ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ПРЕДСИНТАКСИЧЕСКОГО АНАЛИЗА РУССКОГО ТЕКСТА (ИСПА)

Р.Н. Афанасьев
Московский Государственный Институт Стали и Сплавов (Технологический Университет)
Россия, 119991, Москва, Ленинский проспект, д. 4.
e-mail: roman@afanasiev.ru

Т.Ю. Кобзарева
Российский Государственный Гуманитарный Университет
Россия, 125267, Москва, Миусская пл., д. 6
e-mail: Stam@rozenshtein.mccme.ru

Ключевые слова: интеллектуальные системы, анализ естественного языка, автоматический предсинтаксический анализ, типы морфологической омонимии, анализ морфологических омонимов, минимальные грамматические контексты, снятие омонимии по грамматическому контексту.

Предлагается интеллектуальная система предсинтаксического анализа русского текста (ИСПА), включающая в себя аппарат снятия морфологической омонимии (МО) и средства, позволяющие проводить исследования для пополнения базиса системы.
Один из наиболее значимых негативных факторов при автоматическом анализе текста - МО частей речи. Будучи весьма частой, МО уже на этапе сегментации предложения ведет к появлению огромного числа вариантов анализа. Представляется весьма полезной разработка ИСПА, снимающей МО и, соответственно, позволяющей резко сократить число вариантов на всех последующих этапах анализа.
Базисом служит построенная для РЯ классификация типов МО (в объеме словаря А.А.Зализняка). Для каждого типа МО построен алгоритм на основе списка диагностических контекстов - грамматических ситуаций (ДС), позволяющих разрешать неоднозначность, если нет потенциальной синтаксической омонимии. Все ДС объединены в словарь (СДС) и хранятся в Базе Знаний ИСПА.
Созданная ИСПА позволяет без участия программиста редактировать имеющиеся алгоритмы и содержимое СДС, добавлять новые типы омонимии, анализировать работу алгоритмов, накапливать статистические данные для усовершенствования системы. Ядро ИСПА представляет собой dll-библиотеку и может быть использовано в разных системах обработки ЕТ на РЯ..

Введение

Важным препятствием на пути решения задач автоматической обработки естественного текста (ЕТ) является необходимость преодоления языковой неоднозначности [1], т.е. умение исчислять и снимать неоднозначности на всех уровнях анализа ЕТ.

Этап предсинтаксической обработки, позволяющий снимать морфологические неоднозначности, крайне важен для любых систем, работающих с ЕТ, но пока не существует ни лингвистической, ни программной базы, охватывающей представительно проблемы этого уровня [2]. Главное отличие предлагаемого подхода состоит в том, что ЯДРО ИСПА строится на основе грамматических закономерностей РЯ и, в отличие от других систем, использует только морфологические характеристики слов, тривиальную грамматическую модель управления и линейный порядок слов в предложении [3,4,5]. Разрешение морфологических неоднозначностей осуществляется без построения семантических сетей

предложений, их сложного анализа и фильтрации, что существенно ускоряет процесс, делает его более универсальным и удешевляет конечное решение, что особенно важно при решении задач, где семантический анализ вообще не требуется.

Поиск наиболее простых грамматических условий, позволяющих снимать омонимию, показывает, что часто простые ходы дают статистически весьма представительный эффект. Например, тривиальное синтаксическое соображение, что два предиката (предикатива) или относятся к разным сегментам предложения, или сочинены, или, если один из них омоним, то он не предикат (предикатив), имплицирует при снятии такой весьма частой и существенной омонимии как [краткое причастие \ прилагательное (Abr) vs. наречие (D)], условие, что если в тексте между таким омонимом и безусловным предикатом (предикативом) нет ни оператора сочинения, ни границы сегмента – слов или знаков препинания, обеспечивающих возможность появления сочинительной конструкции, - омоним не может быть предикатом (предикативом): ***странно*** *на меня смотрит; мог с тех пор стать* ***совершенно*** *другим;* ***существенно*** *изменив; решение,* ***значительно*** *с этого времени пересмотренное*; и т.д. Это же грамматическое соображение прекрасно работает и при снятии омонимии [существительное vs. личная форма глагола], и - [существительное vs. деепричастие]: *на краю* ***села*** *к тому времени уже построили; немедленно взяв в руки* ***жгут****, нет у меня* ***клея****, поднялась* ***буря*** и т.д.

Однако, сколь бы ни были объемны вводимые условия и сколь подробно ни рассмотрена задача, при работе с естественным текстом практически невозможно a priori учесть все мыслимые в линейной структуре комбинаторные манифестации известных грамматически явлений.

Наибольшие проблемы при разработке подобных систем связаны с тем, что пополнение базисных грамматических «эталонов» сопряжено, как правило, с неизбежными изменениями, и часто существенными, программной реализации. В данной системе осуществлена попытка создания **лингвистически открытого** базиса работы с текстом.

Система предполагает возможность снятия на основе **лингвистически открытого словаря контекстов** всех типов омонимии частей речи в русском языке, исчисленных в построенном нами «Словаре омонимий частей речи» [3] в объеме «Грамматического словаря русского языка» А.А.Зализняка [6,7]

В результате анализа структуры диагностических ситуаций «Словаря диагностических ситуаций», разработанного на первом этапе исследования для самых синтаксически значимых видов омонимии, создан не требующий владения навыками программирования **мета-язык**, обеспечивающий лингвисту возможность пополнения словаря контекстов новыми грамматическими ситуациями, которые система самостоятельно вводит в программу.

Кроме того, система обеспечивает чрезвычайно важную в процессе лингвистической отладки обратную связь – лингвистический и программный контроль результатов построения новых правил и алгоритмов.

Мы сталкиваемся с проблемой неоднозначностей на каждом уровне анализа естественного текста. Несмотря на теоретически весьма вероятные неоднозначности поверхностно-синтаксического анализа, опыт реализации системы [1] показывает, что самым больным вопросом является именно омонимия частей речи. Так как каждый уровень анализа с его собственными неоднозначностями множит количество вариантов, даже небольшое число частеречных омонимов исходного текста приводит на следующих уровнях к порождению необозримого числа вариантов.

Использование разрабатываемой интеллектуальной системы предсинтаксического анализа позволит не только гораздо более продуктивно работать программным модулям следующих уровней поверхностно-синтаксического анализа, но и гораздо эффективнее

обозревать и решать собственно лингвистические проблемы неоднозначности сегментации и построения подчинительных и сочинительных связей.

Структура ИСПА представлена на рис.1.

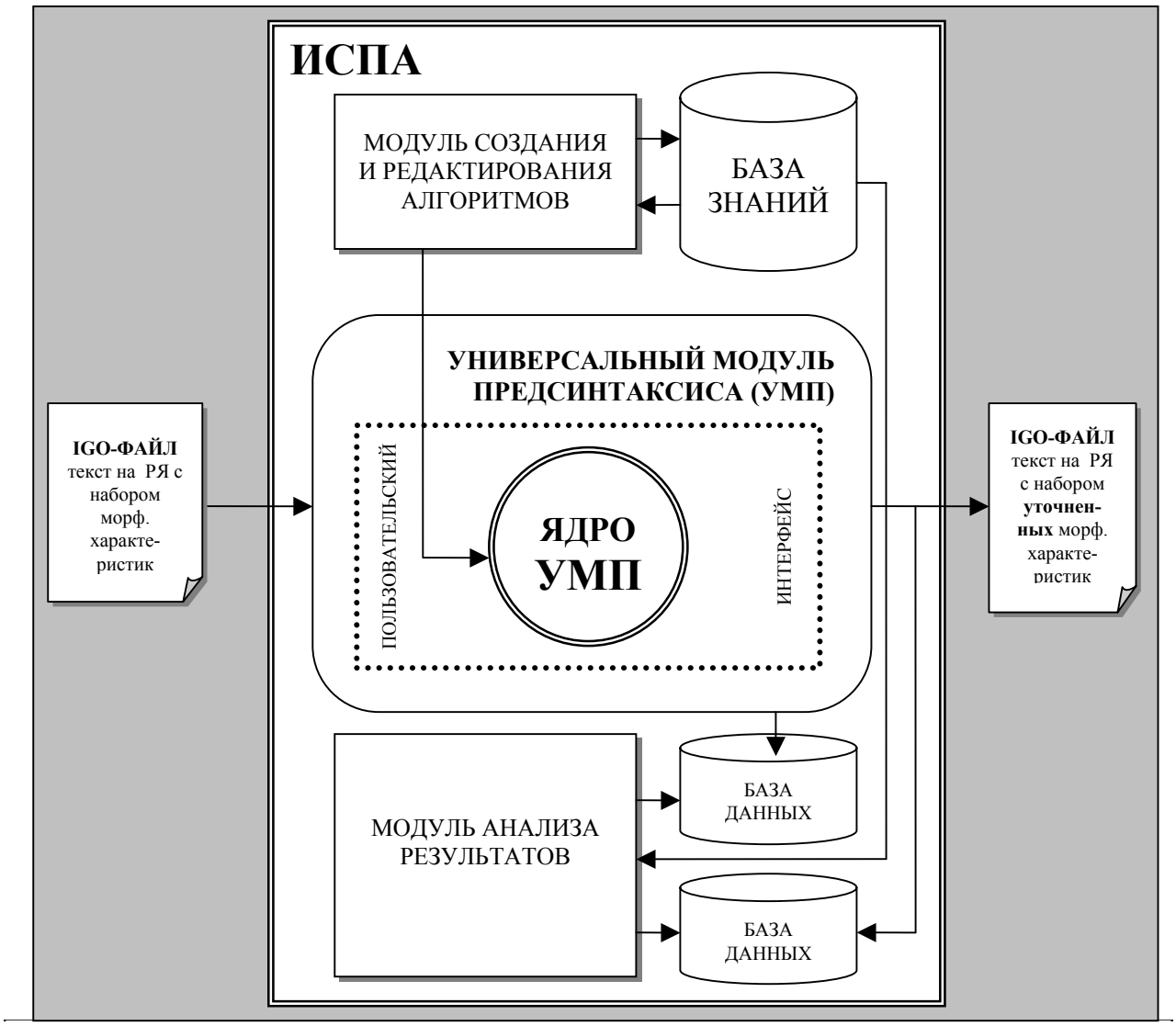

Рис. 1 Структура ИСПА

Далее приведено описание составных частей системы и механизмы их взаимодействия.

База Знаний (БЗ)

База Знаний является хранилищем структурированных и формализованных лингвистических знаний о русском языке в объеме необходимом для решения поставленной задачи. БЗ содержит:

- множество частей речи и их характеристики;

- множество цепочек словоформ, которые удобно задавать списками (сложные предлоги; вводные слова или словосочетания; устойчивые словосочетания, функционирующие как наречия, и т.д.);
- словарь диагностических ситуаций (СДС) [3];

В СДС хранятся сгруппированные по типу омонимии описания линейных структур минимальных грамматических контекстов (диагностических ситуаций), позволяющих идентифицировать часть речи омонима. Ситуации в СДС состоят из двух частей: 1) цепочка (может быть разрывной) компонент предложения - **ярлык ситуации**; 2) морфологические и\или синтаксические **условия,** позволяющие уточнить синтаксический смысл найденной по ярлыку ситуации и определить тем самым функционально-грамматическое значение омонима.

Ярлыки ситуаций задаются и хранятся в СДС в виде фреймов, в слоты которых записываются символы частей речи словоформ – слов и знаков препинания. Поясним некоторые из них: N - существительное, A - полное прилагательное, Vf - глагол в личной форме, D – наречие, Dv – деепричастие, P – предлог, CC – сочинительный союз, Zpt – запятая, T - точка. Слот вида <A/N> обозначает соответствующий омоним, в данном случае – <прилагательное/существительное>. Приведем основные типы фреймов, используемые в представлении ярлыков:

- простые фреймы: <P><A/N>;
- фреймы с разрывами: <A><N>…<P/Dv> (между <N> и <P/Dv> может находится любое количество различных словоформ);
- сложно составные фреймы: <L><GroupA><N/P>, где L – любая словоформа, а GroupA – фрейм вида: [<D>]<A_1>[<Zpt>||<CC>][[<D>]<A_2>]…[<CC>][[<D>]<A_n>] (в квадратных скобках обозначены факультативные слоты);
- различные комбинации выше перечисленных типов: [<T>||<Zpt>]<Dv/N><GroupA><N>…<Zpt>;

Условия, накладываемые на ярлыки и их окружение, представлены в СДС в виде продукционных правил. Приведем некоторые из них в общем виде:

- Если $W \in S$, то … (где W – часть речи или словоформа, S – заданное множество частей речи и словоформ)
- Если между W_1 и W_2 нет W_3, то …
- Если $S_1 = S_2$, то…
- Если непосредственно справа от W_1 $\urcorner \exists$ W_2, то…
- Если непосредственно слева от W_1 $\urcorner \exists$ W_2, то…
- Если $W_1 \underset{упр}{\cap} W_2 \neq \varnothing$ [1], то…
- Если $W_1 \underset{C12}{\cap} W_2 \neq \varnothing$ [2], то…
- Если $W_1 \underset{C6}{\cap} W_2 \neq \varnothing$ [2], то…
- Если $S = \varnothing$, то…

---

[1] Это обозначение является условной записью строимого в процессе анализа пересечения множества, характеризующего способность слова W_1 управлять определенными частями речи в определенных формах и множества соответствующих морфологических характеристик слова W_2. Например, у W_1 = *любуется* есть управление творительным падежом существительного. Если W_2 = *зданием* т.е. имеет творительный падеж в числе своих падежей, то $W_1 \underset{упр}{\cap} W_2 \neq \varnothing$.

[2] Как и в предыдущей сноске – условная запись, но здесь строится пересечение множеств падежей, характеризующих формы слов W_1 и W_2, в первом случае - с точностью до падежа и числа, во втором – с точностью до падежа.

- Если W_1 - единственный оператор между W_2 и W_3, то…
- ...

Пополнение и изменение содержимого БЗ ведется экспертом-лингвистом с помощью модуля создания и редактирования алгоритмов.

Модуль создания и редактирования алгоритмов

Данный модуль является инструментом для создания и редактирования алгоритмов разрешения морфологических неоднозначностей. С помощью специально разработанного предметно-ориентированного языка, он позволяет без участия программистов редактировать алгоритмы и содержимое СДС, добавлять новые типы омонимии и таким образом без участия программиста развивать ИСПА, улучшать, в частности, применительно к языковой специфике области, результаты ее использования.

Предметно-ориентированный язык оперирует терминами предметной области - прикладной лингвистики, а его внешней синтаксической формой являются множества фреймов, шаблонов условий, типов связей, списков исключений и т.д., которые лингвист использует при построении алгоритмов.

После формирования описания алгоритмов на проблемно-ориентированном языке модуль автоматически генерирует соответствующий код на языке Object Pascal и компилирует его в библиотеку DLL. При этом все изменения сохраняются в Базе Знаний.

Одной из важнейших проблем, возникающих при реализации описанного выше подхода является обеспечение генерации синтаксически и семантически правильного кода. Автоматическая проверка семантики в большинстве случаев вряд ли возможна (ошибки можно увидеть только после тестирования при анализе результатов работы ИСПА специалистом), поэтому самое важное - генерировать синтаксически правильный код. При этом условии компиляция всегда будет успешной. Для решения этой непростой задачи был разработан специальный интерфейс к данному модулю, который практически не позволяет пользователю ничего вводить вручную. Пользователь формирует все алгоритмы с помощью нажатия на специальные кнопки, установки флажков, выбора из списков и «перетаскивания» соответствующих объектов, что практически исключает появление синтаксических ошибок.

Универсальный модуль предсинтаксиса (УМП)

УМП стоит из двух частей: ЯДРА и пользовательского интерфейса, разработанного преимущественно для исследовательских целей. Программа, содержащая пользовательский интерфейс, динамически подключает ЯДРО. Достоинство такого подхода в том, что работа с визуальными компонентами сосредоточена во внешнем интерфейсе, а в ЯДРЕ содержится вся алгоритмическая часть УМП. Это позволяет использовать данное ядро внутри систем других разработчиков или напрямую без внешнего интерфейса. ЯДРО обновляется автоматически программой создания и редактирования алгоритмов каждый раз, когда эксперт вносит в БЗ какие-либо изменения.

Исходными данными для ЯДРА УМП является выход автоматического морфологического анализа в виде файла (IGO-файл) - см. рис.1. Каждой словоформе предложения в IGO-файле ставится в соответствие один или - в случае омонимии - несколько наборов грамматических характеристик. Каждая грамматическая характеристика содержит следующую информацию: порядковый номер слова в предложении, часть речи, морфологические характеристики формы (число, падеж, род, время, и т.д.) и грамматическую модель управления. В случае омонимии IGO-файл содержит не верную для данного контекста избыточную информацию, которая и порождает ложные варианты

анализа ЕТ на следующих уровнях анализа. По окончании работы УМП создается файл, по структуре аналогичный входному, но уже с разрешенной, где это возможно, неоднозначностью.

Более подробно алгоритм работы ЯДРА УМП описан в [3].

Все результаты работы УМП автоматически заносятся в Базу Данных Результатов.

База Данных Эталонов (БД-Э) и База Данных Результатов (БД-Р)

БД-Э и БД-Р имеют идентичную структуру и предназначены для автоматизации процесса проверки результатов работы УМП, сбора и хранения статистической информации. В БД-Э хранятся тексты-эталоны, размеченные лингвистом с помощью программы-интерфейса УМП. А также информация о том, по какой ситуации СДС должна быть разрешена та или иная омонимия заданных текстов.

В БД-Р автоматически заносятся результаты работы УМП.

Модуль анализа результатов

Данный модуль позволяет с помощью накопленных статистических данных (количество словоформ в тексте, количество омонимов в тексте, количество случаев снятой омонимии, диагностические ситуации, по которым была разрешена омонимия и т.д.) оценить эффективность работы ИСПА, автоматизировать процесс проверки правильности работы алгоритмов УМП и поиска ошибок на этапе разработки путем сравнения содержимого БД-Э и БД-Р.

Литература

1. Кобзарева Т.Ю., Лахути Д.Г., Ножов И.М. Модель сегментации русского предложения. // Труды Международного семинара Диалог'2001. Аксаково, 2001. Т. 2. С. 185-194.
2. Пащенко Н.А. Об одном подходе к проблеме снятия омонимии при автоматической обработке текстов на естественном языке. НТИ N4 1967.
3. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций. // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог'2002. Протвино, 2002. Т. 2. С. 258-268.
4. Кобзарева Т.Ю., Афанасьев Р.Н. Построение комплекса алгоритмов разрешения морфологических неоднозначностей на базе словаря диагностических ситуаций. // Обработка текста и когнитивные технологии. Казань: Отечество, 2001. Вып. 6. С. 82-87.
5. Kobzareva Tatyana, Afanasyev Roman. An Automatic Analysis of morphologically multivalued words as an Independent Module of the Surface-syntactical Analysis for Russian language (RL). // IV International Conference «Interactive Systems: Problems of Human-Computer Interaction» September 23-27, 2001. Ulyanovsk: USTU, 2001.
6. Зализняк А.А. Грамматический словарь русского языка. М: Русский язык, 1980.
7. Аношкина Ж.Г. Словарь омонимичных словоформ русского языка. М: Машинный фонд русского языка Института русского языка РАН, 2001. (http://irlras-cfrl.rema.ru:8100/homoforms/index.htm)

8. Леонтьева Н.Н. О предмете "прикладная лингвистика" (отвечая Н.В. Перцову). // Московский Лингвистический альманах "Спорное в лингвистике". М: Школа «Языки русской культуры», 1996. Вып. 1. С. 234-244.

Key words: intelligence systems, natural language processing, automatic pre-syntactical analysis, types of morphological homonimy, analysis of morphological homonyms, minimal grammatical contexts, removal of the homonimy on the grammatical context.

AN INTELLIGENT SYSTEM OF PRE-SYNTACTICAL ANALYSIS (ISPA) OF RUSSIAN MORPHOLOGICAL HOMONYMS / Roman Nikolaevich Afanasyev (Moscow State Institute of Steel and Alloys (Technological University), 4 Leninsky prospect, Moscow 117936, Russia, arn@beep.ru), Tatyana Yurjevna Kobzareva (Russian State University for the Humanities, 6 Miusskaya pl., Moscow 125267, Russia, Stam@rozenshtein.mccme.ru).

The multivaluedness of the level of morphological analysis is a rather considerable negative factor in the automatic text analysis. In this connection it is necessary to develop the special Intelligent System of Pre-syntactical Analysis (ISPA) for RL to bridge the morphological and the syntactical stages of analysis. This kind of system is necessary for the tasks of the natural text analysis nearly everywhere.
Homonyms of RL was classified by morphological homonym type (MO). For every type of MO has been constructed algorithms for their solution. ISPA lets add new types of MO, edits and analyses algorithms, accumulates statistical data. The kernel of ISPA is dll-libraries and it can be used in any NLP-system.