Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей

The approach to ontology learning from machine-readable dictionaries

Рубашкин В. Ш. (vrubashkin@yandex.ru),
Бочаров В. В. (victor.bocharov@gmail.com),
Пивоварова Л. М. (lidia.pivovarova@gmail.com),
Чуприн Б. Ю. (boris@vr4591.spb.edu)

Санкт-Петербургский государственный университет

В докладе представлено текущее состояние разрабатываемой авторами технологии пополнения онтологии на основе лингвистического и семантического анализа текстов определений слов (терминов) в энциклопедических и толковых русскоязычных словарях.

1. Положение дел и постановка задачи

Проблема массового пополнения онтологий — с расчетом на достаточно полное покрытие текстов хотя бы в рамках ограниченной предметной области — характеризуется обычно как проблема «бутылочного горлышка» (*bottle neck*). Эта характеристика отражает факт высокой трудоемкости «ручного» (лучше сказать — экспертного) ввода концептов в онтологию и отсутствие широко признанных технологий, обеспечивающих хотя бы частичную автоматизацию процедур ввода. Следует подчеркнуть, что в этом контексте речь идет не о *формировании*, а именно о *массовом пополнении* онтологий. Первоначальное формирование как специализированной, так и в особенности универсальной онтологии предполагает построение системного ядра, состоящего из "высокоуровневых понятий» (*Top-Level Ontology*). Построение такого ядра в любом случае есть дело высококвалифицированного специалиста. В докладе представлен опыт разработки и использования человеко-машинной технологии ввода, опирающейся на использование традиционных словарных ресурсов и ориентированной в основном на пополнение уже сформированного ядра онтологии концептами, именующими объекты («предметные имена»).

В рамках общего дискурса *Ontology Learning* обсуждаются два пути автоматизации пополнения онтологий[1]:

1) на основе анализа дистрибутивных характеристик лексики в корпусе текстов;
2) с использованием традиционной лексикографической информации — энциклопедических и толковых словарей (в англоязычной литературе — *машиночитаемые словари — MRD*).

Изучение результатов, полученных на том и другом направлении, а также наш собст томатического построения таксономий и извлечения других семантически значимых отношений была доведена до технологической реализации, была представлена еще в 1985 г. ([2]). В 1993 г. N. Ide и J. Veronis уже подводят, так сказать, промежуточные итоги ([3]), скептически оценивая как полученные результаты, так и вообще перспективы полностью автоматической «генерации больших баз знаний» из MRD. Их основная аргументация связана с семантическим качеством самих словарных определений. Однако это не остановило дальнейших попыток, и в последнее десятилетие работы в этом направлении развивались широким фронтом — главным образом, на материале англоязычных словарей (ср., напр., [4], [5], [6]).

Другой наш предварительный вывод состоит в том, что при современном уровне развития технологий автоматического понимания текста не следует рассчитывать на то, что распознавание востребованных онтологией словарных характеристик может быть обеспечено исключительно алгоритмическими средствами; оперативное участие администратора онтологии в практически полезной техно-

[1] Подробный обзор сложившихся подходов можно найти в монографии [1].

логии такого рода необходимо.[2] Наш общий подход при этом состоит в том, что средства автоматизации должны обеспечить предварительную структуризацию текстов определений, поддерживать отбор релевантной задаче лексики предметной / проблемной области в машиночитаемом словаре и осуществлять предварительную онтологическую квалификацию содержащегося в словаре термина (не всегда точную и не всегда правильную); роль специалиста, участвующего в этой работе, сводится к тому, что он оценивает предлагаемые системой решения и либо просто подтверждает их, либо вносит необходимые коррективы, т. е., предлагается ориентация на человеко-машинную технологию и на создание обучаемой диалоговой среды пополнения онтологии. Возможности графического редактирования, ставшие стандартом де-факто для современных онторедакторов, существенно облегчают и упрощают эту часть работы.

Основная идея, положенная в основу рассматриваемой технологии, состоит в следующем ([2], [8]). Определения и толкования в энциклопедических и толковых словарях представляют собой частично структурированные тексты. Самое существенное в этом плане то, что определение / толкование в подавляющем большинстве случаев соответствует стандартной логической схеме — «*род — видовые отличия*». Номинальное определение термина в энциклопедическом словаре почти всегда локализовано в первом предложении словарной статьи; в толковых словарях оно вообще, как правило, состоит из одной короткой фразы. При этом логически самая важная часть определения, указывающая родовое понятие, с помощью которого описывается определяемый термин, почти всегда грамматически акцентирована — представлена именной группой с главным словом — существительным в именительном падеже. Несколько примеров. Толковый словарь Ожегова: *МУШКЕТЕР — солдат, вооруженный мушкетом; МУШКЕТ — старинное ружье крупного калибра с фитильным замком.* Энциклопедический словарь [9]: *МАГНЕТРОН — электровакуумный прибор, мощный генератор электромагнитных волн сантиметрового диапазона. Принцип действия магнетрона основан на торможении электронов в скрещенных электрических и магнитных полях. Используется главным образом в устройствах радиолокации, а также в нагревательных установках сверхвысокой частоты.*

Следовательно, первоочередным объектом анализа должна быть указанная именная группа и ее главное слово («опорное слово») прежде всего. Исключения обозримы и обрабатываются спецалгоритмами, позволяющими в большинстве случаев найти «правильное» опорное слово.[3]

Помимо этого в тексте определения, как правило, присутствуют типовые элементы описания, указывающие функциональность, принцип действия, сферу использования, структурно-морфологические характеристики и свойства определяемого объекта/явления, а также, возможно, указаны отношения определяемого с другими понятиями, в частности, с понятиями, уже представленными в онтологии и/или словаре. Конечная цель состоит в том, чтобы они были распознаны и отражены как онтологические характеристики соответствующих концептов.

Понятно, что в такой постановке рассматриваемая задача, сводится к двум подлежащим алгоритмизации процедурам. Первая — структурирование текста определений (*подсистема анализа словарных определений*), вторая — сопоставление таковых с лексическим представлением концептов в онтологии (далее будем именовать ее *подсистема энциклопедического импорта*). При этом вторая процедура может быть организована одним из двух альтернативных способов. Либо для каждого определяемого термина словаря проводится поиск возможных релевантных родовых концептов — для последующего включения термина в таксономию, либо, наоборот, можно указать иерархически организованный фрагмент онтологии, требующий пополнения, и выбирать в множестве структурированных определений те, которые соответствуют указанному фрагменту, уточняя далее тем или иным способом место выбранной группы термина в этом фрагменте онтологии и, если нужно, дополнительные словарные характеристики отдельных терминов. Ясно, что первый подход эффективнее в ситуации, когда соответствующий раздел онтологии уже в первом приближении сформирован, и речь идет о его дополнениях и уточнениях. Второй подход эффективнее может применяться в той ситуации, когда нужно осуществить начальное наполнение определенного раздела.

Имеется в виду, что названные две процедуры работают автономно, интерфейс между ними определяется передаваемой структурой данных — в форме таблиц РБД. Ясно также, что оптимальным

---

[2] Ср. мнение авторитетных японских специалистов T. Morita и T. Yamaguchi, формулируемое в обоснование принятого ими подхода в проекте DODDLE [7]: “*Regarding domain ontology development support, many tools have been done with knowledge engineering, natural language processing and data mining techniques to make possible automatic domain ontology construction from existing information resources… However, as the techniques are not yet mature to achieve the task and domain ontology structure depends on the aspects from human experts (users), full automatic process does not go well with the task. Instead of developing full automatic environment, it is more important to provide refined semi-automatic environment with integrated facilities to construct practical domain ontologies*”.

[3] Обзор типовых случаев невыполнения общего правила дан в [8]. Разумеется, предлагаемым способом не могут быть обработаны «недоброкачественные» определения, — скажем такого рода: *Боразол — B3H6N3* ; им присваивается в позиции опорного слова помета “*unknown*”.

решением с точки зрения организации инструментальных средств здесь является погружение средств импорта в среду онторедактора, с тем, чтобы максимально использовать его функциональность — что и осуществлено в настоящей работе.

2. Онтология и онторедактор.

Базовой структурой любой онтологии является ***таксономия объектов***. В разработанной и используемой нами онтологии (онтология *InTez*) [10] таксономия представлена в форме ***дерева признаков*** [11]. Это структура, которая позволяет наиболее естественным образом отображать в онтологии связи типа *признак — значения признака; применимость признака к классу объектов*; соответственно, вычислять полный набор объемных отношений (*включение, совместимость, несовместимость*) между концептами — классами. Помимо объемных отношений, в онтологии представлены нетаксономические («ассоциативные») отношения — как универсальные (*часть — целое, объект — место, объект — функция*), так и специализированные (*страна — столица, руководитель — организация* и т. д.). Для данной задачи существенно, что онтология является *интерпретируемой*, т. е. поддерживает связь с лексикой естественного языка фиксируя двустороннее соответствие вида

КОНЦЕПТЫ ←→ СЛОВА,

достаточное для фиксации отношений омонимии и синонимии в естественном языке, с одной стороны, и учета многообразия лексической реализации концептов, с другой. Функционально это соответствие представляет собой одновременно толковый словарь и словарь синонимов — разумеется, в пределах наличного концептуального и лексического состава онтологии.

Операционально место концепта в данной онтологии определяется либо указанием классификационного признака, представляющегося основание, по которому выделен именуемый простой класс (*водный транспорт, воздушный транспорт, сухопутный транспорт → по среде перемещения*). либо — в случае многоаспектного определения класса — формальным толкованием, в типовом случае представляющим конъюнкцию простых классов (как в примере с термином *магнетрон: электровакуумный_прибор & генератор_электромагнитных_волн*). В последнем случае эквивалентным построению формального толкования является указание множественного наследования в таксономии. Возможны и другие схемы формальных толкований.

С точки зрения рассматриваемой задачи существенно то, что описание концептов в онтологии не требует столь подробной характеризации определяемого концепта, какая ему дается в энциклопедических словарях. Так, в приведенном выше определении термина *магнетрон* онтологически значимым можно считать лишь первое предложение определения. Второе предложение для формального описания концепта в онтологии избыточно. Поэтому перед средствами лингво-семантического анализа определений не стоит задача полной формализации всего текста. Скорее речь идет о методах целенаправленного поиска и распознавания онтологически значимых элементов.

Оптимальное распределение работы между программными процедурами и администратором онтологии в значительной степени обусловлено **функциональностью онторедактора**. В онторедакторе, поддерживающем работу с онтологией *InTez* [12], имеются, в частности, стандартные средства навигации, графического редактирования таксономии, поиска концептов по вариантам лексического представления, а также средства добавления связей между концептами. Частью онторедактора является также машина ограниченного вывода, являющаяся важным компонентом для подсистемы логического контроля при вводе. Это создает достаточно комфортную среду для быстрого постредактирования добавленных автоматически концептов — если в этом возникает необходимость.

3. Анализ словарных определений

Представляет собой последовательность следующих шагов.[4]

1) лексикографический парсинг словарных статей;
2) выделение опорного слова;
3) уточнение (при необходимости) опорного слова;
4) уточнение формулировки родового понятия присоединением зависимых и других связанных по смыслу с опорным слов.
5) построение формального толкования определяемого термина на основе концептуального содержания, представленного в текущей версии онтологии.

На данном этапе проекта реализованы в достаточно полном объеме п. п. 1, 2 и 3 и в очень ограниченных пределах — п. 5. Функциональность, соответствующая п. 4, частично реализована в подсистеме энциклопедического импорта.

Лексикографический парсинг предназначен для того, чтобы подготовить текст словарной статьи для синтаксического анализа; он включает удаление всех словарных помет, восстановление сокращений, «расклейку» множественных определений.

[4] Более подробно см. в [13].

Можно считать, что пункт 2 при доступных сегодня средствах анализа представляет собой чисто технологическую процедуру и состоит из следующих действий:

- получение грамматических описаний и лемматизация лексики первого предложения словарного определения;
- поиск и извлечение опорного слова (выбирается первое существительное, имеющее признаки: *существительное*, *именительный падеж*);
- частичная синтаксическая разметка первого предложения (анализ контактных связей).

Омонимия грамматических описаний сохраняется; ее частичное разрешение происходит по синтаксическому контексту — по результатам, полученным на этапе синтаксической разметки. Сохраняются также все полученные варианты синтаксической разметки (глобальная синтаксическая омонимия при анализе не рассматривается).

Используемые средства:

- морфологический анализатор АОТ [14];
- синтаксический анализатор АОТ, реализующий алгоритм GLR парсинга, со специально разработанной упрощенной грамматикой, предназначенной для анализа онтологически значимых фрагментов текстов определений;

На выходе процедуры анализа порождаются две таблицы РБД (СУБД *MS SQL Server,* используемая в качестве операционной среды редактора онтологии). Таблица «Термины» содержит термин, опорное слово, определение / толкование (множественные толкования «расклеены» на предыдущем шаге обработки), и дополнительную рабочую информацию. Таблица «*Слова*», связываемая с таблицей «*Термины*», содержит слова первого предложения определения термина, их леммы и части речи, а также элементы синтаксической разметки (*ссылка на синтаксического хозяина, вид связи*). Указанная пара таблиц далее модифицируется процедурами, соответствующими п. 3, после чего передается в подсистему энциклопедического импорта.

Процедура уточнения опорного слова (п. 3) в основном состоит (а) в удалении избыточных слов, с одновременным отысканием «истинного» опорного слова (слова с общим значением '*именование*' и '*принадлежность к классу*'); (б) в логической интерпретации опорных слов с общим значением '*отношение*' (*часть-целое*, *локализация, назначение,…*) и поиском второго объекта, с которым должно быть установлено распознанное отношение.

4. Энциклопедический импорт

Для организации взаимодействия с администратором онтологии в пользовательский интерфейс онторедактора добавлена вкладка *ТерминыСловаря*, отображающая содержимое таблицы *Термины.* Вкладка становится доступна в режиме «Энциклопедический импорт».

Процедура добавления группы терминов состоит из следующих действий.

1. Формирование группы добавляемых терминов: (а) по указанному опорному слову; (б) по лексике куста. Во втором случае администратор должен предварительно указать в онтологии (в дереве признаков) вершину куста, лексика которого должна учитываться при формировании энциклопедической выборки. Все имена концептов (включая синонимы), как непосредственно входящих в куст, так и связанных с концептами куста отношением «*НИЖЕ*» (*конкретизация*), — если они совпали с опорными словами, — включаются в список опорных слов, по которым формируется выборка. Полученная выборка может дополнительно уточняться (ограничиваться) двумя способами — одновременно обоими или любым из них по отдельности:
   1.1. По лексическому составу текста определения. Отбираются только термины, в тексте определения которых найдена заданная комбинация слов (возможно с усечением).
   1.2. По синтаксически зависимым от опорного словам определения. В этом случае строится и предъявляется администратору онтологии частотный словарь всех идентифицированных синтаксической разметкой слов (используются леммы), непосредственно зависимых от выбранного опорного слова (списка опорных слов) Администратору предоставляется возможность указать слова, уточняющие смысл опорного слова, объединив их связкой ***And*** или ***Or***.

   Кроме того, администратор имеет возможность «вручную» исключить из выборки или присоединить к выборке любые термины, представленные в одноименной таблице, устанавливая или снимая в соответствующих записях помету «Выбрано».
2. Автоматический ввод терминов, место которых в онтологии может быть определено самим алгоритмом. По каждому из идентифицированных системой терминов выборки, сформированной в соответствии с п. 1, администратору предъявляется предлагаемое алгоритмом решение, в отношении которого он может:
   (а) согласиться — и тогда система вводит его с определенными системой характеристиками;
   (б) отвергнуть — и тогда термин остается в общей выборке релевантных теме энциклопедических терминов, либо удаляется из выборки;

(в) отредактировать словарное описание по своему усмотрению и завершить редактирование командой ввода.

3. Для остающейся части выборки — выбор администратором базового концепта онтологии. Выбор производится в дереве признаков; может быть выбран концепт — простой класс, либо наименование классификационного признака.
4. Выбор способа добавления.
   4.1. Если администратором в качестве базового указано *наименование классификационного признака*, доступное действие — «*Присоединить к признаку*». В этом случае все термины выборки определяются как несовместимые подклассы, выделяемые по основанию, определяемому выбранным признаком.
   4.2. Если администратором в качестве базового указан *простой класс*, доступные действия:
      4.2.1. «*Добавить как синоним*» — термин добавляется как синоним указанного концепта.
      4.2.2. «*Добавить к новому классификационному признаку*» — система создает классификационный признак, определяющий новое основание деления указанного класса, запрашивая у администратора способ его именования. Все термины выборки подчиняются вновь созданному признаку, образуя разбиение исходного класса на подклассы.
      4.2.3. «*Добавить к новому списочному признаку*» — то же, что в п. 4.2.2. Но при последующей логической обработке вновь образованные подклассы не считаются объемно несовместимыми. Дальнейшая детализация таких подклассов в онтологии не допускается.
5. Если необходимо выполняется — графическое, либо символьное редактирование концептов добавленной группы (напр., перетаскивание некоторых из вновь введенных концептов в другие ветви дерева признаков).

5. Результаты и их обсуждение

Текущее состояние проекта таково. Выполнен анализ словарных определений — в соответствии с описанием в разделе 3 — трех словарей: Российский энциклопедический словарь (свыше 26 тыс. определений), русская Википедия (свыше 42 тыс.; по возможности, исключены персоналии), толковый словарь Т. Ф. Ефремовой (свыше 120 тыс.). Систематически организованных количественных оценок точности определения родового термина пока не делалось. Качественная оценка, полученная путем экспертного оценивания самими разработчиками результатов обработки случайно выбранной 1000 определений в первом из названных словарей дала — для «доброкачественных» определений — около 95 % правильных решений, для всех определений — порядка 85 %. Реализована и находится в процессе содержательной отладки технология энциклопедического импорта, описанная в разделе 4. Опыт, полученный в ходе отладки показал, что уже в своем настоящем виде предлагаемая технология существенно упрощает и ускоряет как отбор терминов для онтологии, так и собственно процесс ввода.

Перспективы дальнейшего развития технологии: использование более точных средств синтаксического анализа, расширяющих возможности точной идентификации онтологических характеристик терминов, в частности, обработка в определенных ситуациях межсегментных связей; формализация терминов других категорий: *наименования признаков и процессы*; поиск и формализация в тексте определения по возможности, всех онтологически значимых характеристик термина; использование наряду с *rule-based* процедурами статистических метрик.

Литература

1. *Buitelaar P., Cimiano P.* Ontology learning and population: bridging the gap between text and knowledge. // Series: Frontiers in artificial intelligence and applications, v. 167. — Amsterdam ; Washington, DC : IOS Press, 2008.
2. *Chodorow M. S., Byrd R. J., Heidorn G. E.* Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics, Chicago, 1985, pp. 299–304.
3. *Ide N., Véronis J.* Extracting knowledge bases from machine-readable dictionaries : Have we wasted our time? // KB&KS'93 Workshop,Tokyo. — 1993. — pp. 257–266
4. *Ponzetto S. P., Strube M.* Deriving a large scale taxonomy from Wikipedia // Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B. C., Canada, 22–26 July 2007, pp. 1440–1445.
5. *Navigli R., Velardi P.* From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions (в [1], pp. 71–87)
6. *Ide N., Veronis J.* Refining Taxonomies Extracted from Machine-Readable Dictionaries. // Hockey, S.,Ide, N. (eds.) *Research in Humanities Computing II*, Oxford University Press, 2003
7. *DODDLE Project*. A Domain Ontology rapiD DeveLopment Environment. URL: http://doddle-owl.sourceforge.net/en/
8. *Рубашкин В. Ш., Капустин В. А.* Использование определений терминов в энциклопедических словарях для автоматизированного пополнения онтологий // XI Всероссийская объединенная конференция «Интернет и современное общество» — СПб., 2008.
9. *Российский* энциклопедический словарь // М.: Большая Российская энциклопедия, 2001
10. *Рубашкин В. Ш.* Онтологии — проблемы и решения. Точка зрения разработчика // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». — М.: Издательский центр РГГУ, 2007
11. *Рубашкин В. Ш.* Представление и анализ смысла в интеллектуальных информационных системах. // М.: Наука, 1989
12. *Рубашкин В. Ш., Пивоварова Л. М.* Онторедактор как комплексный инструмент онтологической инженерии // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008». — М.: Наука, 2008
13. *Бочаров В. В., Пивоварова Л. М., Рубашкин В. Ш.* Логико-лингвистический анализ текстов определений в энциклопедических и толковых словарях. // Мегалинг-2009
14. *Система* автоматической обработки текстов АОТ. http://www.aot.ru/