Интерпретация текстовых документов с использованием формальной грамматики AGFL и компьютерного тезауруса RussNet

Азарова И.В. (Кафедра математической лингвистики СПбГУ),
Секликов Ю. В., Иванов В. Л. (ООО «НЕМО», СПб)

В настоящем докладе обсуждаются вопросы разработки системы анализа и интерпретации тестовых русских предложений, использующей морфосинтаксический парсер на базе формальной грамматики AGFL в сочетании с данными компьютерного тезауруса RussNet, которая разрабатывается сотрудниками Санкт-Петербургского государственного университета совместно с ООО «НЕМО».
Описываемая система демонстрирует возможности формально-грамматического анализа предложе-ний и словосочетаний, при котором производится лемматизация текстовых форм, определяется частеречная принадлежность слов и устанавливаются синтаксические связи между единицами в рамках именных и гла-гольных словосочетаний.
Система интегрирует реализацию формализма AGFL и API для доступа к данным RussNet внутри сис-темы логического вывода. При интерпретации предложения подключается компьютерный тезаурус RussNet, из которого отбираются возможные интерпретации для выделенных лексем в соответствии с полученной морфологической информацией.
Обсуждается метод снятия неоднозначности в рамках синтаксически свя-занных пар слов «прилагательное-существительное» и «глагол-существительное».
Предлагаемая система является частью программного комплекса, ориентированного на поиск и из-влечение данных из электронных документов.

Введение

Данный проект, в рамках которого разрабатывается программная реализация системы обработки массива текстовых документов на естественном языке, является первым этапом сотрудничества для его участников. В рамках проекта предполагается не только реализовать конкретное программное воплощение научных разработок, но и проверить инновационный потенциал взаимодействия научных и коммерческих структур, отработать методы организации совместных проектов в области компьютерной лингвистики.
Участниками исследовательского проекта являются сотрудники Кафедры математической лингвистики Санкт-Петербургского университета и ООО «НЕМО». Задача коллектива Санкт-Петербургского университета – представить необходимые научные достижениями и предметные знания. Задача ООО «НЕМО» – реализовать имеющиеся научные достижения в виде программной технологии и оценить коммерческий потенциал для дальнейшего развития совместного проекта.

Для коллектива университета данный проект – возможность использовать созданные за последние пятилетие формально-грамматическое описание русского языка и компьютерный wordnet-тезаурус RussNet для задач поиска, извлечения и преобразования информации, содержащейся в текстах.
Основная цель ООО «НЕМО» – обеспечить создание коммерческого продукта на основе существующего в Санкт-Петербургском университете потенциала в области компьютерной лингвистики.

Формально-грамматическое описание AGFL

Формализм AGFL[1] является двухуровневой формальной грамматикой, в которой контекстно-свободная порождающая грамматика дополнена решеткой признаков с конечным числом значений. В качестве признаков выступают разные типы лингвистической информации: грамматические категории, параметры лексико-грамматической субкатегоризации основных частей речи, а также любые формальные характеристики, необходимые для описания синтаксических и морфологических конструкций. Данный формализм естественным образом задает объединение в конструкции непосредственных составляющих, отношения согласования и координации на разных уровнях грамматической структуры (от морфологии до синтаксиса).
Помимо формализма AGFL на предыдущих стадиях разработки грамматики была использована система AGFL, разработанная группой К.Костера[2], которая позволяла создавать парсеры, эффективно работающие с большими лексиконами. Однако, поскольку система AGFL первоначально и на протяжении многих лет была ориентирована на использование в операционной среде Unix, вариант системы для MS Windows содержал достаточно большое количество ограничений, которые не давали в полной мере оценить возможности собственно формализма. Кроме того, существуют некоторые проблемы по встраиванию системы AGFL в программные продукты в связи с ориентированностью системы исключительно на анализ текста.
Для решения этих и ряда других проблем подготовлена реализация формализма AGFL, интегрированная в систему логического вывода (LES), которая основана на типизированных структурах признаков (TFS) с ограничениями, разработанную компанией “Немо”. Данная система вывода реализована на языке Java и предназначена для встраивания в веб-ориентированные продукты и сервисы.
Интеграция формализма AGFL и логической системы вывода LES позволяет осуществить в процессе обработки текста запросы к базе RussNet и построить систему правил, описывающую, как должен происходить отбор возможных интерпретаций для лексем в соответствии с полученной морфологической и синтаксической информацией.

Компьютерный тезаурус RussNet

RussNet является русской версией компьютерного словаря типа WordNet. Целью проекта является построение лексико-семантического ресурса для отражения организации лексической системы русского языка в целом (в противоположность терминологическим или частным словарям); для представления ядра общеупотребительной лексики русского языка; для фиксации всех семантических, семантико-грамматических и семантико-деривационных отношений, существенных для лексикона русского языка**.** Ход работ по проекту освещался в

---

[1] Более подробное описание формализма AGFL представлено в работе (Азарова 2003).
[2] URL: http://www.cs.kun.nl/agfl/

ряде публикаций (Азарова и др. 2003; Материалы к компьютерному тезаурусу… 2002; Azarova et al. 2002).
Компьютерный тезаурус задает набор лексикализованных понятий русского языка, т. е. таких, которые имеют регулярное, частотное выражение при помощи лексических единиц. Понятия тезауруса задаются синсетами (синонимическими рядами), в которых выделяется первый элемент, так называемая доминанта синсета – самый частый и нейтральный представитель данного понятия в современных текстах. На синсетах заданы определенные семантические отношения, среди которых можно выделить традиционные тезаурусные связи – родовидовые, меронимические (часть-целое) и ряд других, включая такие сложные отношения, как пресуппозиция и каузация.
Методика построения RussNet опирается на частотные распределения лексем, выявление регулярных контекстов употребления слов, семантический анализ контекстов. Таким образом, RussNet является естественно-языковой онтологией, фиксирующей семантические отношения между понятиями, которые регулярно задаются в русском языке при помощи лексических единиц.
Для использования результатов проекта RussNet в программных продуктах реализован ряд средств для эффективного хранения тезауруса и доступа к его данным. В качестве хранилища данных RussNet использована объектная база данных, содержащая набор индексов, построенных на BTree (Balanced Tree) для быстрого доступа к необходимым синсетам. Разработанный интерфейс API (application program interface) обеспечивает функциональность для получения всей сопутствующей информации для синсета из хранилища данных. Интерфейс реализован на языке Java.
API был использован для построения вэб-интерфейса для просмотра базы RussNet. В настоящий момент он находится по адресу http://russnet.kiberry.ru. В дальнейшем на этом сайте будет опубликован список зеркал этого ресурса.
Для редактирования данных RussNet в рабочей группе использовалось XML представление, разработанное в рамках проекта VisDic[3] сотрудниками Университета им. Масарика (Чехия), которое является общим форматом представления данных в ряде европейских wordnet-проектов. Далее XML файлы, подготовленные различными участниками, добавлялись в объектную базу с помощью разработанных инструментов.

Использование wordnet-тезаурусов для информационного поиска

Эксперименты по использованию компьютерного wordnet-тезауруса для улучшения параметров информационного поиска в основном проводились на материале Принстонского WordNet (Vorrhees 1998).
Wordnet-тезаурус использовался двумя способами. Во-первых, синсеты WordNet (т. е. лексикализованные понятия) представляли содержание документа, в качестве которого могли выступать разные типы данных – от заголовков до полных текстов статей и книг. Во-вторых, WordNet использовался для расширения запроса пользователя. Оказалось, что сложность автоматического разграничения понятий (снятия неоднозначности) не давала возможности эффективно использовать преимущества тезаурусной организации словаря. Однако, сама стратегия применения wordnet-тезауруса заслуживает рассмотрения.
В первом случае результаты стандартного информационного поиска сопоставлялась с результатами, дополнительно учитывающими понятийное описание содержания документа при помощи синсетов существительных. Наборы идущих друг за другом существительных в тексте за исключением стоп-слов, так называемые «лексические цепочки», проецировались

[3] URL: http://nlp.fi.muni.cz/projekty/visdic/

на wordnet-струкутры. В силу многозначности слов проекции были неоднозначными. Далее осуществлялась попытка выбрать наиболее вероятное значение, реализованное в контексте. Анализ начинался с моносемантичных слов, которые задавали предпочтительную область реализации значения. Поскольку wordnet-тезаурусы организованы в структуры при помощи родовидовых отношений, производился поиск дерева или поддерева, включающего максимальное количество понятий из документа. Стратегия поиска таких «лексических цепочек» исходила из предположения, что максимальное количество понятий в документе относится к одной сфере, понятийной области. Очевидно, что такое предположение будет адекватным только для монотематических документов. Как показывают результаты информационного поиска, использующего такую стратегию, она давала довольно плохие результаты при поиске в специальных (терминологических) текстах и довольно стабильные результаты для неспециальных (газетных) текстов, хотя выраженного улучшения параметров поиска все-таки не наблюдалось. Подробный анализ примеров «сбоев», приводящих к ухудшению результатов, показывает, что зачастую они были обусловлены слишком тонкой градацией значений в WordNet, а также включением в словарь метафорических употреблений слов.

Вторая стратегия использования словаря WordNet для информационного поиска была связана с расширением запроса пользователя за счет синонимов для включенных в синсеты слов запроса, а также других синсетов, например, родовых. Расширение запроса тестировалось в двух режимах: (1) нужное значение выбиралось человеком, (2) значение выбиралось автоматически. При «человеческом» выборе значения возрастало значение точности поиска на 35%, при этом значение полноты не менялось. Таким образом, было показано, что распространение запроса пользователя за счет тезаурусной информации дает ощутимые результаты улучшения параметров поиска. При автоматическом выборе значения слов важным являлось, дает ли запрос составить «лексическую цепочку» хотя бы из двух слов, в этом случае наблюдалось незначительное улучшение точности (максимально на 0,7 %).

Таким образом, использование wordnet-тезаурусов для улучшения параметров информационного поиска при стандартной процедуре поиска «лексических цепочек» в родовидовых структурах тезауруса возможно, во-первых, для нетерминологических текстов, во-вторых, при разработке эффективной процедуры разрешения неоднозначности слов, т. е. автоматического определения значения, в котором слово употреблено в контексте.

Стратегия анализа документов с использованием AGFL парсера и компьютерного тезауруса RussNet

Анализ в предлагаемой системе начинается с предварительной обработки текстового документа, так называемого токенизатора, который выделяет в сообщении текстовые и нетекстовые элементы, например, числа, даты, электронные адреса и проч. На вход морфологического блока парсера подается текстовая часть, которая будет анализироваться как набор терминалов, плюс метаописания для нетекстовых элементов (например, чисел) в виде нетерминалов.

В морфологическом модуле задается анализ основных частей речи (существительных, глаголов, прилагательных и наречий) в виде системы продукций, таким же образом представлена система порождения и неосновных частей речи (числительных, местоимений-существительных) за исключением форм, обладающих нестандартными парадигмами (например, личных местоимений *я, мы, ты, вы* и некоторых других слов), которые задаются в специальном модуле готовых форм. В этом же модуле перечисляются служебные слова (предлоги, включая составные; союзы и частицы). Морфологический модуль использует лексикон основ, для которых указаны основные классификационные категории частей речи:

характеристика рода и одушевленности существительных, лексико-грамматический разряд прилагательных, схема управления глаголов и т. д. На первом этапе отработки системы лексикон включает основы 5 тысяч наиболее частотных знаменательных слов и около 10 тысяч наиболее часто встречающихся мотивирующих слов. Грамматика содержит правила для продуктивных типов префиксальных и суффиксальных моделей словообразования основных частей речи. Производное слово, сформированное по правилам морфологического блока, может включать несколько префиксов и суффиксов. Допускается образование слов одной части речи от других. Среди производных слов встречаются как реальные, так и потенциальные слова (которые не были зарегистрированы в корпусе современных текстов, состоящем из 21 млн. словоупотреблений). В дальнейшем предполагается построить модуль обратной связи в блоке морфологического анализа, который обеспечит подсчет частотности употребления основ лексикона с тем, чтобы выявить «пассивные» основы, не используемые с достаточной частотностью в анализируемых текстах, и «активные» образования, которые встречаются с частотностью, превышающей некоторое вычисленное пороговое значение, с тем чтобы вставить такие производные основы в виде готовых форм в лексикон. Такая схема построения лексикона, на наш взгляд, должна обеспечить такой способ владения лексиконом, который характерен для людей-носителей языка.

Результатом работы морфологического блока является приписанная форме слова (возможно аналитической типа *будет работать* или *был построен*) частеречная характеристика и набор значений морфологических категорий: многозначный – в случае омонимии форм. Отдельные словоформы не получат интерпретации, поскольку они не покрываются лексиконом. Эти словоформы сохраняются в лог-файле для того, чтобы впоследствии выявить основы, которые необходимо вставить в лексикон.

Морфологический блок встроен в блок синтаксического анализа так, что локальный синтаксический контекст, указывающий на наличие предлогов, соответствий между значениями грамматических категорий при согласовании, позволит частично разрешить синтаксическую омонимию. Например, словоформа *пути* в конструкции *в пути* получит не 5 интерпретаций в роли существительного, а 2 – местный ед. ч. и винительный мн. ч.

Блок синтаксического анализа включает частотные конструкции словосочетаний, построенных на базе знаменательных частей речи, частотные схемы построения простых предложений, отдельные осложняющие конструкции в простом предложении типа рядов, причастных и деепричастных оборотов. Если некоторое предложение не может быть проанализировано целиком, тогда производится попытка проанализировать словосочетания в составе предложения. Отказ полного анализа обусловлен регулярно изменением порядка слов, вхождением в предложение морфологически неопознанных слов. В режиме отладки блока синтаксического анализа предполагается выявить набор правил, дающий оптимальное соотношение правильно проанализированных структур и минимального количества «ложных» омонимичных структур, которые появились за счет конкуренции вариантов синтаксического описания, но не соответствуют реальной омонимии синтаксических конструкций.

Наиболее детально проработаны структуры именных и глагольных словосочетаний, включающих отношения «прилагательное-существительное» и «глагол-существительное», поскольку такого рода связи дают возможность активным образом подключить данные компьютерного тезауруса RussNet. Эта операция осуществляется в рамках модуля интерпретации, в котором для выделенной на уровне морфосинтаксического анализа лексемы отбираются имеющиеся в тезаурусе значения. Снятие неоднозначности слов в первую очередь производится посредством учета валентной схемы глаголов и прилагательных, которая задана в RussNet при помощи трех типов данных: (1) обобщенной синтаксической позиции, которую занимает валентность в нейтральной конструкции, например, <СУБЪЕКТ> или <НАПРАВЛЕНИЕ> (такие функционально-синтаксические типы соотнесены с

нетерминалами описания в блоке синтаксического анализа); (2) ряда способов поверхностного выражения валентностей, например, <ИМЕНИТЕЛЬНЫЙ> или <"к" + ДАТЕЛЬНЫЙ>; (3) семантического типа существительных, которые занимают валентную позицию, причем такой тип может указывать на конкретную вершину в дереве родовидовой иерархии или на корень дерева/поддерева, например, <ЛИЦО> или <ПРЕДМЕТ>.
Валентности, которые занесены в RussNet описание, обладают характеристикой обязательности/факультативности, которая вычисляется на основании контекстов употребления лексем в определенных значениях в имеющемся корпусе текстов. Выбор реализованного в контексте документа значения слова осуществляется в первую очередь на основании совпадения поверхностной формы реализации синтаксической зависимости в тексте и семантического типа валентной лексемы в wordnet-описании.
Система интегрирует реализацию формализма AGFL и API для доступа к RussNet внутри системы логического вывода LES. Эта система так же, как и AGFL, представляет собой набор продукций, задает определение решетки типов со структурами признаков и обеспечивает использование дизъюнкции, конъюнкции, отрицания в смысле логических языков программирования (таких, как Пролог). Система обладает некоторым количеством **встроенных предикатов**, выводимость (согласованность) которых определяется самой системой. Взаимодействие с продукциями AGFL и доступ к RussNet осуществляется набором встроенных предикатов.

AGFL

Вызов конкретной продукции AGFL для разбора или синтеза осуществляется предикатом *agfl_call*. Возможен вызов как корневой продукции для анализа сегмента текста или предложения, так и любого другого нетерминала продукции AGFL. Например, вызов продукции анализа формы существительного задается следующим образом:

nounf(GENDER, ANIM, NUMBER, CASE, LEMMA)

Для конкретной леммы существительного, например *агроном*, вызов задается следующим образом:

***previous_predicate**, agfl_call(nounf(GENDER, ANIM, NUMBER, CASE, 'агроном'), X),*

***next_predicate**….*

Здесь, в зависимости от связанности переменной X, будет осуществлен синтез всех форм леммы *агроном*, если переменная X не связана, либо проверка существования такой формы леммы, строка которой связана с X. Например, при X=['агроном<u>ной</u>'] сопоставление закончится неудачей.

RussNet

Запросы к RussNet осуществляются набором предикатов, извлекающих и сопоставляющих информацию из объектной базы данных:

- *synset_lemma(SYNSETID, LEMMA, SENSE, LNOTE)* осуществляет поиск или проверку согласованности синсета с идентификатором ***SYNSETID*** и леммы ***LEMMA***;
- *synset_link(SYNSETID, LINKTYPE, TO_SYNSETID)* – наличия связи типа ***LINKTYPE*** синсетов с идентификаторами ***SYNSETID*** и ***TO_SYNSETID***;
- *synset_info(SYNSETID, DEF, POS,…)* – согласованности синсета с идентификатором ***SYNSETID*** и дополнительной информации о синсете.

Например, запрос

findall(LEMMA, synset_lemma('RUS-1601967743', LEMMA, SENSE, LNOTE) , LEMMA_LIST)

вычислит весь набор лемм синсета с идентификатором (id), равным RUS-1601967743 и свяжет полученный список с переменной LEMMA_LIST, вычислит лемму слова *‘агрономами’,* получит для нее SYNSETID, найдет соответствующий гипероним и напечатает его лемму.

```
agfl_call( nounf(GENDER, ANIM, NUMBER, CASE, LEMMA), [агрономами]),
synset_lemma(SYNSETID, LEMMA, SENSE, LNOTE),
sysnset_link(SYNSETID, hyperonym, SYNSETID2),
sysnset_lemma(SYNSETID2, LEMMA2, SENSE, LNOTE),
println(LEMMA2).
```

Кроме вызова продукций AGFL, из системы логического вывода LES предусматривается специальная форма для вызова из AGFL предикатов системы:

verbf (pres, imperf, PERSON, NUMBER, TRANS, non, active, SYNSETID):
verbst (non, VBTYPE, imperf, TRANS, LEMMA), inflsv (pres, VBTYPE, NUMBER, PERSON),
@resolve_synset (LEMMA, SYNSETID) ; verbst_pr (non, VBTYPE, imperf, TRANS, LEMMA),
inflsv (pres, VBTYPE, NUMBER, PERSON), @resolve_synset (LEMMA, SYNSETID).

В этой AGFL продукции вставлен вызов предиката resolve_synset(LEMMA, SYNSETID), определенный в системе.

Литература

1. *Azarova et al. 2002 – Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I.* RussNet: Building a Lexical Database for the Russian Language // Workshop Proceedings: Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002. Las Palmas de Gran Canaria, 2002. P. 60–64.
2. Voorhees 1998 – *Voorhees E. M.* Using WordNet for Text Retrieval // WordNet: an electronic lexical database / Edited by Ch. Fellbaum. Massachusetts Institute of Technology, 1998. P. 285-303.
3. Азарова 2003 – *Азарова И.В.* Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г.) М., 2003. С. 51-55.
4. Азарова и др. 2003 – *Азарова И.В., Митрофанова О.А., Синопальникова А.А.* Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог 2003 (Протвино, 11-16 июня 2003 г.) М., 2003. С. 43–50.
5. Материалы к компьютерному тезаурусу… 2002 – Материалы к компьютерному тезаурусу лексики русского языка / Сост. И.В. Азарова, О. А. Митрофанова. СПб., 2002. 232