УДК 004.89+004.912+004.5+004.93’1+519.7

ПРАГМАТИЧЕСКИЙ АНАЛИЗ С ПРИМЕНЕНИЕМ ПОДХОДОВ К АВТОМАТИЗИРОВАННОМУ СОЗДАНИЮ ОНТОЛОГИЧЕСКОЙ БАЗЫ ЗНАНИЙ*

П.В. Толпегин[1], Д.П. Ветров[2], Д.А. Кропотов[3]

Предлагается подход к автоматизированному пополнению онтологической базы знаний «О Мире» при помощи синтактико-семантического анализа путем «начитывания» естественно-языковых текстов. Полученные данные позволят качественно повысить уровень автоматизированного разрешения анафоры. Подход особенно применим для языков, широко использующих выразительные средства, в частности – русского, нежели чем для языков, опирающихся на строгие морфологические и синтаксические правила.

Введение

Проблема построения онтологии с каждым днем все больше рассматривается в областях, смежных с искусственным интеллектом, в частности, при автоматизированной обработке естественно-языковых (ЕЯ) текстов. В предлагаемой работе рассматриваются подходы к автоматизированному разрешению межклаузной референции в дискурсах текстов при помощи онтологической базы знаний «О Мире» (далее – БЗ), а также подходы к автоматизированному извлечению знаний из массивов текстов (корпусов) (text data mining) и пополнению БЗ.

Достаточно часто яркая цель прагматического анализа затушевывается тем, что внимание авторов разработок фокусируется на технических деталях математических конструкций, а не на конкретных методах составления онтологических баз знаний и анализе свойств предметных областей.

Частота появлений инновационных методов ЕЯ-анализа в последнее время значительно снижается. Джон Бейтман (John Bateman) в исследованиях [Bateman, 2006] отмечает падение количества разработок в

* Работа выполнена при финансовой поддержке РФФИ (проект № 06-06-80464а)
[1] *119991, г. Москва, ГСП-1, ул. Вавилова, 40, ВЦ РАН, pavel@tolpegin.ru*
[2] *119991, г. Москва, ГСП-1, ул. Вавилова, 40, ВЦ РАН, vetrovd@yandex.ru*
[3] *119991, г. Москва, ГСП-1, ул. Вавилова, 40, ВЦ РАН, dkropotov@yandex.ru*

области генерации текстов на ЕЯ в последнее время. «Насыщение» инновационными разработками обуславливается преимущественно используемым экстенсивным подходом. В частности, применяется ручное создание классификаторов и словарей. Интенсивное развитие систем ЕЯ-анализа, как один из вариантов, представляется возможным при тесном взаимодействии с системами онтологического распознавания образов (изображений) и распознавания речи.

Под онтологическим распознаванием изображений понимается система анализа графической информации, способная выделять и различать объекты, их части, формы и их взаимное расположение. В процессе онтологического распознавания будет возможно установление меронимических (часть-целое) отношений между объектами Мира. Полученная информация может лечь в основу БЗ и сыграть немаловажную роль в референциальном (прагматическом) анализе ЕЯ-текстов. Система распознавания речи должна обеспечивать выделение интонации, пауз и интонационных акцентов (ударений).

Достоинством описываемой интеграции систем ЕЯ-анализа и распознавания является репрезентативность получаемых данных (анализируются реальные изображения, объекты, события, речь и явления) и автоматизация процесса пополнения БЗ.

Основным недостатком является отсутствие высокоточных систем распознавания, которые на промышленном уровне могут решить поставленную задачу. С другой стороны, предлагаемый способ не охватит все то, что невозможно увидеть глазом, то есть понятия и некоторые явления.

Прикладной уровень развития систем ЕЯ-анализа ушел далеко вперед по сравнению с онтологическими системами распознавания. Это приводит к тому, что авторам разработок лингвистических систем приходится восполнять отсутствие недостающих знаний (подобных БЗ) ручным созданием словарей (в частности – онтологических баз данных). Близкими или тождественными разработками к БЗ являются онтологический словарь, база знаний «О Мире» и идеографический словарь. Примерами работ по созданию базы данных, содержащей сведения о понятиях и их отношениях, являются: работа [Баранов, 1996], проект WordNet [WordNet, 2006] – лексическая база знаний английского языка, разрабатываемая в Принстонском Университете (Нью-Джерси, США), EuroWordNet [EuroWordNet, 1999] – аналогичный проект для датского, итальянского, испанского, немецкого, французского, чешского и эстонского языков, RussNet [RussNet, 2005] – проект компьютерного тезауруса лексики русского языка (под руководством И.Азаровой, СПбГУ), Русский WordNet [WordNet, 2004] - проект русской версии

WordNet (С. Яблонский, А. Сухоногов, Петербургский Государственный Университет Путей Сообщения).

Прогресс в развитии систем ЕЯ-анализа позволяет строить первичные семантические графы [Сокирко, 2001, Сокирко, 2005], основываясь на графематическом (лексическом), морфологическом, синтаксическом, фрагментационном этапах анализа. На рис. 1 приведен результат работы программы [Сокирко, 2001, Сокирко, 2005] для предложения: «*За последнее время количество исследований в области автоматизированного построения онтологии заметно возросло, но проблема до сих пор остается неразрешенной*».

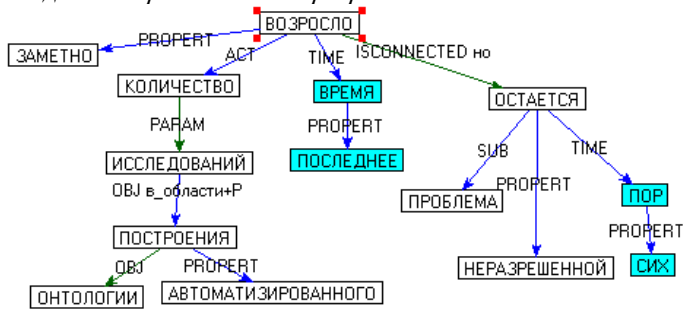

Рис. 1. Пример первичного семантического графа

Актуальным этапом в развитии систем ЕЯ-анализа представляется референциальный (прагматический) анализ (разрешение анафоры). Суть этапа заключается в установлении зависимостей между объектами (существительными и местоимениями), упоминаемыми в простых предложениях (клауз) на протяжении целого дискурса.

Предлагается следующая методологическая классификация межклаузной референции.

1. Классификация межклаузной референции

1.1. Анафора (референция) местоимений

Кореферентные объекты в предложениях могут именоваться по-разному. В частности, с помощью местоимений. Особый интерес представляют личные (он), указательные (этот), вопросительные (который), возвратные (себя) местоимения.

Пример 1. Референция местоимений. Одной цифрой обозначены анафорическое местоимение (анафор) и гипотетические антецеденты. Референтующий антецедент подчеркнут.

*В Атлантическом океане[1] ученые нашли "Затерянный город[1]". **Этот[1]** "город[1,2]" находится у обнаруженного гидротермального источника[2]. **Его[2]** заселяют необычные микроорганизмы[3] и морские существа[3],*

кот орые[3] *выдерж ивают горячую, чрезвычайно богат ую различными минералами морскую воду.*[Aqualife, 2006]

1.2. Синонимия отдельных слов, именных групп и более сложных конструкций

Автор может использовать синонимию, как при описании объектов, так и в названии самих объектов. «*Премьер-минист р*» может быть заменен на слово «*полит ик*», а глагол «*зат рагиват ь*» может употребляться, по меньшей мере, в двух значениях: *задеват ь, касат ься, т рогат ь, прит рагиват ься* и *волноват ь*. Таким образом, проблема понимания ЕЯ-текстов компьютером, в частном случае, сводится к «отождествлению» кореферентных объектов, описанных различными выразительными средствами языка.

1.3. Меронимические, родовые и видовые отношения

Меронимические, родовые и видовые отношения могут встречаться между объектами, упоминающимися в дискурсе.

Пример 2. Меронимические отношения в тексте

Как т олько извест ный ученый[1] *сел под* ***деревом***[2]*, т ак на* ***него***[1] *упало* ***яблоко***[2]*.*

[(яблоко) = часть (дерево)], [(него) = (ученый)]

В [Азарова и др., 2002] приведена более детальная классификация меронимических (реже – партонимических) отношений по логическим связям между понятиями типа «компонент-предмет» (ветка-дерево), «член-множество» (дерево-лес), «материал-предмет» (алюминий-самолет) и другие, как «порция-масса» (кусок-пирог), «место-область» (Москва-Россия).

1.4. Логико-интуиционистские нечеткие правила

Нечеткие правила представляют собой свод закономерностей, накопленных за время наблюдения. В настоящее время представляется сложным говорить об автоматизированном получении таких правил, но они достаточно удобны в использовании со структурированной областью понятий. Пример правила: *Зимой холоднее, чем лет ом.*

В [Клещев и др., 2001] приведены неоспоримые примеры подобных правил, описывающих взаимодействие физических объектов:

- антирефлексивность (нельзя поставить объект сам на себя);
- антисимметричность (если один объект стоит на другом, то второй не может в той же ситуации стоять на первом);
- антитранзитивность (неверно суждение, что объект стоит на другом, если между ними есть третий).

Логико-интуиционистские нечеткие правила могут закладываться в систему как вручную, так и выводиться автоматически на основе собранных сведений о Мире.

1.5. Сложные выразительные средства языка

Референция объектов, упомянутых с применением выразительных языковых средств, включает, в частности, метафоричный перенос, который может быть понятен только человеку при полном анализе контекста. Следующий пример демонстрирует обратную ситуацию, в которой объекты не референтируют между собой. С другой стороны, несложно найти контекст, в котором понятия *ст удент ы* и *хулиганы* будут обозначать одних и тех же лиц.

Пример 3. Референция в случае использования выразительных средств
Под прикрыт ием ***ст удент ов***[1] *в Париж е орудуют банды* ***хулиганов***[2].
[(студенты) ≠ (хулиганы)]

2. Применение машинного обучения к разрешению референции местоимений 3-го лица

Сотрудниками отдела математических проблем распознавания и методов комбинаторного анализа Вычислительного центра им. А.А. Дородницына Российской академии наук (ВЦ РАН) был проведен опыт по ручной разметке корпуса новостных лент на предмет анафоры местоимений третьего лица [Толпегин, 2006]. Предварительно корпус новостей был размечен морфологически, синтаксически и семантически решениями [Сокирко, 2001, Сокирко, 2005, Ножов, 2003]. По результатам разметки была сформирована обучающая выборка для выделения закономерностей в референциальном выборе для местоимений 3го лица с помощью методов машинного обучения [Журавлев, 2006]. Применяя различные методы машинного обучения, удалось добиться эффективности от 62% до 84%. С одной стороны, процент «неудач» в машинном обучении (от 16% до 38%) складывается из погрешностей модулей анализа текстов и погрешностей методов машинного обучения. С другой стороны, анализ ошибочных ситуаций показал, что величина ошибки в обучении в наибольшей степени зависит от выразительных средств языка, которые не подчиняются правилам морфологии, синтаксиса и первичной семантики. Другими словами: результат референциального выбора не всегда однозначно зависит от структуры и других признаков анализируемых предложений. Для одинаковых синтаксических конструкций предложений могут отыскаться совершенно противоположные результаты.

В основу большинства зарубежных решений в области референции положен именно тот подход, который в результате проведенного опыта для русского языка оказался малоэффективным: признаковое пространство морфологии, синтаксиса и первичной семантики [Сокирко, 2001, Сокирко, 2005, Ножов, 2003] не покрывает полностью ответ в референциальном выборе. Рассматриваемый подход по

автоматизированному созданию БЗ позволит сформировать недостающий признак, который сыграет немаловажную роль в разрешении референции в тех языках, которые «отступают» от строгих морфологических и синтаксических правил в пользу выразительных средств языка.

Пример 4. Бивалентная конструкция.

*Маша[?] купила машину[?]. **Она[1] её[2]** любит .*

3. Структура предлагаемого решения

3.1. Автоматизированное составление БЗ

Автоматический способ составления (пополнения) БЗ позволит не только уменьшить трудоемкость задачи, но и повысит репрезентативность собираемой информации. Предлагается подход, который позволит решить задачу автоматизированного пополнения БЗ путем «начитывания» компьютером текстов.

В качестве инструмента для решения был задействован идеографический словарь русского языка О.С. Баранова (http://baranovoc.narod.ru) [Баранов, 1996]. Словарь интересен введенным автором набором тематических рубрик, каждая из которых заполнена соответствующими понятиями и терминами. Рубрикация имеет семь уровней вложенности.

Рассматривая только объекты (существительные), представляется возможным отталкиваясь от морфологических и синтаксических свойств словосочетаний, пополнять древовидную структуру данных.

Пример 5. Извлечение знаний из структуры словосочетаний

ножка стула, состав команды, заместитель председателя, годы работы

Идеографический словарь может сыграть важную роль при классификации объектов (существительных) на понятия, объекты живой, неживой природы, предметы и др. [Каневский и др., 2000, Каневский, 2000].

Путем анализа текстов и представления его в виде графа (рис. 1.), предлагается составить статистическую базу данных (СБД), подобную рассмотренному выше идеографическому словарю и БЗ. Результат обработки текста в виде узлов и дуг декомпозируется в СБД таким образом, чтобы по заданному слову и семантической валентности можно было получить набор слов, связанных с введенным словом указанной валентностью, с соответствующими характеристиками.

$$f\,(\,word_x, valency) = \{word_y : characteristics\}_n. \qquad (3.1)$$

3.2. Разработка нечетких мер

Применение методов интеллектуального анализа данных к данным СБД позволит выявить, в конкретном случае, следующую информацию: для глагола *любить* в значении валентности *СУБЪЕКТ* число одушевленных существительных в несколько раз превышает число неодушевленных существительных. В этой связи формирование и вычисление нечетких мер по различным признакам позволит сделать весомый вклад в разрешение референции. Применение нечетких мер позволит увеличить процент правильных референций для пары «анафор-антецедент». В частности, для примера 4 построенная по предложенному алгоритму система определит, что любить может, как правило, одушевленное лицо, в данном контексте – *Маша*.

3.3. Вероятностный корпусно-ориентированный анализ

Корпусно-ориентированный анализ основывается на поиске в эмпирическом корпусе текстов глагола, связанного с анализируемым анафором и гипотетическим антецедентом. Антецедентом, референтирующим с анафором, выбирается тот гипотетический антецедент, который встречается большее число раз, чем другие вместе с глаголом.

Пример 6. Вероятностный корпусно-ориентированный анализ

В авт омобиль? Иван вст роил блокират ор? коробки переключения передач. Теперь ***его*** *слож но угнат ь.*

По данным поиска в поисковой системе «Яндекс» (www.yandex.ru) «*угнат ь авт омобиль*» встречается в 59 раз чаще, чем «*угнат ь блокират ор*». Однако рекомендуется проводить поиск по синтаксически размеченному корпусу, чтобы избежать ошибок первого и второго рода: случаев, в которых слова стоят рядом друг с другом, но не взаимосвязаны, и случаев, когда слова взаимосвязаны, но разделены одним или несколькими словами.

4. Другие задачи БЗ

Разработка, составление и развитие онтологической базы знаний «О Мире» поможет подойти на качественно новом уровне к практическому решению актуальных задач: тематической классификации текстов, разрешению омонимии семантического поля связи существительного и прилагательного, разрешению омографии (орг*а*н-*о*рган, мук*а*-м*у*ка и др.), т.е. слов, омонимия значения которых разрешима на уровне семантики.

С другой стороны, аккумулируя различным образом информацию путем «начитывания» текстов, БЗ выполняет статистическую функцию. То есть можно будет получить «типичную» информацию, которая характерна для анализируемой среды, а так же просигнализировать об отклонениях в фактологии входных данных: могут быть распознаны «аномалии» в сочетаниях «умный стол» и «третье ухо».

Список литературы

[Aqualife, 2006] Журнал о природе и путешествиях Aqualife – http://dudu.narod.ru/hydro.htm

[Bateman, 2006] Bateman J. Natural Language Generation Systems. – http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/nlg-table-date-sort.html

[EuroWordNet, 1999] EuroWordNet – http://www.illc.uva.nl/EuroWordNet/

[RussNet, 2005] RussNet – http://www.phil.pu.ru/depts/12/RN/index.shtml

[WordNet, 2004] Русский WordNet – http://www.pgups.ru/WebWN/wordnet.uix

[WordNet, 2006] WordNet – http://wordnet.princeton.edu/

[Азарова и др., 2002] Азарова И.В., Митрофанова О.А., Синопальникова А.А., Ушакова А.А., Яворская М.В. Разработка компьютерного тезауруса русского языка типа WordNet // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» СПб., 2002.

[Баранов, 1996] Баранов О.С. Идеографический словарь русского языка –М.: ЭТС 1996.

[Журавлев, 2006] Журавлев Ю.И., Рязанов В.В., Сенько О.В. "РАСПОЗНАВАНИЕ". Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006.

[Каневский, 2000] Каневский Е.А. Атрибуты существительных // Информационные технологии в гуманитарных и общественных науках: семантико-синтаксический анализ текстов. – СПб.: СПб ЭМИ РАН, 2000. Вып. 9.

[Каневский и др., 2000] Каневский Е.А., Клименко Е.Н., Тузов В.А. Об одном подходе к классификации прилагательных // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. - М.: РосНИИ Искусственного Интеллекта, 2000. Т. 2.

[Клещев и др., 2001] Клещев С.А., Артемьева И.Л. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология» // Научно-техническая информация, серия 2 «Информационные системы и процессы», 2001. №2.

[Ножов, 2003] Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы) // Диссертация на соискание ученой степени кандидата технических наук. – М. 2003.

[Сокирко, 2001] Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) // Диссертация на соискание ученой степени кандидата технических наук. – М. 2001.

[Сокирко, 2005] Сокирко А.В. Первичный семантический анализ – http://www.aot.ru/docs/seman.html

[Толпегин, 2006] Толпегин П.В., Ветров Д.П., Кропотов Д.А. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006.