УДК 004.912:303.7

ИСПОЛЬЗОВАНИЕ СЛОВАРНОЙ ИНФОРМАЦИИ ПРИ АНАЛИЗЕ ТЕКСТА

К.К. Боярский, Е.А. Каневский, С.К. Стафеев

Описаны подходы к решению некоторых проблем, возникающих при компьютерном анализе русскоязычного текста. Затронуты вопросы, связанные со снятием лексической и морфологической неоднозначности, с выделением в тексте сложных объектов-словосочетаний и с использованием особенностей контекста для повышения точности разбора текста. Показано, что применение словарной информации может сыграть решающую роль при снятии как морфологической, так и частеречной и лексической омонимии.

Ключевые слова: анализ текста, лексема, морфология, омонимия, семантика, синтаксис, словарь.

Введение

Задаче компьютерного анализа текста на естественном языке посвящено множество теоретических и практических работ. Эти задачи, а именно – поиск документов, рубрицирование и аннотирование документов, диалог с компьютером, машинный перевод и построение баз знаний, – решали и решают различными методами. Однако их решение всегда начинается с морфологического анализа. Вопросам компьютерной морфологии посвящено множество работ, однако, как показал проведенный в 2010 г. форум «Оценка методов АОТ» [1], эта проблема до сих пор не решена окончательно.

Разработанный авторами морфолого-лексический анализатор TextAn [2] обеспечивал получение леммы и всех грамматических характеристик словоформы. В процессе участия в форуме 2010 г. авторами был проведен анализ коллекции текстов общим объемом более 950 тысяч словоформ. Из разобранных словоформ около 82,5% имели только одну лексему, остальным 17,5% соответствовали две и более лексем. В результате анализа неоднозначность по определению лексем уменьшилась в 11,5 раз и составила около 1,5%. В области снятия морфологической неоднозначности успехи несколько ниже. Если считать, что исходная морфологическая неоднозначность составляет около 50% [3], то TextAn снижает ее до 10,5% (конечно, возможные ошибки здесь не учитывались).

Как показало сравнение результатов работы TextAn с эталонным разбором («золотым стандартом»), основной причиной неполного снятия неоднозначности и даже появления ошибок являются ограниченные возможности системы по проверке согласования слов. При этом распространенное мнение о том, что слова в предложениях русского языка связываются в пределах ограниченного участка текста (от 3 до 7 слов), конечно, справедливы во многих случаях, но далеко не всегда. В этом случае для дальнейшего снижения неоднозначности было решено перейти к синтаксическому и частично семантическому анализу текста с максимально полным использованием имеющейся словарной информации. Такой анализ как раз и обеспечивает анализатор SemSin [4], который решает три основные задачи анализа:

1. получение грамматических характеристик словоформ;
2. построение дерева зависимостей;
3. снижение лексической неоднозначности.

Словари анализатора SemSin

Если морфологический анализ еще можно проводить без использования словаря [5], то для синтаксического и семантического анализа словарь крайне необходим. В качестве исходных лексических материалов используются словарь и классификатор В.А. Тузова [6]. Словарь Тузова основан на морфологическом словаре А.А. Зализняка [7], при определении его семантики широко использовался словарь С.А. Кузнецова [8]. К настоящему времени словарь значительно пополнен, для чего с успехом использовалась специально разработанная диалоговая система Adviser [9]. Поскольку при анализе текстов практически всегда обнаруживаются новые слова, отсутствующие в словаре, то он постоянно пополняется. Для внесения в словарь новых слов также используется Adviser.

В процессе разработки на основе исходного словаря была создана морфологическая база данных, содержащая свыше 177 тыс. лексем, распределенных по 1660 семантическим классам. В ней каждой лексеме приписаны морфологические характеристики, а также номер своего класса и актанты или валентности (для подключения зависимых слов) в виде падежей или предлогов с соответствующими падежами – например, !подТв, !наПред и т.д. Часто перед таким аргументом указаны допустимые классы слов, могущих их замещать. Около 14% слов в базе имеют две и более лексем, которые в большинстве случаев относятся к разным классам (классический пример: слову *коса* соответствуют три лексемы – *девичья коса*, *береговая коса* и *острая коса*).

Дополнением к морфологической базе служит база окончаний, насчитывающая около 1040 записей. Каждая запись содержит полный набор окончаний для определенного типа словоизменения с учетом изменения корневых гласных и суффиксов. Хотя типовых вариантов словоизменения не так уж много, но некоторые слова изменяются редким или даже уникальным способом. Например, для схожих слов *отобрать* и *обобрать* невозможно использовать одинаковый тип словоизменения: *от**о**брать – отберу*, *об**об**рать – оберу*.

В анализаторе достаточно эффективно решается проблема словосочетаний. Для этого служит специальная база фразеологизмов, которая обеспечивает разбор трех типов словосочетаний: неизменяемых (*в отличие от*, *а именно*, *в ту пору*), с изменяемым первым словом (*гвоздь программы*) и полностью изменяемых (*белая ворона*). В настоящее время эта база содержит более 4100 фразеологизмов и играет важную роль в снятии неоднозначности, особенно для составных предлогов, союзов и наречий.

Четвертым элементом словарного обеспечения является отдельная база предлогов, хранящая около 2000 сочетаний классов существительных, с которыми взаимодействуют предлоги, и названия связей с хозяевами предложных групп.

Схема работы анализатора SemSin

В состав анализатора SemSin входят 4 блока (рис. 1): словарь, морфологический анализатор, продукционные правила и лексический анализатор.

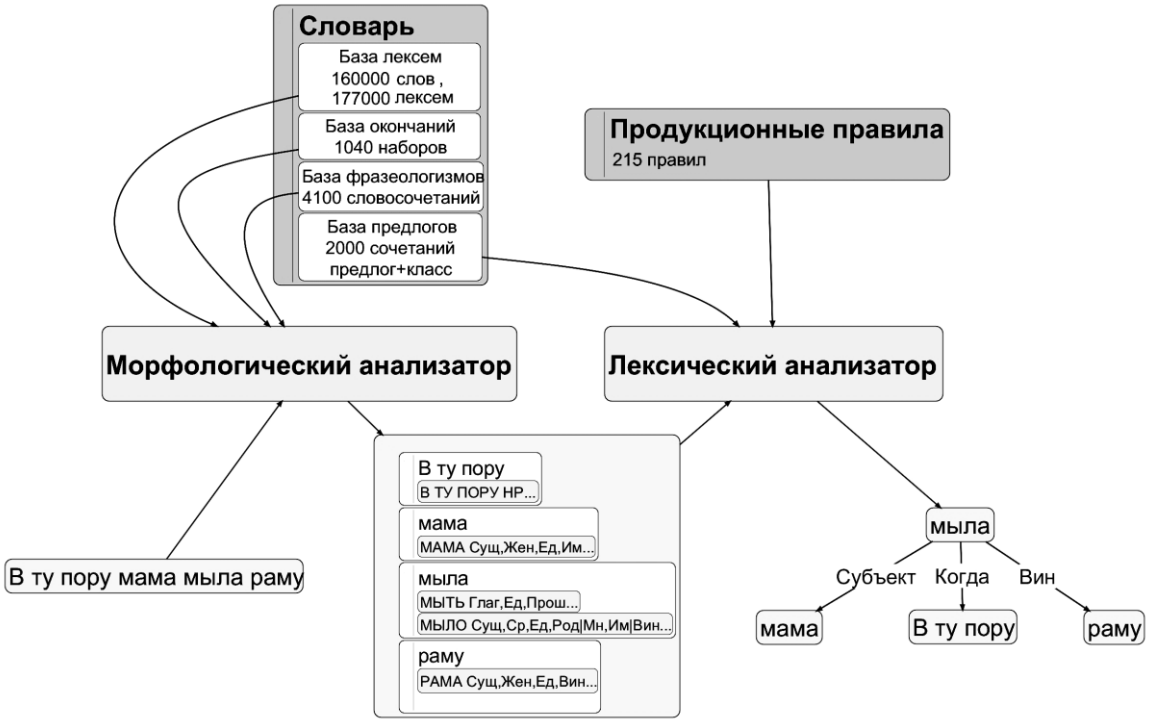

Рис. 1. Схема работы анализатора SemSin

На вход анализатора подается текст на русском языке, который считывается абзацами. Очередной абзац подвергается морфологическому анализу с выделением отдельных токенов (слов, словосочетаний, знаков препинания, чисел и т.д.). При этом каждому токену приписана не только морфологическая, но и синтаксическая (аргументы) и семантическая (класс по классификатору) информации.

Затем цепочка токенов обрабатывается с помощью системы продукционных правил, целью которых является преобразование линейной последовательности токенов в дерево зависимостей. Специфика задачи породила ряд особенностей составления этих правил. В частности, было решено, что снимать морфологическую неоднозначность и строить синтаксическое дерево зависимостей надо одновременно, поскольку эти задачи слишком переплетены между собой.

Система правил не может быть представлена в виде простого набора конъюнкций условий, так как результат сильно зависит от порядка применения правил. Выбранная последовательность исполнения основных групп правил такова: разбиение абзаца на предложения, сборка названий (включая имена собственные), обработка числительных и числовых токенов, сборка групп существительное–прилагательное и предлог–существительное, выделение причастных и деепричастных оборотов и придаточных предложений, нахождение подлежащих, подключение предложных групп к их хозяевам, объединение разделенных частей сегментов. На каждом этапе производится частичное снятие неоднозначности.

Рассмотрим на небольшом примере взаимодействие различных блоков системы. Предположим, что на вход подано предложение *В ту пору мама мыла раму*. Прежде всего, морфологический анализатор определяет леммы и грамматические характеристики слов. При этом выясняется, что простое разделение текста по пробелам не годится, так как словосочетание *в ту пору* должно рассматриваться как единый токен – составное наречие. Обнаруживается также, что токен *мыла* неоднозначен, он может представлять глагол *мыть* или существительное *мыло*, у которого неоднозначно определяется число и падеж.

Последовательность токенов подается на вход лексического анализатора, который управляется набором продукционных правил и использует информацию базы предлогов. В результате неопределенность полностью снимается, определяются типы связей между токенами и строится дерево зависимостей.

Выбранные принципы работы обеспечивают линейность зависимости времени анализа от общего количества слов текста. При этом скорость работы практически не зависит от длины предложения и составляет около 40 слов в секунду (процессор AMD Athlon64 3000+).

Снятие неоднозначности

Как указывалось выше, результат работы морфологического анализатора отличается очень высокой степенью неоднозначности. Многозначность различных характеристик тесно связана с таким понятием, как омонимия. Для ее детального изучения лингвисты выделяют разные типы и подтипы омонимии, которые в данном контексте не вызывают интерес. По этой причине в дальнейшем будем пользоваться термином *неоднозначность*. Процесс снятия или уменьшения неоднозначности часто называют *дизамбигуацией* (word sense disambiguation) [10].

Будем различать три типа неоднозначности.

1. Частеречная неоднозначность возникает при совпадении словоформ разных лексем, принадлежащих различным частям речи. Примерами типичных вариантов являются сущ./личная форма глагола (*жала*, *еду*, *казни*…), сущ./дееприч. (*моря*, *неволя*, *пошив*...), сущ./прилаг. (*больной*, *дорогой*…), дееприч./прилаг. (*скупая*, *строгая*…) и т.д.
2. Морфологическая неоднозначность, возникающая при неоднозначном определении грамматических характеристик словоформы одной лексемы: *точки* (точка, ед. род./мн. им./мн. вин.), *проводите* (проводить, 2-е л. мн. ч./повел.).
3. Лексическая неоднозначность – одна лексема, употребляемая в различных значениях, причем в этих значениях она способна сочетаться с разными словами (*орел*: птица/гордый человек/город/обратная сторона монеты/созвездие/фамилия…).

Для каждого из этих типов используются свои методы обработки. Конечно, зачастую неоднозначность снимается при согласовании слов (сущ. – прил. по роду, числу и падежу, глагол – существительное по числу и т.д.). Такое согласование проводится с помощью правил. Однако в ряде случаев этого недостаточно. Рассмотрим два фрагмента предложений: *Это привело к притоку значительных финансовых средств* и … *в области производства транспортных средств большой грузоподъемности*. В обоих примерах установлено, что слово *средств* стоит в родительном падеже множественного числа и возможности согласования исчерпаны. Однако в первом случае леммой является СРЕДСТВА (денежные), а во втором – СРЕДСТВО. Сделать правильный выбор оказывается возможным только с привлечением словарной информации. В описании аргументов прилагательного *финансовый* указаны семантические классы слов, с которыми возможно образование именной группы. Среди этих классов есть и такой: Физический_объект/Неодуш./Деньги[1], к которому принадлежит лексема *средства*. Во втором предложении оказывается возможным не только найти правильную лемму СРЕДСТВО, но и снять лексическую неоднозначность, установив, что в данном случае это лексема из класса техники (а не способ действия, вещество или пресса).

Возьмем теперь фрагмент *...члены экипажей космических аппаратов*. Также как и выше, анализ аргументов прилагательного позволяет снять неоднозначность и установить, что лексема *аппарат* означает некий прибор, а не совокупность органов организма или сотрудников учреждения. Но к какому слову должна подключаться эта лексема на дереве зависимостей? Дело в том, что и лексема *член* и лексема *экипаж* способны подключать слова в родительном падеже. Однако в словаре имеется информация о том, что преимущественно лексема *член* подключает слова, обозначающие живые объекты. В результате получаем правильный разбор: *члены* → *экипажей* → *аппаратов* → *космических*.

Предложные группы

Сложнее использовать словарную информацию при построении и подключении предложных групп. В качестве примера приведем несколько упрощенный результат разбора предложной группы *в море*:

```
в       В     ПР              $711(!Вин\!Пред)     …<Предлог>
море    МОР   м1  Ед. Пред. $111031($124~!Род)          …<Болезнь>
        МОР   м1о Ед. Пред. $12413/03000()       …<Фамилия>
        МОРЕ с2   Ед. Им. Вин. Пред. $12101(!Род)+$122422($1227\$123~!Род) …<Изобилие+Ландшафт>
```

Здесь «м1», «м1о» и «с2» – грамматические классы существительных (обозначения близки к используемым в [7]), «$» – признак номера класса, а «+» означает наличие двух (и более) лексем с одинаковыми морфологическими характеристиками.

[1] Здесь и далее названия семантических классов приводятся по [6].

Лексема *мор*, соответствующая имени собственному, легко отбрасывается по отсутствию прописной буквы, но остается еще четыре варианта морфологического разбора.

В аргументах предлога *в* указано, что он может подключаться к существительным в винительном и предложном падежах, так что все лексемы, соответствующие словоформе *море*, возможны. Однако подключение к хозяину предложной группы в предложении будет производиться разными связями в зависимости от семантического класса существительного. Информация об этом хранится в специальной базе предлогов. В результате формируются пять вариантов предложной группы (рис. 2).

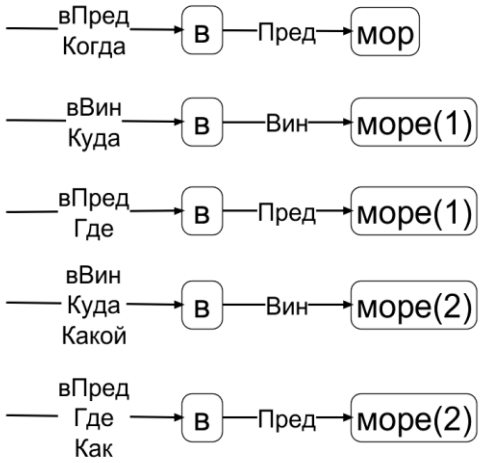

Рис. 2. Формирование предложных групп

Дальнейшее снятие неоднозначности происходит при подключении предложной группы к хозяину на основе анализа аргументов и их классов. Так, в предложении *Корабль находится в море* лексема *находиться* имеет следующие аргументы:

!Им,$15~!подТв,$1~!вПред,!Тв,$1~!Где,!При,!уИмея,$15~!сТв,!наПред[2].

Как видно, с входными связями рассматриваемой предложной группы совпадают аргументы *вПред* и *Где*, которые способны подсоединять слова любых классов. Таким образом, остаются третий и пятый варианты предложной группы. Следовательно, в данной фразе остается слово *море* в предложном падеже. Если же предложение имеет вид *Корабль вышел в море*, то аргументами лексемы *выйти* будут (в упрощенном виде, без указания большинства классов):

!Им,!Дат,!сТв,!поДат,!Откуда, $1224~!Куда,!наВин,!Тв,!заВин,!Изо.

Отсюда однозначно получается, что правилен четвертый вариант предложной группы, причем со связкой Куда. Следовательно, падеж слова *море* – винительный. Снимается также лексическая неоднозначность, остается лексема из класса ландшафтов.

Заключение

Одним из ключевых моментов получения однозначных морфологических характеристик слов и построения дерева зависимостей при автоматическом разборе предложений русского языка является наличие обширной и разноплановой словарной информации. Необходимо, в частности, привлечение семантической информации хотя бы на уровне классов слов и их сочетаемости. Использование этих данных в системе SemSin позволило снизить уровень неоднозначности более чем в 2 раза (по частеречной неоднозначности с 1,5% до 0,7%, по морфологической неоднозначности с 10,5 до 3–4%).

Литература

1. Ляшевская О.Н. и др. Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». – М.: Изд-во РГГУ, 2010. – Вып. 9 (16). – С. 318–326.
2. Каневский Е.А., Боярский К.К. Морфолого-лексический анализатор и классификация текста // Прикладная лингвистика в науке и образовании. Материалы V международной научно-практической конференции 25–26 марта 2010. – СПб: Лема, 2010. – С. 157–163.
3. Боярский К.К., Каневский Е.А. Разработка инструментария для полуавтоматической морфологической разметки текстов // Труды международной конференции «Корпусная лингвистика – 2008». – СПб: СПбГУ, 2008. – С. 83–88.

[2] Здесь $15 обозначает класс действий, $1224 –класс природных зон, $1 – любой класс.

4. Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб: ООО «МультиПроджектСистемСервис», 2011. – С. 233–237.
5. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006. – 304 с.
6. Тузов В.А. Компьютерная семантика русского языка. – СПб: Изд-во СПбГУ, 2004. – 400 с.
7. Зализняк А.А. Грамматический словарь русского языка. – М.: Русский язык, 1980. – 880 с.
8. Кузнецов С.А. Большой толковый словарь русского языка. – СПб: Норинт, 1998. – 1536 с.
9. Боярский К.К., Каневский Е.А. Проблемы пополнения семантического словаря // Научно-технический вестник СПбГУ ИТМО. – 2011. – № 2 (72). – С. 132–137.
10. Турдаков Д.Ю. Методы разрешения лексической неоднозначности // Программирование. – 2010. – № 6. – С. 3–27.

Боярский Кирилл Кириллович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кандидат физ.-мат. наук, доцент, Boyarin9@yandex.ru

Каневский Евгений Александрович – Санкт-Петербургский экономико-математический институт РАН, кандидат технических наук, ведущий научный сотрудник, kanev@emi.nw.ru

Стафеев Сергей Константинович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, доктор техничеких наук, профессор, декан, stafeev@phd.ifmo.ru