Предсинтаксический модуль в анализаторе SemSin

К.К. Боярский[1], Е.А. Каневский[2]

[1]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
[2]Санкт-Петербургский экономико-математический институт РАН
Boyarin@yandex.ru, Kanev@emi.nw.ru

Аннотация

Описаны принципы работы предсинтаксического модуля семантико-синтаксического анализатора SemSin. Использование этого модуля позволяет значительно уменьшить неоднозначность морфологического разбора еще до начала синтаксического анализа. Рассмотрены процедуры токенизации сложных лексических единиц и фразеологических словосочетаний. Приведены примеры снятия омонимии отдельных слов с использованием только ближайшего контекста.

1. Введение

Вопросам компьютерной морфологии посвящено множество работ, однако эта проблема до сих пор не решена окончательно. Известно, что при синтаксическом анализе текста, прежде всего, возникает проблема снятия частеречной и морфологической омонимии, так что вполне обоснована постановка вопроса о снятии омонимии до начала синтаксического анализа. Эта проблема привлекает большое внимание, и даже имеется попытка выделить такой тип анализа особо — как контекстный анализ [7]. Однако большинство считает это предсинтаксическим анализом [6, 12, 13, 16], хотя и пытается с его помощью решать различные задачи.

Наш опыт создания и отладки морфолого-лексического анализатора TextAn [10] и семантико-синтаксического анализатора SemSin [11] подтверждает целесообразность организации подобного предсинтаксического модуля. Будем опираться на анализатор SemSin, осуществляющий глубинный синтаксический анализ русскоязычных текстов. Анализатор использует расширенный вариант словаря В.А. Тузова [17], объем которого доведен до 184 тыс. лексем (около 165 тыс. слов). Имеется классификатор на 1660 классов. Анализ текста осуществляется под управлением продукционных правил. В процессе анализа предложения одновременно выполняются снятие грамматической и частеречной омонимии, сегментация предложения и построение синтаксического дерева зависимостей. Во многих случаях разрешается и лексическая омонимия.

Следует отметить, что, работая со словарем В.А. Тузова, мы используем для каждой лексемы номер ее класса и актанты или валентности (для подключения зависимых слов) в виде падежей (!Им, !Род, !Вин и т. д.) или предлогов с соответствующими падежами (!вВин, !наПред и т. д.). Часто перед таким актантом указаны допустимые классы слов, могущих их замещать. Около 14% слов в **базе лексем** имеют две и более лексемы, которые в большинстве случаев относятся к разным классам (классический пример: слову *коса* соответствуют три лексемы: *девичья коса* — класс «Голова_Волосы», *береговая коса* — класс «Ландшафт_Берег» и *острая коса* — класс «Утварь_Инвентарь»).

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.

Важным элементом словарного обеспечения является отдельная **база предлогов**, хранящая около 2000 сочетаний классов существительных, взаимодействуя с которыми предлоги вырабатывают различные связи с хозяевами предложных групп.

Большинство исследователей считает, что основной функцией предсинтакксического модуля является снятие омонимии. Отдельно отмечаются некоторые дополнительные возможности:

- определение устойчивых словосочетаний [12, 13];
- исправление некоторых ошибок морфологического анализа, обработка несловарных собственных имен и числовых групп [6, 13];
- разбор различных комбинаторных сложностей [13];
- графематический анализ, позволяющий выделить некоторые аббревиатуры, имена с

Интернет и современное общество: сборник научных статей XVI Всероссийской объединенной конференции IMS-2013, Санкт-Петербург, 9 - 11 октября 2013 г.

инициалами, даты и пр., а также разбить текст на предложения [16].

Рассмотрим подробнее функционирование модуля предсинтаксического анализа в системе SemSin. Заметим, что приведенные в статье примеры взяты или из НКРЯ [15], или из корпуса текстов конкурса синтаксических парсеров 2011 г. [14].

2. Фиксация результата морфоанализа

На вход анализатора подается текст на русском языке, который считывается абзацами. Очередной абзац подвергается предварительному анализу с выделением отдельных токенов (слов, знаков препинания, чисел и т. д.). Морфологический анализатор [11, 12], используя базу лексем, осуществляет разбор очередного слова, написанного на русском языке. Результат разбора выдается в виде леммы с морфологическими характеристиками, а также класса (или набора классов) с указанием соответствующих актантов. В случае омонимии выдается последовательность лемм. Следует отметить, что разработанный нами морфологический анализатор, в принципе, разбирает только те слова, которые имеются в словаре.

Сразу убираются крайне редко встречающиеся словоформы деепричастий и глаголов в повелительной форме (*для*, *при*, *моря*…).

Заметим, что работа анализатора SemSin в значительной мере основана на взаимодействии актантов и морфологических характеристик присоединяемых к ним слов. Однако используемый нами морфологический анализатор практически не оперирует с актантами. В связи с этим в ряде случаев актанты после морфологического разбора подлежат уточнению.

В автоматически полученных наречиях типа *географически* остаются параметры прилагательного. Модуль вырабатывает актанты наречия с внешней связью «Как».

Уточняется род для местоимений:

я, ты — "Муж.-Жен.";

кто, никто, кто-то, кто-либо, кто-нибудь, кое-кто — "Муж.";

что, ничто, что-то, что-либо, что-нибудь, кое-что — "Сред.";

его, ее, их —"Муж.-Сред.-Жен."

Для имен существительных:

– имена собственные и слова типа *президентство* (класс «Глава») используются только в единственном числе, если таковое есть (в дальнейшем это поможет разбору фразы типа *во время его президентства*);

– если лексема относится к классу «Титулы» (*король*), то она позволяет подключаться к предикату, стоящему во множественном числе (*его Высочество позволили*…).

Для имен прилагательных в сравнительной степени устанавливаются специальные актанты «!Род, !Сравн"».

Для глаголов:

– для словосочетания *не бывать* к актанту «!Им» добавляется «!Род»;

– для переходных глаголов с отрицанием (не страдательные причастия) к актанту «!Вин» добавляется актант «!Род» с теми же классами (*купить книгу – не купить книги*);

– для непереходных глаголов (не причастие и не имеет актанта «!Инфин») к списку актантов добавляется «!Род» (*я не набирался опыта*);

– для страдательных причастий актант «!Им» заменяется на «!Тв» с теми же классами, что обеспечивает согласование ролей (*Мама испекла пирог – Пирог, испеченный мамой*);

– для всех причастий осуществляется замена актанта «!Вин», обеспечивающая их присоединение к существительным по типу прилагательных.

3. Токенизация

Первым этапом при автоматической обработке текста обычно является токенизация, т. е. выделение минимальных линейных компонент текста, которые в дальнейшем рассматриваются как неделимые единицы. Рассмотрим некоторые вопросы, возникающие при токенизации, и способы их решения, реализованные в семантико-синтаксическом анализаторе SemSin.

Значительную сложность при токенизации представляют слова с дефисами [5]. Некоторые часто встречающиеся помещены в словарь. Это могут быть существительные имена собственные (*Абрау-Дюрсо, Исла-Гранде-Де-Терра-Дель-Фуэго, Исмаил-Заде*) или нарицательные (*контр-адмирал, истребитель-бомбардировщик*), прилагательные (*исправительно-трудовой, пакистано-индийский, светло-бежевый*), наречия (*подобру-поздорову*), местоимения или местоименные прилагательные с частицами (*друг-друга, чем-то*), аббревиатуры (*ун-т*, *ур-ние*) и т. д. Всего на 165-тысячный словарь таких слов около 4,5 тыс. и с ними легко справляется морфоанализатор. Однако словотворчество в данной области настолько распространено, что все варианты охватить невозможно. Рассмотрим некоторые варианты токенизации комплексов с дефисами, используемые в системе SemSin.

3.1. Слова с частицами

Самый простой вариант, если к имеющемуся в словаре слову прибавляется стандартная частица из списка «-ТО -КА -ДЕ -КО -ТА -ТЕ -С -ЛИБО -НИБУДЬ -ТАКИ»:

Пойдём-ка, Гриша, посмотрим, что там в чемодане лежит.

Правда-с, — отвечал Модест.

При морфологическом анализе частица просто отбрасывается, определяются грамматические характеристики основного слова, а затем частица присоединяется обратно. В отличие от НКРЯ, в SemSin из подобного комплекса образуется единый токен, лемма которого представляет собой лемму основно-

го слова плюс частица (*пойти-ка*), а все остальные характеристики берутся из основного слова.

3.2. Алфавитно-цифровые комплексы

При анализе токенов типа *АК-47*, *Ил-14* следует иметь в виду, что их первая часть может не совпадать, а может и совпадать с какой-нибудь словоформой языка. Но это совпадение носит случайный характер и не может служить поводом для заключений о морфологических и семантических параметрах.

Например, морфоанализатор находит, что слово *ил* имеет мужской род, именительный/винительный падеж и относится к классу «Почвы». Однако все составные токены такого типа рассматриваются в системе как несклоняемые существительные, а с точки зрения семантики принято, что они обозначают объекты техники. Конечно, такое допущение не всегда справедливо, особенно в части семантики, например, в предложении «*Сегодня приняты стандарты ТСО-95 и ТСО-90*». В большинстве случаев семантический промах не ведет к ошибкам при построении дерева разбора.

3.3. Сложные прилагательные

Рассмотрим обработку сложных прилагательных с дефисом на примере слова *серо-оливкового*. Морфоанализатор, анализируя это слово по частям, выдает три лексемы: краткое прилагательное среднего рода от леммы СЕРЫЙ, наречие СЕРО и прилагательное мужского/среднего рода в родительном/винительном падежах от леммы ОЛИВКОВЫЙ.

Определяется, что к первой части сложного слова относятся две лексемы, ко второй – одна. Производится согласование частей речи, причем убирается признак краткой формы первой лексемы. Анализируется регистр первой буквы второй части, т. к. возможны имена собственные (*Бур-Комаровский* – существительное, а не прилагательное). Собирается сложная лемма серо-оливковый, к которой приписываются грамматические характеристики второй части, а лишние лексемы удаляются.

3.4. Алфавитно-цифровые числительные

Обрабатываются сочетания цифра – цифра или цифра – буквенное окончание. Для порядковых числительных (*1-я, 2-го* и т. д.) комплекс обрабатывается просто по таблице окончаний с определением возможного рода и падежа. Образованному токену приписывается лемма с окончанием (*1-й*), как и для обычных порядковых числительных.

Сложнее разобраться в ситуации, когда после цифры и дефиса идет какое-то слово (*60-летнему*, *16-разрядным*). Для таких токенов делается попытка найти с помощью морфоанализатора лемму и грамматические характеристики второй половины. Если это удается, найденные характеристики приписываются всему токену в целом, если нет – токен помечается как неизвестный. Последняя ситуация может возникнуть, например, если в тексте упоминаются химические термины типа *5-ГТФ*.

3.5. Сложные существительные

Основная трудность при анализе сложных существительных, написанных через дефис, заключается в определении типа словоизменения и выявлении главной части слова. Изменяться могут как обе части слова (*баба-яга*), так и только первая (*акула-молот*) или только вторая часть (*альфа-лучи*). Кроме того, по семантике *акула-молот* – это акула, а *альфа-лучи* – это лучи. Поэтому для правильности дальнейшего анализа такие слова должны вноситься в словарь. Однако в системе SemSin предприняты некоторые меры для обработки сложных слов с дефисом, отсутствующих в словаре. Рассмотрим несколько примеров.

Наиболее типичный вариант – когда обе части сложного слова имеются в словаре (*крестьянами-кулаками*, *омега-лучами*). Морфоанализатор обрабатывает эти части независимо друг от друга, а затем специальная подпрограмма формирует составную лемму (*крестьянин-кулак*, *омега-луч*), причем грамматические параметры и семантический класс в настоящее время устанавливаются по второй части. То же происходит, если первая часть отсутствует в словаре или вообще записана латиницей (*usb-разъем*).

3.6. Сложные единицы измерения

Объединение сложных единиц измерения типа *куб. м*, *В/см*, *кг/кв. м* и т. д. в единые токены производится с помощью специализированных словарей. Неприятности здесь возникают только из-за омонимии. Так обозначения секунд, вольт, ампер совпадают с предлогами, сокращения *сек* и *мин* – с формами глаголов и существительных. Поэтому приходится рассматривать ближайшее окружение таких слов на предмет обнаружения идентифицирующих признаков: точки посреди предложения, написания слова с прописной буквы. Однако процент ошибок при разборе таких конструкций достаточно велик.

3.7. Составные числительные и даты

Если в тексте контактно стоят несколько слов, образующих составное числительное, для них сразу же выстраивается цепочка синтаксических связей (*двести двадцать второго*). Также на этапе предсинтаксического анализа связываются словосочетания, образующие даты, возможно содержащие аббревиатуры (*9 января 1905 г.*).

3.8. Неизвестные слова

Как бы обширен не был морфологический словарь, в тексте обязательно найдутся неизвестные слова. В системе SemSin сделана попытка спрогнозировать синтаксические и семантические значения таких слов, начинающихся с прописной буквы, на основе анализа их окружения. Приведем некоторые

примеры правил прогноза. Символ «X» будет замещать неизвестное слово.

Контактно слева от X стоит токен, состоящий из прописной буквы с точкой или частиц *фон*, *де*, *ван*, и др. тогда X — фамилия (с инициалом). То же если слева стоят слова *мистер*, *фрау* и т. д. В этом случае удается определить также род слова X. Если же слева стоит слово из особого «географического» списка (*долина*, *залив*, *звезда*), то X — название объекта.

X заключено в кавычки, слева (не обязательно контактно) имеется слово, обозначающее учреждение или предприятие, тогда "X" — название предприятия. Аналогично "X" рассматривается как название и в тех случаях, когда слева стоит слово, обозначающее произведение литературы и искусства или транспортное средство (*крейсер «Варяг»*).

С учетом возможных падежных окончаний X заканчивается на одно из определенного набора окончаний (*-ов*, *-ин*, *-ко*…). Рассматриваем X как фамилию, учитывая однако, что для окончаний типа *-ский*, *-ского*… возможны также прилагательные, выполняющие роль названия (*Зеленоградский*). Путем анализа окончаний производится определение возможных падежей [3].

X записано латиницей, рассматриваем его как название фирмы.

После токенизации производится разбивка текста на предложения [2].

4. Обработка фразеологизмов

Одним из путей уменьшения омонимии является широкое использование стандартных сочетаний слов — фразеологизмов [12, 13]. Используемые нами фразеологизмы можно разделить на 3 группы: неизменяемые (их в словаре 1730), с изменяемым первым словом (700) и полностью изменяемые (2070).

4.1. Обычные фразеологизмы

Слова, входящие в состав изменяемых фразеологизмов (*выйти замуж*, *гвоздь программы*, *антонов огонь*, *белая ворона*), в единый токен не объединяются, но сразу соединяются в лексическую группу подходящей связкой. При этом группе приписывается соответствующий семантический класс по классификатору В.А. Тузова [17]. Так, *гвоздь программы* не будут забивать в стену, а *белая ворона* получит класс «Человек_Личность». По сути, изменяемые фразеологизмы состоят из сложных для разбора словосочетаний и идиом. Известно более 8000 идиом [1].

Неизменяемые фразеологизмы по существу являются составными оборотами, в состав которых входят:

предлоги — *в зависимости от*, *в знак*, *в соответствии с*, *в течение*, *несмотря на*, *по мнению*, *по сравнению с* и др.;

наречия — *в конце концов*, *в основном*, *время от времени*, *в свое время*, *в то же время*, *до смерти*, *на самом деле* и др.;

союзы — *а также*, *в том числе и*, *вместо того чтобы*, *если бы*, *как только*, *не только*, *потому что*, *так как* и др.;

частицы — *все же*, *как бы*, *как раз*, *к тому же*, *едва не*, *вроде бы*, *вроде как*, *все ж* и др.;

вводные обороты — *к слову сказать*, *по вашему мнению*, *мягко говоря*, *главным образом*, *другими словами* и др.;

предикативные обороты — *не дай бог*, *лыка не вяжет*, *нечего было делать*, *не может быть*, *не составит труда* и др.

Достаточно полные их списки приведены в НКРЯ [15]: наречия и предикативные обороты — 2222, предлоги — 317, союзы — 159, частицы — 35 и вводные обороты — 217. При разборе в анализаторе SemSin они объединяются в один токен.

4.2. Фразеологизмы с двойным значением

Как известно, некоторые обороты могут играть в тексте двоякую роль — в частности, роль предлога или роль наречия в зависимости от следующих за ними слов. Таковы, например, обороты *в конце*, *в ожидании*, *в результате*, *в сторону*, *до конца*, *к концу*, *по пути* [2]. Как легко заметить, все эти словосочетания, выступающие в роли предлога, требуют после себя существительного (или местоимения) в родительном падеже и этим отличаются от наречия. Так, например, у Гончарова:

*Они прошли **до конца** аллеи молча*.

*Сказка не над одними детьми в Обломовке, но и над взрослыми **до конца** жизни сохраняет свою власть*.

*Прочитайте **до конца**, и вы увидите, что мне иначе поступить нельзя*.

Достаточно простое правило легко решает задачу распознавания, хотя данный предлог имеет два значения: места — Куда (Где) и времени — доКогда (Когда). После предлога стоит существительное или прилагательное в родительном падеже, иначе это словосочетание является наречием.

4.3. Особые фразеологизмы

Более сложная ситуация возникает в том случае, когда сочетание нескольких слов может быть оборотом, а может и не быть им.

Так, например, сочетание слов *в лице* может быть предлогом, если за ним следует существительное или прилагательное в родительном падеже:

***В лице** грозного родителя Викентьев представлял Нила Андреича*.

В этом случае при разборе данное словосочетание следует объединить в один токен. Это же сочетание может быть и предложной именной группой:

*Она мгновенно оставила его руку и изменилась **в лице***.

В таком случае объединения в один токен не происходит, каждое слово сохраняет свои индивидуальные параметры.

Аналогично анализируется словосочетание *в области* — в первом предложении это словосочетание является предлогом, во втором — это два отдельных слова (предлог *в* и существительное *область*):

*Он жил **в области** красоты и любви.*

*Размер пенсии ниже прожиточного минимума, установленного **в области**.*

Несколько более сложно анализируется словосочетание *в деле*, как это видно из приведенных примеров:

*Принять участие **в деле** спасения полярников.*

*Система успешно показала **себя в деле**.*

*Факт преступления был налицо…, например, **в деле** В. И. Засулич*

В первом предложении данное словосочетание, несомненно, является предлогом, а в двух последних — это два отдельных слова. Отсюда вытекает такое правило: словосочетание *в деле* является предлогом в том случае, если за ним следует прилагательное в родительном падеже или существительное в родительном падеже, которое не является фамилией (не относится к классу «Личность_ФИО»).

Словосочетание *в среднем* тоже имеет два значения, как это видно из приведенных ниже примеров:

*Крестьяне составляли здесь **в среднем** менее трети присяжных.*

*Как и следовало ожидать, к Пасхе поднялись цены на яйца — **в среднем**, на 2–4 рубля.*

*Это произошло **в среднем** течении реки Волга.*

В первых двух предложениях данное словосочетание, является наречием, в последнем — это два отдельных слова. Отсюда следует такое правило: словосочетание *в среднем* является наречием в том случае, если за ним следует знак препинания или отсутствует прилагательное или существительное в предложном падеже.

Рассмотрим теперь словосочетание *к счастью*. Это словосочетание может быть самостоятельным вводным оборотом или в его состав может быть включена предложная именная группа с предлогом *для*:

***К счастью**, нам это не грозит: у нас всегда найдется, что изменить к лучшему.*

***К счастью** для нашей авиации, я этим почти не занималась.*

Во всех случаях этот вводный оборот должен быть выделен запятыми в середине предложения или одной запятой, если он расположен в начале или конце предложения.

Однако в предложении

*Для многих людей жизнь с правильно выбранным партнером — это ключ **к счастью** и гармонии.*

— это два отдельных слова.

Следует отметить, что количество подобных фразеологизмов довольно велико, и они требуют специального исследования.

Таким образом, в результате использования фразеологизмов обеспечивается снятие неоднозначности и более правильный синтаксический разбор предложения, поскольку во многих случаях связи между словами во фразеологизмах определяются более верно, чем в случаях их отсутствия.

5. Обработка отдельных слов

Как уже отмечалось выше, снятие грамматической и частеречной омонимии производится в процессе анализа предложения при построении синтаксического дерева зависимостей. При возможности разрешается и лексическая омонимия.

Однако в ряде случаев возникает необходимость снять или хотя бы уменьшить омонимию до начала синтаксического разбора. Это становится особенно важным в том случае, когда попадаются словоформы, обладающие высокой степенью морфологической или частеречной омонимии. В ряде случаев анализ ближайшего окружения позволяет частично или полностью убрать эту омонимию и тем самым упростить дальнейший разбор. Естественно использовать для этого предсинтаксический модуль. В этом случае улучшается работа правил и повышается точность построения дерева зависимостей.

Следует отметить, что мы не претендуем на такой глобальный подход к анализу омонимов, который используется, например, в [12]. Мы пытались в этом модуле снять неоднозначность только у тех омонимов, которые попадались нам при экспериментальном разборе текстов, учитывая, что основная работа по снятию омонимии выполняется в процессе синтаксического анализа. Рассмотрим ряд примеров снятия омонимии, которые можно разделить на 4 группы.

5.1. Анализ омонимии по словоформам

Словоформа *перед* может быть предлогом, если после нее стоит слово в творительном падеже, иначе это существительное:

*Ставя **перед** собой задачу, можно работать без усилий и добиваться большего…*

*Юрасов стискивает зубы и, принуждая себя к неподвижности, соображает: спрыгнуть при такой быстроте нельзя, до ближайшей остановки ещё далеко; нужно пройти на **перёд** поезда и там ждать.*

Словоформа *нашей* может быть прилагательным, если спереди стоит предлог или сзади согласованное существительное, иначе это повелительная форма глагола:

*Всегда можно улучшить что-то в **нашей** жизни…*

***Нашей** метку на рубашку.*

Словоформа *порядка* – это предлог, если за ним стоит число или числительное, иначе это существительное:

*По официальным данным, в настоящее время в Китае насчитывается **порядка** 113 тыс. Internet-кафе.*

*Булыжная формула изменения **порядка** суммирования несколько более мудреная.*

5.2. Анализ омонимии по леммам

Слово *сорок* анализируется по лемме, поскольку числительное *сорок* имеет две словоформы, совпадающих со словоформами слова *сорока*:

*Но судно не тонуло: на нём был старый, опытный капитан, **сорок** матросов да представитель фирмы…*

*Мы, к счастью, были гарантированы от этой опасности, потому что промежуток между двумя дождями длится около **сорока** часов.*

*Возможно, ответ к задаче принесла на хвосте **сорока**, а может быть и нет…*

Очевидно, что в двух первых предложениях анализируемое слово является числительным, поскольку за ним следует существительное (или прилагательное) в родительном падеже. Если таковых нет, то это существительное. Если числительное *сорок* входит в составное числительное, то такая омонимия снимается раньше (см. раздел 3.7):

*Дозировка: приблизительно 150–300 мг за **сорок** пять минут до сна.*

*К сентябрю **сорок** четвёртого года клиника Павла Алексеевича вернулась в Москву.*

5.3. Анализ омонимии по сегменту

В некоторых случаях для снятия омонимии можно использовать и не расположенные рядом слова, если они обладают четко выраженными характеристиками.

Такая ситуация возникает, например, при анализе словоформы *по**т**ом* (пот) и наречия *пото**м***. Дело в том, что словоформа *по**т**ом* встречается только с теми глаголами, которые содержат актант «!Тв» с классом «Выделения». В противном случае это наречие *пото**м***:

*От этих людей пахло **потом** и нестиранными носками.*

*Индульгенция! Бенкендорфа обметало крупным, как градины, **потом**.*

*Успех приходит к тем, кто мыслит, а **потом** действует…*

Как видно, во втором предложении анализируемое слово и глагол расположены не контактно, однако программа находит глагол и осуществляет правильный выбор — это существительное.

Аналогичная ситуация имеет место при анализе краткого прилагательного *намерен*, прилагательного *намеренный* и глагола *намерить*. Их некоторые словоформы совпадают. Полностью снять омонимию не удается, но если контактно или поблизости имеется инфинитив, то это краткое прилагательное *намерен*:

*Этой политики «Парус» **намерен** придерживаться и впредь…*

*Как в Google, так и в NASA **намерены** попутно решить и ряд технических задач…*

5.4. Анализ омонимии по графематике

Анализ слова *по* характерен тем, что тут нельзя использовать ближайший контекст — надо использовать графематику. Если это слово не находится в начале предложения, то *по* (предлог), *По* (река) и *ПО* — аббревиатура (программное обеспечение).

Так при разборе предложения

***Перед** Новым Годом **по** решению правления компания закупила двадцать две лицензии на **ПО**.*

правила этой группы должны сработать 3 раза для разрешения омонимии:

Перед → предлог, так как следом стоит слово в творительном падеже;

по → предлог, так как слово начинается со строчной буквы;

ПО → аббревиатура, так как данное слово написано прописными буквами.

Аналогично разбирается слово *про*, которое имеет два значения: *про* (предлог) и *ПРО* — аббревиатура (противоракетная оборона).

*Поразившись и смутившись **про** себя, я ничего не ответила…*

*Первый шаг **—** создание **ПРО** Москвы.*

6. Заключение

Для тестирования было взято 372 предложения, содержащих 5385 слов. Предложения брались из корпуса текстов конкурса синтаксических парсеров 2011 г. [14] Специального отбора примеров, характеризующих работу именно модуля предсинтаксического анализа, не проводилось.

Исходно морфологической неоднозначностью из этих предложений обладало 1652 (30,1%) слова, для которых вырабатывалось 5300 вариантов морфологии. Лемма неоднозначно определялась для 673 (12,5%) слов.

Модуль предсинтаксического анализа уменьшил число неоднозначных лемм до 428, т. е. примерно на треть, а количество неоднозначных вариантов больше чем на четверть.

Наличие подобного модуля существенно повышает точность работы анализатора при построении дерева зависимостей. С другой стороны, за счет анализа нелокальных синтаксических и семантических отношений происходит дальнейшее снятие омонимии. В конечном итоге в этих предложениях остается только 42 (0,8%) неоднозначно определенные леммы и 144 (2,8%) слова с неоднозначной морфологией.

Литература

[1] Баранов А.И. Словарь-тезаурус современной русской идиоматики. Аванта+, 2004. 1135 с.

[2] Боярский К.К., Каневский Е.А. Разбиение текста на предложения // Дискуссия теоретиков и практиков. Научно-практический журнал. 2010. №1 (3). С. 135—137.

[3] Боярский К.К., Каневский Е.А. Автоматическое выявление фамилий в тексте // Интернет

и современное общество: Материалы XV Всероссийской объединенной конференции «Интернет и современное общество». СПб.: МультиПроджектСистемСервис, 2012. С. 195—198.

[4] Боярский К.К., Каневский Е.А., Клименко Е.Н. Морфологический анализ текста в системе MAZE-32 // Информационные технологии в гуманитарных и общественных науках. СПб.: СПб ЭМИ РАН, вып. 11, 2001. С. 1—8.

[5] Дорохина Г.В., Журавлёв А.О., Бондаренко Е.А. Исследование алгоритма морфологического анализа слов с дефисным написанием // Системы и средства искусственного интеллекта. ССИИ-2012: материалы международной научной молодёжной школы. Донецк: ИПИИ «Наука і освіта», 2012. С. 17—24.

[6] Епифанов М.Е., Антонова А.Ю., Баталина А.М., Кобзарева Т.Ю., Лахути Д.Г. Итеративное применение алгоритмов снятия частеречной омонимии в русском тексте // Труды международной конференции Диалог'2002. М., 2002. С. 119–123.

[7] Зинькина Ю.В., Пяткин Н.В., Невзорова О.А. Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды международной конференции Диалог'2006. М., Наука, 2006. С. 399–402.

[8] Каневский Е.А. Некоторые вопросы автоматической обработки текстов // Экономико-математические исследования: математические модели и информационные технологии. СПб.: Нестор-История. 2009. Вып. 7. С. 274—284.

[9] Каневский Е.А., Колпакова Н.В. К вопросу построения морфологического анализатора // Труды Международного семинара Диалог’99 по компьютерной лингвистике и ее приложениям. РосНИИ ИИ, 1999. Т.2. С. 98—106.

[10] Каневский Е.А., Боярский К.К. Морфолого-лексический анализатор и классификация текста // Прикладная лингвистика в науке и образовании. Материалы V Международной научно-практической конференции 26–27 марта 2010. СПб.: ЛЕМА, 2010. С. 157—163.

[11] Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор SemSin. URL: http://www.dialog-21.ru/digest/2012/?type=doc (дата обращения: 22.05.2013).

[12] Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды международной конференции Диалог'2002. М., 2002. С. 258—268.

[13] Кобзарева Т.Ю. Морфоанализ in vivo // Труды Международной конференции Диалог’2004. М., 2004. С. 286—291.

[14] 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». Вып. 9 (16). М.: РГГУ, 2010. С. 318—326.

[15] Национальный корпус русского языка URL: http://www.ruscorpora.ru/ (дата обращения: 9.07.2013).

[16] Сокирко А. В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) / Диссертация на звание к.т.н. М., 2001. URL: http://www.aot.ru/docs/sokirko/sokirko-candid-4-1.html (дата обращения: 9.07.2013).

[17] Тузов В.А. Компьютерная семантика русского языка. СПб.: СПбГУ, 2004.

Pre-syntactical module of the parser SemSin

K. K. Boyarsky, E. A. Kanevsky

The principles of work pre-syntactical module of the semantic-syntactic analyzer SemSin are described. Using of this module allows to significantly reduce the ambiguity of morphological parsing before the beginning of the syntactical stage. Tokenization procedures for complex lexical units and phraseological expressions are considered. The paper contains examples of removing ambiguity of certain words using only the nearest context.