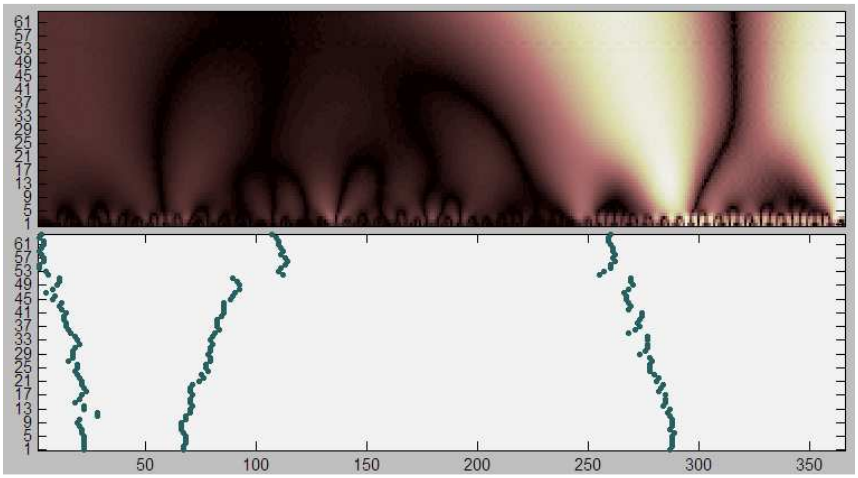


Рис. Часть VI.11. Результат вейвлет-анализа (непрерывное вейвлет-преобразование): сверху – вейвлет-скейлограмма; снизу – линии локальных максимумов (скелетон)

Безусловно, финансово-экономические факторы имеют непосредственное влияние на общественные процессы. На рис. 6.11 приведена динамика изменения курса продажи наличного доллара США в банках Украины в течение 2008 года.

Глава 3. Сложные информационные сети

Информационные системы могут быть представлены как сетевые структуры, так называемые динамические сети [30, 31]. Текущее состояние информационной системы может быть представлено в виде графа $< M, L >$, где M – это множество компонент (например, документов) информационной системы, а L – множество ребер, например, связей подобия, цитирования, ссылок и т.д. В настоящее время наряду с традиционным теориями графов, систем и сетей масоового обслуживания активно развивается теория сложных сетей (от англ. – *Complex Networks*), в рамках которой предлагаются подходы к решению вычислительно сложных задач, характерных для современных сетей.

§ 3.1. Основы концепции сложных сетей

Основной причиной актуальности теории сложных сетей являются результаты современных работ по описанию реальных компьютерных, биологических и социальных сетей. Такие сети имеют характеристики, не свойственные сетям с равновероятной связностью узлов, а строятся на основе связных структур, степенных распределений и узлов-концентраторов.

Представляющие интерес сети чаще всего разрежены – присутствует лишь малая часть возможных ребер, соединяющих отдельные узлы. Поэтому сегодня особую актуальность приобретают методы работы с разреженными матрицами.

Действительно, практически все современные сети можно считать сложными. Так, например, известная задача синтеза топологии сети допускает комбинаторный подход, опирающийся на представление сети в виде конечного графа без петель и кратных ребер, вершины которого соответствуют узлам сети, а ребра – линиям связи.

Вместе с тем, использование методов перечисления графов для решения задачи топологической оптимизации считается неперспективнымм, так как необходимо иссследовать огромное количество возможных вариантов соединения узлов линиями связи. Например, в сети из 10 узлов существует 2^{45} вариантов размещения линий связи (для 10 узлов теоретически возможно $C_{10}^2 = \frac{10 \cdot 9}{2} = 45$ линий соединений. Каждая из этих возможных линий связи может реально существовать – состояние «1», или не существовать – состояние «0», то есть всего возможностей 2^{45}).

Для меншего количества узлов (например, $n = 3$) линии связи, могут быть реально посчитаны ($2^{\frac{3 \cdot 2}{2}} = 8$) вариантами (рис. 6.12).

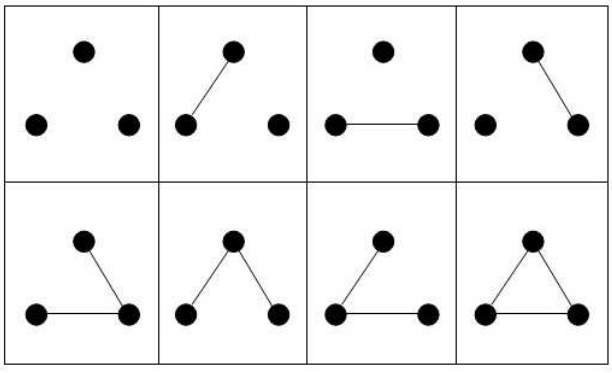


Рис. Часть VI.12. Варианты размещения линий связи при $n = 3$

Сложные сети обычно рассматриваются в абстрактном пространстве, в котором расположение вершин не имеет значения. Для некоторых реальных типов сетей такое рассмотрение оправдано.

Однако, существует множество систем, в которых расположение компонент весьма важно, поскольку влияет на эволюцию сети. Такие сети называются географическими или пространственными. В географических сетях существование прямого соединения между вершинами может зависеть от многих ограничений, таких как расстояние между ними, географический рельеф, территориальные ограничения и т.д. Модели, предназначаемые для представления таких сетей, должны учитывать эти ограничения.

§ 3.2. Параметры сложных сетей

Теория сложных сетей как область дискретной математики изучает характеристики сетей, учитывая не только их топологию, но и статистические феномены, распределение весов отдельных узлов и ребер, эффекты протекания, просачивания, проводимости в таких сетях тока, жидкости, информации и т.д. Оказалось, что свойства многих реальных сетей существенно отличаются от свойств классических случайных графов. Изучения таких параметров сложных сетей, как кластерность, посредничество или уязвимость напрямую относятся к теории живучести, так как именно от этих свойств зависит способность сетей сохранять свою работоспособность при деструктивном воздействии на их отдельные узлы или ребра (связи).

Несмотря на то, что в рассмотрение теории сложных сетей попадают различные сети – электрические, транспортные, информационные, наибольший вклад в развитие этой теории внесли исследования социальных сетей. Термин «социальная сеть» обозначает сосредоточение социальных объектов, которые можно рассматривать как сеть (или граф), узлы которой – объекты, а связи – социальные отношения. Этот термин был введен в 1954 году социологом из «Манчестерской школы» Дж. Барнсом (J. Barnes) в работе «Классы и сборы в норвежском островном приходе». Во второй половине XX столетия понятие «социальная сеть» стало популярным у западных исследователей, при этом в качестве узлов социальных сетей стали рассматривать не только представителей социума, но и другие объекты, которым присущии социальные связи. В теории социальных сетей получило развитие такое направление, как анализ социальных сетей (Social Network Analysis, SNA). Сегодня термин «социальная сеть» обозначает понятие, оказавшееся шире своего социального аспекта, оно включает, например, многие информационные сети, в том числе и веб-пространство или социальные интернет-сети.

В рамках теории сложных сетей рассматривают не только статистические, но динамические сети, для понимания структуры которых необходимо учитывать принципы их эволюции.

В теории сложных сетей выделяют три основных направления: исследование статистических свойств, которые характеризуют поведение сетей; создание модели сетей; предсказание поведения сетей при изменении структурных свойств. В прикладных исследованиях обычно применяют такие типичные для сетевого анализа характеристики, как размер сети, сетевая плотность, степень центральности и т.п.

О «структуре сообщества» в сложной сети можно говорить тогда, когда существует фрагмент сети – группа узлов, которые имеют высокую плотность ребер между собой, при том, что плотность ребер между отдельными фрагментами – низкая. Традиционный метод для выявления структуры сообществ – кластерный анализ. Существуют десятки приемлемых для этого методов, которые базируются на разных мерах расстояний между узлами, взвешенных путевых индексах между узлами и т.п. В частности, для больших социальных сетей наличие структуры сообществ оказалось неотъемлемым свойством.

К потере живучести информационной системы может привести разрыв связей между ее компонентами, например, при устранении из информационного пространства наиболее весомых компонент, то есть таких, которые имеют, допустим, наибольший коэффициент посредничества (betweenness). Этот коэффициент для конкретного узла сети определяется как сумма по всем парам узлов сети соотношений количества кратчайших путей между ними, проходящими через заданный узел, к общему количеству кратчайших путей между ними.

При анализе сложных сетей как и в теории графов исследуются параметры отдельных узлов; параметры сети в целом; сетевые подструктуры.

Параметры узлов сети

Для отдельных узлов выделяют следующие параметры:

- входная степень связности узла – количество ребер графа, которые входят в узел;
- выходная степень связности узла – количество ребер графа, которые выходят из узла;
- расстояние от данного узла до каждого из других;
- среднее расстояние от данного узла до других;
- эксцентричность (eccentricity) – наибольшее из геодезических расстояний (минимальных расстояние между узлами) от данного узла к другим;
- посредничество (betwetnness), показывающее, сколько кратчайших путей проходит через данный узел;
- центральность – общее количество связей данного узла по отношению к другим;
- уязвимость, рассматриваемая как уровень спада производительности сети в случае удаления вершины и всех смежных ей ребер.

Степень связности k_i узла i – это количество ребер, соединенных с этой вершиной.

Соответственно, средняя степень всей сети рассчитывается как среднее всех k_i для всех узлов сети.

Как отмечено выше, в случае ориентированных сетей имеется две разновидности степеней связности узла: выходная, соответствующая количеству исходящих из данного узла ребер, и входная, равная количеству заходящих в данный узел ребер.

Общие параметры сети

Для расчета индексов сети в целом используют такие параметры, как: число узлов, число ребер, геодезическое расстояние между узлами, среднее расстояние от одного узла к другим, плотность – отношение количества ребер в сети к возможному максимальному количеству ребер при данном количестве узлов, количество симметричных, транзитивных и циклических триад, диаметр сети – наибольшее

геодезическое расстояние в сети, уязвимость, рассчитываемая как максимальная уязвимость всех вершин сети, ассортативность как мера корреляции между степенями узлов и т.д.

Существует несколько актуальных задач исследования сложных сетей с точки зрения живучести, среди которых можно выделить следующие основные:

- определение фрагментов сети (клик, кластеров), в которых узлы связаны между собой сильнее, чем с членами других подобных фрагментов;
- выделение фрагментов сети (компонент связности), которые связаны внутри и не связаны между собой;
- нахождение перемычек, т.е. узлов, при изъятии которых сеть распадается на несвязанные части.

Распределение степеней связности узлов

Важной характеристикой сети является функция распределения степеней узлов $P(k)$, которая определяется как вероятность того, что узел i имеет степень $k_i = k$ То есть распределение степеней $P(k)$ отражает долю вершин со степенью k.

Для ориентированных сетей существует распределение выходящей полустепени $P^{out}(k^{out})$, и полустепени входной $P^{in}(k^{in})$, а также распределение общей степени $P^{io}(k^{in}, k^{out})$. Последнее задает вероятность нахождения узла с входной полустепенью k^{in} и выходной полустепенью k^{out}.

Сети, характеризующиеся разными $P(k)$, демонстрируют весьма разное поведение. $P(k)$ в некоторых случаях может быть распределениями Пуассона ($P(k) = e^{-m} m^k / k!$, где m – математическое ожидание), экспоненциальным ($P(k) = e^{-k/m}$) или степенным ($P(k) \sim 1/k^{\gamma}, \quad k \neq 0, \quad \gamma > 0$).

Важной особенностью многих реальных сетей является распределение степеней узлов $P(k)$ по степенному закону.

Сети со степенным распределением степеней связности узлов называются безмасштабными (*scale-free*). Именно безмасштабные распределения часто наблюдаются в реально существующих сложных сетях. При степенном распределении возможно существование узлов с очень высокой степенью, что практически не наблюдается в сетях с пуассоновым распределением.

Путь между узлами

Если два узла i и j можно соединить с помощью последовательности из m ребер, то такую последовательность называют маршрутом (walk) между узлами i и j, а m называю длиной маршрута.

Говорят, что узлы i и j связны, если существует маршрут между ними. Отношение связности транзитивно, т.е. если узел i связн с узлом j, а j связн с k, то i связн с k. При этом маршрут, у которого начало, и конец находятся в одном и том же узле, причем все остальные вершины используются ровно один раз, называется циклом.

Расстояние между узлами определяется как длина маршрута от одного узла до другого. Естественно, узлы могут быть соединены прямо или опосредованно. Путем между узлами d_{ij} назовем кратчайшее расстояние между ними. Для всей сети можно ввести понятие среднего пути, как среднее по всем парам узлов кратчайшего расстояния между ними:

$$l = \frac{2}{n(n+1)} \sum_{i \geq j} d_{ij},$$

где n – количество узлов, d_{ij} – кратчайшее расстояние между узлами i и j.

Венгерскими математиками П. Эрдешем (P. Erdős) и А. Реньи (A. Rényi) было показано, что среднее расстояние между двумя вершинами в случайном графе растет как логарифм от числа вершин [32, 33].

На практике живучесть сети связи определяют как вероятность наличия пути между любой парой узлов.

Некоторые сети могут оказаться несвязными, т.е. в них найдутся узлы, расстояние между которыми является бесконечным. Соответственно, средний путь может оказаться также равным бесконечности. Для учета таких случаев вводится понятие глобальной эффективности сети как среднего инверсного пути между узлами, рассчитываемое по формуле:

$$E = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}},$$

где сумма учитывает все пары узлов. Эта характеристика отражает эффективность сети при пересылке информации между узлами (предполагается, что эффективность в пересылке информации между двумя узлами i и j обратно пропорциональна расстоянию между ними).

Обратная величина глобальной эффективности – среднее гармоническое геодезических расстояний:

$$h = \frac{1}{E}.$$

Так как данная формула снимает проблему расхождения при определении среднего пути, то эта характеристика лучше подходит для графов с несколькими компонентами связности.

Эффективное расстояние между двумя узлами в общем случае больше, чем кратчайшее расстояние.

Сети также характеризуются таким параметром как диаметр или максимальный по длине путь, то есть равный максимальному значению из всех d_{ij}.

Коэффициент кластерности

Д. Уаттс (D. Watts) и С. Строгатц (S. Strogatz) в 1998 году определили такой параметр сетей, как коэффициент кластерности [34], который соответствует уровню связности узлов в сети. Этот коэффициент характеризует тенденцию к образованию групп взаимосвязанных узлов, так называемых клик (*clique*). Кроме того, для конкретного узла коэффициент кластеризации показывает, сколько ближайших соседей данного узла являются также ближайшими соседями друг для друга.

Коэффициент кластерности для отдельного узла сети определяется следующим образом. Пусть из узла выходит k ребер, которые соединяют его с k другими узлами, ближайшими соседями. Если предположить, что все ближайшие соседи соединены непосредственно друг с другом, то количество ребер между ними составляло бы $\frac{1}{2}k(k-1)$. То есть это число, которое соответствует максимально возможному количеству ребер, которыми могли бы соединяться ближайшие соседи выбранного узла. Отношение реального количества ребер, которые соединяют

ближайших соседей данного узла к максимально возможному (такому, при котором все ближайшие соседи данного узла были бы соединены непосредственно друг с другом) называется коэффициентом кластерности узла i – $C(i)$. Естественно, эта величина не превышает единицы.

Существует еще один способ вычисления коэффициента кластерности (транзитивности), базирующийся на такой формуле:

$$C = \frac{3N_\Delta}{N_3},$$

где N_Δ – количество 3-циклов в сети, а N_3 – количество связных 3-компонент.

3-цикл определяется при этом как множество трех узлов с ребрами между каждой парой узлов. Связная 3-компонента – множество, состоящее из трех узлов, в котором каждый узел достижим из другого узла, непосредственно или опосредованно. Таким образом, в 3-компоненте центральный узел должен быть инцидентен двум другим. Множитель 3 введен из учета вариантов различных 3-компонент для каждого 3-цикла, этот множитель обеспечивает выполнение неравенства $0 \le C \le 1$. Тогда мы получаем:

$$N_\Delta = \sum_{k>i>j} a_{ij} a_{ik} a_{jk};$$

$$N_3 = \sum_{k>i>j} \left(a_{ij} a_{ik} + a_{ji} a_{jk} + a_{ki} a_{kj} \right),$$

где a_{ij} – элементы матрицы смежности A, соответствующей сети, сумма берется по всем компонентам различных узлов i , j и k только один раз.

Коэффициент кластерности может определяться как для каждого узла, так и для всей сети. Соответственно, уровень кластерности всей сети определяется как нормированная по количеству узлов сумма соответствующих коэффициентов отдельных узлов.

Разница между двумя подходами к определению кластерности состоит в том, что, усреднив по вершинам, мы получаем во втором случае одинаковое влияние для каждого треугольника в сети, а в первом случае учитывается равный взнос для каждого узла.

Это приводит к разным значениям коэффициента кластерности, потому что узлы с большими степенями с большей вероятностью входят в состав большего количества треугольников, чем вершины с меньшими степенями.

Рассмотренный ниже феномен «малых миров» непосредственно связан с уровнем кластерности сети.

Посредничество

Значение узла для сети тем больше, чем в большем количестве путей он задействован. Поэтому, полагая, что обмен данными происходит по кратчайшим путям между двумя вершинами, можно измерить количественно значение узла с точки зрения посредничества (betweenness), определяемого количеством кратчайших путей проходящих через узел. Эта характеристика отражает роль данного узла в установлении связей в сети. Узлы с наибольшим посредничеством играют главную роль в установлении связей между другими узлами в сети. Посредничество b_m узла m определяется по формуле:

$$b_m = \sum_{i \ne j} \frac{B(i,m,j)}{B(i,j)},$$

где $B(i,j)$ – общее количество кратчайших путей между узлами i и j, $B(i,m,j)$ – количество кратчайших путей между узлами i и j, проходящих через узел m.

Если учитывать, что кратчайшие пути могут быть неизвестны, и вместо этого для навигации в сети используются поисковые алгоритмы, то посредничество (промежуточная центральность) узла может быть выражена вероятностью его нахождения поисковым алгоритмом.

Уровень преобладания наибольшего посредника в этом случае определяется в соответствии с формулой:

$$CPD = \frac{1}{n-1}\sum_i (B_{\max} - B_i),$$

где $B_{\max}$ – самое большое в сети значение уровня посредничества.

Преобладание центрального узла будет равно 0 для клики и 1 для звезды, в которой центральный узел входит во все пути.

Эластичность и уязвимость сети

Противоположные свойства эластичности и уязвимости сетей относятся к распределению расстояний между узлами при изъятии отдельных узлов. Эластичность сети зависит от ее связности, т.е. существовании путей между парами узлов. Если узел будет изъят из сети, типичная длина этих путей увеличится. Если этот процесс продолжать достаточно долго, сеть перестанет быть связной. Р. Альберт (Réka Albert) из университета штата Пенсильвания (США) при исследовании атак на интернет-серверы изучала эффекты, возникающие при изъятии узла из сети, представляющей собой подмножество WWW из 326000 страниц [35].

Среднее расстояние между двумя узлами, как функция от количества изъятых узлов, почти не изменилось при случайном удалении узлов (высокая эластичность). Вместе с тем целенаправленное удаление узлов с наибольшим количеством связей приводит к разрушению сети. Таким образом, Интернет является высоко эластичной сетью по отношению к случайному отказу узла в сети, но высокочувствительной к намеренной атаке на узлы с высокими степенями связей с другими узлами.

Один из способов найти критичные компоненты сети – поиск самых уязвимых узлов [36]. Если производительность сети связана с ее глобальной эффективностью, уязвимость узла может быть определена как спад производительности в случае удаления узла и всех смежных емк ребер из сети:

$$V_i = \frac{E - E_i}{E},$$

где E – глобальная эффективность исходной сети, а E_i – глобальная эффективность после удаления узла i и всех смежных ему ребер.

Упорядоченное распределение узлов относительно их уязвимостей связано со структурой всей сети, таким образом, наиболее уязвимый узел занимает наивысшую позицию в сетевой иерархии. Мера уязвимости сети – максимальная уязвимость среди всех ее узлов:

$$V = \max_i V_i .$$

Коэффициент элитарности

В наукометрии на протяжении длительного времени исследуются сети цитирований. Известно, что влиятельные исследователи определенных областей

формируют сообщества сетевого типа, выражающиеся, например, в публикации совместных работ. Такая закономерность наблюдается также в других реальных сетях и отражает такую тенденцию, как хорошая связность между узлами-концентраторами. Это явление, известное под названием элитарность (или феномен «клуба богатых» – rich-club phenomenon), может быть охарактеризовано коэффициентом элитарности, введенным в работе [37].

Анализ топологии веб, проведенный Ши Жоу (S. Zhou) и Р. Дж. Мондрагоном (R.J. Mondragon) из Лондонского университета, показал, что узлы с большой степенью исходящих гиперссылок имеют больше связей между собой, чем с узлами с малой степенью, тогда как последние имеют больше связей с узлами с большой степенью, чем между собой. Исследование показало, что 27% всех соединений имеют место между всего 5% наибольших узлов, 60% приходится на соединение других 95% узлов с 5% наибольших и только 13% – это соединение между узлами, которые не входят в лидирующие 5%.

Элитарность степени k у сети G – это некое множество узлов со степенью, большей k, $\Re(k) = \{v \in N(G) \mid k_v > k\}$. Коэффициент элитарности степени k выражается следующим образом:

$$\phi(k) = \frac{1}{|\Re(k)|(|\Re(k)|-1)} \sum_{i,j \in \Re(k)} a_{ij},$$

где сумма соответствует удвоенному количеству ребер между вершинами в «элите». Эта характеристика подобна коэффициенту кластерности, она определяет долю связей, существующих между узлами со степенью превышающей k.

Корреляция степеней связанных вершин

Значительное количество структурных и динамических свойств сети определяется с помощью оценки корреляции между степенями соседних узлов. Такая корреляция может быть выражена через совокупное распределение $P(k,k')$, т.е. как вероятность того, что произвольно выбранное ребро соединяет узел степени k с узлом степени k'. Зависимость между степенями вершин можно выразить в терминах условной вероятности того, что произвольно выбранный сосед вершины степени k имеет степень k' [38]:

$$P(k'|k) = \frac{\langle k \rangle P(k,k')}{kP(k)}.$$

При этом $\sum_{k'} P(k'|k) = 1$. В случае неориентированных сетей $P(k,k') = P(k',k)$ и $k'P(k|k')P(k') = kP(k'|k)P(k)$.

Если сеть ориентированная, то k – это степень предшествующего узла, k' – степень последующего узла, значения k и k' могут быть входными , выходными или полными степенями. В общем случае $P(k,k') = P(k',k)$.

Значения $P(k,k')$ и $P(k|k')$ формально описывают корреляции степеней узлов, однако их сложно вычислять экспериментальным путем, что связано с размером сети и малой мощностью выборки узлов с высокими степенями. Эту проблему можно решить, вычислив среднюю степень ближайших соседей узлов с заданной степенью k по формуле:

$$S(k) = \sum_{k'} k' P(k'|k).$$

Показатель корреляции степеней связности позволяет выделить отдельные классы сетей. Если корреляция отсутствует, то $S(k)$ не зависит от значений k, $S(k)=\langle k^2\rangle/\langle k\rangle$. Если $S(k)$ возрастает при увеличении k, то узлы больших степеней тяготеют к соединениям с узлам больших степеней, и сеть относят к ассортативным (отсюда и феномен «клуба богатых»), тогда как если $S(k)$ – убывающая функция от k, то вершины больших степеней тяготеют к соединениям с вершинами малых степеней, и сеть называют дизассортативной [39].

В работе [39] был подсчитан коэффициент корреляции Пирсона для некоторых реальных и смоделированных сетей. Было обнаружено, что, несмотря на отображение моделями специфических особенностей структуры (степенное распределение степени связности узлов, свойство «малого мира»), большинство из них не воспроизводит ассортативность реальной сети. Например, для модели построения сети со степенным распределением степеней связности узлов Барабаши-Альберта [40] значение $r=0$. Тем не менее, социальные сети склонны к ассортативности, а биологические и технологические часто дизассортативны.

Известно, что ассортативные сети менее уязвимы к равновероятным атакам, а дизассортативные менее уязвимы к целенаправленным атакам на узлы-концентраторы. Также, например, синхронизация состояния компонент сети происходит быстрее в ассортативных сетях. Например, при распространении заразной болезни социальные сети в идеальном случае должны быть ассортативны: при контроле малой доли узлов-концентраторов сеть разбивается на изолированные компоненты связности, что позволяет эффективно контролировать распространение инфекции.

§ 3.3. Сложные сети и задачи компьютерной лингвистики

Первым шагом при применении теории сложных сетей к анализу текста является представление этого текста в виде совокупности узлов и связей, построение сети языка (language network) [41].

Существуют различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям сети языка. Узлы могут быть соединенны меду собой, если соответствующие им слова стоят рядом в тексте [42; 43], принадлежат одному предложению [44], соединены синтаксически [45; 46] или семантически [47; 48].

Сохранение синтаксических связей между словами приводит к изображению текста в виде направленной сети (directed network), где направление связи соответствует подчинению слова.

Поставим в соответствие каждому слову узел сети. Соединим каждые два узла связью, если соответствующие им слова стоят в предложении рядом. Такое представление называют L-пространством. В L-пространстве, равно как и в других приведенных ниже представлениях, при возникновении кратных связей принято сохранять лишь одину из них.

a. L-пространство. Связываются соседние слова, которые принадлежат к одному предложению. Количество соседей для каждого слова (окно слова) определяется радиусом взаимодействия *R*, чаще всего рассматривается случай *R* = 1.

b. B-пространство. Рассматриваются узлы двух видов, соответствующие предложениям и словам, которые им принадлежат.
c. P-пространство. Все слова, которые принадлежат одному предложению, связываются между собой.
d. C-пространство. Предложения связываются между собой, если в них употреблены одинаковые слова.

В случае L-пространства связи могут учитывать не только «ближайших соседей», но и группы из нескольких слов, которые находятся на определенном расстоянии друг от друга. Для этого вводится поняте «радиуса действия» R: при R = 1 связь существует лишь между ближайшими соседями, при R = 2 – между ближайшими и следующими близкими соседями и т. д. Переменная R может принимать значения от R = 1 до $R_{\max}$, где $R_{\max}$ + 1 – общее количество слов в предложении.

Еще один способ представить текст в виде сети заключается в использовании двудольных (bipartite) графов. В таком предсталенни (B-просторанство) рассматриваются узлы двух видов. Один вид соответствует предложениям, второй – словам. Связь между различными узлами означает, что слово принадлежит предложению.

В P-просторанстве все слова, принадлежащие одному предложению, считаются связанными между собой. В C-просторанстве узлы соответствуют предложением, а связь между узлами-предложениями устанавливается в том случае, если у них есть общие слова.

В случае рассмотрения L-пространства языка количество соседних слов, между которыми строятся связи, определяется параметром R: при R = 1 связаны между собой лишь ближайшие соседи, при R = 2 связи строятся между узлом-словом, его ближайшими и предшествующими близкими соседями и т. д. Рост «радиуса взаимодействия» R приводит к росту количества связей, достигая насыщения при $R = R_{\max}$.

Для сети, построенной на основании Британского национального корпуса, оказалось, что данная сеть английского языка безмасштабна, а поведение степени $P(k)$ характеризуется двумя режимами степенного распределения со значением степенного показателя $\gamma = 1{,}5$ для k < 2000 и $\gamma = 2{,}7$ для k > 2000 соответственно.

Согласно определению, если средняя длина кратчайшего пути растет с размером (количеством узлов) сети медленнее любой функции степени, то сеть является «малым миром». Сети малого мира чрезвычайно компактны. Для упомянутой выше сети английского языка длина кратчайшего пути составляет всего $< l >= 2{,}63$. Поскольку рост R приводит лишь к добавлению новых связей, то $< l >$ уменьшаются с ростом R.

Специфической формой корреляции в сетях является образование кластеров. Коэффициент кластерности C характеризует склонность сети к образованию соединенных троек узлов. Известно, что для полного графа C = 1, а для сети в форме дерева C = 0. Отношение среднего коэффициента кластерности исследуемых сетей к коэффициенту кластерности классического случайного графа свидетельствует о том, что сети языков являются хорошо коррелированными структурами. Такие корреляции растут с ростом «радиуса взаимодействия» R. Для Британского национального

корпуса на основании анализа текстов, которые содержали $\approx 10^7$ слов, получено значение коэффициента кластености $< C >= 0,687$. hCi = 0:687 [42].

В случае рассмотрения P –пространства каждое слово-узел связано со всеми другими словами, которые принадлежат общему предложению. Таким образом, каждое предложение текста входит в сеть как полный граф – клика взаимосвязанных узлов. Разные предложения-клики объединяются в сеть благодаря общим словам. В L-просторанстве слова связываются в пределах окна, размеры которого характеризуются переменной R. Когда размер этого окна становится ровным размеру предложения, то представление этого предложения в L- и в P -пространствах совпадают. Соответственно, когда размер окна становится ровным размеру самого длинного предложения текста ($R = R_{\max}$), то представление всего текста в L- и в P-пространствах совпадают.

Полученны результаты, которые убедительно свидетельствуют о том, что сеть языка яляется сильно коррелирующим безмасштабным малым миром (scale - free small world).

Существует ряд трудов, в которых сделана попытка объяснить свойства сетей языка с помощью сценария подавляющего присоединения (preferential attachment [49]), рассматривая их как результат процесса роста, когда новые узлы-слова с большей вероятностью присоединяются к узлам-хабам, имеющим много связей.

§ 3.4. Моделирование сложных сетей

Модель слабых связей

Существует класс сложных сетей, которым присущи так называемые «слабые» связи. Аналогом слабых социальных связей являются, например, отношения с далекими знакомыми и коллегами. В некоторых случаях эти связи оказываются более эффективными, чем связи «сильные». Так, группой исследователей из Великобритании, США и Венгрии, был получен концептуальный вывод в области мобильной связи, заключающийся в том, что «слабые» социальные связи между индивидуумами оказываются наиболее важными для существования социальной сети [50].

Для исследования были проанализированы звонки 4.6 млн. абонентов мобильной связи, что составляет около 20% населения одной европейской страны. Это был первый случай в мировой практике, когда удалось получить и проанализировать такую большую выборку данных, относящихся к межличностной коммуникации.

В социальной сети с 4.6 млн. узлов было выявлено 7 млн. социальных связей, т.е. взаимных звонков от одного абонента другому и обратно, если обратные звонки были сделаны на протяжении 18 недель. Частота и продолжительность разговоров использовались для того, чтобы определить силу каждой социальной связи.

Было выявлено, что именно слабые социальные связи (один-два обратных звонка на протяжении 18 недель) связывают воедино большую социальную сеть. Если эти связи проигнорировать, то сеть распадется на отдельные фрагменты. Если же не учитывать сильных связей, то связность сети нарушится (рис. 6.13).

1 2 3

Рис. Часть VI.13. Структура сети:
1) полная карта сети социальных коммуникаций; 2) социальная сеть, из которой изъятые слабые связи; 3) сеть, из которой изъятые сильные связи: структура сохраняет связность

Оказалось, что именно слабые связи являются тем феноменом, который связывает сеть в единое целое. Надо полагать, что данный вывод справедлив и для веб-пространства, хотя исследований в этой области до сих пор не проводилось.

Модель малых миров

Несмотря на огромные размеры некоторых сложных сетей, во многих из них (в веб-пространстве, в частности) существует сравнительно короткий путь между двумя любыми узлами – геодезическое расстояние. В 1967 г. психолог С. Милгран в результате проделанных масштабных экспериментов вычислил, что существует цепочка знакомств, в среднем длиной шесть, практически между двумя любыми гражданами США [51].

Д. Уаттс и С. Строгатц обнаружили феномен, характерный для многих реальных сетей, названный эффектом малых миров (Small Worlds) [52].

Сетевые структуры, соответствующие свойствам малых миров обладают следующими типичными свойствами: малая средняя длина пути относительно диаметра сети (что характерно также для случайных сетей) и большой коэффициент кластеризации (что присуще сетям с регулярной структурой).

При исследовании этого феномена ими была предложена процедура построения наглядной модели сети, которой присущ этот феномен.

Чтобы построить сеть «малого мира», следует начать с регулярной циклической решетки с N вершинами, каждая из которых соединена с k (в частности, k = 2) ближайшими соседями в каждом направлении. Для каждой вершины задается $2k$ связей, где $N >> \log_2(N) >> 1$. Затем каждое ребро пересоединяется со случайной парой вершин с вероятностью p .

При условии $p = 0$ получается упорядоченная решетка с большим количеством циклов и большими расстояниями, а при условии $p \to 1$сеть становится случайным графом с короткими расстояниями, и малым количеством циклов. В некоем среднем случае присутствуют и короткие расстояния, и большое количество циклов.

Три состояния этой сети представлены на рис. 6.14: регулярная сеть – каждый узел которой соединен с четырьмя соседними, та же сеть, у которой некоторые «ближние» связи случайным образом заменены «далекими» (именно в этом случае возникает феномен «малых миров») и случайная сеть, в которой количество подобных замен превысило некоторый порог.

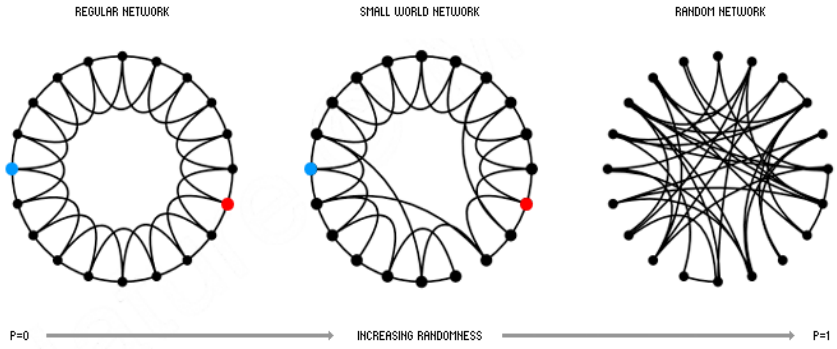


Рис. Часть VI.14. *Модель Уаттса-Строгатца*

На рис. 6.15 приведены графики изменения средней длины пути и коэффициента кластеризации искусственной сети Д. Уаттса и С. Строгатца от вероятности установления «далеких связей» (в полулогарифмической шкале).

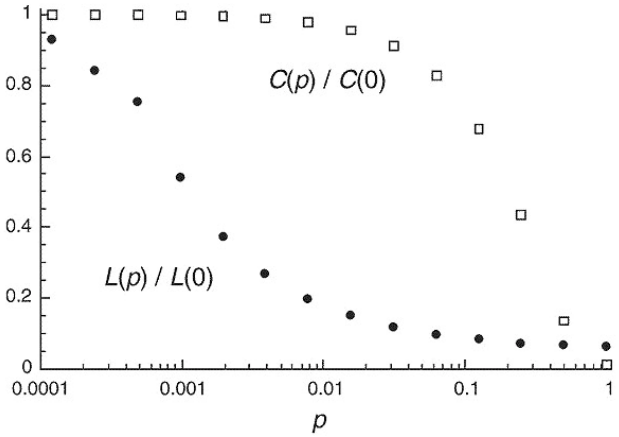


Рис. Часть VI.15. Динамика изменения длины пути и коэффициента кластерности в модели Уаттса-Строгатца в полулогарифмической шкале (горизонтальная ось – вероятность замены ближних связей далекими)

В реальности оказалось, что именно те сети, узлы которых имеют одновременно некоторое количество локальных и случайных «далеких» связей, демонстрируют одновременно эффект малого мира и высокий уровень кластеризации. Веб-пространство также является сетью, для которой также подтвержден феномен малых миров.

Эти исследования дают основания полагать, что зависимость веб-пространства от больших узлов значительно существеннее, чем предполагалось ранее, т.е. она еще более чувствительна к злонамеренным атакам. С концепцией «малых миров» связан также практический подход, называемый «сетевой мобилизацией», которая реализуется над структурой «малых миров». В частности, скорость распространения информации благодаря эффекту «малых миров» в реальных сетях возрастает на порядки по сравнению со случайными сетями, ведь большинство пар узлов реальных сетей соединены короткими путями.

Кроме того, сегодня довольно успешно изучаются масштабируемые, статические, иерархические "малые миры" и другие сети, исследуются их фундаментальные свойства, такие, как стойкость к деформациям и перколяция. Недавно было показано, что наибольшую информационную проводимость имеет особый класс сетей, называемых "запутанными" (англ. – *entangled networks*). Они

характеризуются максимальной однородностью, минимальным расстоянием между любыми двумя узлами и очень узким спектром основных статистических параметров. Считается, что запутанные сети могут найти широкое применение в области информационных технологий, в частности, в новых поколениях веб, позволяя существенным образом снизить объемы сетевого трафика.

Модель случайной сети Эрдоша-Рени

Существует две модели классического случайного графа: в первой считается, что M ребер распределены произвольно и независимо между парами из N вершин графа; во второй модели фиксируется вероятность m, с которой может объединяться каждая из пар врешин. При $m \to \infty$ и $N \to \infty$ для обоих вариантов распределение степеней узлов k определяется формулой Пуассона:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!},$$

где среднее значение степени узла: $\langle k \rangle = 2M / N$ для первой модели и $\langle k \rangle = mN$ для второй. При этом средняя длина кратчайшего пути для сети Эрдоша-Рени составляет

$$\langle l \rangle = \ln(N) / \ln(\langle k \rangle),$$

а коэффициент кластерности:

$$C \sim \langle k \rangle / N.$$

Построение случайного графа Эрдоша-Рени выполняется следующим

Образом. Пусть в начале имеестя N изолированных вершин, к которым

Последовательно добавляются ребра, которые случайным образом соединяют пары вершин. В результате такого процесса доля связанных вершин определяется выражением:

$$G = 1 - \sum_{n=1}^{\infty} \frac{n^{n-1}}{n!} \langle k \rangle^{n-1} e^{-n\langle k \rangle},$$

где n – номер шага процесса добавления ребер.

Таким образом часть связанных вершин монотонно возрастает с увеличением средней степени $\langle k \rangle$, переходя от степенной зависимости к экспоненциальной.

Процедура преимущественного присоединения Барабаши-Альберта

Наибольшее количество реальных сетей соответствуют степенному закону распределения, который является, как известно, признаком самоподобия. Благодаря дальним корреляциям система не имеет масштаба изменения параметров (в связи с эти сложные системы, которые характеризуются степенным распределением, называются безмасштабными).

Сценарий построения сетей Барабаши-Альберта базируется на двух механизмах – росте и преимущественном присоединении (preferential attachment). Данная модель использует такой алгоритм: рост сети происходит начиная с небольшого количества узлов n_0, к которым на каждом временном шагу добавляется новый узел с $n \le n_0$ связями, которые присоединяются к уже существующим узлам; преимущественное присоединение состоит в том, что вероятность присоединения $P(k_i)$ нового узла к уже существующему узлу i зависти от степени k_i узла i:

$$P(k_i) = \frac{k_i}{\sum_j k_j}.$$

Здесь в знаменателе суммирование ведется по всем узлам. Как компьютерные модели, так и аналитические решения модели Барабаши-Альберта дают степенную асимптотику распределения степеней узлов с показателем $\gamma = 3$.

Фазовые переходы при распределении доходов

Рассмотрим модель распределения доходов между людьми (узлами социальной сети), основанную на модифицированной процедуре Барабаши-Альберта. Доход в рамках предлагаемой модели рассматриваются как величина, пропорциональная степени узла, т.е. количеству связей, которыми обладает тот или иной узел. Таким образом, в данной модели узел сети будет считаться тем богаче, чем выше его степень. Справедливость данного подхода может быть проиллюстрирована, например, деятельностью предпринимателей, коммивояжеров, туристических фирм, успешность которых, как правило, зависит от количества партнеров и т.п.

При этом, если рассматривать связи между узлами в динамике, в частности, возможность появления новых связей, то, очевидно, распределение степеней узлов будет в значительной степени зависеть от некоторого порога, который необходимо преодолеть для установления связей (аналог в экономике – порог выхода предпринимателя или компании на тот или иной рынок).

Рассмотрим модель эволюции динамической сети, состоящей из N узлов. В начальном состоянии сеть содержит N ненаправленных ребер, вес каждого из которых – 1. В этом состоянии 1-й узел связан ребром со 2-м, 2-й – с 3-м, …, i-й – с ($i+1$)-м,…, N-й – с 1-м.

В рамках этой модели определяется весовое значение для каждого узла – его нормированная степень, пропорциональная количеству ребер, смежных с данным узлом (величина, находящаяся в интервале [0,1]).

Эволюция сети заключается в формировании новых ребер по следующему алгоритму: в течение каждого цикла (их количество – M задается заранее) для всех узлов по очереди, начиная с 1-го формируется ребро, если выполняется условие, что степень данного узла превышает некоторый заданный заранее порог p ($0 < p < 1$). В этом случае данное ребро соединяет исходный узел с некоторым другим, номер которого определяется случайным образом.

Пример сформированной таким образом сети для значений N=25, M=10, p=0.05 приведен на рис. 6.16.

На рис. 6.17 показано распределение степеней узлов при N=200, M=50, p=0.002-0.007. При этом номера узлов ранжированы по значениям их степеней (чем выше степень, тем меньший порядковый номер).

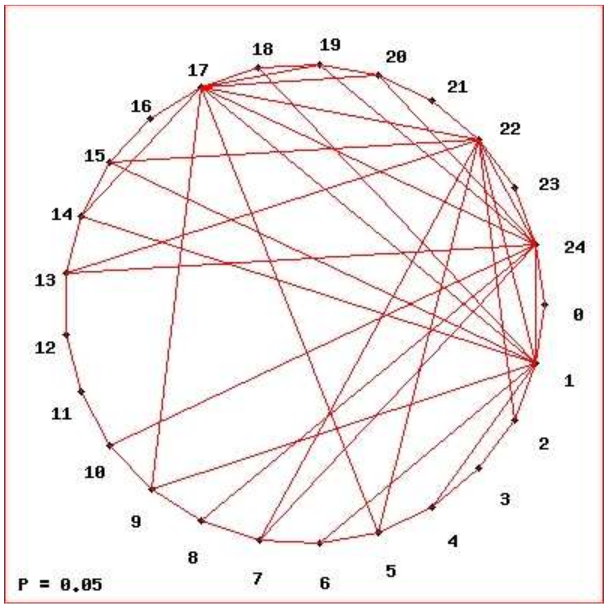


Рис. Часть VI.16. Сеть из 25 узлов, формируемая за 10 шагов алгоритма при p=0.05

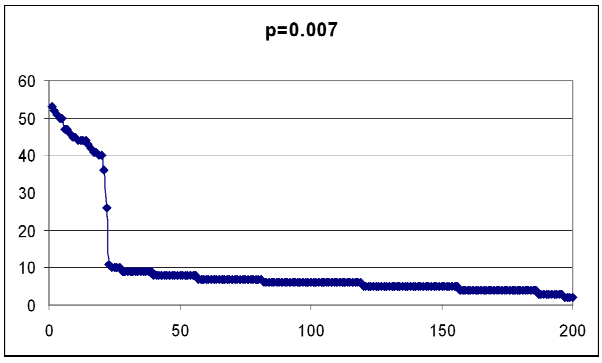


Рис. Часть VI.17. *Распределения степеней узлов при N=200, M=50, p=0.007*

Распределение, получаемое в результате моделирования, не совпадает с общеизвестной закономерностью распределения доходов В. Парето. Вместе с тем, закономерность Парето не может объяснить эффект отсутствия «среднего класса» (Middle Class) [53] в социальных системах с относительно высоким порогом переходов между уровнями доходов, что наглядно демонстрируется в рамках предлагаемого подхода.

Как показали эксперименты, небольшом пороге *p* наблюдается относительно равномерное распределение степеней узлов (по условиям модели – доходов). При p=0.001–0.003 можно предположить наличие «среднего класса». Однако при повышении порога (p>0.003) узлы после прохождения *M* циклов формирования ребер разделяются на две группы, между которыми наблюдается явный разрыв. Так как степень узла в рамках данной модели – это уровень «богатства», то происходит жесткое расслоение на богатых и бедных. Скачек при больших значениях как раз и

характеризует отсутствие «среднего класса», т.е. плавного перехода при распределении степеней узлов.

Аналогия – нехватка капитала в период начального накопления – группа узлов становится «богатой», получая большую степень, недосягаемую для остальных.

На рис. 6.18. приведен усредненный по график плотности вероятности распределения степеней узлов для значений *N=50, M=2, p=0.02.*

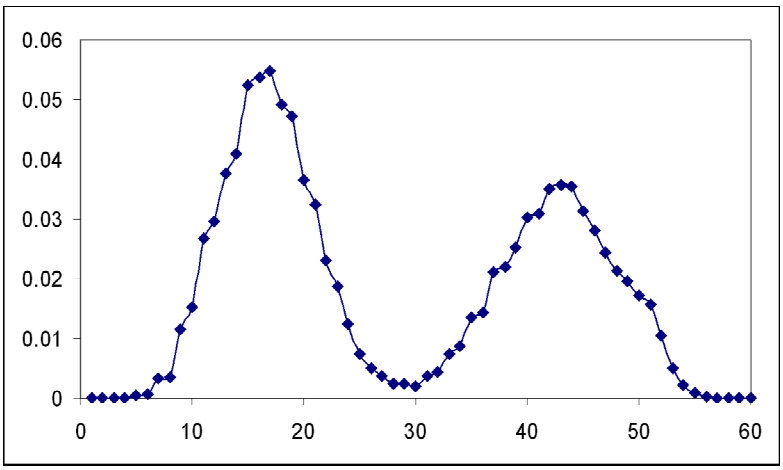


Рис. Часть VI.18. Эмпирическая плотность вероятностей распределения степеней узлов для значений N=50, M=20, p=0.02, усредненная по 1000 реализациям

Явно выраженное наличие двух «колоколов» в графике плотности вероятности также свидетельствует о четком разделении значений на два класса.

Трехмерный график распределения степеней узлов в зависимости от номера (ранга) узла и значения порога *p* приведен на рис. 6.19. На этом графике видно, что значение скачка является возрастающей функцией от *p*, что можно трактовать как увеличение абсолютного разрыва между классами при возрастании порога.

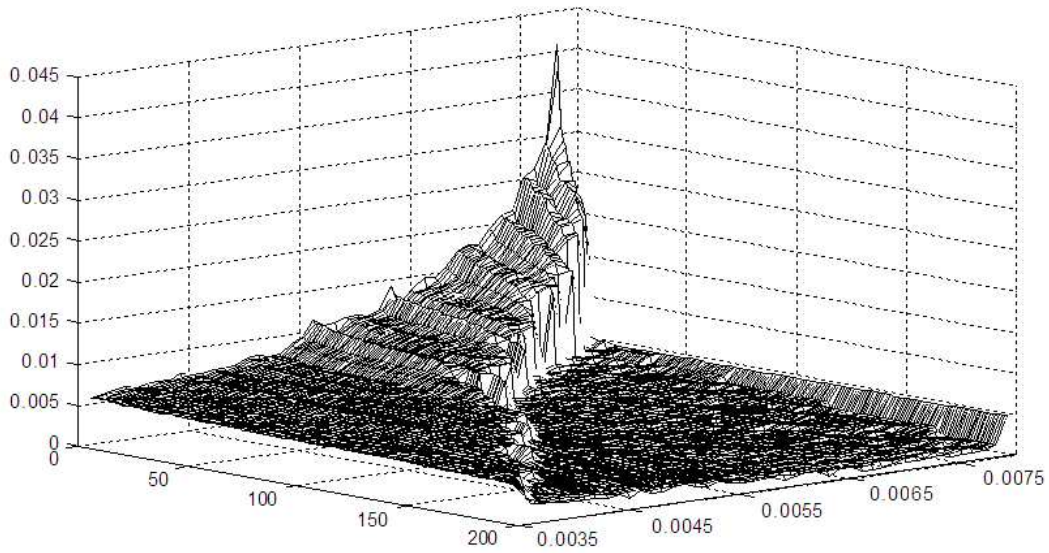


Рис. Часть VI.19. Распределение степеней узлов (ось 0Z) при N=200, M=50 в зависимости от значения порога p (ось 0X) и ранга узла (ось 0Y)

В результате анализа описанной в работе модели получено распределение степеней узлов сети со скачкообразным переходом, объясняющее отсутствие «среднего класса» в рамках предлагаемого подхода и предметной области.

Список используемой литературы

[1] Salton G, Wong A, Yang C. A Vector Space Model for Automatic Indexing. // Communications of the ACM, 18(11):613-620, 1975.

[2] Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. - 240 с.

[3] Broder A. Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. – P. 1-10.

[4] Иванов С.А. Мировая система научной коммуникации как информационное пространство // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: 8-я Междунар. конф. "Крым 2002": Материалы конф., Судак, 9-17 июня, 2001 г. – М., 2001. – Т.1. – С. 1123-1126.

[5] Брайчевский С.М. Современные информационные потоки: актуальная проблематика / С.М. Брайчевский, Д.В. Ландэ // Научно-техническая информация. – Сер. 1. – Вып 11. –2005. – С. 21-33.

[6] Арнольд В.И. Аналитика и прогнозирование: математический аспект // Научно-техническая информация. - Сер. 1. - Вып. 3. - 2003. - С. 1-10.

[7] Нейман Дж. Теория самовоспроизводящихся автоматов. - М.: Мир, 1971. – 382 с.

[8] Wolfram S. A New Kind of Science. - Champaign, IL: Wolfram Media Inc., 2002. – 1197 p.

[9] Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы – М.: Либроком (Editorial URSS), 2009.

[10] Гарднер М. Математические досуги. – М.: Мир, 1972.

[11] Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion // Technological forecasting and social change, 1993. - Vol. 44. - № 1. - pp. 87-97.

[12] Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений // Научно-техническая информация. Серия 2. – №1, 1975. – С. 9 - 20.

[13] Herdan G. The Advanced Theory of Language as Choice and Chance. Berlin–Heidelberg–New York, 1966.

[14] Яглом А.М., Яглом И.М. Вероятность и информация. Изд. 3-е, перераб. и дополн. М., «Наука», 1973.511с.

[15] Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing - Cambridge, Massachusetts: The MIT Press, 1999.

[16] Bell A., Fosler-Lussier E., Girand C., Raymond W. Reduction of English function words in Switchboard // Proceedings of ICSLP-98. - Vol 7. – 1998. - pp. 3111-3114.

[17] Simon H. A. Biometrika 42, 425 (1955).

[18] Bradford, S.C. "Sources of Information on Specific Subjects". Engineering: An Illustrated Weekly Journal (London), 137, 1934 (26 January). – P. 85-86.

[19] Алексеев Н.Г. Применение закона Бредфорда при комплектовании фонда научной библиотеки // Тезисы докладов конференции "Библиотечное дело-1996". URL: http://libconfs.narod.ru/1996/4s/4s_p1.html

[20] Heaps H.S. Information Retrieval - Computational and Theoretical Aspects. Academic Press, 1978.

[21] Grootjen F.A., Van Leijenhorst D. C., van der Weide T.P. A formal derivation of Heaps' Law // Inf. Sci. – Vol. 170(2-4). - pp. 263-272. - 2005. URL: http://citeseer.ist.psu.edu/660402.html

[22] Шредер М. Фракталы, хаос, степенные законы. Миниатюры из бесконечного рая. – М.: Регулярная и хаотическая динамика, 2001. – 528 с.

[23] Столлингс В. Современные компьютерные сети. 2-е изд. – СПб.: Питер, 2003. – 783 с.

[24] Иванов С.А. Стохастические фракталы в Информатике // Научно-техническая информация. Сер. 2. — 2002. — № 8. — С. 7–18.

[25]. Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A.L. Mosaic organization of DNA nucleotides. // Phys Rev E. 1994.— 49 (2).— P. 1685−1689.

[26] Ландэ Д.В., Снарский А.А. Динамика отклонения элементов ряда измерений от локальных линейных аппроксимаций // Реєстрація, зберігання і оброб. даних. — 2009. —Т. 11, № 1. — С. 27–32.

[27] Федер Е. Фракталы.— М.: Мир, 1991.— 254 с.

[28] Чуи К. Введение в вэйвлеты. — М.: Мир, 2001. — 416 с.

[29] Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. — 1996. — Т. 166, № 11. — С. 1145–1170.

[30] Newman M.E.J. The structure and function of complex networks // SIAM Rev. — 2003. — 45.— P. 167–256.

[31] Dorogovtsev S.N., Mendes J.F.F. Evolution of networks: from biological networks to the Internet and WWW. — Oxford University Press, 2003.

[32] Erdős, P., Renyi A. On Random Graphs. I // Publ. Math. — 1959. — **6**.— P. 290–297.

[33] Erdős P., Renyi A. On the evolution of random graphs // Publ. Math. Inst. Hungar. Acad.—1960. —Sci. 5. — P. 17–61.

[34] Watts D.J., Strogatz S.H. Collective dynamics of «smallworld» networks // Nature.— 1998.— 393.— P. 440–442.

[35] Albert R., Jeong H., Barabasi A. Attack and error tolerance of complex networks // Nature. — 2000.— 406.— P. 378–382.

[36] Gol’dshtein V., Koganov G.A., Surdutovich G.I. Vulnerability and hierarchy of complex networks // Phys. Rev. Lett.— 2004.

[37] Zhou S., Mondragon R.J. The rich-club phenomenon in the internet topology // Commun. Lett. IEEE. — 2004. — 8. — P. 180–182.

[38] Boguna M., Pastor-Satorras R., Vespignani A. Statistical mechanics of complex networks / Lecture and notes in physics. — Springer Berlin. —2003.— P. 127–147.

[39] Newman M.E.J. Assortative mixing in networks // Phys. Rev. Lett.— 2002. — 89 (208701).

[40] Barabasi A., Albert R. Emergence of scaling in random networks // Science.— 1997.— 286. — P. 509–512.

[41] Ю. Головач, В. Пальчиков, Лис Микита і мережі мови, Журн. Фіз. Досл. 10 (2006) 247-291.

[42] Ferrer-i-Cancho R., Sole R. V. The small world of human language // Proc. R. Soc. Lond. B 268, 2261 (2001).

[43] Dorogovtsev S.N., Mendes J. F. F. Language as an evolving word web // Proc. R. Soc. Lond. B 268, 2603 (2001).

[44] Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V. The network of concepts in written texts // Preprint physics/0508066 (2005).

[45] Ferrer-i-Cancho R., Sole R.V., Kohler R. Patterns in syntactic dependency networks // hys. Rev. E 69, 051915 (2004).

[46] Ferrer-i-Cancho, R. The variation of Zipf's law in human language. // Phys. Rev. E 70, 056135 (2005).

[47] Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P. Topology of the conceptual network of language // Phys. Rev. E 65, 065102(R) (2002).

[48] Sigman M., Cecchi G. A. Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 99, 1742 (2002).

[49] Albert R., Jeong H., Barabasi A.-L. Diameter of the world wide web // Nature (London) 401, 130 (1999).

[50] Bjorneborn L., Ingwersen P. Toward a basic framework for webometrics // J. Amer. Soc. Inform. Sci. and Technol. — 2004. — 55(14). — P. 1216−1227.

[51] Milgram S. The small world problem // Psychology Today. — 1967. — 2.— P. 60–67.

[52] Watts D.J., Strogatz S.H. Collective dynamics of «smallworld» networks // Nature.— 1998.— 393.— P. 440–442.

[53] Cashell B.W: Who Are the “Middle Class”?, CRS Report for the Congress, March 20, 2007.

БОЛЬШАКОВА Елена Игоревна
КЛЫШИНСКИЙ Эдуард Станиславович
ЛАНДЭ Дмитрий Владимирович
НОСКОВ Алексей Анатольевич
ПЕСКОВА Ольга Вадимовна
ЯГУНОВА Елена Викторовна

Автоматическая обработка текстов на естественном языке
и компьютерная лингвистика

Изд. лиц. ИД № 06117 от 23 октября 2001 г. Подписано в печать .06.11.
Формат 60x84/16 . Бумага типографская № 2. Печать - ризография.
Усл.печ. л. 17,1 Уч.-изд. л. 15,4. Тираж 500 экз. Заказ Изд. № .
Московский государственный институт электроники и математики
109028 Москва, Б. Трехсвятительский пер., 3/12.