22.

ОНТОЛОГИИ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ: ОПИСАНИЕ ПОНЯТИЙ И ЛЕКСИЧЕСКИХ ЗНАЧЕНИЙ

ONTOLOGIES FOR NATURAL LANGUAGE PROCESSING: DESCRIPTION OF CONCEPTS AND LEXICAL SENSES

Б.В.Добров *(dobroff@mail.cir.ru)*

Н.В. Лукашевич *(louk@mail.cir.ru)*

Научно-исследовательский вычислительный центр МГУ
АНО Центр информационных исследований

Проблема соотношения понятия и значения приобретает практический характер при разработке онтологий для автоматической обработки текстов. В статье рассматриваются существующие подходы к проблеме описания понятий и значений в различных онтологиях, а также решения, принимаемые в процессе разработки компьютерного тезауруса русского языка РуТез.

Введение

Для того, чтобы сделать автоматическую обработку текстов более качественной и надежной, необходимо использовать знания и о языке, и об окружающем мире.

Знания о мире могут быть представлены с помощью онтологий - систем понятий, для которых описаны отношения и заданы правила вывода.

Чтобы применить онтологию для автоматической обработки текстов, необходимо понятиям онтологии сопоставить набор языковых выражений (слов и словосочетаний), которыми понятия могут выражаться в тексте.

Соответственно для лексической единицы ссылка на то или иное понятие становится в той или иной мере описанием значения этой единицы.

В процессе сопоставления онтологии и лексических значений, или разработки онтологии непосредственно для обработки текстов становится ясно, что многократно обсуждаемый в литературе вопрос о соотношении понятия и лексического значения приобретает совершенно практический характер.

На каждом шаге такой работы приходится решать:

- насколько понятия, вводимые в онтологию, должны быть связаны со значениями лексических единиц, терминов данной области;
- должны ли разработчики (и до какой степени) руководствоваться системой существующих языковых значений;
- насколько и в каких своих элементах значения должны быть сходны и могут отличаться лексические единицы, отнесенные к одному и тому же понятию, ставшие, таким образом, синонимами относительно онтологии – онтологическими синонимами.

Проблема соотношения описаний понятийных единиц и значений активно обсуждается как разработчиками онтологий [10, 12, 13], так и разработчиками лингвистических компьютерных ресурсов WordNet [14], EuroWordNet, FrameNet [8] . Эта тема является центральной для международного семинара OntoLex, проводимого с 2000 года.

В статье мы рассмотрим существующие подходы к проблеме описания понятий и значений в различных онтологиях, а также решения, принимаемые в процессе разработки компьютерного тезауруса русского языка РуТез [3].

1. Понятие и значение: практические аспекты

Проблеме соотношения семантического значения и понятия посвящено много литературы. Наиболее существенными для практики разработки онтологии для автоматической обработки текстов являются следующие положения [1, 2, 5].

Подчеркивается, что и понятие, и лексическое значение относятся к категориям мышления, при этом между ними есть существенные различия.

Значение включает в себя помимо понятийного содержания (сигнификативно-денотативного компонента значения), такие компоненты как оценочный, стилистический, сочетаемостный.

Значение включает лишь различительные черты объектов, иногда относительно поверхностные, а понятия охватывают их наиболее глубокие существенные свойства.

Описать значения многих слов как совокупности общих и одновременно существенных признаков может быть очень трудно. Обсуждая значения, лингвисты часто употребляют такие термины, как «прототипическое

значение», «оттенки значения», «вероятностный компонент значения» и т.п., что подчеркивает значительную нечеткость границ лексических значений. В связи с этим сигнификативный компонент значения слова называется наивным, бытовым понятием.

Считается, что значение и понятие совпадают лишь в сфере терминологии [5].

Бытовые понятия сложно представить в виде формальной системы, пригодной для логического вывода, например, описать таксономические связи, по следующим причинам:

1) из-за их нечеткости, расплывчатости;
2) контекстной зависимости, когда реализация некоторых компонентов значения существенно зависит от контекста;
3) существования значительных рядов синонимов, отличающихся оттенками значений, что затрудняет разбиение таких рядов на совокупность взаимосвязанных понятийных единиц. Например, сколько понятий онтологии оптимально (и на основе каких принципов) сопоставить следующему ряду слов со значением ОШИБКА: *ошибка, погрешность, недосмотр, просмотр, ляп, промах, оплошность, осечка, прокол, упущение, недочет,* а также *ослышка, описка, опечатка, оговорка.* Таким словам обычно трудно найти точные слова-соответствия в других языках, то есть слова, имеющие такой же оттенок значения и такие же особенности употребления.

Несмотря на описанные проблемы, разработка моделей представления знаний о мире и о языке в рамках онтологий имеет смысл, поскольку, как мы покажем далее:

- далеко не все бытовые понятия нечетки и расплывчаты, многие из них достаточно близки понятиям различных профессиональных сфер;
- многие из указанных проблем преодолимы;
- онтологическая модель дает возможность создания больших компьютерных ресурсов, сочетающих знания о мире и о языке, что чрезвычайно существенно для современных задач автоматической обработки текста и реальных предметных областей.

2. Способы сопоставления понятий и лексических значений в компьютерных ресурсах

Процедура сопоставления понятий онтологий и языковых выражений может быть осуществлена различными способами:

Во-первых, онтология может быть сделана заранее, путем логической классификации, а затем к ее единицам могут быть приписаны языковые единицы [10, 16]. Так, например, Doug Lenat [13], руководитель известного проекта в области представления знаний CYC, в рамках которого предполагалось формализовать знания здравого смысла (common sense) и использовать их, в частности, для обработки текстов на естественном языке, считает, что учет значений слов может только запутать ("words are often red herrings"), что значения слов делят мир неоднозначно, а линии деления происходят из самых различных причин: исторических, физиологических и т.п.

Предлагается создавать онтологию путем логического анализа, «сверху-вниз». При этом, имена вводимых понятий (желательно) должны отражать те признаки, которые заложены в основу деления. В результате получаются имена понятий достаточно громоздкие, неестественные, с ними трудно оперировать как разработчикам, так и возможным пользователям.

Другой проблемой такого подхода является то, что при приписывании языковых выражений к логически обоснованной системе понятий получается, что одно и то же слово может соответствовать слишком большому количеству таких «правильных» понятий в зависимости от контекста, возникает излишняя многозначность лексической единицы.

Второе направление, которое обычно обсуждается, это установление соответствий между иерархическими лексическими ресурсами из «семьи wordnet'ов» [12] и некоторой онтологией. Единицей в таких системах является значение отдельной лексемы или совокупности синонимов - синсет. Между такими единицами устанавливаются различные типы отношения, прежде всего иерархические, как, например, родовидовые отношения для существительных.

Предполагается, что разработчики такого рода ресурсов разрабатывают иерархию лексических значений естественного языка, а для более строгого описания знаний о мире необходимо сопоставить такие ресурсы с какими-либо формальными онтологиями [15].

Так, содержанием одного из проектов является установление отношений между WordNet и EuroWordNet с одной стороны и формальной онтологией SUMO - Standartized Upper Merged Ontology [6] с другой стороны. Проект состоит в том, чтобы установить соответствие между синсетами WordNet и понятиями онтологии, при котором каждый синсет WordNet либо напрямую сопоставляется с понятием онтологии, либо является гипонимом для некоторого понятия, либо примером понятия онтологии.

Участники другого проекта OntoWordNet [9] считают, что недостаточно провести формальную склейку ресурса типа WordNet и формальной онтологии, необходима значительная реструктуризация исходного лексического ресурса.

Третий путь – попытаться разработать единый ресурс, в котором были бы сбалансированы обе части: система понятий – и система лексических значений [11], что заключается в разумном разделении этих единиц в создаваемом ресурсе и аккуратном описании их взаимосвязей.

3. Сбалансированное описание онтологической структуры и значений лексических единиц

Приверженцем сбалансированного описания понятий онтологии и лексических значений является С. Ниренбург [6]. В онтологиях MicroKosmos и OntoSem (5 тысяч понятий) проводится четкое разделение онтологии и словаря. Онтология, по мнению Ниренбурга, должна быть максимально независимой от конкретного языка, в то же время понятия онтологии должны иметь свое непосредственное отражение в языковых значениях.

Словарная статья языкового значения в онтологии Ниренбурга может иметь и достаточно простую структуру, представляя собой ссылку на понятие онтологии, и достаточно сложную структуру, содержащую и ссылку на понятие онтологии и особенности конкретной лексической единицы.

Например, все глаголы изменения в онтологии приписаны одному и тому же понятию Change-event. Особенности слов описываются в словарной статье, например, для глагола увеличить (increase) указывается, что в семантической роли ТЕМА этого глагола должна выступать СКАЛЯРНАЯ_ВЕЛИЧИНА (например, цена или высота) и указывается, что значение этой величины меняется на большее.

Грэм Хирст [7, 11] указывает, что в устоявшуюся структуру ресурсов реляционной семантики, разделяемую на два уровня (концептуально-семантический и семантико-синтаксический), должен быть введен третий уровень внутрипонятийный/стилистико- семантический**.**

Понятийно-семантический уровень задает относительно грубую понятийную иерархическую систему, которая основывается на денотативных, независимых от контекста, свойствах значений слов.

Каждому понятию поставлен в соответствие набор синонимов, а их особенности (стилистическое употребление, отношение говорящего, коннотации и т.п.) описываются в дополнительных, внутрипонятийных структурах. Например, предполагается, что все слова, синонимичные слову *ошибка,* должны быть отнесены к одному и тому же понятию, а их различия должны быть отражены с помощью формализованного языка.

Хирст подчеркивает, что часто может оказаться, что определить, какие близкие по смыслу слова лучше описать в рамках внутренней структуры понятия, а какие разнести в разные понятия, очень непросто. Он, с одной стороны, надеется на интуицию лингвиста, с другой стороны, подчеркивает, что взгляд на понятийную структуру с точки зрения другого языка может лучше проявить границы понятий.

4. Отношение понятие-значение в тезаурусе русского языка РуТез

Наиболее точно «жанр» тезауруса РуТез можно охарактеризовать как лингвистическая онтология для автоматической обработки текстов, то есть это онтология, большинство понятий которой вводится на основе значений реально существующих языковых выражений.

Тезаурус русского языка РуТез представляет собой иерархическую сеть понятий. Каждое понятие имеет имя, отношения с другими понятиями, набор языковых выражений – текстовых входов (слов, словосочетаний, терминов), значения которых соответствуют этому понятию. Текущий объем тезауруса РуТез - 48 тысяч понятий, 122 тысячи текстовых входов (в том числе более 65 тысяч отдельных слов).

Имя понятия – это однозначное слово, слово с пометой или словосочетание, значение которого наиболее точно отражает суть понятия и при этом в большинстве случаев реально употребляется носителями русского языка.

С одной стороны, имя понятия лишь «этикетка», а понятие описывается своим местом в сети тезауруса. С другой стороны, понятность и однозначность имени существенно облегчает анализ качества описания понятия в тезаурусе и результатов автоматической обработки текстов на основе тезаурусных знаний.

Концепция тезауруса РуТез как онтологии определяет рассмотрение слов, относящихся к разным частям речи, но выражающие один и тот же смысл (деривативы), как онтологических синонимов (ср. концепцию WordNet [14]).

Развитие тезауруса русского языка РуТез началось с тематико-терминологического уровня, называемого Общественно-политическим тезаурусом. Общественно-политическая область описывает сферу общественной жизни современного общества и включает терминологию, относящуюся к таким сферам, как политика, экономика, военная сфера, промышленность, сельское хозяйство, социальная сфера, культура и др.

При разработке Общественно-политического тезауруса проблема различия понятия и значения практически не возникала, поскольку расхождения между понятием и значением минимальны не только в терминологии, но и в тематической лексике общезначимого языка, имеющим отношения к тем профессиональным областям, которые непосредственно контактируют с повседневной жизнью населения, таким как транспорт, строительство, банки, право и многие другие [4].

Это связано, во-первых, с тем, что в этой лексике наиболее значимым становится сигнификативно-денотативный компонент, и снижается доля других компонентов значения. Во-вторых, эта лексика постоянно взаимодействует с профессиональной терминологией и, таким образом, частично приобретает ее черты.

При работе с терминологией в рамках Общественно-политического тезауруса мы видели, что обычно для каждого понятия предметной области существует его однозначное, точное наименование (собственно термин в

терминоведении, дескриптор в информационно-поисковых тезаурусах). Часто таким однозначным наименованием является словосочетание.

В реальных текстах предметной области для ссылки на понятие помимо основных терминов может использоваться множество разнообразных языковых выражений: словообразовательные и синтаксические варианты термина, многозначные слова, разговорные слова. Все такие возможности описываются как текстовые входы понятия.

Например, анализируя толкование второго значения многозначного слова *позиция – «2. Место расположения войск в бою»,* можно сделать вывод, что его можно описать как текстовый вход понятия *ВОИНСКИЕ ПОЗИЦИИ* с текстовыми входами *позиция, воинские позиции, позиции войск* и др.

Таким образом, выработались следующие принципы работы с языковым материалом:

- если слово многозначное, то для лучшего представления его значений в тезаурусе подбирается однозначное выражение, имеющее тот же смысл. Например, для слова *полип* в толковом словаре два значения, одно из которых однозначно можно выразить словосочетанием *морской полип*;
- если значение слова расплывчатое, например, имеет варианты значения в толковании, то для каждого из этих вариантов по возможности подбирается однозначно именующее выражение.

Так, для описания значения лексемы *покрывало* («1. *Кусок ткани, предназначенный для покрывания чего-либо, покрывающий что-либо // легкое одеяло, обычно служащее для покрывания постели днем»)* вводятся два понятия *ПОКРЫВАЛО (покрывающая ткань*) и *ПОСТЕЛЬНОЕ ПОКРЫВАЛО*, как вид первого понятия, а сама лексема *покрывало* описывается как текстовый вход к обоим понятиям.

С помощью словосочетаний с четким значением и соответствующей системой понятий можно описать и совокупность близких по смыслу слов, как, например, слов-синонимов к слову ошибка. Мы можем описать сигнификаты этих слов с помощью такого набора понятий: *ОШИБКА, ГРУБАЯ ОШИБКА, МЕЛКАЯ ОШИБКА, СЛУЧАЙНАЯ ОШИБКА*.

Как неоднократно указывалось в литературе [11, 12], онтология должна минимально зависеть от конкретного языка.

Видно, например, что если сопоставить точный вариант на другом языке каждому слову-синониму слова *ошибка* невозможно, то значения вышеупомянутых словосочетаний можно хорошо и точно передать на других языках, что означает, что предложенное решение удовлетворяет критерию независимости от конкретного языка, предъявляемому к онтологиям.

Это решение многократно опробовано при переводе Общественно-политического тезауруса, который сейчас практически полностью переведен на английский язык, то есть теперь каждое понятие имеет имя на английском языке и набор английских текстовых входов.

Кажущиеся многочисленными лакуны в русском или английском языке во многих случаях удалось устранить с помощью реально употребляемых словосочетаний, сочинительных конструкций.

Интересным примером здесь является передача понятийного содержания понятия *ВЕКСЕЛЬ* на английском языке. Дело в том, что векселя делятся на простые векселя (promissory notes или просто notes) и переводные (bills of exchange или просто bills). Значению термина вексель соответствует конструкция *bills and notes* (80000 употреблений в Google).

Таким образом, мы стремимся организовать РуТез именно как сеть понятий, сигнификатов, а не как чисто лексическую иерархию. При этом стараемся максимально полагаться на сигнификаты существующих в языке лексем и словосочетаний, по мере возможности не вводя искусственные понятия, чтобы обеспечить большую понятность каждой понятийной единицы тезауруса.

Заключение

Теоретическая проблема соотношения понятия и лексического значения приобретает практический характер при разработке онтологий для автоматической обработки текстов.

При создании онтологических ресурсов, предназначенных для обработки текстов, предложены различные подходы. Наиболее перспективным нам представляется подход сбалансированного описания онтологической структуры и значений лексических единиц (MicroKosmos, OntoSem, G.Hirst). Аналогичный подход используется нами при разработке тезауруса РуТез и позволил построить большой ресурс (48 тысяч понятий, 122 тысяч текстовых входов).

Основным источником понятийных единиц тезауруса являются сигнификаты значений слов и выражений русского языка. На нашем опыте мы убедились, что сигнификаты многозначных слов или слов с расплывчатым значением можно четче выделить и описать, подобрав соответствующие, однозначные по смыслу словосочетания, выражающие не зависящие от контекста свойства сигнификатов.

Поэтому мы позиционируем тезаурус русского языка РуТез как лингвистическую онтологию, иерархическую сеть понятийно-сигнификативных единиц.

Список литературы

1. *Апресян Ю.Д. Лексическая семантика. Синонимические средства языка. М.: Восточная литература, 1995.*
2. *Гак В.Г., Лексическое значение слова – Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990.*

3. *Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002. – М.: Наука, 2002. Т.2. С.338-346.*
4. *Лукашевич Н.В, Добров Б.В. Взаимодействие лексики и терминологии в общезначимой сфере языка // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004, 2004. С.172-178.*
5. *Степанов Ю.С. Основы общего языкознания. Учебное пособие. М.: Просвещение, 1975.*
6. *Atserias J., Clament S., Rigau G. Toward the Meaning Top Ontology: Sources of Ontological Meaning. - Proceedings of International conference Language Resources and Evaluation (LREC-2004), 2004. V.1, p. 11-14.*
7. *Edmonds P., Hirst G. Reconciling fine grained lexical knowledge and coarse-grained ontologies in representation of near-synonyms. Proceedings of workshop on Semantic Approximation, Granularity and Vagueness, Breckenridge, Colorado, 2000.*
8. *Fillmore C.J., Miriam R.L. Petruck J.R., Abby W. Framenet in Action: The Case of Attaching, International Journal of Lexicography, 2003. Vol 16.3: 297-332.*
9. *Gangemi A., Navigli R., Velardi P. The OntoWordNet project: extension and axiomatisation of conceptual relations in wordnet // International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), Catania (Italy). 2003.*
10. *Gruber T.R. A translation approach to portable ontologies. Knowledge Acquisition, 1993. 5(2):199-220.*
11. *Hirst G. Ontology and the Lexicon. - Handbook on Ontologies in Information Systems, Berlin - Springer, 2003.*
12. *Hovy E., Nirenburg S. Approximating an interlingua in a principled way. Proceedings of the DARPA Speech and Natural Language Workshop, Hawthorne, NY, 1992.*
13. *Lenat D., Miller G., Yokoi T. CYC, WordNet, and EDR: critiques and responses // Communications of the ACM. Volume 38 , Issue 11 (November 1995), p. 45 - 48.*
14. *Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Five papers on WordNet. - CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990.*
15. *Prevot L., Borgo S., Oltramari A. Interfacing Ontologies and Lexical Resources. In the proceedings of OntoLex-2005, 2005. p. 91-102.*
16. *Reed S., Lenat D. Mapping ontologies into Cyc // AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada, 2002.*