Квазисинонимы в лингвистических онтологиях

Near-synonyms in linguistic ontologies

Лукашевич Н. В. (louk@mail.cir.ru)

Научно-исследовательский вычислительный центр МГУ
им. М. В. Ломоносова; АНО Центр информационных исследований

Одной из важных проблем разработки лингвистических онтологий является вопрос представления значений квазисинонимов посредством набора дискретных понятий онтологии. В статье кратко рассматриваются подходы к представлению квазисинонимов в тезаурусе WordNet и онтологии МикроКосмос, а также подробно описываются принципы описания квазисинонимов в тезаурусе русского языка РуТез.

1. Введение

Для обработки текстов на естественном языке разрабатываются такие ресурсы как лингвистические онтологии, то есть онтологии, понятия которых в значительной мере связаны со значениями языковых единиц, терминов предметной области.

Одной из серьезных проблем разработки лингвистической онтологии является выработка принципов формирования единиц (понятий) онтологии. Понятия онтологии должны соответствовать понятиям предметной области [9], а, как известно, взаимоотношения между понятиями и языковыми значениями достаточно сложны.

Общие рекомендации по вводу понятий в онтологии заключаются в том, что понятие онтологии должно отчетливо отличаться от близких понятий в понятийной сети (родовых понятий, видовых понятий, понятий-сестер). Кроме того, нужно различать понятие и его названия: не стоит заводить отдельные понятия для синонимов [1, 6, 8]. Эти рекомендации не так просто выполнять, если онтология создается на основе реальных языковых значений [3].

Во-первых, непросто отличать понятия и его названия, работая с языковыми значениями. Во-вторых, серьезную проблему представляют совокупности близких по смыслу слов — квазисинонимов, значения которых различаются по нескольким характеристикам (понятийному содержанию, отношению говорящего, коллокациям и др.), видоизменяются в зависимости от контекста. Для многих таких совокупностей квазисинонимов чрезвычайно трудно установить однозначное соответствие на других языках, поскольку, чаще всего, на другом языке данной совокупности квазисинонимов соответствует другая совокупность квазисинонимов, которая характеризуется своей системой параметрических различий и соответственно своими особенностями.

Проблема представления значений квазисинонимов в лингвистической онтологии состоит в том, что не всегда ясно, как наилучшим образом представить совокупность близких значений квазисинонимов набором отдельных понятий. Лингвистическая онтология, которая хоть и учитывает существующие лексические значения, все же должна оставаться онтологией. По общим принципам организации онтологической иерархии основные элементы онтологии — понятия должны иметь четкие, независимые от контекста отличия от соседних понятий.

Создание различимых понятий в лингвистической онтологии важно еще и тем, что четкое понимание различий близких понятий позволяет аккуратно описать их отношения между собой и с соседними понятиями. Кроме того, различимые понятия онтологии легче описать средствами других языков.

В данной статье мы кратко рассмотрим подходы к представлению квазисинонимов в таких лингвистических онтологиях как WordNet [5] и MikroKosmos [6], а также подробно опишем принципы описания квазисинонимов в Тезаурусе русского языка РуТез [11], который мы также рассматриваем как лингвистическую онтологию.

2. Квазисинонимы в лингвистической онтологии WordNet

Наборы синонимов — синсеты — являются основными структурными элементами тезауруса

WordNet [5]. Авторы данного тезауруса считают два выражения синонимичными и относят их к одному и тому же синсету, если замена одного из них на другое в большинстве предложений не меняет значения истинности этого высказывания.

Этот основной принцип устройства WordNet приводит к тому, что не выполняется один из важнейших принципов разработки онтологий — это различение собственно понятия и способов его называния, то есть вводятся разные синсеты для разных способов наименования одной и той же сущности.

Имеется несколько типов смешений понятий и их названий в ресурсах типа WordNet.

Во-первых, смешение понятий и их названий проявляется в поддержке разных иерархий для разных частей речи. Действительно, с помощью, какой бы части речи в тексте не было бы упомянуто понятие ПРИВАТИЗАЦИЯ (*приватизировать, приватизационный, приватизация*) — это всегда ссылка на одно и то же понятие разными лексическими средствами, от изменения части речи не должны меняться отношения этого понятия с другими понятиями.

Первые разработчики ворднетов для других языков в рамках проекта EuroWordNet рассматривали возможность соединения всех дериватов в одной иерархии, поскольку такое разделение противоречит принципам разработки онтологических ресурсов. Однако, в конце концов, решение о соединении частей речи принято не было [2].

Вторым типом проявления смешения понятия и его названия является использование разных синсетов для описания старых и новых названий, названий понятия в разных диалектах языка, в разных текстовых жанрах и т. п.

Следствием принципа синонимичной подстановки является то, что WordNet имеет значительное количество синсетов, которые плохо отличимы друг от друга, что также нарушает онтологические принципы описания понятий. Так, например, имеется четыре различных синсета, обозначающие *сходство, подобие*, каждый следующий из которых является гипонимом для предыдущего и при этом является практически не отличимым от своего гиперонима:

sameness — *(the quality of being alike; "sameness of purpose kept them together")*

similarity — *(the quality of being similar)* — сходство

likeness, alikeness, similitude — *(similarity in appearance or character or nature between persons or things; ``man created God in his own likeness'')* — сходство по внешности, характеру или природе между людьми или объектами).

resemblance — *(similarity in appearance or external or superficial details)* — сходство во внешности или во внешних или поверхностных деталях.

3. Квазисинонимы в онтологии МикроКосмос

В лингвистической онтологии МикроКосмос [6] собственно онтология и лексикон разделены. Лексикон системы описывает значения слов и словосочетаний, устанавливая ссылки на понятия онтологии. Проблема квазисинонимов решается за счет объединения квазисинонимов к одному и тому же понятию онтологии, особенностей конкретных лексем описываются в словарных статьях словаря.

Авторы онтологии приводят пример, что все глаголы изменения в онтологии приписаны одному и тому же понятию CHANGE-EVENT [7]. 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.

Если мы обратимся к сайту ресурса, то мы увидим, что ситуация с реализацией изложенных принципов достаточно сложная. Так, понятию CHANGE_EVENT сопоставлен в лексиконе большой список слов, которые, по мнению авторов, онтологии соответствуют этому понятию, например: *acclimatization* (акклиматизация — приспособление к другому климату), *commerzialization* (коммерциализация), *contamination* (загрязнение), *damage* (повреждать), *deteriorate* (ухудшать), *improve* (улучшать) и многие другие — для этих слов не было заведено отдельных понятий.

В то же время среди нижестоящих по иерархии понятий можно увидеть следующие: ADJUST (адаптировать, приспособить), CORRECT-EVENT (исправление, коррекция), DIVIDE (делить), INTEGRATE (интегрировать), RESTRUCTURE (реструктуризация) и др. Непонятно, почему для одних значений слов были заведены отдельные понятия, а для других нет. Почему значение слова *acclimatization* не заслуживает отдельного понятия, хотя есть важное отношение к климату, биологическим процессам, а значение слова *adjust* такой концепт получило?

Помимо вопросов последовательности/непоследовательности описания имеются и явные последствия для процедур автоматической обработки текстов.

Так, сложной становится процедура установления, какие все-таки слова из большего списка словарных входов к понятию CHANGE-EVENT, могут рассматриваться как синонимы, каковы соотношения между этими словами. Кроме того, относительно небольшая величина онтологии приводит к тому, что при работе в конкретном приложении и конкретной предметной области многое придется доделывать и вводить дополнительные понятия даже для слов, которые уже учтены в онтологии.

Таким образом, на наш взгляд, в приведенных примерах MikroKosmos проблема квазисинонимов

решается путем чрезмерного переобобщения, что может привести к проблемам в реальных предметных областях. Необходимо выделить дополнительный уровень понятий, который поможет более четко разделить слова, не сваливая их в единый, большой мешок.

4. Понятия и значения 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.

Значения языковых выражений, которые могут породить отдельное понятие в тезаурусе РуТез, относятся не только к общей лексике, но и могут являться терминами конкретных предметных областей, относящихся к сфере общественной жизни (экономика, право, международные отношения, политика), к сферам обслуживания населения (транспорт, банки и др.). Это связано с тем, что жизнь конкретных групп населения значительно связана с некоторыми профессиональными сферами деятельности, многие понятия из этих сфер легко перетекают в сферу общего языка, могут начать обсуждаться в общезначимых источниках информации (газетах, новостных сообщениях).

В качестве источников понятий в тезаурусе РуТез также активно используются словосочетания. Основным принципом введения такого рода понятий является необходимость фиксации некоторой дополнительной информации, которую невозможно описать в понятиях, соответствующих значениям слов — компонентов словосочетания.

В тезаурусе РуТез единицей является не множество синонимичных слов или терминов как в тезаурусе WordNet, а понятие — как единица мышления, обобщающая предметы и явления действительности [9]. Для ссылки на понятие в тексте могут использоваться несколько синонимичных текстовых выражений. Слова и словосочетания, значения которых могут быть представлены как ссылки на одни и те же понятия тезауруса, будем называть онтологическими синонимами.

Таким образом, онтологическими синонимами могут являться:

- слова, являющиеся разными частями речи (*стабилизация, стабилизироваться, стабилизационный),*
- языковые выражения, относящиеся к разным языковым стилям (*коммунальная квартира, коммуналка*),
- отдельные слова, устойчивые выражения, свободные словосочетания, значения которых соответствуют данному понятию (*аэропорт- воздушные ворота, газ — газообразное вещество)*.

Каждое понятие в тезаурусе имеет понятное, однозначное имя, что важно для ведения тезауруса, анализа результатов автоматической обработки текстов. В качестве названия может выступать однозначное словосочетание, однозначное слово, или пара синонимов, пересечение значений которых однозначно идентифицирует данное понятие.

В настоящее время тезаурус РуТез включает более 52 тысяч понятий, к которым приписано более 160 тысяч слов и словосочетаний. Тезаурус используется для таких видов автоматической обработки текстов как автоматическое концептуальное индексирование, автоматическое рубрицирование, аннотирование, кластеризация.

5. Принципы представления значений квазисинонимов в тезаурусе РуТез

Поскольку в настоящее время понятия тезауруса РуТез не имеют внутренней структуры в виде фреймовых элементов или атрибутов, то отличительные свойства понятий могут проявляться в наборе отношений с другими понятиями или в особенностях ассоциированных с понятием онтологических синонимов.

Для описания набора близких по смыслу значений квазисинонимов посредством набора различимых понятий лингвистической онтологии в РуТез применяется следующая процедура, которую мы рассмотрим на примере синсетов из WordNet, отражающих значение сходства (см. п. 2)

На первом шаге необходимо выделить наиболее существенные для тезаурусного описания признаки квазисинонимов, то есть такие признаки, в зависимости от которых требуется установление разных отношений с другими понятиями тезауруса.

В совокупности английских слов со значением сходства (similarity), таким признаком, например, является способность выражать сходство по внешним характеристикам. Значения некоторых квазисинонимов этой группы часто применяются именно к внешним характеристикам предметов, то есть в тезаурусе должно быть обозначено отношение к соответствующему понятию:

likeness, alikeness, similitude — *(similarity in appearance or character or nature between persons or things; "man created God in his own likeness")* — сходство по внешности, характеру или природе между людьми или объектами.

resemblance — *(similarity in appearance or external or superficial details)*.

Это означает, что в языке, жизни людей значимым является сходство по внешним характеристикам и нужно отразить этот факт соответствующим понятием.

На втором шаге необходимо подыскать подходящее название такому понятию

В случае квазисинонимов к слову *similarity,* таким названием понятия может служить словосочетание *Similarity in appearance* (34 700 страниц в поисковой системе Google). Понятие вводится в тезаурус с таким названием.

На третьем шаге необходимо найти разные способы выражения этого же понятия в виде словосочетаний и отдельных слов, например, *resemblance in appearance, similarity of appearance, external resemblance* и др. Все эти варианты добавляются в качестве текстовых вариантов к понятию (рис. 1).

На четвертом шаге для отражения значений слов, которые часто выражают именно это понятие, но могут использоваться и для выражения сходства вообще, например, *resemblance*, такое слово указывается как текстовый вход к понятию SIMILARITY IN APPEARANCE и как текстовый вход к более общему понятию SIMILARITY.

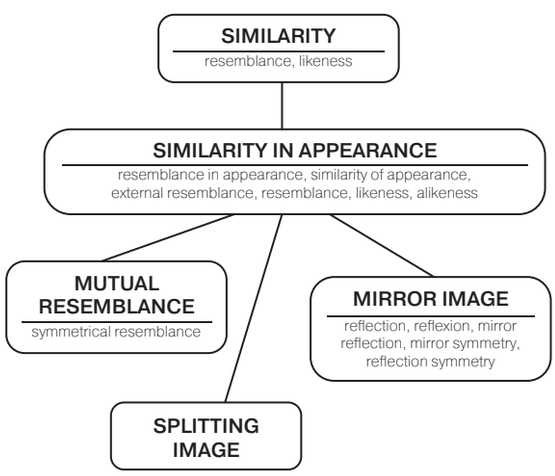

Рис. 1. Фрагмент совокупности отличимых понятий, отражающих значения квазисинонимов слова similarity. Заглавными буквами указаны имена понятий, прописными — соответствующие текстовые входы понятия. Текстовые входы, совпадающие с именем понятия, не указаны

В случае, если независимых от контекста характеристик для различения значений квазисинонимов, найти не удается, то необходимо представить их в виде одного понятия. Для большей ясности имя такого понятия может быть составлено как пара соединенных в этом понятии синонимов.

В качестве примера анализа значений квазисинонимов на русском языке возьмем синонимические ряды, представленные в синонимическом словаре НОСС [12]. Этот словарь интересен тем, что его словарная статья содержит подробный перечень сходных черт и различий синонимов. На основе такой словарной статьи разбора удобно показать, какие различия приводят к представлению синонимического ряда словаря в виде онтологических синонимов одного и того же понятия, а для значений каких слов, представленных в данном словаре, как синонимы, введены несколько понятий, и, таким образом, в рамках тезауруса РуТез они синонимами не являются.

В качестве первого примера рассмотрим пару синонимов *памятник, монумент.*

В словаре НОСС [12, стр. 257] указываются пять различий употребления этих слов (по величине, форме, увековечиваемому объекту и др.). Анализ примеров употребления этих синонимов показывает, что указанные различия выполняются лишь по умолчанию, имеется достаточное число примеров употребления обоих синонимов в связи со всеми возможными типами увековечиваемых сущностей. Так, авторы словаря утверждают, что «в память о конкретном человеке обычно ставится памятник, о группе людей — и памятник, и монумент, о событии — монумент; идеи воплощаются в монументах».

Между тем, в память о конкретном человеке может быть установлен монумент:

Монумент выдающемуся исследователю севера Западной Сибири, лесоводу, этнографу Александру Дунину-Горкавичу торжественно откроется в Ханты-Мансийске. (http://ural.rian.ru/culture/20070614/81566803.html).

В память события может быть установлен памятник:

На Пролетарской площади вновь оборудован сквер, в котором установлен памятник Победы (http://www.megatula.ru/site/tulskii_krai/raionnye_centry/67/)

Памятник может быть поставлен идее:

Он сказал, что это не первая акция вандалов в отношении памятника русско-армянской дружбы (http://www.patriarchia.ru/db/text/56928.html)

Кроме того, авторы словаря указывают, что различия «нейтрализуются при повторной, сокращенной номинации того же сооружения». Таким образом, у слов *монумент* и *памятник* не нашлось ни одного четкого различающего свойства, которые привели бы к установлению разных отношений с другими понятиями тезауруса, поэтому эти два слова должны рассматриваться как онтологические синонимы.

В качестве второй пары синонимов, которую мы проанализируем с помощью словаря НОСС [12], рассмотрим пару слов *водитель, шофер.*

При рассмотрении этих слов авторы словаря указывают следующее различие: «шофер управляет только автомобилем или автобусом, водитель и другими транспортными средствами [12, стр.53]». Из этого замечания понятно, что *шофер* и *водитель* не могут быть онтологическими синонимами, поскольку водитель должен иметь отношения с понятиями, соответствующими словам *вагоновожатый, судоводитель*, а *шофер* — нет. Это означает, что для отражения значений этих слов необходим ввод, по крайней мере, двух понятий с названиями *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* и *ВОДИТЕЛЬ АВТОМОБИЛЯ*. Видовыми понятиями для понятия *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* будут такие понятия как *ВАГОНОВОЖАТЫЙ*, *СУДОВОДИТЕЛЬ*.

В то же время, носители языка ощущают эти слова как синонимы (см. также [10]). Чтобы отразить и это ощущение, и способность расширительного употребления, необходимо слово *водитель* представить как текстовый вход к двум понятиям *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА и ВОДИТЕЛЬ АВТОМОБИЛЯ*.

Сначала представляется, что слово *шофер* должно быть отнесено как текстовый вход к понятию *ВОДИТЕЛЬ АВТОМОБИЛЯ,* но можно заметить, что водители автомобилей могут быть любителями, и профессиональными работниками, а слово *шофер* все-таки относится к профессиональным водителям. Таким образом, онтологический анализ пары синонимов показал, что для адекватного отражения системы понятий, скрывающихся за близкими по смыслу словами *водитель* и *шофер*, нужно использовать три понятийные единицы: *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА, ВОДИТЕЛЬ АВТОМОБИЛЯ, ШОФЕР (ПРОФЕССИОНАЛЬНЫЙ ВОДИТЕЛЬ)* (см. рис. 2).

Рис. 2. Понятийная структура, соответствующая близким по значению словам ***водитель*** и ***шофер***

Необходимость принятия решений о представлении значений близких по смыслу языковых выражений посредством совокупности понятий возникает и в конкретных предметных областях.

Так, ситуации кредитования соответствуют такие слова и словосочетания как: *кредитование, кредит, кредитная услуга, кредитное обслуживание, кредитная операция, выделение кредита, выдача кредита, выделение кредитных средств, предоставление кредита* и др. Имеется специфика употребления конкретных выражений из этого списка. Однако неправильным является введение дополнительных понятий онтологии для отражения именно специфики употребления. И в данном случае каждое вводимое понятие должно иметь четкий набор отличительных отношений. До тех пор, пока такие отличия не выделены, все такие выражения должны представляться как онтологические синонимы.

Заключение

Развивая тезаурус РуТез как лингвистическую онтологию, мы пытаемся следовать двум, вообще говоря, противоречивым критериям.

С одной стороны, мы формируем понятия тезауруса максимально близко к значениям языковых выражений, поскольку считаем, что чрезмерное обобщение, кластеризация значений ведет к искажению системы отношений, проблемам в приложениях автоматической обработки текстов.

С другой стороны, мы стараемся, чтобы понятие тезауруса было действительно понятием, то есть было отличимо от близких по смыслу понятий. Во многих случаях использованием реально существующих многословных выражений позволяет нам смягчить эти противоречивые требования. Введение понятия на базе значения многословного выражения не меняет суть лингвистической онтологии, но во многих случаях позволяет ввести более отчетливо отделимые понятия.

Литература

1. *Bouaud J., Bachimont B., Charlet J., Zweigenbaum P.* Methodological principles for structuring an "ontology"// Proceedings of IJCAI-95 Workshop "Basic Ontological Issues in Knowledge Sharing", 1995.
2. *Climent S., Rodriguez H., Gonzalo J.* // Definitions of the links and subsets for nouns of the EuroWordNet project. — Deliverable D005, WP3.1, EuroWordNet, LE2-4003, 1996.
3. *Hirst G.* Ontology and the Lexicon // Staab S., Studer R. (eds.) Handbook on Ontologies in Information Systems. Berlin: Springer, 2003. P. 209–230.
4. *Magnini B., Speranza M.* Merging Global and Specialized Linguistic Ontologies // Proceedings of OntoLex 2002, 2002. С. 43–48.
5. *Miller, G., Beckwith, R., Fellbaum, C., Gross D., Miller K.* Five papers on WordNet // CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.
6. *Nirenburg S., Raskin V.* Ontological Semantics. MIT Press: 2004.
7. *Nirenburg S., McShane M., Beale S.* The Rationale for Building Resources Expressly for NLP. // Proceedings of the 4st International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal, 2004. P. 3–6.
8. *Noy N. F., McGuinness D.* Ontology Development 101: A Guide to Creating Your First Ontology // Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.
9. *Smith B.* Beyond Concepts: Ontology as Reality Representation // Proceedings of International Conference on Formal Ontology and Information Systems FOIS-2004, 2004.
10. *Александрова З. Е.* Словарь синонимов русского языка // М.: Русский язык, 1999.
11. *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 / Под ред. А. С. Нариньяни. М.: Наука: 2002. Т. 2. С. 338–346.
12. *НОСС.* Новый объяснительный словарь синонимов русского языка. Третий выпуск. Под общим руководством акад. Ю. Д. Апресяна. М.: Языки славянской культуры, 2003.