OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ЛИНГВИСТИЧЕСКАЯ ОНТОЛОГИЯ – ТЕЗАУРУС РУТЕЗ

Алексеев А.А., Добров Б.В., Лукашевич Н.В.

НИВЦ МГУ им. М.В. Ломоносова,
г. Москва, Российская Федерация
a.a.alekseevv@gmail.com
dobrov_bv@mail.ru
louk_nat@mail.ru

В статье рассматривается современное состояние тезауруса русского языка РуТез, сравнивается его структура с другими концептуальными компьютерными ресурсами, рассмотрены типы используемых в различных приложениях путей тезаурусных отношений и окрестностей понятий тезауруса. В качестве примера применения тезауруса РуТез в автоматической обработке текстов рассматривается технология выявления тематической структуры текста путем построения лексических цепочек на основе нескольких факторов близости между языковыми выражениями.

Ключевые слова: лингвистическая онтология; автоматическая обработка текстов; тезаурус русского языка; лексические цепочки

Введение

Одним из наиболее известных лексических ресурсов в сфере компьютерной лингвистики и автоматической обработки текстов является компьютерный тезаурус WordNet [Miller, 1998], который в последней версии (3.0) включает приблизительно 155 тысяч различных слов и словосочетаний, организованных в 117 тысяч понятий, или совокупностей синонимов (synset).

Появление этого тезауруса в свободном доступе в Интернет вызвало всплеск исследований по его использованию в различных компьютерных приложениях автоматической обработки текстов, породил большое число последователей в разных странах, создающих такие ворднеты для своих языков [Vossen, 1998], а также стал базой для многоплановых дискуссий и исследований того, на основе каких принципов должны строиться большие лингвистические ресурсы, пригодные для разнообразных приложений в области компьютерной лингвистики.

Разработчиков новых ворднетов можно разделить на две категории. Часть разработчиков считает, что важным делом является точное воспроизведение структуры и состава англоязычного WordNet (обычно называемого Принстонский WordNet по месту работы его авторов), поскольку предполагается, что таким образом обеспечивается более тесная связь с англоязычным ресурсом и лексической системой английского языка. Другие разработчики полагают, что для создания качественного ресурса собственного языка необходимо учесть специфику его лексической системы, а также учесть критику и проблемы Принстонского WordNet. При таком подходе разработчики развивают собственную структуру синсетов, руководствуясь общими принципами построения ворднетов.

На текущий момент в России отмечен большой интерес к ресурсам типа WordNet. Однако начатые проекты разработки русских ворднетов [Азарова, Синопальникова, 2003; Сухоногов, Яблонский, 2005; Гельфенбейн и др., 2003] не достигли объема и/или качества, необходимых для их использования в приложениях.

В данной статье будут описаны структура и современное состояние тезауруса русского языка РуТез, который относится к тому же классу лингвистико-онтологических ресурсов, что и WordNet, но при этом обладает значительными отличиями [Лукашевич, 2011]. В ближайшее время мы планируем начать готовить тезаурус РуТез к публикации и, таким образом, данный ресурс может стать источником семантических знаний в

различных системах автоматической обработки текстов подобно ресурсам типа WordNet.

1. Принципы разработки тезауруса РуТез

Разработка тезауруса РуТез началась в 1998 году на базе Общественно-политического тезауруса [Лукашевич, 2011]. Основным назначением тезауруса является использование его как источника знаний о мире и языке в приложениях информационного поиска и автоматической обработки текстов.

По принципам своей разработки тезаурус РуТез объединяет три существующие традиции разработки компьютерных ресурсов:

- традиции разработки традиционных информационно-поисковых тезаурусов;
- традиции разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- традиции создания формальных онтологий.

Так как предполагается использовать лингвистический ресурс в автоматическом режиме обработки текстов, то необходимо использовать методологию разработки лексических ресурсов типа WordNet, в которой важны следующие положения:

- единицы тезауруса РуТез создаются на основе значений реально существующих языковых выражений;
- многоступенчатое иерархическое построение лексико-терминологической системы понятий;
- принципы описания значений многозначных слов и выражений.

Вместе с тем единица тезауруса РуТез рассматривается не как совокупность синонимов – синсет, а как понятие, имеющее уникальные свойства и уникальное имя в иерархически организованной системе понятий, как это рассматривается в онтологических исследованиях. Кроме того, для описания отношений между понятиями используются отношения с формальными свойствами и аксиомами.

Наконец, практика принятия решений ввода понятий на основе значений многословных выражений близка к практике разработки традиционных информационно-поисковых тезаурусов.

Таким образом, тезаурус РуТез принадлежит к особому классу онтологий, так называемым лингвистическим онтологиям, поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса. В то же время он является тезаурусом, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте, – такой набор текстовых входов понятий необходим для использования онтологий для автоматической обработки текстов.

В настоящее время тезаурус РуТез включает в себя 53 тысячи понятий, более 210 тысяч отношений между понятиями, 155 тысяч слов и словосочетаний русского языка

2. Единицы тезауруса РуТез

Единицей описания в тезаурусе РуТез является не множество синонимичных слов или терминов, как в тезаурусе WordNet, а понятие, отражающее значимые классы сущностей, различаемых людьми в мире, в современной общественной жизни, в психической жизни людей. При этом предполагается, что значения слов и выражений, существующие в современных естественных языках, позволяют выделить главное, существенное для современной жизни людей.

При такой ориентации на значения реально существующих выражений естественного языка важным принципом ввода понятий в тезаурус является отличимость каждого вводимого понятия от других понятий в системе понятий. Такая цель иногда трудно достижима, но, на наш взгляд, она должна ставиться, поскольку использование в качестве единиц тезауруса таких отличимых понятий позволяет единым образом представлять лексические значения литературного языка и значения терминов предметной области, более последовательно описывать систему отношений между понятиями и тем самым облегчает формальный вывод на отношениях, позволяет отображать единым образом систему значений разных языков

В тезаурусе РуТез каждое понятие должно иметь однозначное имя, которое построено на базе его текстовых входов, и должно быть понятным носителю языка. Понятие может иметь комментарий, который пишется в случае необходимости и не является частью имени понятия.

Каждое вводимое понятие должно быть снабжено списком слов и словосочетаний, с помощью которых можно сослаться в тексте на вводимое понятие – текстовых входов. В качестве таких текстовых входов могут быть отдельные слова (существительные, прилагательные, глаголы, наречия), а также именные и глагольные группы. Текстовый вход может быть многозначным (иметь другие значения), тогда он должен быть помечен как многозначный. Для лучшего распознавания в тексте текстовые входы тезауруса РуТез снабжаются последовательностью нормализованных форм всех составляющих многословного выражения (мужской род, именительный падеж, единственное число).

Языковые выражения (слова, словосочетания, термины), которые были описаны как текстовые входы одного и того же понятия, становятся

неразличимыми с точки зрения тезауруса РуТез – онтологическими синонимами.

Понятия в тезаурусе РуТез могут иметь достаточно большие ряды онтологических синонимов. Приведем пример синонимического ряда, включающего несколько типов синонимов для понятия *ДУШЕВНОЕ СТРАДАНИЕ* (по алфавиту): *боль (м – многозначное), боль в душе, в душе наболело, душа болит, душа саднит, душевная боль, душевная рана, душевное страдание, душевный недуг, рана в душе, рана в сердце, рана души, саднить, саднить на душе, саднить на сердце, сердечная рана*

Как видно, синонимический ряд понятия может содержать значительно количество синтаксических вариантов словосочетаний, некоторые словосочетания образуются заменой слова-компонента на синоним. Установление соответствия таких текстовых входов понятию является наиболее простым способом обнаружения понятия в тексте.

В тезаурусе РуТез существуют два основных способа представления значений многозначных терминов.

Первым способом представления многозначности является задание одного и того же текстового входа разным понятиям тезауруса (М-многозначность). Например, текстовый вход *пилот* сопоставлен двум разным понятиям понятию *ЛЕТЧИК* и понятию *АВТОГОНЩИК*. Такое представление полностью соответствует представлению разных значений слов, принятому в тезаурусах типа WordNet.

Второй способ представления многозначности используется в тех случаях, когда слово представлено в тезаурусе в одном значении, но если известно, что оно может употребляться и в других значениях в целевых текстах, то ему ставится специальная пометка многозначности (А-многозначность).

3. Отношения в тезаурусе РуТез

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений, должны выполнять разнообразные функции, включая расширение поискового запроса и вывод рубрики; разрешение лексической многозначности; тематического анализа текстов с учетом их лексической связности.

Для реализации любой из этих функций необходимо осуществление своеобразного логического вывода: встретив вхождение некоторого понятия в тексте, нужно делать многошаговые проходы по отношениям.

Для стабильной работы на разных типах текстов в тезаурусе РуТез используется небольшой набор надежных отношений, т.е. отношений, которые не исчезают, не меняются в течение всего срока существования любого или подавляющего большинства экземпляров понятия.

В результате исследований и экспериментов мы пришли к набору отношений ресурса, предназначенного для эффективной автоматической работы в информационно-поисковых приложениях и приложениях автоматической обработки текстов.

В тезаурусе РуТез имеется четыре основных типа отношений. Первый тип отношений – родовидовое отношение *ниже-выше*, представляет собой отношение класс-подкласс, обладает свойствами транзитивности и наследования.

Второй тип отношений – отношение *часть-целое*. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого, и не относиться к чему-либо другому.

В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения *часть-целое*, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого *несимметричной ассоциацией асц*2 - *асц*1, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, но когда одно из которых не существовало бы без существования другого. Например, понятие *КИПЕНИЕ* требует существования понятия *ЖИДКОСТЬ*. В онтологических исследованиях такое отношение называется отношением онтологической зависимости.

Последний тип отношений – *симметричная ассоциация* – связывает, например, понятия, очень близкие по смыслу, но которые разработчики не решились соединить в одно понятие. Таким примером до недавнего времени были понятия *ПОЛИЦИЯ* и *МИЛИЦИЯ,* отношение между ними было естественным образом симметричным. После переименования российской милиции было решено все-таки оставить только одно понятие *ПОЛИЦИЯ,* а слово *милиция* сделать текстовым входом к этому понятию.

Отношения *выше-ниже, часть-целое* и *несимметричная ассоциация* являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии – так называемое нижняя

полуокрестность понятия (O^-), а также может быть определена совокупность понятий, которые являются для него вышестоящими по иерархии – так называемое верхняя полуокрестность понятия (O^+).

Наиболее часто используются следующие виды путей между понятиями:

- *Путь по иерархии вверх* P_{up} *(*C_0*,* C_{00}*)*: От понятия C_0 к понятию C_{00} существует путь по иерархии вверх, если $C_{00} \in O^+(C_0)$;

- Путь по иерархии вниз $P_{down}(C_0, C_{00})$: От понятия C_0 к понятию C_{00} существует путь по иерархии вниз, если $C_{00} \in O^-(C_0)$;

Введенные типы концептуальных путей используются в процедурах автоматического разрешения лексической неоднозначности, расширения поискового запроса, тематического анализа текстов, вывода рубрик по тексту.

4. Тематический анализ текста на основе тезауруса РуТез

Существуют две основные проблемы применения лингвистических онтологий в автоматической обработке текстов. Во-первых, ни один ресурс не может быть исчерпывающе полон, поэтому для эффективной обработки текстов его применение должно сопровождаться извлечением недостающей информации из текстов.

Во-вторых, обычно требуется применение технологий, которые могут разрешать лексическую неоднозначность слов и выражений, значения которых представлены в лингвистической онтологии. Однако качество разрешения лексической неоднозначности может быть достаточно низким для некоторых типов общеупотребительных слов и нестабильным для других типов слов.

В данном разделе мы опишем новую развиваемую нами технологию работы с текстами, которая преодолевает вышеуказанные проблемы и направлена на построение тематического представления текста в виде совокупности так называемых тематических узлов – близких по смыслу слов и выражений, соответствующих отдельному участнику ситуации, описываемой в тексте. Ранее мы строили тематическое представление на основе исключительно информации, описанной в тезаурусе предметной области. Новая технология учитывает в этой процедуре ряд разнородных факторов, только одним из которых является информация из тезауруса.

4.1. Тематическая структура текста и лексические цепочки

Как известно, тематическая структура текста представляет собой иерархическую структуру тем и подтем, причем темы рассматриваются как предикаты над некоторым множеством участников [Алексеев, Лукашевич, 2011].

Темы и подтемы, находящиеся в иерархических отношениях друг с другом, связываются между собой общими участниками или группами участников, имеющих отношение друг к другу. В тексте эти группы участников, а также варианты их наименования выглядят как так называемые лексические цепочки [Hirst, St-Onge, 1998]. Так, в тексте про вывод американской авиабазы из Киргизии такой цепочкой может быть следующая: *Киргизия, Киргизстан, киргизский, кыргызский, Бишкек* и т.п. Поэтому качественное распознавание таких лексических цепочек в тексте выявляет совокупность взаимодействующих между собой участников описываемой ситуации, и, следовательно, эксплицирует структуру текстового содержания.

Особенностью нашего моделирования лексических цепочек является использование для их построения комбинации нескольких факторов и одного разделяющего фактора, а именно: два языковых выражения не могут быть включены в одну лексическую цепочку, если их частота употребления в одних и тех же предложениях текста более чем в два раза превышает их встречаемость в соседних предложениях (для обоснования см. [Алексеев, Лукашевич, 2011]). Действительно, если мы посредством лексических цепочек пытаемся выявить основных участников описываемой в тексте ситуации, то частая встречаемость отдельных слов (не входящих в устойчивое словосочетание) в предложении свидетельствует в пользу их отнесенности к различным, взаимодействующим между собой участникам.

В качестве единиц, из которых строятся лексические цепочки в нашем случае, выступают: отдельные слова знаменательных частей речи, словосочетания, представленные в тезаурусе РуТез, а также новые словосочетания, извлеченные из текста. Новое словосочетание, полученное из текста, должно иметь частоту в этом тексте не менее 2, и обладать некоторой устойчивостью употребления в данном тексте. Мы проверяем устойчивость словосочетания путем сопоставления совместной встречаемости слов в тексте друг за другом и их отдельной встречаемости в заданных фрагментах предложения.

Отметим, что в предлагаемом методе построения лексических цепочек мы не делаем предварительного разрешения многозначности неоднозначных слов, предполагая, что разные учитываемые нами факторы должны усиливать отношения между теми значениями слов, которые более подходят текущему текстовому контексту.

4.2. Факторы, используемые для построения лексических цепочек

Для построения лексических цепочек используется совокупность факторов, каждый из

которых дает вес близости между языковыми выражениями (словами или словосочетаниями). Эти факторы могут являться контекстно-независимыми и контекстно-зависимыми.

Первый **фактор TS (T**hesaurus **S**imilarity**)** – это контекстно-независимый фактор, вычисляемый на основе тезаурусных отношений. Максимальный вес связи по тезаурусу (равный 1.0) придается паре выражений, который являются однозначными и указаны в тезаурусе как синонимы. Вес связи по тезаурусу снижается обратно пропорционально длине пути между понятиями, текстовыми входами к которым являются данные выражения. Многозначность слов учитывается следующим образом: если многозначное слово относится к некоторому понятию С, и в тексте не было ни одного однозначного текстового входа этого понятия, то вес тезаурусного отношения умножается на некоторый снижающий коэффициент (сейчас установлен = 0.9). Если оба выражения в оцениваемой паре являются многозначными, то снижающий коэффициент применяется второй раз.

Второй контекстно-независимый **фактор BS** (**B**eginning **S**imilarity) оценивает формальное сходство выражений по написанию. Максимальный вес сходства составляет 1 для слов с одинаковым началом (5 букв: *киргиз – Киргизия*). Для многословных выражений, включающих схожие слова, за каждое различающееся слово в паре выражений берется штраф 0.1.

Используются два фактора сходства, вычисляемые на основе пословных контекстов внутри предложений. **Фактор SPS** (**S**calar **P**roduct **S**imilarity) оценивает сходство между выражениями на основе скалярного произведения пословных векторов их контекстов в предложении (максимум – 0.5). **Фактор SC** (**S**trong **C**ontext) оценивает количество одинаковых контекстов величиной 4 слова (2 слова влево и вправо), нормированное на максимум величины SC, достигнутому в данном тексте.

Важнейшим контекстным фактором является **фактор NSF** (**N**eighboring **S**entence **F**eature)**,** который представляет собой оценку превышения совместной встречаемости в соседних предложениях по сравнению с встречаемостью внутри одних и тех же предложений (см. [Алексеев, Лукашевич, 2011]).

Вычисляется этот фактор следующим образом:

$$NSF=Min[1,\ \frac{Fext - F\,\mathrm{int}}{Max\,(Fext - F\,\mathrm{int})}\ ],$$

где *Fext* – это совместная встречаемость выражений в соседних предложениях текста, *Fint* – внутри одних и тех же предложений текста.

Этот же фактор служит ограничителем соединения выражений в одну лексическую цепочку. Если величина NSF<0, то выражения не будут объединены в одну лексическую цепочку при любой суммарной величине сходства между выражениями.

Для сравнения сходства более длинных лексических цепочек служит **фактор EO (E**mbedded **O**bjects**)**, который носит булевский характер и указывает, что у цепочек есть общий элемент. Этот фактор позволяет быстрее объединять в единую цепочку фрагменты похожих цепочек, начавшие формироваться отдельно.

Веса всех факторов для каждой пары выражений суммируются, и такие пары могут быть упорядочены по мере снижения веса сходства. Примеры наиболее сходных пар выражений для текста об американской авиабазе в Киргизии показаны в Табл. 1.

Таблица 1. Наиболее сходные по нескольким факторам пары языковых выражений в тексте про американскую авиабазу в Киргизии

Выражение1	Выражение2	Вес сходства
Авиабаза	*Авиабаза Манас*	2.78
Киргизия	*Киргизстан*	2.75
Арендная плата	*Плата за аренду*	2.49
Аль-Каида	*Аль-Каеда*	2.47
База	*Военная база*	2.45
Кыргызская республика	*Кыргызстан*	2.45

4.3. Формирование лексических цепочек

После упорядочения пар выражений начинается итеративный процесс соединения выражений в лексические цепочки. Мы формируем лексические цепочки в форме тематических узлов, т.е. в цепочке выделяется главный элемент (центр), который принадлежит только этой цепочке (тематическому узлу). Обычно главный элемент тематического узла является наиболее частотным среди других элементов этого узла. Остальные элементы тематического узла могут входить в еще один тематический узел.

Процесс начинается с рассмотрения пары выражений с максимальным весом. Эта пара соединяется в один узел так, что более частотное выражение становится центром этого узла, а второе элементом. Центр тематического узла становится представителем элементов узла со всеми их вхождениями и контекстами.

Склеивание в единый тематический узел тематических узлов с несколькими элементами происходит похожим образом. Центр наиболее частотного тематического узла становится центром нового объединенного тематического узла.

Итеративный процесс останавливается при достижении между очередной парой выражений порогового веса сходства, равного 1.1. Эта

пороговая величина имеет прозрачный смысл и отражает высокую степень сходства по тезаурусу (однозначные синонимы) плюс небольшое подтверждение от других факторов. Данный вес сходства может быть достигнут и на основе других факторов, при отсутствии в тезаурусе какой-либо связи между двумя выражениями.

Примерами автоматически построенных лексических цепочек могут служить следующие цепочки:

(Киргизия): Киргизстан, Кыргызстан, Киргизский, столица Киргизии, Киргизская республика, Киргизская столица, киргиз, парламент Киргизии..

(Авиабаза): авиабаза Манас, база, база ВВС, авиационная база, база ВВС, база Манас, закрытие базы, военная база США.

В настоящее время мы тестируем возможности использования такого рода лексических цепочек для автоматического аннотирования отдельных текстов и новостных кластеров.

Заключение

В данной статье мы рассмотрели современное состояние тезауруса русского языка РуТез, сравнили его структуру с другими концептуальными компьютерными ресурсами. Мы также рассмотрели типы используемых в различных приложениях путей и окрестностей в тезаурусе.

В настоящее время мы развиваем технологии, которые позволяют применять тезаурус РуТез без применения процедуры разрешения многозначности, совместно с техниками извлечения недостающих знаний из самих текстов.

Библиографический список

[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. М., С. 43-50

[Алексеев, Лукашевич, 2011] Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. 2011. N 4. С. 95-103

[Гельфенбейн и др., 2003] Гельфенбейн И.Г., А.В. Гончарук, В.П. Лехельт, А.А. Липатов, В.В. Шило. Автоматический перевод семантической сети WORDNET на русский язык // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003. М., 2003.

[Лукашевич, 2011] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского университета, 2011.

[Сухоногов, Яблонский, 2005] Сухоногов А.М., Яблонский С.А. Автоматизация построения англо-русского WordNet. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2005 / Под ред. А.С.Нариньяни. – М.: Наука, 2005.

[Hirst, St-Onge, 1998] Hirst G., St-Onge D. Lexical Chains as representation of context for the detection and correction malapropisms // WordNet: An electronic lexical database and some of its applications /C. Fellbaum, editor. Cambrige, MA: The MIT Press, 1998.

[Miller, 1998] Miller G. 1998. Nouns in WordNet // WordNet – An Electronic Lexical Database / Fellbaum, C (ed). – The MIT Press, pp. 23-47.

[Vossen, 1998] Vossen P. EuroWordNet: A multilingual Database with Lexical Semantic Network. – Dodrecht. 1998.

LINGUISTIC ONTOLOGY – RUTHES THESAURUS

Alekseev A.A. , Dobrov B.V. , Loukachevitch N.V.

НИВЦ МГУ им. М.В. Ломоносова,
г. Москва, Российская Федерация

a.a.alekseevv@gmail.com
dobrov_bv@mail.ru
louk_nat@mail.ru

Introduction

One of well-known lexical resources in the sphere of computational linguistics and natural language processing is WordNet thesaurus.

In this paper we describe the structure and the current state of thesaurus of Russian language RuThes, which belongs to the same class of conceptual linguistic resources (linguistic ontologies) as WordNet. At the same time RuThes has specific features of knowledge representation motivated by its initial purpose to be a resource for natural language processing in information retrieval applications. In the near future we plan to begin the preparation of RuThes thesaurus to publication and therefore this resource can become a useful instrument of Russian text analysis similar to Wordnet-like resources.

Main Part

In principles of its development RuThes thesaurus comprises three existing traditions of computer resources such as traditional information-retrieval thesauri, WordNet-like thesauri and formal ontologies.

A unit of RuThes thesaurus is a concept reflecting significant classes of entities in the world and the internal life of people (not synonym sets as in WordNet). The small set of conceptual relations in the thesaurus describes the most reliable relations of concepts. Formal properties of the relations provide a basis for formal logical inference in natural language processing.

As an example of application of RuThes thesaurus to natural text analysis we consider the thematic structure analysis of texts including lexical chain construction based on several factors.

Conclusion

In this paper we present the structure and the current state of Thesaurus of Russian Language RuThes and describe the example of its application to the thematic analysis of texts.