12.

СРЕДСТВА СЕМАНТИЧЕСКОГО ПОИСКА[1]

SEMANTIC SEARCH METHODS

А.Л. Воскресенский *(avosj@yandex.ru)*

Г.К. Хахалин *(gkhakhalin@yandex.ru)*

Показано, что локальные контексты недостаточны для снятия омонимии при переводе со словесного на жестовый язык. Рассматриваются методы сопоставления концептов на основе анализа синтаксиса и семантики. Предлагается способ автоматизированного поиска новых (для пользователя) документов в сети Интернет.

«… Нам нужно задать себе вопрос не о том, едина ли
природа, а вопрос: каким образом она является единой?»
А. Пуанкаре

1. Введение

Как указывается в [1], за последние несколько лет большинство крупных российских госпредприятий и частных фирм перешли на автоматизированные системы управления. Неотъемлемой частью любой такой системы является СУБД — база товаров, клиентов, услуг и т. п. Соответственно, необходимо обрабатывать большие массивы жестко структурированных данных. Однако гораздо чаще возникает проблема поиска и выборки необходимой информации из большого неструктурированного массива, содержащего как текстовые, так и мультимедийные данные.

В [2] семантический поиск определяется как поиск по содержательным аспектам всех компонентов документов электронной библиотеки. Однако, в связи с отсутствуем в настоящее время единого понимания «содержательных аспектов» аудио, видео и других графических компонентов, в данной работе мы ограничиваемся следующим определением: «Семантический поиск — вид автоматизированного полнотекстового информационного поиска с учетом смыслового содержания слов и словосочетаний запроса пользователя и предложений текстов проиндексированных информационных ресурсов» [3].

Обращение к данной тематике вызвано нуждами проекта по разработке системы перевода русского текста в изображения жестов жестового языка, используемого глухими России [4, 5]. Как и в любой системе перевода, здесь также возникает задача снятия омонимии для выбора жеста, правильно отображающего смысл текстового высказывания.

Но в случае перевода со словесного на жестовый язык проблемы снятия омонимии отличаются от подобных задач при переводе с одного словесного языка на другой. Некоторые понятия, однозначно воспринимаемые в словесном языке, в жестовом языке приобретают несколько значений, которые должны быть выделены и разделены для генерации правильного перевода. Методики, базирующиеся на использовании тезаурусов [6], в данном случае применимы лишь частично, так как определение жеста, передающего нужное значение, требует более глубокой обработки контекста.

Использование локальных контекстов (см., например, [7]) явно недостаточно для снятия омонимии при переводе со словесного языка на жестовый язык. В связи с тем, что жесты передают не значения слов, а концепты, часто описываемые группой слов, нами предложен метод, расширяющий положения [7]. Описание метода изложено ниже.

Положения, лежащие в основе данного метода, прошли экспериментальную проверку [8] и могут использоваться при поиске в сети Интернет документов, содержащие новые знания.

2. Проблемы понимания текста

Одной из наиболее мощных современных систем, использующих технологии семантического поиска, является RetrievalWare компании Convera (ранее Excalibur), позиционирующей свою систему как «первую платформу по извлечению знаний» [9].

Основой семантического поиска в RetrievalWare является использование семантических сетей, описывающих смысл слов языка и связи между обозначаемыми ими понятиями. В [1] отмечается, что в данном случае под термином «семантическая сеть» понимается тезаурус, позволяющий не только найти слова, связанные

[1] * Работа осуществляется при финансовой поддержке фонда «Научный потенциал» (http://hcfoundation.ru, договор на получение гранта № 67 от 30.12.2005).

по смыслу с данным, но и определить количественно «семантическое расстояние между ними». Однако, как указывается в [6], «часто различные словарные источники дают различный набор значений многозначных слов, выделяют оттенки значений, причем один и тот же тип многозначности может быть по-разному описан для различных слов даже в одном и том же словаре».

Нужно заметить, что в ряде случаев под многозначностью маскируются *разные* слова, обозначающие разные понятия и имеющие разные, но частично совпадающие парадигмы. Так, например, в [10] отмечается «относительное равноправие форм *дире́кторы* и *директора́* в текстах разных типов», при этом замечается, что «форма мн. ч. *дире́кторы* давно признается архаичной».

Справедливо отмечая, что эти слова встречаются в «текстах разных типов», [10] не указывает типы текстов. Очевидно, в первом случае слово *дире́кторы* встречается в технических текстах, где это слово обозначает элемент антенны (например, телевизионной) или монтажной оправки. Во втором случае слово *директора́* встречается в текстах, где упоминаются административные должности и именно в этом случае форма мн. ч. *дире́кторы* практически не применяется.

Указанные два типа текстов относятся к различным предметным областям, и рассмотренные слова являются не разными формами одного и того же слова, а внешне похожими представителями разных миров, имеющими общего предка (лат. rectus).

Хотя общепринятым мнением является то, что существует единая научная картина мира, вероятнее всего, что это лишь идеал, к которому стремится наука. У каждого человека имеется собственная внутренняя картина мира, причем она конечна. Это доказывается тем, что всегда найдется вопрос, на который данный человек не сможет ответить. Единство человеческого общества (компании друзей, народности, цивилизации) достигается путем усвоения (в той или иной степени) в результате обучения единых (или близких) пониманий явлений окружающей действительности, образующих объекты внутренней картины мира. Но перечень этих одинаково с другими понимаемых явлений разный для каждого человека.

Совокупность совпадающих по значениям объектов внутренних картин мира составляет картину мира (предметную область) сообщества людей, например, профессионального. Но следует еще раз отметить, что внутренняя картина мира любого члена этого сообщества в целом будет отлична (пусть по некоторым элементам) от общей картины мира сообщества, в частности за счет разного понимания даже общепринятых толкований фактов и явлений.

Этим объясняются отмеченные в [6] отличия в словарях, составленных разными людьми.

Помимо отличий в картинах мира разных людей, внутренняя картина мира человека, вероятно, представляет совокупность отдельных предметных областей, слабо связанных (или не связанных) друг с другом. По всей видимости, этим объясняется сложность переноса известных в одних условиях взаимодействий объектов в другие условия (в новую предметную область), что является основой методологии ТРИЗ [11], используемой при целенаправленном создании новых изобретений.

Одно и тоже слово в разных контекстах приобретает разный смысл. Значения слова *море* в примерах (1) и (2) на рис. 1 очевидно различны. При этом предметные области для примеров (1) и (2) на рис. 1 также различны и не связаны друг с другом (если в примере (2) не подразумевается *море продуктов моря*). В то же время значения слова *море* в примерах (2), (3), (4) одинаковы, но предметная область для примера (2) отлична для предметных областей примеров (3) и (4). Наконец, предметная область примера (4) поглощает предметную область примера (3), что отображено на рисунке.

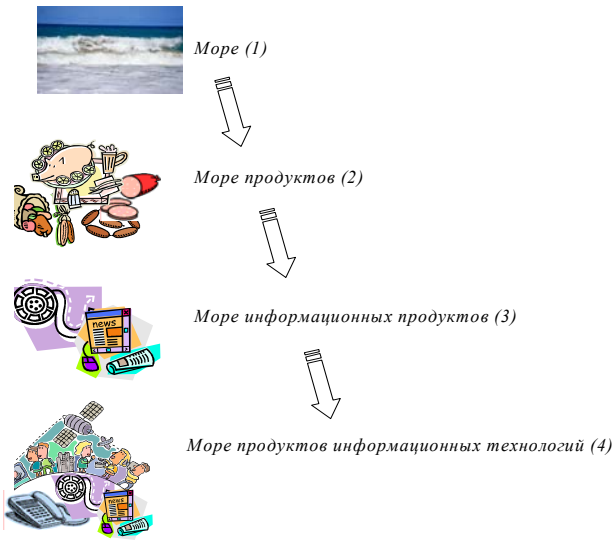

Рис. 1. Разные значения слов в разных предметных областях

Можно констатировать, что понимание одного и того же текста разными людьми в той или иной степени отличается. Соответственно, это влияет как на формулировку запроса, так и на оценку результатов его выполнения.

Учитывая наблюдения психологов, что «единицы мысли и единицы речи не совпадают» [12], а также изменчивость значений слов в зависимости от контекста, ясно, что результаты семантического поиска в тексте не могут быть абсолютно точными. Значение результата поиска может лишь с некоторой вероятностью соответствовать смыслу запроса, причем оценка соответствия не является абсолютной, а зависит от позиции оценивающего.

3. Методика семантического поиска

В приведенных выше примерах мы не выходили за пределы *локального* контекста, который обычно ограничен одним предложением.

В [7] сформулировано положение, что «Two *different* words are likely to have *similar* meanings if they occur in *identical* local contexts». Это положение используется для выбора нужного значения слова в словаре, используя пояснительные тексты словарных статей.

Однако при переводе текста на жестовый язык использование локального контекста оказывается недостаточным. Например, слово *забор* передается разными жестами, в зависимости от того, внутри или вне территории, окруженной забором, находится субъект [13]. Эта информация может быть почерпнута лишь из общего содержания текста.

Расширим положение D. Lin [7], сформулировав его следующим образом: *Два различных слова (или группы слов) могут иметь схожее значение, если они встречаются в идентичных (или схожих) контекстах*.

Но в этом случае слово (или группа слов), значение которого определяется путем сопоставления окружающего его контекста и контекста справочного слова (или группы слов), является лишь «заглушкой», резервирующей место в контексте. Эта «заглушка» может быть опущена, если предусмотрены меры, обеспечивающие соответствующий «зазор» в контексте.

Известен способ поиска схожих по семантике документов на основе сопоставления их лексических векторов [14]. Но при задании поискового запроса, включающего все (или наиболее значимые) лексические вектора документа-прототипа (или обобщенные вектора группы документов), документы, относящиеся к другим (хотя и сравнительно близким) предметным областям, будут отсеяны.

В то же время известен так называемый «сленговый метод» [15], используемый, например, при установлении авторства литературных произведений, по которому при сопоставлении документов исключаются наиболее частотные слова.

Предлагаемый метод семантического поиска является модификацией «сленгового метода» и использует замену «заглушкой» наиболее значимого лексического вектора запроса. В этом случае в результате поиска находятся документы, принадлежащие к разным (но сравнительно близким) предметным областям, описывающие разные (в том числе и ранее неизвестные для автора запроса) значения понятия, соответствующего исключенному при запросе лексическому вектору.

Описанный метод использован при поиске в сети Интернет и получил экспериментальное подтверждение [8]. В результате проведенных экспериментов получена индикативная математическая модель релевантности запроса:

$$Y_{2t} = 0{,}517 + 0{,}157A + 0{,}207AB - 0{,}264AC \; + + \; 0{,}122BC - 0{,}246ABC$$

Здесь символами *B* и *C* соответственно обозначены учет синтаксиса (порядка слов в запросе) и морфологии. Характерно, что значимым оказалось лишь взаимодействие этих факторов. Это свидетельствует, что при семантическом поиске необходимо одновременно учитывать как морфологию слов запроса, так и его синтаксис.

Очевидно, что для этого необходимо проводить анализ текстов полученных в результате поиска документов, так как индексы поисковых машин, сохраняя информацию о положении слов в документе (при координатном индексе), не хранят знаки препинания, играющие существенную роль при синтаксическом анализе текстов [16].

Роль знаков препинания обычно в системах синтаксического анализа принижается, отводя им лишь роль разделителей текста.

Однако учет знаков препинания в ряде случаев облегчает семантический анализ текста. В жестовом языке жест *«одежда»* имеет и значение *«мне все равно»* или *«мне плевать»*. Возникает задача разделения значений глагола *плевать*. Рассмотрим примеры:

Доктор, мне плевать. (5)
Доктор, мне плевать? (6)
Доктор, неужели Вы думаете, что мне плевать? (7)

Ситуация в примере (6) соответствует, вероятнее всего, посещению стоматологического кабинета и выражение *мне плевать* должно означать указанное действие, тогда как это выражение в примере (5) должно рассматриваться как выражение равнодушия. Разделение этих значений может быть осуществлено на основе анализа знаков препинания без сложного анализа контекста. Однако пример (7) показывает, что это не всегда возможно.

4. Другие случаи применения

Предлагаемый способ семантического поиска рассматривался выше применительно к задаче перевода текста в жесты. Но он может использоваться не только для этого.

Для преобразования WWW в Семантическую сеть необходимо в миллионах существующих документов расставить тэги, задающие значения URI [17] хотя бы основных терминов документа. Очевидно, что этот процесс может быть осуществлен только соответствующими программными агентами.

Метод его работы в общем случае тривиален: поиск онтологии, в которой есть описание данного термина; при наличии нескольких таких онтологий, уточнение значения термина по контексту (снятие омонимии) и выбор соответствующей онтологии; определение и запись в документ значения URI термина, что в дальнейшем обеспечивает однозначное определение этого термина.

Но что делать, если нужно задать URI термина, для которого не удается установить связь с какой-либо из онтологий, значение которого отсутствует в доступных словарях синонимов?

Представляется, что формирование запроса на основе предложения, содержащего такой термин, но в котором этот термин заменен пропуском, позволит найти в Сети синонимы этого термина, имеющие такой же окружающий контекст. Тогда значение этого термина (URI) может быть определено на основе значения синонима.

Результаты экспериментов показывают, что предлагаемый способ поиска оказывается достаточно эффективным и при поиске новой информации по заданной тематике. При этом возможен следующий алгоритм работы агента, ищущего для пользователя новую информацию в Сети WWW:

1. Имеющиеся документы по определенной тематике объединяются в одну группу.
2. В этой группе документов определяются наиболее часто встречающиеся слова и словосочетания, а также определяются мера близости документов и критерий отнесения документов к указанной группе.
3. Выделяются наиболее часто встречающиеся предложения, содержащие указанные наиболее часто встречающиеся слова и словосочетания.
4. Эти предложения используются в качестве поисковых запросов, причем указанные наиболее часто встречающиеся слова и словосочетания исключаются и заменяются операторами, указывающими поисковой машине, что в запросе имеется пропуск величиной n слов.
5. Результаты запроса сравниваются с документами группы (см. п. 1). Из результатов исключаются документы, совпадающие с уже имеющимися в группе, а также те, которые по заданному критерию не относятся к данной группе.
6. Релевантные результаты запроса добавляются в соответствующую группу документов.
7. Пп. 1-6 периодически повторяются.

5. Заключение

Предложенная методика семантического поиска позволяет отбирать близкие по общему контексту документы, даже если они принадлежат к разным предметным областям. Тем самым можно собирать и обобщать знания, рассредоточенные в различных областях.

Предполагается, что предложенная методика может быть полезной при преобразовании сети Интернет в Семантическую сеть.

Показано, что для создания программы перевода текста в жесты языка глухих России, необходимо использовать средства семантической обработки текста, при этом необходимо учитывать весь предшествующий контекст, использование локальных контекстов недостаточно для выявления правильных значений, которые нужно представить жестами.

Список литературы

1. *Бойцов И. Системы поиска по массивам неструктурированной информации // РЕЛИБ, 2003. (http://www.relib.com/articles/article.asp?id=216)*
2. *Зацман И.М. Семантический поиск научной информации: неоднородные коммуникативные компоненты и цветовая палитра объектов поиска // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. семинара Диалог'2002. М.: Наука, 2002. Т.2, С. 214-227.*
3. *http://www.itpromotion.ru/glossary/?word=9*
4. *Воскресенский А.Л. Компьютерный банк жестовой речи // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конференции Диалог'2003. М.: Наука, 2003. С. 688-691.*
5. *Voskressenski A. Signs and speech: two forms of human communication // Proceedings of the Ninth International Conference «Speech and Computer» SPECOM'2004. Saint-Petersburg, Russia, 2004. P. 666-669.*
6. *Лукашевич Н.В., Добров Б.В. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. семинара Диалог'2002. М.: Наука, 2002. Т.2, С. 338-346.*
7. *Lin D. Using syntactic dependency as local context to resolve word sense ambiguity // Proceedings of the 35th annual meeting on Association for Computational Linguistics. Madrid, Spain, 1997. P. 64-71.*
8. *Воскресенский А.Л., Хахалин Г.К. Формирование запросов к поисковой машине для извлечения знаний из Интернета // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конференции "Диалог'2005". М.: Наука, 2005. С. 86 – 91.*
9. *Papadopoullos A., Van Winkle J. RetrievalWare 8 — the Knowledge Discovery Platform // A Convera technical overview. (http://www.convera.com/whitepapers. TO-RW8-031111.pdf).*
10. *Беликов В.И. Yandex как лексикографический инструмент // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конференции Диалог'2004. М.: Наука, 2004. С. 39-46.*
11. *Альтшуллер Г.С. Творчество как точная наука: Теория решения изобретательских задач // М.: Сов. радио, 1979.*

12. *Выготский Л.С. Психология // М.: ЭКСМО-Пресс, 2000.*
13. *Фрадкина Р.Н. Говорящие руки: Тематический словарь жестового языка глухих России // М., 2001.*
14. *Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Тр. Пятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Санкт-Петербург, 2003. (http://rcdl2003.spbu.ru/proceedings/B1.pdf)*
15. *Хайтун С.Д. Наукометрия. Состояние и перспективы // М.: Наука, 1983.*
16. *Хахалин Г.К. Лингвистическая трансляция сложных и эллиптических ЕЯ-предложений // Тр. VIII Национальной конференции по искусственному интеллекту «КИИ-2002». Коломна, 2002. С. 251-256.*
17. *Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American, May 17, 2001.*