Министерство образования и науки Российской Федерации
Институт систем информатики им. А.П. Ершова СО РАН

Т. В. Батура, М. В. Чаринцева

ОСНОВЫ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ

Учебное пособие

Новосибирск
2016

В пособии рассмотрены методы задания синтаксической структуры предложений и основные принципы построения синтаксических анализаторов. Перечислены проблемы автоматической обработки текстов, которые до сих пор не удается решить в полной мере. Обсуждаются способы создания систем автоматического анализа эмоциональной окраски текстов и алгоритмы, положенные в их основу. В пособие включены задания, которые предлагались на различных олимпиадах по математической и компьютерной лингвистике в период 2002–2009 годы. Задания снабжены решениями и указаниями.

Пособие соответствует части курса лекций по дисциплине «Математическая лингвистика и обработка текстов на естественном языке», который читается аспирантам Института систем информатики им. А.П. Ершова СО РАН.

Оглавление

Введение

С распространением Интернета количество информации, в том числе текстовой информации на естественном языке, стало расти чрезвычайно быстро. Стремительное развитие современного общества и компьютерных технологий требует постоянного совершенствования методов обработки информации. В настоящее время выделяют следующие основные направления компьютерной лингвистики: информационный поиск, извлечение информации, машинный перевод, автореферирование, корпусная лингвистика, построение экспертных и вопросно-ответных систем, создание тезаурусов и онтологий и некоторые другие.

Компоненты, составляющие структуру систем анализа текстов, – это лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора, как правило, является выходом другого. Выделяют следующие компоненты систем обработки текстов.

Основные этапы построения систем автоматической обработки текстов:

1) графематический анализ (осуществляется на уровне символов);
2) морфологический анализ (осуществляется на уровне слов);
3) фрагментационный анализ (осуществляется на уровне фраз, частей предложения);
4) синтаксический анализ (осуществляется на уровне предложений);
5) семантический анализ (осуществляется на уровне текста).

Часто графематический анализ сводится к выполнению только лексического анализа и совмещен с морфологическим анализом. *Лексический анализ (токенизация)* – выделение в тексте слов, цифровых комплексов, знаков препинания, формул и т. д. Все эти выделенные элементы называют *лексемами (токенами)*.

Иногда на этапе графематического анализа осуществляют деление текста на более крупные, чем отдельные слова, единицы (абзацы, предложения). На этом этапе ориентиром для разбивки являются пробелы, отступы, знаки препинания. Подзадачей графематического анализа является *сегментация* – поиск границ между словами в тексте без пробелов (например, на китайском или японском языках).

Морфологический анализатор может состоять из четырех компонент: стемматизации, лемматизации, приписывания граммем и получения парадигм.

Стемматизация (или *стемминг* от англ. stemming) – это процесс нахождения основы слова для заданного исходного слова. Отметим, что получаемая псевдооснова не обязана совпадать с грамматической основой рассматриваемой словоформы; достаточно, чтобы словоформы,

соответствующие одной парадигме, получали в результате работы алгоритма одну и ту же псевдооснову.

Лемматизация – процесс приведения словоформы к лемме. *Лемма* – нормальная (словарная) форма слова.

Приписывание словоформе множества наборов граммем – приписывание словоформе грамматических характеристик (грамматических признаков). *Граммема (грамматическая характеристика)* – это элементарный морфологический описатель, относящий словоформу к какому-то морфологическому классу.

Получение парадигм – процесс построения всех форм слова по начальной форме.

Ясно, что морфологического анализа явно недостаточно для выбора одной конкретной морфологической интерпретации слова. К тому же, выбор одной интерпретации может повлиять на выбор интерпретации для соседних слов. Поэтому программы работают с целым набором возможных морфологических интерпретаций, постепенно выделяя наиболее вероятные на следующих этапах анализа.

Далее в пособии описаны методы задания синтаксической структуры предложений, более подробно рассмотрены этапы фрагментационного и синтаксического анализа в системах автоматической обработки текстов.

Раздел 1. Методы задания синтаксической структуры предложений

После того как произведен морфологический анализ слов текста, система автоматической обработки переходит к этапу синтаксического анализа, в результате которого определяются взаимосвязи между отдельными словами и частями предложения.

Как правило, синтаксическая структура предложения – это граф, узлами которого выступают слова предложения. При этом, если два слова связаны каким-либо образом, то соответствующие им вершины графа связаны дугой с определенной окраской. Возможные окраски дуг зависят от языка, на котором написано предложение, а также от выбранного способа представления синтаксической структуры предложения.

При синтаксическом анализе предложений русского языка для окраски дуг можно использовать вопросы, задаваемые от одного слова к другому. Некоторым словам (например, предлогам) вообще не соответствует ни одна из вершин графа, но эти слова влияют на согласования слов друг с другом.

В общем понимании **синтаксический анализ** – это процесс сопоставления линейной последовательности лексем (слов, токенов) естественного или формального языка с его формальной грамматикой.

Результатом синтаксического анализа является синтаксическая структура предложения, представленная в виде системы составляющих (дерева составляющих) или дерева зависимостей (дерева подчинения).

Остановимся подробнее на этих двух методах задания синтаксической структуры предложений.

1.1. Системы составляющих

Пусть x – произвольная непустая цепочка в словаре V. Множество C отрезков цепочки x называется **системой составляющих** этой цепочки, если оно удовлетворяет следующим двум условиям:

1) C содержит отрезок, состоящий из всех точек цепочки x, и все одноточечные отрезки x;
2) любые два отрезка из C либо не пересекаются, либо один из них содержится в другом.

Элементы C мы будем называть **составляющими**. Одноточечные отрезки называются **точечными составляющими**. Отрезок, состоящий из всех точек цепочки, называется **полной составляющей**. Полную и точечные составляющие называют **тривиальными**, остальные составляющие – **нетривиальными**.

Для наглядного изображения системы составляющих можно пользоваться следующим простым способом: заключать каждую нетривиальную составляющую в скобки, причем левую и правую скобки, отвечающие одной составляющей, помечать одной и той же меткой так, чтобы разные пары скобок были помечены разными метками; в качестве меток можно использовать натуральные числа. Можно обойтись и без меток, так как для каждой левой скобки можно однозначно указать соответствующую ей правую.

Если цепочку интерпретировать как предложение естественного языка, то система составляющих может быть использована в качестве способа выражения информации о его синтаксической структуре.

Пример

Предложение *Онегин, добрый мой приятель, родился на брегах Невы* допускает следующую «естественную» систему составляющих: *(Онегин, (добрый (мой приятель))),(родился (на (брегах Невы))).*

Среди многочисленных систем составляющих, которые имеет предложение естественного языка, лишь весьма немногие «правильны», т. е. адекватно отражают синтаксическое строение предложения. При этом понятие «правильной» системы составляющих не абсолютно. Оно зависит от соглашений лингвистического характера, отражающих определенные содержательные представления о синтаксической структуре предложения данного языка.

При фиксированной системе соглашений предложение также может иметь несколько «правильных» систем составляющих. Нередко эти системы соответствуют различным толкованиям смысла предложения. В лингвистике такое явление – наличие у одного предложения двух или более правильных «синтаксических анализов» – известно под названием синтаксической омонимии.

Пример

Допустим, есть две возможных системы составляющих для одного и того же предложения:

(Все (эти известия)) (произвели (на меня) (удручающее впечатление));

(Все (эти известия)) произвели (на меня) (удручающее впечатление).

Первая система отвечает традиционному представлению о подлежащем и сказуемом как о главных членах предложения. Вторая – соответствует взгляду на подлежащее и дополнение как на «актанты», равноправно подчиненные сказуемому (*произвели* – главное слово, к нему равноправно относятся *все эти известия, удручающее впечатление* и *на меня*).

Если A и B – составляющие некоторой системы C, то выражение «B **непосредственно вложена в** A» (B **непосредственная составляющая** A, будем писать в этом случае $B \subset\subset A$) означает, что $B \subset A$ и в C нет составляющей, вложенной в A, содержащей B и отличной от A и от B.

Нетрудно видеть, что граф $\langle C, \subset\subset \rangle$ является деревом, корнем которого служит полная составляющая, а висячими узлами – точечные составляющие.

В ранее приведенных примерах была использована скобочная запись. На рис. 1 представлена система составляющих в виде бинарного дерева. Система составляющих C – это целое множество отрезков, в том числе отдельные слова и предложение целиком.

A и B – непосредственные составляющие.

A и D – нет, т. к. $(\exists B \in C : B \subset A \wedge D \subset B \wedge A \neq B \wedge D \neq B)$.

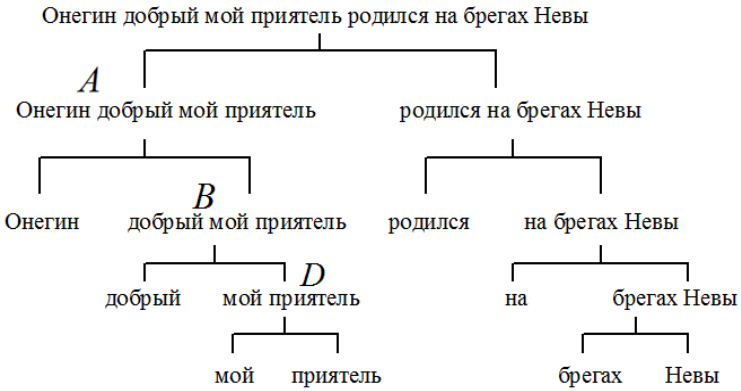


Рис. 1. Пример системы составляющих, представленной в виде дерева

Другими словами, основная идея грамматики составляющих заключается в следующем: всякая сложная грамматическая единица складывается из двух более простых и не пересекающихся единиц, называемых ее непосредственными составляющими.

Рассматривая составляющую как узел дерева составляющих, можно определить ее различные характеристики так же, как для обычных деревев (в теории графов). В частности, **ранг** – максимальное число дуг пути, соединяющих начальную вершину с выбранной. Ранги всегда определены, если граф направленный. Систему составляющих, имеющую бинарное дерево, называют **бинарной**.

Важными характеристиками составляющей являются также степень левого ветвления и степень правого ветвления. Известна гипотеза В. Ингве, согласно которой в естественных языках имеются специальные механизмы, обеспечивающие для большинства предложений ограниченность степеней левого ветвления, в то время как степени правого ветвления ничем в принципе не ограничены. Имеются ввиду конструкции следующего типа:

Вот кот,
Который пугает и ловит синицу,
Которая часто ворует пшеницу,
Которая в темном чулане хранится
В доме,
Который построил Джек.

Существуют, однако, языки, для которых это предположение об ограниченности степеней левого ветвления заведомо не подтверждается, например, венгерский.

Чтобы добавить в систему составляющих дополнительную информацию, рассматривается отображение этой системы составляющих во множество всех подмножеств некоторого конечного множества, элементы которого называются метками.

Упорядоченную тройку $\langle C, W, \varphi \rangle$ будем называть **размеченной системой составляющих**, где

C – система составляющих,

W – множество меток,

φ – отображение C в 2^W.

Составляющая, включающая более одного слова, называется **группой**, а слово, соответствующее корневому узлу в дереве, которое описывает группу, – **вершиной группы**.

Метки содержательно интерпретируются как разновидности синтаксических групп слов и словосочетаний. Обычно выделяют следующие виды синтаксических групп (фразовые категории):

- именная группа (группа существительного, ИГ; англ. noun phrase, NP) – возглавляется существительным;
- группа прилагательного (ГПрил; англ. adjectival phrase, AP) – возглавляется прилагательным;
- наречная группа (НарГ; англ. adverbial phrase, AdvP) – возглавляется наречием;
- предложная группа (ПрГ; англ. prepositional phrase, PP) – возглавляется предлогом;

– глагольная группа (ГГ; англ. verb phrase, VP) – возглавляется глаголом;
– предложение (П; англ. sentence, S).

Некоторые фразовые категории, в частности, именная группа и предложение, обладают свойством рекурсивности – способностью включать в себя составляющие той же фразовой категории.

Одна из проблем грамматики составляющих заключается в снятии неоднозначностей (синтаксической омонимии). В русском языке существует относительно свободный порядок слов в предложении, поэтому одному и тому же предложению будут соответствовать несколько синтаксических деревьев. Для таких случаев грамматика составляющих подходит не очень хорошо, гораздо удобнее использовать грамматику зависимостей.

1.2. Деревья зависимостей

В грамматике зависимостей порядок слов в предложении не важен. Главное – знать, от какого слова зависит каждое слово в предложении и каким типом связи обозначена эта зависимость.

Пусть x – произвольная непустая цепочка в словаре V, X – множество всех элементов цепочки x.

Произвольное бинарное отношение $\rightarrow$ на X такое, что граф $\langle X, \rightarrow \rangle$ является деревом, называется **отношением зависимости** (или **отношением синтаксического подчинения**) для x.

Другими словами, если считать слова элементами цепочки, то цепочка x – это последовательность слов в предложении, а X – это множество всех слов предложения.

Дерево зависимости можно естественно изобразить графически (рис. 2). Для дерева зависимости обычным образом определяются зависимость, группа зависимости, ширина дерева, ранг узла и дерева.

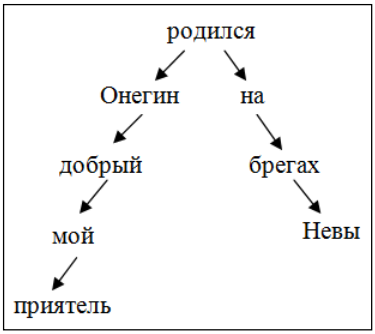


Рис. 2. Дерево зависимостей для предложения

Для анализа предложений естественного языка используются размеченные деревья зависимостей.

Размеченное дерево зависимостей для цепочки x – это четверка $\langle X, \to, Z, \psi \rangle$, где

$\langle X, \to \rangle$ – дерево зависимостей для x,

Z – конечное множество (его элементы называются **метками**),

ψ – отображение множества дуг дерева $\langle X, \to \rangle$ в Z.

В классе всевозможных деревьев зависимостей можно выделить подкласс, который содержит подавляющее большинство деревьев для предложений реальных языков. Это класс так называемых проективных деревьев. В них наблюдается некоторая упорядоченность слов.

Дерево зависимостей $\langle X, \to \rangle$ для цепочки x называется **проективным**, если для любых трех точек α, β, γ цепочки x из того, что $\alpha \to \beta$ и γ лежит между α и β, следует, что γ зависит от α.

Еще один важный класс деревьев зависимостей (являющийся расширением предыдущего) – класс слабо проективных деревьев.

Дерево зависимостей $\langle X, \to \rangle$ для цепочки x называется **слабо проективным**, если для любых четырех точек $\alpha, \beta, \gamma, \delta$ цепочки x из $\alpha \to \beta$ и $\gamma \to \delta$ следует, что пары α, β и γ, δ не разделяют друг друга.

При графическом способе изображения деревьев зависимостей слабая проективность равносильна возможности провести все стрелки так, чтобы никакие две из них не пересекались.

Содержательный смысл условий проективности и слабой проективности может быть охарактеризован приблизительно так: при выполнении этих условий слова, близкие синтаксически, близки и по

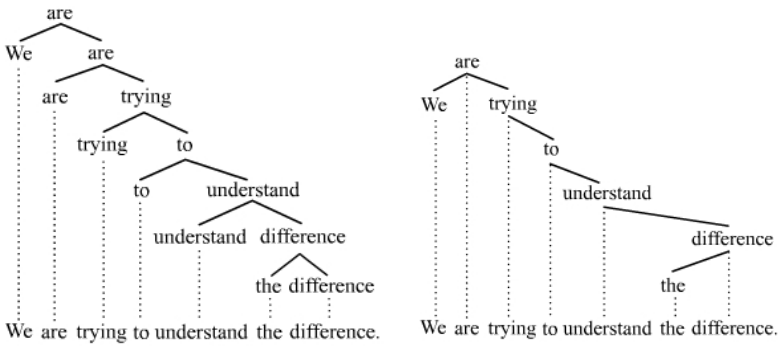
положению в тексте. При этом проективность обеспечивает «более тесную» текстовую близость.

В художественной литературе, особенно в поэзии, отклонения от слабой проективности и тем более от проективности вполне обычны. В русской разговорной речи возможен свободный порядок слов, и при определенной интонации предложение все равно может быть понятным и допустимым, т. е. условие слабой прективности также не соблюдается. Напротив, в научной и деловой прозе (по крайней мере, русской) деревья подчинения подавляющего большинства предложений слабо проективны и даже проективны.

Программы синтаксического анализа включают в себя косвенно или в явном виде фильтр на непроективность. Требование проективности синтаксической структуры предложения универсально для большинства индоевропейских языков.

Для приведенного на рис. 3 дерева зависимостей выполняются условия проективности и слабой проективности. К примеру, условие проективности выполняется, если возьмем α = *Онегин*, β = *добрый* или β = *мой*, γ = *приятель*; условие слабой проективности выполняется, если возьмем α = *Онегин*, β = *родился*, γ = *брегах*, δ = *Невы*.

На рис. 3 приведено сравнение деревьев составляющих и деревьев зависимостей.



Дерево составляющих Дерево зависимостей

Рис. 3. Сравнение деревьев составляющих и деревьев зависимостей

Основные отличия заключаются в следующем.

1. Деревья зависимостей минимальны по сравнению с деревьями составляющих, т. к. обычно содержат меньшее число вершин.

2. Деревья зависимостей отражают не реальный порядок слов в предложении, а только иерархию связей.
3. Грамматика составляющих опирается на бинарное разбиение (выделяется именная и глагольная группы), а в грамматике зависимостей глагол является основой всей структуры предложения.

Таким образом, грамматика зависимостей имеет два главных преимущества. Во-первых, условия проективности и слабой проективности в некоторых случаях помогают «предсказывать» связи между словами или исключать невозможные варианты разбора. Во-вторых, грамматика зависимостей хорошо отражает специфику языков с произвольным порядком слов и позволяет формально описать строение языковых конструкций.

Раздел 2. Принципы построения синтаксических анализаторов

Синтаксический анализ – один из важных этапов в системах автоматической обработки текстов. Рассмотрим подробнее работу синтаксических анализаторов для русского (на примере системы «АОТ») и английского языков (на примере системы «Link Grammar Parser»).

2.1. Проект «Автоматическая Обработка Текста» (АОТ)

Более детальная информация имеется на сайте проекта [5]. В проекте «Автоматическая обработка текста» обратим внимание на этапы фрагментационного и синтаксического анализа.

На этапе фрагментационного анализа

- осуществляется определение типа фрагмента;
- применяются правила снятия омонимии при определении типа фрагмента;
- применяются правила, устанавливающие иерархию.

Синтаксический анализ основан на последовательном применении синтаксических правил для построения синтаксических групп и соблюдении принципа проективности для синтаксических групп. Принцип проективности был сформулирован в предыдущем разделе. Суть его в следующем: из того, что две синтаксические группы пересекаются, следует, что одна лежит в другой. При графическом способе изображения деревьев зависимостей слабая проективность равносильна возможности провести все стрелки так, чтобы никакие две из них не пересекались.

Задача фрагментационного анализа состоит в том, чтобы выделить в предложении фрагменты (синтаксические группы) и установить иерархию на множестве этих единств, не используя семантической информации и информации о модели управления. Иерархия отражает тот факт, что в предложении некоторые фрагменты синтаксически зависимы от других.

На вход фрагментационного анализа поступает текст, разбитый на предложения. Каждое предложение разбито на слова и знаки препинания. Каждому слову приписана морфологическая информация (все возможные пары <грамматическая характеристика, лемма>, которым удовлетворяет слово).

На выходе имеется текст, состоящий из предложений, разбитых на линейно неразрывные фрагменты. На множестве фрагментов установлена иерархия, т. е. про каждый фрагмент известно, какие фрагменты в него непосредственно вложены и в какие он непосредственно вложен. Каждому фрагменту приписано множество типов.

1. Определение типа фрагмента

Типом фрагмента может быть ровно одно значение из списка: глагол в личной форме, краткое причастие, краткое прилагательное, предикативное слово, причастие, деепричастие, инфинитив, вводное слово. Если ни одно из предыдущих значений не подходит, в качестве типа фрагмента выбирается пустое значение.

Начиная с первого значения из списка, по порядку проверяется, есть ли в данном фрагменте слово этой части речи. Если такое слово найдено и у него нет омонимов других частей речи, то дальнейшие поиски прекращаются, и тип фрагмента – это значение, на котором сделана остановка. Если для данного значения из списка не нашлось неомонимичных (с точностью до части речи) подходящих слов, но есть омонимичные, тогда для фрагмента не устанавливается однозначно тип, а постулируется несколько вариантов, которые либо уничтожатся на уровне семантики, либо останутся в выходной структуре.

Пример

на этот раз она не права

Для этого фрагмента есть два варианта. Тип фрагмента – «краткое прилагательное» (*права* – ж.р., ед.ч. от *правый*). Тип фрагмента – «пусто» (*права* – и.п./в.п., мн.ч.; р.п., ед.ч. от *право*).

Пример

мои права забрали в милиции

Для этого фрагмента тип определяется однозначно, т. к. *забрали* – неомонимичный глагол в личной форме. Глагол в личной форме стоит в списке на первом месте, дальнейшие поиски возможных вершин фрагмента не ведутся.

2. Правила снятия омонимии при определении типа фрагмента

В том случае, если тип фрагмента не определяется однозначно или если имеется несколько слов, которые могли бы быть вершиной фрагмента, в системе АОТ применяется определенный набор правил для снятия омонимии. Для наглядности рассмотрим несколько примеров.

Первое правило

Если есть слово, один из омонимов которого – краткое прилагательное или краткое причастие, и во всем предложении нет существительного или местоимения в именительном падеже того же числа и рода (если число единственное), то этот омоним уничтожается.

Пример

Права он получил только с пятой попытки.

Так как во фрагменте нет существительного женского рода в единственном числе и в именительном падеже, значит, у слова *права* уничтожается омоним – краткое прилагательное женского рода единственного числа от *правый*.

Второе правило

Если во фрагменте есть неомонимичная предикация (глагол в личной форме, краткое прилагательное, краткое причастие, предикативное слово, причастие или деепричастие), то во всех остальных словах данного фрагмента уничтожаются омонимы этих частей речи.

Пример

Мыла на кухне она не нашла.

Так как *нашла* – неомонимичный глагол, то у слова *мыла* уничтожается омоним – прошедшее время, единственное число, женский род от глагола *мыть*.

3. Правила, устанавливающие иерархию – это правила, вкладывающие один фрагмент в другой или объединяющие фрагменты на основании стандартной информации о структуре фрагмента и сведений об отдельных словах. Рассмотрим пример правила, применяемого к фрагментам с типом «причастие».

Пример

1. ***Написанная в спешке*** *программа выполнила недопустимую операцию.*

Проверяем, не стоит ли причастие – вершина фрагмента – перед определяемым словом: осуществляем поиск существительного, совпадающего с причастием по роду, числу и падежу. Если такое существительное найдено, то данный фрагмент (причастный оборот) вкладывается в соседний правый, и осуществляется выход из правила.

2. *Программа,* ***написанная в спешке****, выполнила недопустимую операцию.*

Поиск в соседнем левом фрагменте существительного или местоимения, совпадающего с причастием в роде числе и падеже. Если такое слово найдено, то данный фрагмент (причастный оборот) вкладывается в соседний левый, и осуществляется выход из правила.

Любое простое предложение может быть «разорвано» причастными или деепричастными оборотами, или придаточными предложениями, которые, в свою очередь, тоже могут быть «разбиты» другими оборотами и придаточными. Иногда части цельного высказывания находятся на значительном расстоянии друг от друга, а глубина вложения теоретически не ограничена. Поэтому для определения семантики высказывания

особенно важен корректно проведенный процесс сбора воедино фрагментов предложения.

На этапе синтаксического анализа осуществляется построение синтаксических групп при помощи последовательного применения синтаксических правил (около 40 правил, все правила упорядочены). Причем соблюдается принцип проективности для синтаксических групп.

Пример

рубить дрова; есть кашу; читать книгу

Правило для построения групп: глагол + прямое дополнение (ПРЯМ_ДОП).

Рассматриваются цепочки: глагол + существительное в винительном падеже.

Главная группа: глагольная группа.

Пример

радостно сообщил; тихо и смирно сидит; хорошо знаю

Правило для построения групп: наречие + глагол (НАРЕЧ_ГЛАГОЛ).

Рассматриваются цепочки: одиночное наречие + одиночный глагол; группа наречий + одиночный глагол; одиночное наречие + группа инфинитивов.

Главная группа: глагольная группа.

Пример

краше тебя; уютнее твоего дома

Правило для построения отсравнительной группы (ОТСРАВН).

Рассматриваются цепочки: две группы, у первой группы главное слово – сравнительное прилагательное, у второй группы главное слово – существительное в родительном падеже.

Главная группа: группа прилагательного.

2.2. Синтаксический анализатор LINK

Синтаксический анализатор Link Grammar Parser [6] был разработан в 1990-х годах в университете Карнеги-Меллона. Данный подход отличается от классической теории синтаксиса. Система приписывает предложению синтаксическую структуру, которая состоит из множества помеченных связей (коннекторов), соединяющих пары слов. Link Grammar Parser использует информацию о типах связей между словами.

Анализатор имеет словари, включающие около 60000 словарных форм. Он позволяет анализировать большое число синтаксических конструкций, включая многочисленные редкие выражения и идиомы. Link Grammar Parser довольно устойчив, он может пропустить часть предложения, которая ему непонятна, и определить некоторую структуру оставшейся

части предложения. Анализатор способен работать с неизвестной лексикой и делать разумные предположения (на основе контекста и написания) о синтаксической категории неизвестных слов.

Анализ в системе проходит в два этапа.

1. Построение множества синтаксических представлений одного предложения. На этом этапе рассматриваются все варианты связей между словами, и выбираются среди них те, которые удовлетворяют
   - **критерию проективности** (связи не должны пересекаться);
   - **критерию минимальной связности** (получившийся граф должен содержать наименьшее число компонент связности. Компонента связности графа – некоторое множество вершин графа такое, что для любых двух вершин из этого множества существует путь из одной в другую, и не существует пути из вершины этого множества в вершину не из этого множества.).
2. Постобработка предназначена для работы с уже построенными альтернативными структурами предложения.

Получаемые диаграммы по сути являются аналогом деревьев подчинения. В деревьях подчинения от главного слова в предложении можно задать вопрос к второстепенному. Таким образом, слова выстраиваются в древовидную структуру. Синтаксический анализатор может выдать две или более схемы разбора одного и того же предложения. Это явление называется синтаксической синонимией.

Главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей (около 100 основных, причем некоторые из них имеют 3–4 варианта). В некоторых случаях тщательная работа над разными контекстами привела авторов системы к переходу к почти семантическим классификациям, построенным исключительно на синтаксических принципах. Так, выделяются следующие классы английских наречий: ситуационные наречия, которые относятся ко всему предложению в целом (clausal adverb); наречия времени (time adverbs); вводные наречия, которые стоят в начале предложения и отделены запятой (openers); наречия, модифицирующие прилагательные и т. д.

Из достоинств системы нужно отметить, что организация самой процедуры нахождения вариантов синтаксического представления очень эффективна. Построение идет не сверху вниз (top-down) и не снизу вверх (bottom-up), а все гипотезы отношений рассматриваются параллельно: сначала строятся все возможные связи по словарным формулам, а потом выделяются возможные подмножества этих связей. Это, во-первых, приводит к алгоритмической непрозрачности системы, поскольку очень трудно проследить за всеми отношениями сразу, а во-вторых – не к

линейной зависимости скорости алгоритма от количества слов, а к экспоненциальной, поскольку множество всех вариантов синтаксических структур на предложении из N слов в худшем случае равномощно множеству всех основных деревьев полного графа с N вершинами.

Последняя особенность алгоритма заставляет разработчиков использовать таймер для того, чтобы вовремя останавливать процедуру, которая работает слишком долго. Однако все эти недостатки с лихвой компенсируются лингвистической прозрачностью системы, в которой с одинаковой легкостью прописываются валентности слова, причем порядок сбора валентностей внутри алгоритма принципиально не задается, связи строятся как бы параллельно, что полностью соответствует нашей языковой интуиции.

Правила соединения слов описаны в наборе словарей. В таблице 1 приведены примеры различных типов словарей, с которыми работает анализатор.

Таблица 1

Примеры типов словарей, с которыми работает Link Grammar Parser

words.n.1	исчисляемые существительные в единственном числе (*book*)
words.n.2.s	существительные во множественном числе, оканчивающиеся на «s» (*books*)
words.n.2.x	существительные во множественном числе, не оканчивающиеся на «s» (*man – men*)
words.n.3	неисчисляемые существительные (*air*)
words.n.4	существительные, которые могут быть исчисляемыми или неисчисляемыми (*tea, coffee*)

Для каждого слова в словаре записывается, какими коннекторами оно может быть связано с другими словами предложения. Коннектор состоит из имени типа связи, в которую может вступать рассматриваемая единица анализа. Например, пометка S соответствует связи между субъектом и предикатом, O – между объектом и предикатом. Только основных, наиболее важных связей имеется более 100. Для обозначения направления связи справа к коннектору присоединяется знак «+», слева – знак «–». Левонаправленный и правонаправленный коннекторы одного типа образуют связь (link).

Например, если слову W1 приписан коннектор A+, а слову W2 – коннектор A–, то в синтаксической структуре предложения, состоящего из двух слов W1 W2, будет проведена связь A между словами W1 и W2. Предложение же W2 W1 не получит никакой интерпретации, поскольку

W2 приписан коннектор A–, который образует связь только влево, а слову W1 приписан A+, который образует связь только вправо (рис. 4).

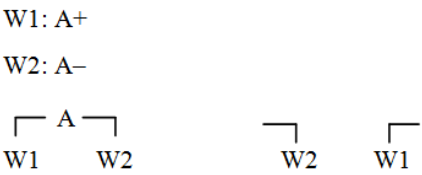


Рис. 4. Приписывание коннекторов словам в словаре

Одному слову может быть приписана формула коннекторов, составленная с помощью определенных связок.

& – несимметричная конъюнкция. Например, если слову W приписана формула A+ & B+ (обозначается «W: A+ & B+»), то некоторое слово X, с которым слово W образует связь A, должно стоять раньше по тексту, чем слово Y, с которым слово W образует связь B.

or – дизъюнкция. Если W: A+ or B-, то слово W может образовывать либо связь A вправо, либо связь B влево.

{} – факультативность. Если W: A+ & {B+}, то после того, как слово W образовало правую связь A, оно может образовывать или не образовывать связь B.

@ – неограниченность означает, что связь может строиться неограниченное число раз.

Пример

На рис. 5 изображен пример разбора предложения *I am sitting on a chair* анализатором Link Grammar Parser.

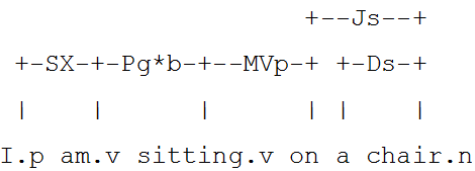


Рис. 5. Результат разбора предложения анализатором Link Grammar Parser

Использованы следующие обозначения:

.p – местоимение (pronoun);

.v – глагол (verb);

.n – существительное (noun).

В разборе предложения участвуют связи, перечисленные в таблице 2.

Таблица 2

Типы связей Link Grammar Parser

SX	соединеняет местоимение «I» с формами глагола «to be»
Pg*b	соединяет форму глагола «to be» с предлогом, прилагательным или причастием наст. вр. (**Pg**) или прош. вр. (**Pv**)
MV	соединяет глаголы и прилагательные с модифицирующими фразами, такими как именные группы, предложные группы, наречия, временные выражения и т. д. (**MVp** соединяет предлоги с глаголами; **MVa** соединяет наречия с глаголами и др.)
Js	соединяет предлог с его дополнением
D	соединяет определитель («a», «the», «some», «this», «each», «many», «much») с существительным (**Ds** соединяет сущ. с артиклем, **Dmc** – с «many», **Dmu** – с «much» и др.)

Отметим также недостатки Link Grammar Parser.

1. Практическое тестирование системы показывает, что при анализе сложных предложений, длина которых превышает 25–30 слов, возможен комбинаторный взрыв. В этом случае результатом работы анализатора становится «панический» граф, как правило, случайный вариант синтаксической структуры, с лингвистической точки зрения неадекватной.

2. Применение описанных выше идей затруднено для флективных языков типа русского, ввиду значительно возрастающего объема словарей, которые возникают в силу морфологической развитости флективных языков. Каждая морфологическая форма должна описываться отдельной формулой, где нижний индекс входящего в нее коннектора должен обеспечивать процедуру согласования. Это приводит к усложнению набора коннекторов и к увеличению их количества.

Раздел 3. Проблемы автоматической обработки текстов

В рамках синтаксического анализа предложения на сегодняшний день успешно решена и уже нашла применение в производстве задача автоматического выделения именных групп. В частности, при извлечении информации из текста нередко возникает необходимость распознать именованные элементы (имена людей, названия организаций, географические названия, события, временные и денежные обозначения и пр.) или определить отношения между выделенными сущностями. Однако созданию качественного автоматического синтаксического парсера препятствует ряд проблем. К ним относятся:

- морфологическая омонимия (part-of-speech tagging, word-category disambiguation);
- синтаксическая омонимия (syntactic homonymy);
- лексическая многозначность (word sense disambiguation);
- синтаксическая синонимия (syntactic synonymy);
- разрешение кореферентности (coreference resolution).

Для анализа предложения и тем более текста каждая из них в настоящее время остается не решенной. Познакомимся с ними подробнее.

1. Морфологическая омонимия – совпадение одной или нескольких грамматических форм слов, принадлежащих разным частям речи, в написании и произношении.

Пример

Я траву косил *косой*,
Дождик вдруг пошел *косой*.
Бросил я тогда косить
И на Стешу стал косить.
Ну а Стеша, ох, краса,
Как огонь ее коса!

Явление морфологической омонимии весьма негативно отражается на скорости работы программы синтаксического анализа. На длинных предложениях количество комбинаторных вариантов иногда достигает нескольких сотен, поэтому используются разного рода математические и лингвистические ухищрения, позволяющие избежать анализа всех комбинаторно возможных вариантов.

Для сравнения в системе АОТ, которая упоминалась ранее, скорость морфологического анализатора составляет 6000 слов в секунду, синтаксического – 300 слов в секунду.

Существует два подхода для снятия морфологической омонимии: детерминированный и вероятностный.

Детерминированный подход (развивается с 60-х годов) основан на локальном и глобальном синтаксическом разборе, синтаксических словарях и на правилах согласования слов.

Например, в предложении *Кошка пила молоко* есть три существительных, стоящих в именительном падеже, и одно из них может быть глаголом. Как быть? Есть правило: для слова, имеющего признаки и существительного, и глагола (у нас это *пила*) мы должны найти существительное, согласованное с глаголом в роде и числе или только в числе (в случае составного подлежащего или простого подлежащего во множественном числе). Таких правил имеется целый набор.

Вероятностный подход (развивается последние 20 лет) использует статистику совместной встречаемости грамматических признаков слов в больших корпусах, омонимия в которых снята заранее. Здесь предполагается использование методов машинного обучения системы (с учителем или без).

Практически все существующие алгоритмы снятия омонимии включаются в состав синтаксического анализа, что создает трудноразрешимое противоречие, когда для успешного снятия омонимии необходимы точные результаты синтаксического анализа, для получения которых, в свою очередь, нужно предварительно снять омонимию.

Приведем наглядный пример, демонстрирующий, как омонимы могут быть использованы для создания сложных конструкций.

Пример

Рассмотрим грамматически правильное предложение:

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.

Фразу можно перевести так: «Баффальские буйволы, запуганные (другими) баффальскими буйволами, запугивают баффальских буйволов».

Припишем каждому слову часть речи:

Buffalo.a buffalo.n Buffalo.a buffalo.n buffalo.v buffalo.v Buffalo.a buffalo.n, где

«a» – имя прилагательное, Buffalo = город в США, штат Нью-Йорк (здесь прил. баффальский по названию местности);

«n» – существительное, buffalo = буйвол;

«v» – глагол, buffalo = запугивать, озадачивать.

Данное грамматически верное предложение на английском языке не может корректно разобрать ни один из существующих в наши дни синтаксических парсеров.

2. Синтаксическая омонимия – неоднозначность, возникающая из-за неясности синтаксических связей между словами в предложении.

Пример

Преподаватель предложил прийти на зачет во вторник (*предложил во вторник* или *прийти во вторник?*).

Мать любит дочь (кто кого любит?).

Ему нужно отдать долг (*ему должны отдать* или *он должен отдать?*).

3. Лексическая омонимия и полисемия – совпадение одной или нескольких грамматических форм слов в написании и произношении; слова при этом принадлежат одной части речи, но имеют несколько различных значений.

Пример

Ключ – инструмент для открывания; *ключ* – источник воды.

Ключ подошел, дверь открылась.

Я напился из ключа.

Жизнь бьет ключом.

Пример

Bark – кора; *bark* – лаять.

The dogs bark at the tree.

В процессе работы над проблемой разрешения лексической многозначности было обнаружено большое количество трудностей, чаще всего обусловленных свойствами человеческой психологии и речи. Кроме того, значения слов сильно зависят от контекста.

Существует два подхода для снятия лексической омонимии: глубокий и поверхностный.

Глубокий подход обычно использует отнологии и предполагает наличие доступа к знаниям о мире. Несмотря на то, что при данном подходе знания о мире возможно хранить в удобном для компьютерной обработки формате, глубокий подход считается не слишком результативным на практике. Связано это с тем, что онтологии, как правило, содержат информацию о небольших областях жизни и не могут применяться к любого рода исследованиям.

При **поверхностном подходе** не ставится задача понимания текста, но производится анализ близлежащих слов. Например, рассмотрим слово *bass*, имеющее по крайней мере два значения: *рыба окунь* и *низкий голос*. Если рядом со словом *bass* в тексте присутствуют слова *sea* или *fishing*, скорее всего, что в данном случае имеет место первое значение, относящееся к рыбам. Подобные правила могут быть автоматически

извлечены при использовании корпуса текстов с размеченными значениями слов. Пусть этот подход и не покрывает по мощности предыдущий, но по эффективности на практике легко его обгоняет.

4. Синтаксическая синонимия – явление, при котором синтаксические конструкции имеют близкие значения и способны в определенных контекстах заменять друг друга.

Пример

John fell silent not knowing what to say.
John fell silent as he didn't know what to say.
John fell silent without knowing what to say.

The Byron poems.
The poems by Byron.
Byron's poems.
The poems of Byron.

5. Разрешение кореферентности – определение отношений между компонентами высказывания, в котором имена ссылаются на один и тот же объект. Благодаря кореферентности текст получается связным.

Пример

Книга лежит на столе. *Она* тяжелая.
Миша был зол на Диму, потому что *он (Дима)* украл *его (Мишин)* обед.
Миша был зол на Диму, поэтому *он (Миша)* украл *его (Димин)* обед.
Carol told *Bob* to attend the party. *They* arrived together.
If *they* are angry about the music, *the neighbors* will call the cops.

Существует несколько способ выражения референции:

- графический: *фотоконтент – фото-контент – фото контент*;
- транслитерация: *Yandex – Яндекс*;
- аббревиация: *ФМШ – физико-математическая школа – СУНЦ*;
- синонимия: *больница – госпиталь*.

Определение отношений между сущностями опирается на использование онтологий. Факты можно представлять как строки в таблице, а в столбцах размещать объекты и отношения между ними. Определение отношений между сущностями часто осуществляется одновременно с разрешением кореференций, либо с распознаванием именованных элементов.

Среди методов решения задачи распознавания именованных сущностей (Named Entity Recognition) особую популярность получили методы машинного обучения с учителем. В некоторой степени точность

обучающихся систем выделения сущностей зависит от наличия дополнительных ресурсов: словарей сущностей, коллекций размеченных текстов и т. д. Главный минус машинного обучения в том, что если ошибиться с выбором критериев, по которым происходит обучение, то полученный результат будет неудовлетворительным. Система будет часто делать ошибки, точность будет очень низкая, а усилий затрачено уже слишком много. Кроме того, инструменты для автоматической разметки русскоязычных текстов пока не очень развиты, а существующие не всегда легко доступны. Следует учитывать, что обучающая выборка должна быть достаточно объемная, размечена верно, единообразно и полностью. А это достаточно трудоемкий процесс.

Другим перспективным направлением в решении задачи распознавания именованных сущностей является использование онтологий и словарей. В частности, хорошим ресурсом имен известных людей и названий компаний является Википедия, в которой, помимо прочего, содержатся возможные варианты написания сущностей. Неоспоримым преимуществом данного направления является то, что словарные методы исключают ложные срабатывания и дают максимальную точность.

В настоящее время активно ведется работа над созданием гибридных методов, опирающихся на машинное обучение с учителем и одновременно использующих Википедию в качестве источника дополнительной информации о сущностях.

Раздел 4. Анализ тональности текстов

Анализ тональности текстов (определение эмоциональной окраски текстов, сентимент-анализ, Sentiment analysis, Opinion mining) – область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах.

В изучении этого вопроса следует понимать, что любое высказывание несет оттенок субъективности, является отражением реальности именно с точки зрения говорящего, т. е. присутствует некоторая модальность (модальность – семантическая категория, выражающая отношение говорящего к содержанию его высказывания). Автор всегда вкладывает в высказывание свой жизненный опыт, оперирует своими образами. Здесь можно вспомнить выражение: «Все в мире относительно».

Анализ тональности находит свое применение в

- социологии – сбор разного рода данных из соц. сетей (например, о религиозных взглядах);
- медицине и психологии – выявление психических особенностей и особенностей поведения конкретных пользователей и групп пользователей. Например, также можно определять тенденцию к возникновению депрессии у пользователей соцсетей в связи с политическими, экономическими трудностями;
- маркетинге – например, можно оценить фильм, ресторан, гостиницу и пр.;
- политологии – сбор свежих данных из блогов о политических взглядах населения.

Целью является нахождение мнений в тексте и определение их свойств, при этом каждое высказывание можно представить в виде набора признаков:

- **субъект тональности (автор)** – кому принадлежит высказывание;
- **объект тональности и его свойства (тема)** – о чем говорится в высказывании;
- **тональная оценка** – позиция автора относительно упомянутой темы.

В современных системах автоматического определения эмоциональной оценки текста чаще всего используется одномерное эмотивное пространство: позитив или негатив (хорошо или плохо). Конечно, гораздо более интересны случаи использования многомерных пространств.

Одним из минусов данного подхода является то, что эмоциональную составляющую документа не всегда можно однозначно определить, т. е.

документ может содержать как признаки позитивной оценки, так и признаки негативной. Поэтому в некоторых случаях удобнее считать, что необходимо осуществить классификацию полярности выбранного документа (высказывания) на три класса: позитивный, негативный или нейтральный.

Пример

Сообщение в Твиттере:

Программа «Глобальное образование»: наши студенты получат гранты на обучение в ведущих университетах мира.

Автор: Дмитрий Ливанов @DmitryLivanov

Тема: Выплаты российским студентам

Тональность: положительная

Например, у нас есть подборка рецензий на художественный фильм, и стоит задача определить, какие это рецензии – положительные или отрицательные. Эту задачу можно решить с помощью автоматической системы оценки тональности текста: система определяет характер рецензии, анализируя языковые средства.

Несмотря на то, что тональность является лишь одной из характеристик мнения, именно задача классификации тональности является наиболее часто изучаемой в наши дни. Это можно объяснить несколькими причинами.

1. Определение автора и темы являются гораздо более трудными задачами, чем классификация тональности, поэтому имеет смысл сначала решить более простую задачу, а затем уже переключиться на остальные.

2. Во многих случаях достаточно лишь определить тональность, т. к. другие характеристики уже известны. Например, если мы собираем мнения из блогов, обычно авторами мнений являются авторы постов, следовательно, определять автора нам не требуется. Также зачастую нам уже известна тема: например, если мы производим в Твиттере поиск по ключевому слову «Windows 8», то затем нужно лишь определить тональность найденных твитов.

Существуют различные подходы к определению тональности: основанные на правилах; основанные на словарях; использующие машинное обучение с учителем; использующие машинное обучение без учителя. В контексте задачи анализа тональности текстов методы машинного обучения без учителя, вообще говоря, показывают очень низкую точность. По этой причине их редко используют в системах сентимент-анализа несмотря на то, что они автоматизированы и не требуют данных для обучения. Рассмотрим оставшиеся три подхода.

1. Системы, состоящие из набора правил

Например, для предложения «*Я люблю шоколадные конфеты*», можно применить следующее правило: если слово («люблю») входит в положительный набор глаголов («люблю», «обожаю», «нравится» и т. д.), и в предложении не имеется отрицаний, то классифицировать тональность как «положительная».

Ввиду того, что этот метод наиболее точный, многие коммерческие системы используют данный подход несмотря на то, что он требует больших затрат, т. к. для хорошей работы системы необходимо составить большое количество правил. Зачастую правила привязаны к определенной теме (например, «ресторанная тематика»), и при смене темы («обзор фотоаппаратов») требуется заново составлять правила. Тем не менее, этот подход является наиболее точным при наличии хорошей базы правил, но совершенно неинтересным для исследования.

Преимущества подхода

– Наиболее точный.

Недостатки подхода

– Требует больших затрат для создания правил и поддержки их в актуальном состоянии.
– Классификатор привязан к определенной теме.
– Неинтересен для исследования (в нашем случае это недостаток).

2. Подходы, основанные на словарях, используют тональные словари, которые представляют собой список слов со значением тональности для каждого слова, как например, в таблице 3.

Таблица 3

База ANEW, переведенная на русский язык

слово	**тональность (1–9)**
счастливый	8,21
хороший	7,47
скучный	2,95
сердитый	2,85
грустный	1,61

Чтобы проанализировать текст, можно воспользоваться следующим алгоритмом: сначала для каждого слова в тексте определяется его тональность из словаря (если оно присутствует в словаре), а затем вычисляется общая тональность всего текста. Вычислять общую

тональность можно разными способами. Самый простой из них – среднее арифметическое всех значений.

Основной проблемой словарных методов считается трудоемкость процесса составления словаря: чтобы получить метод, классифицирующий документ с высокой точностью, термины словаря должны иметь вес, адекватный предметной области документа. Например, слово «большой» является положительной характеристикой по отношению к объему памяти жесткого диска, но отрицательной по отношению к размеру мобильного телефона.

Существует ряд тезаурусов, специально размеченных с учетом эмоциональной составляющей. Такие словари, описанные далее, необходимы компьютерным программам при анализе тональности текста.

WordNet-Affect (http://wndomains.fbk.eu/wnaffect.html) – это семантический тезаурус, в котором понятия, связанные с эмоциями («эмоциональные концепты», англ. «affective concepts») представлены с помощью слов, обладающих эмоциональной составляющей («эмоциональные слова», англ. «affective words»). WordNet-Affect состоит из такого подмножества синсетов WordNet, где каждый синсет, соответствующий «эмоциональному концепту», может быть представлен с помощью «эмоциональных слов».

WordNet-Affect является расширением WordNet Domain, где каждому синсету приписано не менее одной пометы предметной области (например, спорт, политика, медицина). Всего в иерархически организованную структуру было включено около двухсот предметных помет. Таким образом, WordNet-Affect был создан на основе WordNet для английского языка путем выбора и отнесения наборов синонимов (синсетов) к различным эмоциональным понятиям. В частности, синсеты глаголов, существительных, прилагательных, наречий, которые представляют собой описание эмоций, были вручную размечены с помощью специальных эмоциональных меток (affective labels, A-labels). Эти эмоциональные метки характеризуют различные состояния, выражающие настроения, эмоциональные отклики или ситуации, которые вызывают эмоции. Примеры таких эмоциональных меток на русском и английском языках приведены в таблице 4. Существуют версии WordNet-Affect для других языков.

Подобный словарь, позволяющий разделить эмоции только на три класса (положительную, отрицательную, нейтральную), SentiWordNet (http://sentiwordnet.isti.cnr.it/) широко используется в англоязычных автоматических системах сентимент-анализа.

Таблица 4

Примеры меток в словаре для описания эмоций

Эмоциональная метка	**Пример**
Эмоция (emotion)	сущ. гнев#1, гл. бояться#1 (fear)
Настроение (mood)	сущ. враждебность#1 (animosisy), прил. любезный#1J (amiable)
Особенность (trait)	сущ. агрессивность#1 (aggressiveness), прил. конкурентный#1 (competitive)
Когнитивное состояние (cognitive state)	сущ. замешательство#2 (confusion), прич. изумленный#2 (dazed)
Физическое состояние (physical state)	сущ. хворь#1 (illness), прич. утомленный#1 (exhausted)
Гедонический сигнал (gedonic signal)	сущ. боль#3(hurt), сущ. страдание#4 (suffering)
Ситуации, вызывающие эмоции (emotion-eliciting situation)	сущ. неловкость#3 (awkwardness), сущ. безопасность#1 (out of danger)
Эмоциональные отклики (emotional responses)	сущ. холодный пот#1 (cold sweat), гл. дрожать#2 (tremble)
Поступки (behaviour)	сущ. преступление#1 (offense), прич. замедленный#1 (delayed)
Отношение, позиция (attitude)	сущ. нетерпимость#1 (intolerance), сущ. оборона#1 (defensive)
Чувство (sensation)	сущ. холод#1 (coldness), гл. чувствовать#3 (feel)

Для русского языка создан словарь эмоциональной лексики (http://www.cir.ru/SentiLexicon/ProductSentiRus.txt), представляющий собой список из 5000 оценочных слов, извлеченных из коллекций отзывов в нескольких предметных областях (фильмы, книги, игры, телефоны, камеры). Словарь был разработан прежде всего для выявления предпочтений в маркетинговых исследованиях, а не для анализа общественного мнения в текстах социально-политической направленности.

Преимущества подхода

– Простой в применении.

Недостатки подхода

– Сильная зависимость терминологии от контекста.
– Трудоемкость составления словарей.

3. Системы, использующие машинное обучение с учителем

Методы классификации, использующие машинное обучение с учителем, рассмотрены в предыдущей главе. Напомним, что процесс классификации состоит из следующих этапов.

1. Индексация документов.

2. Уменьшение размерности пространства признаков.

3. Построение и обучение классификатора (с помощью методов: KNN, SVM, NB, DT и др.).

4. Оценка качества классификации.

Для каждого документа из обучающей выборки нужно указать «правильный» ответ, т. е. тип тональности (например, положительная или отрицательная); по этим ответам и будет обучаться классификатор.

На этапе индексации документов необходимо осуществить выбор признаков. В качестве признаков можно рассматривать кобинации *n*-грамм слов или *n*-грамм символов.

Пример (n-граммы слов)

Предложение «*Мне нравятся шоколадные конфеты*».

Набор униграмм: (*Мне, нравятся, шоколадные, конфеты*)

Набор биграмм: (*Мне нравятся, нравятся шоколадные, шоколадные конфеты*)

Их комбинация: (*Мне, нравятся, шоколадные, конфеты, Мне нравятся, нравятся шоколадные, шоколадные конфеты*)

Обычно униграммы и биграммы слов дают лучшие результаты, чем *n*-граммы более высоких порядков (триграммы и выше), т. к. обучающая выборка в большинстве случаев недостаточно большая для подсчета *n*-грамм высших порядков.

Пример (n-граммы символов)

Предложение «*Мне нравятся шоколадные конфеты*».

Набор четырехсимвольных *n*-грамм: (*«Мне », «не н», «е нр», « нра», «нрав», …*)

Символьные *n*-граммы могут быть полезны:

- при наличии орфографических ошибок в тексте – набор символов у текста с ошибками и набор символов у текста без ошибок будет практически одинаков в отличие от слов;
- для языков с богатой морфологией (например, для русского) – в текстах могут встречаться одинаковые слова, но в разных вариациях (разные род или число), но при этом не изменяется корень слов, а следовательно, и общий набор символов.

Несмотря на то, что такой способ может показаться слишком примитивным, т. к. на первый взгляд набор символов не несет в себе никакой семантики, этот метод иногда дает результаты даже лучшие, чем *n*-граммы слов. Если присмотреться, то можно увидеть, что *n*-граммы

символов соответствуют в какой-то мере морфемам слов, а в частности, корень слова («люб») несет в себе его смысл.

Также можно использовать дополнительные признаки, такие как части речи, пунктуация (наличие в тексте смайлов, восклицательных знаков), наличие в тексте отрицаний («не», «нет», «никогда»), междометий и т. д.

Символьные *n*-граммы применяются гораздо реже, чем *n*-граммы слов, но иногда они могут улучшить результаты. Для коротких сообщений, например, в Твиттере, больше подходят символьные *n*-граммы. Для текстов больше подходят *n*-грамм слов. В некоторых случаях полезна их комбинация.

После того как документы проиндексированы, следует осуществить выбор функции взвешивания. В качестве весовой функции часто используют TF-IDF. Строится она следующим образом.

Шаг 1. Вычисляется ***TF*** (*term frequency*) **частота термина** – оценка важности слова t в пределах одного документа d.

$$TF = \frac{C_{t,d}}{C_d}, \text{ где}$$

$C_{t,d}$ – сколько раз слово t встречается в документе d;

C_d – общее число слов в документе d.

Шаг 2. Вычисляется ***IDF*** (*inverse document frequency*) **обратная частота документа** – инверсия частоты, с которой слово t встречается в документах коллекции. IDF уменьшает вес общеупотребительных слов.

$$ITF = \log \frac{|D|}{D_t}, \text{ где}$$

$|D|$ – общее количество документов в коллекции;

D_t – количество всех документов, в которых встречается слово t.

Шаг 3. Итоговый вес термина t в документе d относительно всей коллекции документов вычисляется по формуле:

$$V_{t,d} = TF \cdot IDF .$$

Таким образом, большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Однако для задачи оценки тональности текстов в отличие от задачи поиска не слишком важны слова, которые часто повторяются в документе (т. е. слова с высоким TF). Поэтому для анализа тональности стандартная функция TF-IDF не дает хороших результатов. Гораздо лучше сюда

подходит бинарная функция взвешивания или модификация функции TF-IDF – так называемая дельта TF-IDF. Основная идея состоит в том, чтобы придать больший вес словам, которые имеют не нейтральную тональность. Первые два шага остаются прежними. На третьем шаге итоговый вес слова t в документе d вычисляется по формуле:

$$V_{t,d} = C_{t,d} \cdot \log\left(\frac{|N| \cdot P_t}{|P| \cdot N_t}\right), \text{ где}$$

$C_{t,d}$ – количество раз слово t встречается в документе d;

$|P|$ – количество документов положительной тональности;

$|N|$ – количество документов отрицательной тональности;

P_t – количество документов положительной тональности, в которых встречается слово t;

N_t – количество документов отрицательной тональности, в которых встречается слово t.

Преимущества подхода

- Простой в применении.
- В случае использования символьных n-грамм независимость от языка и терминологии.

Недостатки подхода

- Зависимость результата от выбора параметров метода.
- Трудоемкость подготовки данных для обучения и тестирования.

Задачи

Задача № 1

Экспериментальная компьютерная программа строит следующие русские словосочетания:

1. сосновый лес
2. это книжное содержание
3. цена оборудования
4. содержание пьесы тебя
5. этот новый гоночный автомобиль
6. чайная цена
7. дружеский совет молодого учителя
8. ученик того колдуна
9. гоночный автомобильный владелец
10. гаечный ключ меня
11. взмах крыла

Задание 1. Исправьте ошибки и сформулируйте их причину.

Задание 2. Одно из приведенных выше грамматически правильных словосочетаний может также являться результатом ошибки. Найдите его.

Задача № 2

Представьте себе, что вы услышали следующие предложения:

1. Анна слумная и будястая.
2. Ольга шотная и птиная.
3. Маргарита и Елена будястые, но лакимные.
4. Учитель лакимный и морчный.
5. Марина чапастая, но морчная.
6. Евгений не только фузый, но и бумсный.
7. Несмотря на то, что Петр был куластый, он все-таки порой становился совсем фузым.
8. Мария всегда такая толная и тачная, что на нее приятно смотреть.
9. Несмотря на то, что Андрей был немного бумсным, его все-таки смело можно было назвать толным.
10. Хотя Катерина чапастая, но с близкими людьми она всегда тачная.
11. Сегодня дети были шотные и не будястые.

Задание 1. Укажите, какое сочетание слов вы услышите вероятнее всего:

a) Алена чапастая и птиная.
b) Актеры были не только слумными, но и шотными.
c) Ребята привели домой собаку, которая была лакимная, но фузая.

Задание 2. Определите, какое из прилагательных имеет положительную коннотацию:

a) чапастый
b) бумсный
c) фузый
d) куластый

Объясните свои ответы. Примите к сведению, что звукосочетания в словах не влияют на их значения.

Задача № 3

Прочитайте предложение с необычным глаголом.

Когда чудовище *гапнуло* добычу, оно утащило ее в пещеру.

Задание 1. Дополните предложения различными формами, образованными от этой лексической единицы.

Раньше мы часто _____ с детьми по лесу.
Долгое время он работал _____.
Нас угощали _____ пирогом.
_____ – очень захватывающее занятие.

Задание 2. Есть ли альтернативные способы образования подобных слов? Приведите примеры других возможных вариантов и поясните свой ответ.

Задача № 4

Для изучения языка с практической или научной целью необходимо иметь точные сведения о том, как изменяются слова по падежам, числам и т. д. Эти сведения можно представить в виде специальных помет, отсылающих к соответствующим образцам склонения и спряжения. Существует особый тип словарей, в которых собрана информация такого рода – грамматические словари.

Ниже приводятся некоторые русские существительные в том виде, как они представлены в Грамматическом словаре русского языка:

в'едьма	жо 1 a	*похвал'а*	ж 1 b
дождь	м 2 b	*п'уля*	ж 2 a
кар'ась	мо 2 b	*таб'ун*	м 1 b
н'яня	жо 2 a	*черт'а*	ж 1 b

оде'яло	с 1 а	*шеф*	мо 1 а
ол'ень	мо 2 а		

Задание. Установите, что означает каждая из указанных помет, и определите, какие пометы имеют в Грамматическом словаре следующие слова: *блин, вещество, вождь, карта, княгиня, миля, множество, панцирь, слон, ярус*.

Задача № 5

Для наименования некоторых объектов в русском языке мы используем сочетания слов, состоящие из нескольких согласованных элементов, например, *почтовый ящик, полевые цветы, шоколадные конфеты, песочный замок.* В японском языке тоже существует такой способ образования лексических единиц (сочетаний прилагательного и существительного, в которых прилагательное уточняет описание существительного). В этом случае два слова, соединяясь вместе, претерпевают определенные изменения согласно правилам словообразования японского языка:

ikebana («живые цветы»): ike – жизнь, hana – цветы;

asagiri («утренний туман»): asa – утро, kiri – туман;

hoshizora («звездное небо»): hoshi – звезда, sora – небо.

Такие сложносоставные существительные в японском языке также в свою очередь могут быть частью других составных существительных, состоящих из трех или более частей.

Посмотрите внимательно на слова, приведенные ниже. Вы можете заметить, что порядок построения таких существительных изменяет их значение и форму слова в целом.

Задание. Ниже приведен список нескольких слов из японского языка с их значениями.

sakura – цветок вишни
shiru – суп
iro – цвет, цветной
kami – бумага
tana – полка
tanuki – енот
nise – поддельный
tsukuri – создатель
hako – коробка

Используя данные слова, дайте перевод следующих японских сложносоставных существительных:

nisetanukijiru
nisedanukijiru
irogamibako
irokamibako
nisezakuradana
nisesakuradana

Объясните свои ответы.

Задача № 6

Посмотрите на последовательности символов. Символ, находящийся в левой части строки, может переходить в один из символов в правой части строки. То есть трансформации одного символа в другой осуществляются согласно заданным правилам. Правила могут применяться до тех пор, пока это возможно. Последовательность символов в строке менять нельзя.

S → AB
A → ab
A → aAb
B → bcd
B → bBc

Задание 1. Какие из этих последовательностей можно закончить, используя эти правила, если начинать ее с “S”?

1. abdc
2. abbcd
3. aabbbcd
4. aaabbbcd
5. abbbbcdcc

6. aabbccdcc
7. aabbbbcdc
8. aaabbbbcd
9. aaabbbbcdc
10. aabbbbbcdcc
11. aaabbbbbbcdcc

Задание 2. Следующая строка не может быть сгенерирована согласно этим правилам: bbbbcdccc

К списку, приведенному выше, добавьте еще одно правило, которое сделает возможным создание этой последовательности.

Ответы, указания и решения

Решение задачи № 1

Задание 1. Ошибки содержатся в пунктах 2, 4, 6, 9 и 10. Исправить их можно следующим образом:

2. содержание этой книги
4. содержание твоей пьесы
6. цена чая
9. владелец гоночного автомобиля
10. мой гаечный ключ

Программе неизвестны притяжательные местоимения – вместо них она всегда ставит личные в родительном падеже. Кроме того, если программе удается обнаружить прилагательное, образованное от зависимого существительного, она ставит это прилагательное. Если такого прилагательного нет, ставится существительное в родительном падеже.

Задание 2. Результатом ошибки может являться также словосочетание под номером 7: «дружеский совет молодого учителя» – вместо «совет друга молодого учителя».

Решение задачи № 2

Прилагательные, используемые в задаче в качестве примеров, не существуют в русском языке. Но все их можно разделить на две группы: прилагательные с положительной коннотацией и с отрицательной. Это свойство прилагательных можно назвать «полярностью».

Каждое предложение связано с двумя или более прилагательными следующим образом: «X и Y» обозначает, что X и Y имеют одинаковую полярность. X, но Y обозначает, что они имеют противоположную полярность. Более того, «X и не Y» показывает противоположные полярности, «несмотря на то что / хотя X, Y» также указывает на противоположные полярности. «не только X, но и Y» или «такая X и Y, что» говорит о том, что оба прилагательных имеют одинаковую полярность.

Так, читая предложение о Марии, которая «всегда такая *толная* и *тачная*, что на нее приятно смотреть», мы понимаем, что оба этих прилагательных имеют положительную коннотацию. Структура предложения 8 подсказывает нам, что оно содержит прилагательные бесспорно одной полярности. Сопоставляя другие случаи употребления этих же прилагательных в других предложениях, мы можем сделать вывод об их положительной или отрицательной полярности.

В задаче присутствуют семь прилагательных с положительной коннотацией (*толный, тачный, шотный, лакимный, куластый, морчный, птиный)* и пять с отрицательной (*бумсный, чапастый, слумный, будястый, фузый*).

Задание 1. Только третье предложение содержит прилагательные правильных полярностей, которые сочетаясь, образуют логически верное предложение.

Задание 2. Только прилагательное *куластый* имеет положительную коннотацию.

Решение задачи № 3

Задание 1. Одно из возможных решений – дополнить предложения следующим образом.

Раньше мы часто *гапали* с детьми по лесу.

Долгое время он работал *гапщиком*.

Нас угощали *гапным* пирогом.

Гапание – очень захватывающее занятие.

Это наиболее вероятные формы слов, образованные от глагола «*гапать*» путем добавления различных аффиксов по аналогии с образованием подобных словоформ в русском языке. Так, для образования формы глагола прошедшего времени мы добавляем к основе глагола суффикс *-л-* (*ходили, играли*), для образования отглагольного имени существительного, обозначающего профессию, в русском языке к основе глагола часто добавляется суффикс *-щик / -чик* (*грузчик, носильщик*). Чтобы образовать отглагольное прилагательное, мы используем суффикс *-н-* (*копченый, печеный*).

Задание 2. Возможно большое количество различных вариантов ответов на данное задание. Например, «Долгое время он работал *гапником*» (по аналогии с «*охранником*») или «Нас угощали *гапочным* пирогом» (по аналогии с *креветочным*, *печеночным*).

Отглагольное существительное «*гапание*» также может быть заменено на *гапование* (ср. *фехтование*, *маринование*) или *гапка* (ср. *копка, стирка*).

Решение задачи № 4

Значение большинства помет установить просто: *ж* – женский род, *м* – мужской, *с* – средний, *о* – одушевленность; 1 – конечный согласный основы твердый, 2 – конечный согласный основы мягкий. Труднее всего установить значение помет *a* и *b*. Но сравнивая, например, слова *в'едьма* и *черт'а*, можно догадаться, что помета *a* относится к словам, у которых

ударение во всех формах стоит на основе – *в'едьма, в'едьмы, в'едьмами* и т. д., а помета *b* – к словам, у которых оно стоит на окончании (если оно не нулевое) – *черт'а, черт'ы, черт'ами* и т. д.

Ответ на задание:

блин	м 1 b	*м'иля*	ж 2 a
веществ'о	с 1 b	*мн'ожество*	с 1 a
вождь	мо 2 b	*п'анцирь*	м 2 a
к'арта	ж 1 a	*слон*	мо 1 b
княг'иня	жо 2 a	*'ярус*	м 1 a

Решение задачи № 5

Задание. Перевод данных японских существительных будет следующий:

nisetanukijiru – поддельный суп, сделанный из енотов
nisedanukijiru – суп, сделанный из поддельных енотов
irogamibako – коробка для цветной бумаги
irokamibako – цветная коробка для бумаги
nisezakuradana – полка для поддельных цветков вишни
nisesakuradana – поддельная полка для цветков вишни

В японском языке, когда мы соединяем вместе два слова, первое слово изменяет форму второго слова. Например, если слово hashi («палочки») идет до слова hako («коробка»), то образуется лексема, имеющая значение «коробка для палочек». Рассмотрим другой пример: добавление дополнительного компонента nuri («лакированный») перед существительным hashi изменит значение этого слова на «лакированные палочки».

Каждое простое (монокомпонентное) слово имеет две формы: основную форму, которая используется при употреблении данного слова отдельно от каких-либо других слов, и вариативную форму, которая иногда используется в составных существительных.

Основная форма	**Вариативная форма**
hako	bako
hana	bana
hashi	bashi
kami	gami
kiri	giri

sakura	zakura
shiru	jiru
sora	zora
tana	dana
tanuki	danuki
tsukuri	dzukuri

В вариативной форме изменяется первая буква, которая соответствует определенной букве в основной форме этого слова. Так, начальная буква *h* заменяется на *b*, *k* на *g*, *s* на *z*, *sh* на *j*, *t* на *d*, и *ts* на *dz*.

В дополнение к выше сказанному, необходимо обратить внимание на то, что при образовании сложносоставных существительных в японском языке можно вывести правила комбинирования базовых форм слов (обозначим их **a**, **b**, **c**) и соответствующих им вариативных (обозначим их **a**, **b**, **c**). Мы замечаем, что существительные, состоящие из двух элементов, строятся согласно следующему правилу:

a + b = ab

Слова, состоящие из трех элементов, имеют два различных варианта сочетания, каждый из которых изменяет значение слова. Если мы сначала составляем слово из **a** и **b**, и потом прибавляем к получившейся словоформе третий элемент **c**, то сложносоставное существительное образуется согласно следующему правилу:

(a + b) + c → ab + c → abc

Если мы сначала составляем слово из **b** и **c**, а потом добавляем элемент **a**, то сложносоставное существительное образуется по следующему правилу:

a + (b + c) → a + bc → abc

Таким образом, когда мы соединяем два слова, состоящих из одного или нескольких элементов в одно, мы пользуемся следующими правилами:

– первое слово используется в его начальной форме;

– если второе слово состоит из одного компонента (монокомпонентно), мы используем его вариативную форму;

– если второе слово является сложносоставным, оно не изменяет свою форму.

Решение задачи № 6

Задание 1. Могут быть получены строки 2, 3, 5, 7, 8, 10 и 11.

Полученные строки должны удовлетворять следующей формуле:

Na + Nb + Mc + bcd + Md, где

N – произвольное число повторений символов a, b;

M – произвольное число повторений символов c, d.
Задание 2. Есть два возможных варианта ответа.
К списку можно добавить следующее правило: S → B
Другой возможный вариант: A → Ø

Список литературы

Основная литература

1. Bing Liu. Sentiment Analysis and Opinion Mining. Morgan and Claypool Publishers, May 2012. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.9480&rep=rep1&type=pdf
2. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, 2008. 1024 p.
3. Link Grammar Parser. URL: http://www.abisource.com/projects/link-grammar/.
4. Автоматическая обработка текста. URL: http://www.aot.ru.
5. Батура Т. В. Математическая лингвистика и автоматическая обработка текстов: учеб. пособие / Новосиб. гос. ун-т. Новосибирск: РИЦ НГУ, 2016. ISBN 978-5-4437-0548-4. 166 с.

Дополнительная литература

1. Bo Pang, Lillian Lee. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. № 2. P. 1–135.
2. Воронцов К. В. Машинное обучение. Курс лекций. URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_%28курс_лекций%2C_К.В.Воронцов%29.
3. Кобзарева Т. Ю. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию. М.: РГГУ, 2015. 371 с.
4. Лукашевич Н.В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам // Электронные библиотеки. 2015. Т. 18. № 3-4. С. 88-119.