Автоматизированная обработка научно-технических текстов с помощью Онтологии по естественным наукам и технологиям

Б.В. Добров, Н.В. Лукашевич

Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова
dobroff@mail.cir.ru, louk@mail.cir.ru

Аннотация

В статье представлен метод автоматической обработки научных документов на основе онтологии по естественным наукам и технологиям ОЕНТ. При разработке онтологии учитывались такие языковые явления, как лексическая многозначность и синонимия. Обработка текстов базируется на свойствах связного текста – лексической и тематической связности. В результате автоматической обработки для научных публикаций строится концептуальный индекс, производится автоматическая рубрикация и автоматическое аннотирование.

1. Введение

В настоящее время большое внимание уделяется фиксации научного знания в формализованном виде посредством создания онтологических ресурсов, использования формальных средств дескриптивной логики. Особенно большое внимание созданию таких онтологий уделяется в медицинской области, в которой создаются сверхбольшие онтологии SNOWMED-CT [12], GALEN [9], FMA [10].

В то же время отмечается, что при использовании таких сверхбольших ресурсов имеются сложности с работой современных программ логического вывода (reasoners). Отмечается, что стандартные средства вывода Racer, Pellet, FaCT++ не позволяют выполнять многие сценарии логического вывода на реальных биомедицинских онтологиях из-за большого их большого объема и сложности [6, 7]. В качестве актуальной обсуждается задача выделения относительно небольших полезных для приложений сегментов таких онтологий [11].

Если рассматривать задачи использования научных онтологий в приложениях автоматического анализа текста таких, как автоматическая рубрикация, аннотирование, концептуальное индексирование, то в таком случае онтология должна быть очень большой величины [4]. Кроме того, сопоставление с текстом требует специальных методов работы с такими языковыми явлениями, как синонимия и многозначность, которые распространены и в научных текстах.

В статье будут рассмотрены этапы обработки научных текстов на основе Онтологии по естественным наукам и технологиям, специально создаваемой ОЕНТ [2, 3] как ресурс для автоматической обработки неструктурированной научно-технической информации.

Прикладная цель создания ОЕНТ – создать ресурс для автоматизации наукометрического анализа (определение фронта исследований, ключевых работ в области), проведения патентных исследований, а в сочетании с ранее созданным [4] информационно-поисковым тезаурусом для концептуальной обработки общественно-политических текстов – ресурс для поддержки автоматизации экспертизы инновационных проектов.

2. Онтология ОЕНТ

Онтология ОЕНТ создается для широкой области естественных наук, включая химию, физику, геологию, географию, математику, элементы биологии

По структуре онтология представляет собой иерархическую сеть понятий. Каждое понятие имеет имя. Для сопоставления с текстом каждое понятие снабжается набором текстовых выражений («текстовых входов», «терминов»), значения которых соответствуют данному понятию. В качестве таких текстовых входов могут выступать однословные существительные, прилагательные, глаголы, именные и глагольные группы. Количество таких текстовых входов понятий может быть достаточно велико - при вводе нового понятия делаются специальные усилия, чтобы максимально подробно перечислить его возможные текстовые входы.

Широта предметной области онтологии ОЕНТ связана с тем, что имеется достаточно большое число многозначных (для различных предметных областей) терминов, что в большинстве случаев указывается отнесением одного и того же текстового входа к разным понятиям онтологии. Например, текстовый вход *дерево* сопоставлен трем разным

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

понятиям: *ДЕРЕВО (РАСТЕНИЕ) – ДЕРЕВО (МАТЕРИАЛ) – ДЕРЕВО (ГРАФ)*, текстовый вход *белок* соответствует понятиям: *БЕЛОК (ПОЛИМЕР) – БЕЛОК ЯЙЦА – БЕЛОК ГЛАЗА*, текстовый вход *атмосфера* соответствует понятиям *АТМОСФЕРА НЕБЕСНОГО ТЕЛА* и *АТМОСФЕРА (ЕДИНИЦА ДАВЛЕНИЯ)*.

Кроме того, текстовые входы, которые имеют не только терминологическое значение, но и общеупотребительное могут быть отмечены специальной пометкой многозначности, например, это такие слова, как *отражать, отражение, последовательность, аргумент*, одно из терминологических значений которых входит в состав текстовых входов онтологии ОЕНТ, а второе значение является общеупотребительным.

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений, должны выполнять разнообразные функции.

Во-первых, эти отношения должны использоваться в классических функциях информационно-поисковых тезаурусов для расширения поискового запроса или вывода рубрики документа. Во-вторых, отношения важны для разрешения многозначности языковых единиц, включенных в ресурс, поскольку естественным методом реализации автоматической процедуры разрешения многозначности является сопоставление контекста употребления многозначной единицы в тексте и контекста соответствующего понятия в онтологическом ресурсе. В-третьих, отношения в онтологическом ресурсе могут использоваться для выявления лексической связности в текстах с целью применения выявленной структуры текста для улучшения качества обработки текстов.

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

В онтологии ОЕНТ 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 (подробнее см. [2, 3]).

В настоящее время онтология ОЕНТ включает 55 тысяч понятий, 140 тысяч терминов из таких областей, как математика, физика, химия, геология, экология, биология. Между понятиями установлено более 200 тысяч прямых отношений, что в соответствии с алгеброй сочетания отношений позволяет установить суммарно более двух миллионов отношений. Описанные термины в значительной мере покрывают терминологию этих областей, вводимую в средней школе и на начальных курсах ВУЗов.

3. Обработка научных текстов на основе онтологии ОЕНТ

Этапы обработки текстов на основе онтологии ОЕНТ (далее Онтология) включают в себя следующие этапы:

- графематический и морфологический анализ,
- автоматическое сопоставление с Онтологией,
- автоматическое разрешение многозначности слов,
- формирование концептуального индекса документа,
- автоматическое построение тематического представления текста,
- автоматическая рубрикация по заданному рубрикатору,
- автоматическое аннотирование.

Результаты обработки отдельного текста используются для построения компонентов различных информационных систем – информационного поиска, агрегирования новостных потоков (кластеризация), составления аналитических отчетов (различные типы обзорных рефератов) и т.д.

3.1 Построение концептуального индекса текста

На первом этапе обработки текстов на основе Онтологии производится сравнение единиц текста с единицами Онтологии. Сравнение текста и текстовых входов Онтологии происходит на основе морфологического представления единиц текста и единиц Онтологии. Последовательности лемм, сопоставленные текстовым входам Онтологии, сопоставляются с последовательностями лемм документа.

При сопоставлении Онтологии и текста специальное внимание уделяется обработке многозначных слов и выражений для выбора одного из значений слова. Основой для применяемого алгоритма разрешения многозначности служит оценка семантической близости между возможными значениями, с одной стороны, и окружающим текстовым контекстом, с другой стороны. При этом рассматривается как локальный контекст, который задается в виде некоторого окна – линейной окрестности многозначного вхождения слова, так и глобальный контекст, в который входят все слова текста [5].

В результате сопоставления текста с Онтологией и разрешения многозначности текстовых единиц строится концептуальный индекс текста – набор понятий онтологии, упомянутых в тексте. Для выявленных в тексте понятий из онтологии выгружаются отношения, как установленные между понятиями непосредственно, так и выводимые по свойствам отношений. В результате создается онтологическая проекция текста.

Проиллюстрируем степень текущего покрытия текстов научных новостей единицами Онтологии, для этого рассмотрим следующий фрагмент текста (в тексте жирно выделены текстовые входы понятий Онтологии, многозначные текстовые входы дополнительно подчеркнуты):

Разработан одношаговый процесс ***производства наноцепей*** *из* ***графена*** *Согласно* ***статье****, опубликованной совместной* ***американо-французской группой ученых*** *в* ***журнале*** **Science, электрические схемы наномасштаба** *можно* ***создавать*** *«в один проход» при помощи модифицированной* ***техники атомной силовой микроскопии. Методика*** *подразумевает «****печать»*** *проводящих* ***нанопроводов*** *на* ***поверхности оксида графена. Графен*** *- самый тонкий из известных на сегодняшний день материалов. Он представляет собой* ***лист атомов углерода****, образующих* ***двумерную гексагональную кристаллическую решетку****. Мы уже ни раз писали о том, что за счет своих уникальных* ***физических свойств****, данный материал потенциально мог бы* ***заменить кремний*** *в разнообразных* ***«электронных»*** *приложениях. (http://sci-lib.com/article745.html).*

В данном небольшом фрагменте обозначено шесть многозначных слов. В частности, слово *техника* может относиться к понятиям *ТЕХНИЧЕСКОЕ УСТРОЙСТВО* или *МЕТОД;* а слово *лист* – к трем понятиям *ЛИСТ (ПЛАСТ), ЛИСТ РАСТЕНИЯ, ЛИСТ БУМАГИ.* Также здесь проявляется и синонимия: слово *наномасштаб* является текстовым входом к понятию *НАНОРАЗМЕР, а* словосочетание *группа ученых* – к понятию *ИССЛЕДОВАТЕЛЬСКИЙ КОЛЛЕКТИВ.*

3.2 Построение тематического представления текста

Любой связный текст обладает свойством тематической связности, т.е. связный текст должен иметь единую тему, и каждое предложение связного текста должно иметь то или иное отношение к этой теме. Кроме того, связный текст содержит большое количество близких по смыслу слов, терминов (лексическая связность текста), что является проявлением тематической связности. Распознавание цепочек близких по смыслу терминов – тематических узлов – позволяет лучше определить основную тему текста [4].

Для построения тематических узлов конкретного текста сначала выделяются потенциальные центры тематических узлов. Сначала тематические узлы собираются вокруг понятий заголовка и первого предложения текста. Затем тематические узлы собираются для остальных понятий, начиная с самых частотных. В очередной узел включаются понятия Онтологии, которые упомянуты в тексте и связаны с центральным понятием тематического узла (непосредственно или выводятся на основе свойств отношений). Те понятия, которые уже попали в тематический узел некоторого понятия, свой тематический узел не образуют. В тематический узел понятия попадают такие упомянутые в тексте понятия, которые связаны с центральным либо прямыми отношениями по OEHT, либо наследуемыми отношениями.

После построения очередного тематического узла выбирается следующее по частотности (заголовку) понятие Онтологии, еще не включенное в тематические узлы, и образует следующий тематический узел.

Чтобы оценить связанность между понятиями в тексте, вводится понятие “текстовая связь’: данное понятие считается связанным по тексту с теми понятиями, которые находятся на расстоянии не более *n* понятий от очередного вхождения данного понятия безотносительно к порядку следования понятий в тексте.

В соответствии с моделью [3] предполагается, что основными тематическими узлами в первую очередь являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями;
- сумма частот текстовых связей между ними максимальна.

Для текста примера были выделены следующие основные тематически узлы:

- *ГРАФЕН, ОКСИД ГРАФЕНА;*
- *НАГРЕВАНИЕ;*
- *НАНОЦЕПЬ, НАНОПРОВОД;*
- *ИССЛЕДОВАТЕЛЬСКИЙ КОЛЛЕКТИВ, НАУЧНЫЙ РАБОТНИК;*
- *НАНОРАЗМЕР, НАНОМЕТР.*

Локальные тематические узлы представляют собой некоторые важные характеристики основных тематических узлов. Тематический узел считается локальным, если этот узел имеет текстовую связь с частотностью большей единицы с одним из основных тематических узлов. Понятия, не вошедшие в

состав основных и локальных тематических узлов, объявляются "упоминавшимися" в тексте.

Таким разбиением тематических узлов на основные и локальные задается разбиение понятий, упомянутых в тексте, на следующие пять классов по их важности для анализируемого текста:

- главные понятия основных тематических узлов (основные темы);
- другие понятия основных тематических узлов;
- главные понятия локальных тематических узлов (локальные темы);
- другие понятия локальных тематических узлов;
- упоминавшиеся понятия.

Для использования такой структуры в дальнейшей обработке каждый уровень тематического представления ставится в соответствие некоторые вес. Результирующий вес понятия является линейной функцией от веса его позиции в тематическом представлении и относительной частотности понятия в документе [4].

4. Автоматическая рубрикация текстов на основе тематического представления

Процедура рубрикации базируется на автоматически построенном тематическом представлении документов, которое моделирует основную тему и подтемы документа наборами (тематическими узлами) близких по смыслу понятий, упомянутых в документе. Такая основа рубрикации дает возможность обрабатывать тексты разных типов и размеров: нормативные акты, газетные статьи, новостные сообщения, научные публикации в области гуманитарных наук, социологические опросы [1].

При создании лингвистического профиля рубрикатора каждая рубрика R описывается дизъюнкцией альтернатив, каждый дизъюнкт представляет собой конъюнкцию:

$$R = \bigcup_i D_i \; ; \qquad D_i = \bigcap_j K_{ij} \; ,$$

Конъюнкты в свою очередь описываются экспертами с помощью так называемых «опорных» понятий ОЕНТ. Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее, каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия: без расширения (обозначается символом «N»), полное расширение по дереву иерархии ОЕНТ (символ «E»), расширение только по родовидовым связям (символ «L»), расширение по всем видам отношений на один уровень иерархии (символ «W»), расширение на один уровень иерархии, не включая отношения *ниже* (символ «V»).

Опорное понятие может быть как «положительным», т. е. добавлять нижерасположенные понятия в описание конъюнкта, так и «отрицательным», т. е. вырезать из описания рубрики свои подчиненные понятия. Результатом применения расширения опорных понятий является совокупность понятий Онтологии, полностью описывающая конъюнкт:

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn}) = \bigcup_k d_{ijk}$$

Вес конъюнкта зависит от максимального веса входящего в него понятия Онтологии. Вес дизъюнкта предназначен учитывать не только сумму весов составляющих его конъюнктов, но и меру близости конъюнктов в тексте.

Вес рубрики представляет собой максимум весов входящих в описание рубрики альтернатив. В случае имеющихся иерархических связей между рубриками оценка релевантности нижестоящих рубрик переносится на вышестоящие, так что при запросе по вышестоящей рубрике будут выходить и документы, к которым были приписаны нижестоящие рубрики.

Алгоритм рубрицирования работает следующим образом. Для всех понятий ОЕНТ, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

В настоящее время научные тексты рубрицируются по рубрикатору, представляющему собой модификацию рубрикатора РФФИ, в который добавлен ряд рубрик по современным технологиям производства.

Так для текста примера автоматически получены следующие рубрики:

- *Нано- и мембранные технологии*,
- *Нано и микроструктуры*,
- *Наноматериалы*.

5. Автоматическое аннотирование документа

Тематическое представление документа помогает формированию краткого реферата (аннотации) документа. (21.1)

Знания человека о тематической связности между терминами вытекают из знаний о предметной области, в рамках которой написан текст. Таким образом, то новое и важное, что несет в себе текст и что должна отразить в себе аннотация, это именно то, каким образом взаимодействуют между собой разные основные темы текста.

Отсюда следует первый принцип составления аннотаций: важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те предложения текста, которые содержат. по крайней мере, два термина, входящих в состав разных основных тем текста. При этом для каждой пары выявленных основных тем текста в аннотацию выбираются те предложения, в которых эта пара обсуждалась первый раз, следуя по порядку текста. Повторы основных тем в предложениях повышают связность аннотации. Аннотация для текста примера получилась такой (алгоритм выбрал заголовок текста и предложения 1 и 15):

Разработан одношаговый процесс производства ***наноцепей*** *из* ***графена****. Согласно статье, опубликованной совместной американо-французской* ***группой ученых*** *в журнале Science, электрические схемы* ***наномасштаба*** *можно создавать <в один проход> при помощи модифицированной техники атомной силовой микроскопии. Для локального* ***нагревания*** *поверхности* ***оксида графена****, коллеги предложили использовать нагретое острие атомно-силового микроскопа.*

5. Заключение

В статье представлена технология автоматической обработки текстов на основе онтологии по естественным наукам и технологиям ОЕНТ. При разработке онтологии учитывались такие языковые явления как многозначность слов и синонимия. Обработка текстов базируется на свойствах связного текста – лексической и тематической связности. В результате автоматической обработки, для научных публикаций строится концептуальный индекс, производится автоматическая рубрикация и автоматическое аннотирование.

Литература

[1] Агеев, М.С. Автоматическая рубрикация текстов: методы и проблемы / Агеев М.С., Добров Б.В., Лукашевич Н.В. // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2008. Том 150. Кн. 4. С. 25 – 40.

[2] Добров, Б.В. Онтология по естественным наукам и технологиям ОЕНТ: структура, состав и современное состояние / Добров Б.В., Лукашевич Н.В. // Электронные библиотеки (электронная версия: [Электронный ресурс]. — Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2008/part1/DL).

[3] Добров, Б.В. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам / Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Седьмой Всероссийской научной конференции (RCDL’2005) (Ярославль 4-6 октября 2005 г.). Ярославль: ЯрГУ им. П.Г. Демидова, 2005. С.70 – 79.

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

[5] Лукашевич, Н.В. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний / Лукашевич Н.В., Чуйко Д.С. // Интернет-математика 2007: Сборник работ участников конкурса. Екатеринбург: Изд-во Урал. ун-та, 2007. С.108 – 117.

[6] Подколодный, Н.М. Онтологическое моделирование в биоинформатике и системной онтологии // Труды Второго симпозиума «Онтологическое моделирование». Казань, 2011. С. 233 – 269.

[7] Keet, C.M. A survey of requirements for eutomated reasoning services for bio-ontologies in OWL / Keet C.M., Roos M., Marshall M.S. // Workshop on OWL: Experiences and Directions. Insbruck, 2007.

[8] Noy, N. Specifying ontology views by traversal / Noy N., Musen M. // Proceedings of International Semantic Web Conference, LNCS-3298. 2004. C. 713 – 725.

[9] Rector, A. Ontological Issues in Using a Description Logic to Represent Medical Concepts: Expierence from GALEN / Rector A., Rogers J. // Proceedings of IMIA WG6 Workshop. 1999.

[10] Rosse, C. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy / Rosse C., Mejino J. // Journal of Biomedical Informatics. 36 (6). 2003. C.478 – 500.

[11] Seidenberg, J. Web Ontology Segmentation: Analysis, Classification, Use / Seidenberg J., Rector A. // Proceedings of Conference WWW-2006. 2006. C. 13 – 22.

[12] Stearns, M.Q.SNOWMED clinical terms: overview of the development process and project status / Stearns M.Q., Price C., Spackman K., Wang A. // Proceedings of AMIA Symposium. 2001. C. 662 – 666.

Automated Text Processing of Scientific Documents Based on Ontology on Natural Sciences and Technology

B. Dobrov, N. Loukachevitch

The paper is devoted to description of document processing stages based on linguistic and ontological knowledge. Such language phenomena as ambiguity, synonymy, thematic and lexical cohesion of texts are taken in account.