Компьютерная морфология в контексте анализа связного текста

Ермаков А.Е., Плешко В.В.
ООО “Гарант-Парк-Интернет”

Доклад посвящен ключевым проблемам морфологического разбора слов в тексте на русском языке. Затронуты вопросы, связанные с анализом неизвестных слов, омонимией, выделением в тексте сложных объектов-словосочетаний, и показано, как использовать формальные особенности текста и контекст для повышения точности разбора. Изложены принципы построения морфоанализатора, способного генерировать гипотезы о словоизменении с учетом различных допущений, в том числе эффективное кодирование словаря, реализация быстрого поиска, алгоритмы анализа неизвестных слов на основе правил и по аналогии с другими словами. В заключение обсуждается словарь словоизменения.

Введение

Компьютерная морфология необходима в прикладных системах, ведущих поиск и анализ информации на естественном языке. Основные функции, обеспечиваемые модулем морфоанализа: получение всех словоформ слова, постановка слова в заданную форму и получение грамматических характеристик словоформы. Вопросам компьютерной морфологии посвящено множество работ, однако все известные нам касаются лишь анализа отдельно взятого слова. При применении морфоанализатора к разбору связного текста возникает комплекс проблем, которые выдвигают дополнительные требования к модулю морфоанализа, на чем мы бы хотели заострить внимание. Хотя наш опыт касается русского языка, о котором далее и пойдет речь, основные из затронутых проблем являются универсальными.

Большая часть слов текста представляет неизменный фундамент языка и охватывается словарем в пределах 100 тысяч слов. Другая, более редкая, но не менее важная составляющая лексикона, постоянно пополняется и в принципе не имеет четко очерченных границ, прежде всего в части имен собственных и словообразовательных вариантов известных слов. К счастью, общие правила словообразования и словоизменения обладают регулярностью, что позволяет во многих случаях достаточно точно идентифицировать не только модель словоизменения, но и лексико-семантический разряд неизвестного слова. Возможность анализа неизвестных слов – необходимое качество морфоанализатора.

Реальность текста такова, что даже наличие сколь угодно “умного” морфоанализатора не всегда позволяет точно идентифицировать отдельное слово из-за присутствия в тексте омонимии. Для этого зачастую необходим учет контекста, как в рамках предложения, так и всего текста в целом: учет формальных особенностей написания, синтаксической организации фразы, кореферентных имен и правил их введения в текст. Отдельную проблему представляет выделение многословных единиц, таких как полные наименования организаций, которые должны обрабатываться как единое целое. В связи с этим модуль морфоанализа должен иметь гибкие настройки, которые позволяют эффективно использовать его в составе обработчика текста, порождая множество правдоподобных гипотез о словах при возможности омонимии, часть из которых подтверждается, а часть отвергается впоследствии на основании контекста.

1. Принципы построения компьютерной морфологии русского языка

Разработанный нами модуль RCO Morphology, выпускаемый под торговой маркой RCO в компании “Гарант-Парк-Интернет” (http://www.rco.ru/), предоставляет три возможности: точный анализ известного

слова по словарю, высоко достоверный анализ неизвестного слова на основе комплекса правил, вероятностный анализ посредством соотнесения с моделями словоизменения часто встречающихся слов.

В ходе разработки модуля были исследованы различные подходы к организации словаря словоизменения и правил морфоанализа с соответствующими способами поиска информации, в результате чего была найдена эффективная система кодирования русской морфологии, которая обеспечивает минимальный размер словаря при максимальном быстродействии. Так, общий объем словаря в 115 тысяч слов (более 3-х миллионов словоформ) и данных, необходимых для анализа неизвестных слов, не превышает 3-х Мбайт, обеспечивая при этом разбор 100 тысяч известных слов в секунду или около 20 тысяч неизвестных на процессоре P2-400.

1.1. Система кодирования словаря словоизменения

Неизменяемая часть слова, общая у всех его форм, представляется графической основой, возможно пустой (*идти–шел*). Вся оставшаяся часть слова описывается набором присоединяемых к основе окончаний. Список окончаний, упорядоченных в соответствии с грамматическими формами, образует парадигму (модель) словоизменения. В русском языке существует четыре типа парадигм: парадигма существительного (14 грамматических форм, включая два родительных и предложных падежа), парадигма прилагательного (31 грамматическая форма) и парадигма глагола (146 возможных прямых форм и 86 возвратных). К четвертому типу относятся все неизменяемые слова.

Большинство слов языка изменяется стандартным образом, т.е. имеет одинаковые окончания в одинаковых грамматических формах. Вследствие этого все различные парадигмы компактно представляются в виде строк в трех таблицах размером 14, 31 и 146 столбцов, а при основе слова хранится ссылка на соответствующую строку таблицы. Большинство окончаний в парадигмах также является стандартным, и для их хранения используется общая таблица окончаний, а в таблицах парадигм хранятся только ссылки на окончания. Таким образом, каждое слово словаря описывается основой и кодом парадигмы словоизменения - типом парадигмы и номером парадигмы в соответствующей типу таблице парадигм. Для построения заданной грамматической формы достаточно выбрать ссылку на соответствующее окончание из парадигмы и получить его строку из таблицы окончаний, после чего приписать к основе.

Для быстроты поиска при анализе все основы хранятся в виде дерева. Корневой узел дерева соответствует нулевой основе, каждый дочерний узел – возможной однобуквенной основе, каждый из следующих дочерних узлов – своей двухбуквенной основе с первой буквой, соответствующей узлу-родителю, и так далее. Структура дерева задается масками переходов, которые хранятся в каждом узле и определяют возможные последующие символы в основе. Дополнительно узел может содержать коды парадигм и частей речи, если соответствующая основа присутствует у каких-либо слов. Помимо основ, в форме деревьев хранятся все прочие дополнительные наборы строк, используемые в алгоритмах, которые описаны далее.

1.2. Точный морфологический анализ

Для быстрого поиска словоформы в словаре используется дополнительная структура – дерево окончаний, которое дублирует все окончания, представленные в таблице окончаний в форме строк. В каждом узле дерева хранятся коды всех парадигм, в которые входит соответствующее окончание.

Поиск словоформы реализуется следующим образом. Слово анализируется с конца на совпадение с деревом окончаний. Для каждого совпавшего окончания (включая пустое) оставшаяся часть слова ищется в дереве основ. В случае полного совпадения остатка слова с некоторой основой происходит сравнение кодов парадигм, хранимых при окончании и основе. Если обнаруживается общий код, значит, данное окончание возможно при данной основе и словоформа распознается, а из узла дерева основ извлекается код части речи. После этого продолжается поиск в дереве окончаний, так как возможно несколько вариантов разбора слова с разными окончаниями и основами. Если необходимо определить, каким именно грамматическим формам соответствует распознанная словоформа, достаточно провести поиск окончания в найденной парадигме – выполнить серию операций сравнения строк. Для оптимизации поиска в узлах дерева окончаний на каждую парадигму хранится номер первого вхождения окончания в парадигму и число повторений окончания в ней.

1.3. Морфологический анализ на основе правил

Основные правила могут быть применены, если слово имеет характерный аффикс. В состав словаря словообразования входит несколько сотен часто употребимых префиксов и постфиксов. Постфикс слова может включать неизменяемый суффикс и изменяемое окончание, вследствие чего хранение постфиксов аналогично хранению общего словаря, только вместо дерева основ используется дерево суффиксов.

Префикс задается с указанием части речи тех слов словаря, к которым он может присоединяться, что позволяет распознавать слова, образованные от известных, например, *трех* или *трех-* может присоединяться к прилагательному, а *авиа* – как к существительному, так и к прилагательному.

Постфикс задается с указанием парадигмы словоизменения и части речи, к которой будет отнесено слово, содержащее постфикс. Например, наличие суффиксов *енко* или *штейн* вместе с соответствующими окончаниями при слове, написанном с заглавной буквы, позволяет однозначно идентифицировать его как фамилию, наличие суффикса *горск* – как географическое наименование мужского рода, а *ович* и *евич* – как фамилию или отчество.

Отдельно проверяются специальные правила словообразования, например, образование наречий по шаблону *по-*ки*.

1.4. Вероятностный морфологический анализ

Алгоритм вероятностного морфоанализа отличается от точного тем, что вместо дерева основ используется дерево суффиксов, автоматически сформированное на этапе компиляции словаря. В дерево суффиксов включаются концы основ, встречающиеся не менее 30-ти раз в словах с одинаковой парадигмой изменения и имеющие длину не более 4-х символов при наличии остатка в основе не менее 3-х символов. Эмпирически определено, что эти величины обеспечивают наибольшую точность анализа.

При поиске для каждого окончания находится самое длинное совпадение конца оставшейся части слова с одним из суффиксов при условии наличия при суффиксе и окончании одинаковой парадигмы с частотой встречаемости не менее 30-ти. В качестве наиболее вероятной парадигмы изменения принимается та, при которой суммарная длина суффикса (возможно, нулевая) и окончания оказывается наибольшей. Эта величина используется в дальнейшем в качестве численной характеристики достоверности разбора. При наличии нескольких кандидатов равной длины приоритет имеет парадигма с большей частотой встречаемости в словаре.

2. Использование морфоанализа при разборе текста

Использование морфоанализа предполагает этап формирования правдоподобных вариантов разбора слов, которые могут быть подтверждены или опровергнуты на последующих этапах анализа текста, например, за счет использования алгоритмов полного синтаксического анализа или локального снятия омонимии.

2.1. Настройки морфоанализатора

Использование алгоритмов вероятностного морфоанализа позволяет определить морфологические характеристики неизвестного слова с высокой степенью достоверности в том случае, если известен его лексико-семантический разряд. Гипотезы о наиболее часто встречающихся разрядах неизвестных слов можно сформировать на основании формальных признаков. К таковым относятся:

- особенности написания (верхний/нижний регистр, русские/латинские/специальные символы) в сочетании с формальной позицией слова в тексте и типом текста (заголовок или начало предложения текста, поисковый запрос);
- повторение слова в тексте, когда в одном из упоминаний омонимия отсутствует.

В зависимости от комбинации указанных факторов, морфоанализатор обязан определять возможные варианты словоизменения с учетом следующих параметров:

- Среди каких категорий слов – имен собственных или нарицательных - производить поиск вариантов;
- Среди каких типов парадигм словоизменения производить поиск;
- Предполагать, что словоформа стоит в единственном числе;
- Предполагать, что словоформа стоит в нормальной форме.

Комбинации первых двух параметров позволяют задавать основные лексико-семантические разряды слов – например, имя собственное, изменяемое по парадигме прилагательного, соответствует фамилии. Последние два параметра могут использоваться прежде всего при анализе слов в поисковом запросе.

2.2. Формальные факторы, определяющие варианты разбора слова

Как показали эксперименты, морфоанализ слов в тексте наиболее корректно производится в соответствии со схемой, представленной ниже в виде таблицы.

Таблица 1. Схема генерации гипотез морфологического разбора

	Разряд \ Написание	***Все прописные***	***Первая прописная***	***Все строчные***
1	**Известное нарицательное**	всегда	всегда	всегда
2	**Известное собственное**	всегда	всегда	если не 1 и L > 5
3	***Неизвестная фамилия***	-	если (не 1 и не 2) или (не 1 и не первое слово предложения)	-
4	***Неизвестное прилагательное***	если не 1 и не 2 и L > 5	если не 1 и не 2	если не 1 и не 2
5	**Неизвестная организация**	если не 1 и не 2 и L > 5	если не 1 и не 2 и не 3 и не 4	-
6	***Неизвестное существительное (нарицательное)***	если не 1 и не 2 и L > 5	если не 1 и не 2 и не 3 и не 4 и не 5	если не 1 и не 2
7	***Неизвестное неизменяемое***	если не 1-6 (назв. организации или аббревиатура)	если не 1–6 (назв. организации или фамилия)	если не 1–6 (неизвестное существительное)

В каждой ячейке таблицы указаны условия, при выполнении которых для слова с написанием, указанным в заголовке столбца, делается попытка сформировать его вариант разбора в предположении о лексико-семантическом разряде, указанном в заголовке строки. L означает число символов в слове. Например, для слова, все буквы которого прописные, всегда делается попытка найти известное слово в словаре среди имен нарицательных и имен собственных, а в случае неудачи, если длина слова превышает пять символов, вероятностный морфологический анализ пытается построить три независимых варианта разбора слова в предположении, что оно является названием организации, нарицательным существительным или прилагательным. Если не один из достоверных вариантов разбора не был построен – предполагается, что слово является неизменяемым – названием организации или аббревиатурой (соответственно может иметь любой род и падеж плюс единственное число). Фамилия может быть целиком написана прописными буквами только в заголовке, однако малоизвестные фамилии (отсутствующие в словаре) включать в заголовок не принято.

Слово считается известным не только в том случае, если его удалось найти в словаре, но и если его удалось идентифицировать по правилам. В связи с этим в таблицу не включен разряд слов - названий географических мест, так как большинство из них точно идентифицируется на основе характерных постфиксов типа *ево*, *олье*, *евск*. По статистике большинство неизвестных слов, имеющих нехарактерные постфиксы и префиксы, относится к фамилиям, названиям организаций, существительным и прилагательным, образованным от названий географических мест. Неизвестные глаголы практически не встречаются.

Три отдельных класса представляют:

- слова смешанного написания, часть из которых представляет названия организаций (типа *ИНКОМе*);
- дефисные слова (не считая дефисных префиксов и наречий), обработку которых следует производить в предположении, что это либо два слова, относящиеся к одной части речи (утонение-приложение), либо одно известное слово, разделенное символом переноса;
- латинские слова, которые при наличии в предложении других русских слов следует рассматривать как неизменяемые имена собственные или части многословных наименований.

2.3. Учет контекста при разборе слова

Учет контекста необходим не только для отвержения неадекватных вариантов разбора слова, но и для порождения достоверных вариантов в тех трудных случаях, когда информации о написании слова и алгоритмов морфоанализа оказывается не достаточно для определения лексико-семантического разряда.

Способы учета контекста можно разделить на три группы.

Во-первых, это учет ближайшего формального контекста – имен, аббревиатур, сокращений и других знаков, позволяющих установить разряд слова в случаях типа: *А.Е. Ростов*, *г-н Банк*, *г. Шуцкий*, *ООО "Мелкора"* и др.

Во-вторых, это учет повторных (кореферентных) употреблений слова в тексте. Так, персоны и организации обычно вводятся в текст полным наименованием (*В.В.Волков*, *инвестиционная компания "Инкаструм"*, *город Нужкунь*), когда контекст позволяет точно идентифицировать разряд слова, и лишь потом именуются той или иной частью полного имени (фамилия, имя-отчество, название). Иногда слово само по себе стоит не в омонимичной форме (*Волковым=Волков*), хотя в другом месте текста может возникнуть омонимия (*Волков=волк*). На практике при повторном употреблении одной из словоформ имени собственного даже при наличии известных омоничичных словоформ следует отдавать предпочтение имени (совместные употребления типа *Волков охотится на волков* в норме не встречаются).

В-третьих, это учет грамматики языка, что требует локального или полного синтаксического разбора предложения. Обсуждение этих методов выходит за рамки доклада, однако стоит заметить, что ввиду высокой сложности они должны быть отделены от этапа морфоанализа и генерации гипотез, будучи применены впоследствии исключительно для отвержения неадекватных гипотез.

Задача учета формального контекста тесно пересекается с задачей выделения в тексте объектов – самостоятельных единиц-словосочетаний, выступающих как единое семантическое и синтаксическое целое. К таковым относятся разнообразные наименования, строевые элементы (многословные союзы, предлоги, вводные) и др. В большинстве случаев правила написания подобных объектов в тексте выходят за рамки общей грамматики русского языка.

Для выделения подобных объектов нами разработан модуль RCO Pattern Extractor, который сопоставляет цепочки лексем с образцами, заданными на формальном языке [2]. Мощный язык описания объектов позволяет оперировать как формальными особенностями написания слов, используя, в частности, язык регулярных выражений, так и всеми их грамматическими атрибутами, формируемыми модулем морфоанализа. Образцы сложных объектов могут строиться иерархически, включая образцы более простых, а грамматика языка описания образцов обеспечивает как бесконтекстное, так и контекстно-зависимое распознавание объектов. Выделенному объекту приписываются грамматические атрибуты, которые могут наследоваться от заданного слова в словосочетании, а также множество семантических атрибутов. Например, словосочетание *за границей*, после которого не следует слово в родительном падеже, будет выделено как объект, получающий грамматические атрибуты наречия. Объекту *инвестиционная компания ООО "Инкаструм"* будут присвоены все грамматические атрибуты, присутствующие у слова *компания*, и семантические атрибуты: Форма = *ООО*, Отраслевой тип = *компания* или *инвестиционная компания*, Название = *Инкаструм*. На основании семантических атрибутов можно определить возможные кореферентные наименования объекта в тексте и их лексико-семантический разряд, например, *Инкаструма.*

Обработка текста производится по предложениям, каждое из которых обрабатывается вначале лексико-морфологическим анализатором, а затем – модулем выделения объектов, причем в ходе морфоанализа используется информация о семантических атрибутах объектов, выделенных в предыдущих предложениях, - полный текстовый процессор содержит обратную связь.

Заключение

Мы попытались показать всю сложность проблем, встающих при морфологическом разборе. Эти проблемы вызваны омонимией и их решение упирается в разработку текстового процессора, работающего в тесном взаимодействии с модулем морфоанализа, учитывающего разнообразные особенности написания слов и их контекст, а потому весьма сложного даже без применения алгоритмов синтаксического анализа. Закладываемая в этот процессор система правил и соответствующая ей система принятия решений является разнородной и тяжело структурируемой, ее настройка может быть произведена только эмпирически, вследствие чего написание хорошего лексико-морфологического анализатора текста относится скорее к сфере искусства программирования, чем к науке. Много лет работая в этой области, мы и наши клиенты все

еще часто “натыкаемся” на ошибки при разборе даже корректного текста. На сегодняшний день не существует русскоязычного процессора, в котором все известные проблемы были бы решены до конца.

С другой стороны, задачи, связанные собственно с морфоанализатором отдельной словоформы, на сегодняшний день следует признать решенными. Основные из них мы здесь кратко описали - оптимальное кодирование словаря, реализация быстрого поиска, алгоритмы анализа неизвестных слов на основе правил и по аналогии с другими словами.

В завершение коснемся самого словаря словоизменения, который часто является предметом дискуссий. Существуют два крайних взгляда на предпочтительный объем используемого словаря, и оба они ошибочны.

В первом случае постулируется достаточный объем в пределах десятка тысяч слов-исключений из “нормальных” правил словоизменения, а остальные слова предлагается обрабатывать по этим самым “нормальным” правилам. Опыт показывает, что это не так. Только глаголов в русском языке мы насчитываем около двадцати пяти тысяч, причем для них бессловарный морфоанализ невозможен, так как большинство глагольных окончаний омонимично окончаниям других частей речи. В частности, массово присутствуют случаи, когда формы глаголов на *-ся* и *-сь* возвратными не являются, и выступают не в значении пассива прямой формы (*производиться*), а в самостоятельном значении (*торговаться*).

Во втором случае считается, что чем больше объем словаря, тем выше качество морфоанализа. Это действительно так, но только теоретически. Прообразом всех сегодняшних словарей послужил грамматический словарь Зализняка [1] объемом около 100 тысяч слов. Клоны этого словаря используются сегодня в различных коллективах, однако все они еще содержит ошибки среди редко употребимых слов. Так, у Зализняка информация о формах глаголов на “*ся*” отсутствует, вследствие чего во всех морфоанализаторах эти формы изначально строились как возвратные от прямых форм. Из восьми тысяч наиболее употребимых глаголов, на проверку которых пока хватило сил у наших лингвистов, около трех тысяч потребовало разведения прямой и возвратной форм в отдельные слова. Поэтому, когда разработчиками анонсируются словари объемом в 200 тысяч слов, это означает, что либо вторая половина слов изменяется регулярно и может быть проанализирована по правилам, либо в этой части словаря содержится много ошибок. Работы, сопоставимой с работой Зализняка, еще никто не делал, да это навряд ли нужно. Оставшаяся за рамками 100 тысяч слов часть лексикона является в принципе бесконечной, неустойчивой и может анализироваться только по правилам с учетом контекста употребления. Изредка появляющиеся устойчивые неологизмы и специальные термины, конечно же, должны пополнять словарь, но их количество в тексте не соизмеримо с количеством имен собственных.

Литература

1. А.А. Зализняк. Грамматический словарь русского языка. Словоизменение. - М., “Русский язык”, 1977, 880 с.
2. Ермаков А.Е., Плешко В.В., Митюнин В.А. Выделение объектов в тексте на основе формальных описаний. // Информационные технологии. - 2003. – N 12. – С. 1-6.