АВТОМАТИЗАЦИЯ ОНТОЛОГИЧЕСКОГО ИНЖИНИРИНГА В СИСТЕМАХ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТА

AUTOMATIZATION OF AN ONTHOLOGICAL ENGINEERING FOR SYSTEMS OF KNOWLEDGE MINING IN TEXT

Ермаков А.Е. *(ermakov@rco.ru)*

ООО "ЭР СИ О", Москва

Компьютерная лингвистика и интеллектуальные технологии:

труды Международной конференции Диалог'2008

(Бекасово, 4-8 июня 2008 г.)

Аннотация

Доклад посвящен вопросам использования онтологий в системах извлечения знаний из текста. Рассматриваются особенности онтологий, используемых в таких системах. Предлагается методика автоматизированного построения онтологии, когда термины предметной области и связи между ними первоначально выделяются при помощи методов компьютерного анализа текста.

Онтологии в системах извлечения знаний

Перед системами извлечения знаний из текста сегодня встают насущные практические задачи, появление которых стимулировано развитием Интернета, содержащего огромное количество текстовой информации - реальные элементы утилитарного знания, полученные людьми в результате не только их профессиональной, но и бытовой деятельности. К таковым задачам, по мнению автора, относятся:

- Поиск и извлечение элементов знания, явно присутствующих в текстовой коллекции в виде: а) утверждения (*лекарство Антипилин – полная ерунда*; *наиболее вероятная причина свиста под капотом автомобиля в сырую погоду – слабое натяжение ремня генератора*); б) факта (*после принятия Антипилина может подниматься давление*; *летом 2006 фирма Пежо отозвала 20000 автомобилей из-за возможного возгорания в системе электроусилителя руля*).

- Порождение сложного знания путем обработки элементов знания следующими способами: а) генерация нового знания как цепочки логического вывода из элементарных утверждений и/или фактов, например: *продукт X некачественный* (утверждение), *X - продукт компании Y в 1997* (факт), *Z - технический директор компании Y с 1996 по 1998 годы* (факт), следовательно, *Z - плохой руководитель* (знание); б) эксплицирование

обобщенного знания, скрытого в совокупности частных утверждений и/или фактов, например, порождение выводов типа *препарат Антипилин имеет меньше побочных эффектов, чем Глипирон* (на основании анализа отзывов больных) или *Автомобили Форд Фокус ломаются чаще, чем Мицубиси. Типичная причина поломок автомобиля Форд Фокус – засорение бензонасоса* (на основании анализа отзывов владельцев автомобилей).

Согласно определению Т. Грубера, онтология - это спецификация концептуализации предметной области [1]. Это формальное и декларативное представление, которое включает словарь понятий и соответствующих им терминов предметной области, а также логические выражения (аксиомы), которые описывают множество отношений между понятиями. Для описания отношений в онтологиях используются весь арсенал формальных моделей и языков, разработанных в области искусственного интеллекта – исчисление предикатов, системы продукций, семантические сети, фреймы и т.п. Таким образом, модный сегодня термин “онтология” оказался близок по значению к термину “искусственный интеллект”, а термин “онтологический инжиниринг” явился синонимом термина “инженерия знаний”. На сегодняшний день существует не менее десятка зарубежных систем, относимых к классу инструментов онтологического инжиниринга, которые поддерживают различные формализмы для описания знаний и используют различные машины вывода из этих знаний. Наиболее известные из них – это Protégé (http://protege.stanford.edu), CYC (http://www.cyc.com), KAON2 (http://kaon2.semanticweb.org), OntoEdit (http://www.ontoprise.de/products/ontoedit), KADS22 (http://hcs.science.uva.nl/projects/kads22/index.html). Хороший обзор таких систем представляет собой работа [2]. Среди уже разработанных онтологий наиболее известными и объемными являются CYC (http://www.cyc.com) и SUMO (http://www.ontologyportal.org/).

Перекочевав в смежные с искусственным интеллектом области, термин онтология стал популярен в области систем машинного анализа текста, где в большинстве случаев используется в узком значении – в качестве синонима термина “тезаурус” или “классификатор” – и представляет собой просто словарь понятий (концептов), каждому из которых соответствует синонимический ряд терминов, плюс иерархическую структуру взаимосвязей между ними типа “часть-целое” или “общее-частное”. Такие “онтологии в слабом смысле” используются для формулировки запросов к поисковой машине, для автоматической классификации (категоризации) текстов, и, пожалуй, на этом все. Работающих прикладных программ, относимых к классу систем извлечения знаний из текста и использующих “онтологии в сильном смысле”, т.е. методы искусственного

интеллекта, способные нетривиально перерабатывать извлеченные из текста элементы знаний (интерпретировать, обобщать, выявлять зависимости, прогнозировать и т.п.), сегодня не существует, во всяком случае, для русского языка. Такое ограниченное использование онтологий обусловлено, на взгляд автора, двумя факторами. Во-первых, слабым распространением систем лингвистического анализа текста, способных интерпретировать синтаксические отношения между словами и потому действительно извлекать знания как некие нетривиальные элементы, обладающие внутренней структурой, пригодные для нетривиальной смысловой обработки искусственным мозгом – такие системы на мировом и российском рынках только начали появляться в последние несколько лет (Net Owl, Attensity, RCO Fact Extractor) и еще не успели "обрасти" приложениями. Во-вторых, относительно низкой достоверностью автоматически извлекаемых из текста утверждений и фактов, что обусловленно как несовершенством алгоритмов анализа текста, так и качеством источников информации, поскольку практически интересно извлечение знаний не из научной литературы, которая уже представляет конгломерат знания, а из текстовых “помоек”, к каковым относятся социальные сети Интернет, современные СМИ, и даже архивы научно-технических отчетов.

Другая особенность применения онтологий в системах извлечения знаний из текста – необходимость иметь дополнительную лингвистическую составляющую как для распознавания различных способов обозначения понятий (синонимичные термины), так и для семантической интерпретации разнообразных языковых конструкций в отношения между этими понятиями (синонимичные лексико-грамматические конструкции).

В итоге, для систем извлечения знаний из текста наиболее типичной является онтология “в слабом смысле” с относительно бедной концептуальной, но чрезвычайно богатой лингвистической составляющей.

Онтологический инжиниринг как объект автоматизации

Объединяя стандартные операции, выполняемые при формировании концептуальной составляющей онтологии [3,4,5], с теми операциями, которые диктуются требованиями к лингвистической составляющей, можно сформулировать перечень действий, подлежащих выполнению экспертом в ходе онтологического инжиниринга:

1. Формирование концептуальной схемы онтологии на основании профессиональных знаний в предметной области:

а) отбор базовых понятий-концептов. Например: *автомобиль, узел, тип кузова, год выпуска, пробег, техническое обслуживание, надежность, проходимость, экономичность.*

б) классификация базовых понятий с формированием абстрактных понятий – имен классов: типов объектов, их характеристик, ситуаций с их участием. Например: понятия - типы объекта: *автомобиль*, *узел автомобиля*; понятия - типы атрибутов объекта: *год выпуска*, *пробег*, *производитель*; понятия - типы характеристик объекта: *внешний вид*, *комфорт*, *ходовые качества*, *надежность*, *безопасность*; понятия - типы ситуаций (включая роли участников): *поломка (автомобиль, узел, причина)*, *техническое обслуживание (автомобиль, место, причина, стоимость, время ожидания)*,

в) определение возможных отношений понятий. Например: *автомобиль->{описывается}->атрибут, автомобиль->{содержит в себе}->узел, узел->{содержит в себе}->узел, объект->{характеризуется}->характеристка, характеристка->{характеризуется}->характеристка, объект->{ситуация}->объект;* и т.д.

2. Формирование фактического терминологического наполнения онтологии – соотнесение всех терминов предметной области с понятиями в концептуальной схеме, в ходе чего:

а) расширяется словарь понятий за счет наращивания онтологии "в глубину", если онтология предполагает родо-видовые связи (общее->частное, часть->целое) между понятиями одного класса, например: *узел автомобиля->двигатель->система зажигания->траблер->бегунок*, *ходовые качества->управляемость->склонность к сносу передней оси*;

б) для каждого понятия формируется словарь возможных терминов-значений: *производитель автомобиля* = {*АвтоВАЗ*, *Шевроле США*, *Шевроле Украина*, *Шевроле Корея*, …}, сила двигателя={*сильный*, *слабый*}

3. Формирование лингвистической составляющей:

а) фиксируются синонимичные обозначения каждого понятия или значения (термины): *Митсубиси = Мицубиши = Mitsubishi*, *двигатель = мотор = движок, маломощный = слабый = хилый*

б) описываются способы выражения отношений из онтологии в языке – типовые лексико-грамматические конструкции, для чего используется соответствующий лингвистическому анализатору формализм, например [6]. Так, отношение *объект->{характеризуется}->характеристка* может выражаться в тексте из Интернета такими конструкциями: *слабый двигатель, мотор – слабак, малая мощность двигателя, движок*

имеет небольшую мощность, движок еле тянет, автомобиль с трудом разгоняется, *тачка не прет*, и многими другими.

Автоматизация онтологического инжиниринга предполагают такую организацию этого процесса, при которой первоначальный перечень терминов предметной области и структура их взаимосвязей автоматически выявляются программными средствами на основании статистической обработки результатов лингвистического анализа коллекции текстов, после чего верифицируются и структурируются экспертом в соответствии с его имплицитной моделью знаний и прагматическими требованиями прикладной системы, для которой разрабатывается онтология.

С теоретической точки зрения, эффективность такой автоматизации онтологического инжиниринга обуславливается следующими факторами:

- В ходе просмотра конкорданса предметной области (частотного лексикона со взаимосвязями и контекстом) у эксперта активизируются соответствующие элементы его персональной модели знаний, что стимулирует эксплицирование и вербализацию этой модели;
- Концептуальная модель, формируемая с учетом фактического текстового материала, является актуальной, так как индивидуальная модель знаний эксперта в ходе эксплицирования автоматически верифицируется и стандартизируется в соответствии с общепринятыми представлениями;
- Легко формируется актуальное терминологическое наполнение, в том числе профессиональный сленг.

Алгоритмический арсенал для обработки текста

Технические решения, предлагаемые здесь к использованию при автоматизации формирования онтологии, основываются на следующей алгоритмической базе:

- способе генерации всех грамматически правильных словосочетаний – элементов смысла текста – на основании синтаксического анализа предложений с последующим обходом сети синтактико-семантических отношений. Соответствующие правила описаны в работе [7].

- способе установления ассоциативно-статистических связей между терминами, который основан на подсчете частоты их совместной встречаемости в рамках одной структурной единицы текста, обычно предложения. При этом в качестве вероятности наличия смысловой связи между терминами A и B можно рассматривать как абсолютную

частоту их совместной встречаемости F(A,B), так и ее отношение к максимальной из полных частот встречаемости F(A) или F(B), поскольку отношение F(A,B) / F(A) есть условная вероятность появления термина A совместно с термином B.

- синтаксическом способе установления связей, который предполагает выявление терминов, связанных с другими терминами на основе определенных типов связей в предложении или даже целых лексико-синтаксических конфигураций, определяемых требуемыми шаблонами [6]. Сложность используемых лексико-синтаксических шаблонов определяется наличием априорных знаний о типовых способах языкового описания отношений в предметной области. В наиболее простом и типичном случае возможен анализ на основании самых общих синтаксических шаблонов:

- Согласованное определение (прилагательное, причастие) выражает атрибут, качество объекта: *мощный двигатель*, *стучащая подвеска*;

- Признаковое существительное, при котором объект упоминается в позиции несогласованного определения, выражает атрибут, качество объекта: *мощность двигателя*, *мягкость подвески*;

- Событийное (обычно отглагольное) существительное, при котором объект упоминается в позиции несогласованного определения, выражает ситуацию, в которой участвует объект: *работа двигателя*, *стук подвески*;

- Существительное или прилагательное, связанное с объектом глаголом-связкой или стоящее в позиции субстантивного сказуемого, выражает атрибут, качество объекта: *двигатель – (был) слабак*, *подвеска является мягкой*;

- Полнозначный глагол или событийное существительное, при котором объект выступает в роли актанта, представляет ситуацию (действие, процесс, состояние), в которой участвует объект: *двигатель тянет*, *перебирать двигатель*, *стук в подвеске*.

- Наречие при глаголе, при котором объект упоминается в позиции субъекта, косвенно выражает характеристику объекта через его действие: *двигатель тянет слабо*, *подвеска мягко покачивается*.

Как видно, достоинством синтаксического способа является высокая точность выявления связей. Достоинством ассоциативно-статистического способа является его универсальность, которая заключается в отсутствии необходимости априорных предположений о структуре возможных синтаксических связей между терминами, и устойчивость к стилю текста, позволяющая выявить ассоциативные связи даже на

грамматически некорректном тексте или тексте особого стиля, к каковым часто относятся сообщения из Интернета.

Методика автоматизированного построения онтологий

Описываемая далее методика автоматизации операций, выполняемых экспертом в ходе разработки онтологии, базируется на идее итерационного выделения из коллекции текстов вначале наиболее простых и часто упоминающихся "сущностей" предметной области, а затем все более сложных, на основании определенных критериев их связи (сочетаемости) с более простыми сущностями, зафиксированными экспертом в ходе обработки результатов предыдущих итераций.

Методика состоит из следующих шагов:

Этап 1. Построение словаря терминов – обозначений “сущностей” предметной области.

1. Для каждого предложения текста производится синтаксический анализ с получением дерева синтаксических зависимостей между составляющими предложения. Дерево зависимостей преобразуется в сеть синтактико-семантических отношений. На основе обхода сети синтактико-семантических отношений производится синтез термино-подобных словосочетаний [7].

2. Для всего корпуса текстов составляется словарь термино-подобных словосочетаний, обозначающих неодушевленные и/или одушевленные предметы – именных групп, в которых главным словом являются предметные существительные. На этом этапе в словарь не включаются глаголы, прилагательные и образованные от них существительные, которые могут представлять ситуации и свойства, связанные с объектами предметной области. Для каждого словосочетания запоминается небольшой набор ссылок на предложения текста – цитат.

3. Фильтрация и сортировка словаря. Для каждого термина словаря - подсчет его полной и независимой частоты встречаемости. Отношение полной и независимой частот встречаемости позволяет учесть иерархию смыслов, которая выражается в уровне синтаксической зависимости одних элементов словосочетаний от других. Например, смыслы, входящие в состав словосочетания *натяжение ремня генератора*, не равнозначны: речь идет в первую очередь о *натяжении*, затем о *ремне*, и лишь опосредованно затрагивает *генератор*. В то же время цельный смысл *натяжение ремня генератора* более информативен, чем *натяжение ремня*, а *ремень генератора* информативнее, чем *ремень*, так как включает в себя конкретизирующие элементы. В итоге, те слова и словосочетания, для которых отношение величин "частота независимой

встречаемости" (не в составе других словосочетаний) и "полная частота встречаемости" оказывается близко к нулю, могут быть отброшены как неполные части устойчивых терминов.

4. Иерархическая группировка элементов словаря на основе лексической вложенности слов и словосочетаний. Например, два подмножества из множества словосочетаний *коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *автоматическая коробка*, *механическая коробка*, *задняя передача*, *высокая передача*, *низкая передача* могут быть сгруппированы как по общему существительному *коробка*, так и по общему существительному *передача*, в результате чего для эксперта определяются два возможных входа в словарь: *коробка* = { *коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *автоматическая коробка*, *механическая коробка* } и *передача* = { *коробка передач*, *автоматическая коробка передач*, *механическая коробка передач*, *задняя передача*, *высокая передача*, *низкая передача* }

5. Верификация/уточнение/пополнение построенного словаря терминов (обозначений объектов) экспертом в предметной области, в том числе фиксация синонимичных обозначений одних и тех же объектов.

Этап 2. Расширение словаря терминов именами ситуаций и свойств объектов предметной области.

1. Для каждого ранее зафиксированного термина-объекта предметной области - поиск слов (словосочетаний), связанных связями типа “объект-атрибут” и “объект-ситуация”, на основании шаблонов, задающих соответствующие конфигурации синтаксических связей.

2. Формирование общего словаря терминов – объектов, их атрибутов и ситуаций с их участием, группировка элементов словаря на основе взаимосвязей, выделенных на шаге 1, установление ссылок на предложения текста (цитат). Результирующий словарь представляет собой семантическую сеть взаимосвязанных сущностей трех классов, вход в которую возможен от частотного словаря имен объектов, атрибутов или ситуаций, а переход по связям между сущностями сопровождается возможностью просмотра текста, в котором связь раскрывается.

3. Исследование семантической сети экспертом в предметной области и окончательное формирование концептуальной составляющей онтологии (этап 1 процесса онтологического инжиниринга) - определение абстрактных понятий (классов объектов, их свойств и ситуаций) с определением типизированных отношений между сущностями этих классов; окончательное формирование фактического наполнения онтологии (этап 2

процесса онтологического инжиниринга) - соотнесение всех терминов словаря с понятиями в схеме онтологии, в том числе фиксация синонимичных обозначений свойств и ситуаций, определение возможных иерархических отношений между сущностями одного класса.

4. Если словарь объектов еще не был полностью сформирован (на этапе 1 экспертом была проанализирована только высокочастотная часть словаря), то возможно повторение шагов 1-3 с выделением новых упоминающихся объектов, связанных с уже известными свойствами и ситуациями из предметной области.

Этап 3. Описание способов выражения отношений из онтологии в языке – типовых лексико-грамматических конструкций.

1. Выявление множества ассоциативно-статистических связей между всеми терминами предметной области, для которых существует связь в онтологии. Ассоциативно-статистическая связь устанавливается между терминами, совместно упоминавшимися в предложениях текста не менее заданного числа раз.

2. Построение списков цитат из текста для каждого типа связей из онтологии, с предварительным отсевом статистически малодостоверных связей и тех связей, которые выражаются уже известными способами и могут быть выделены на основании синтаксических шаблонов (Шаг 1 Этапа 2).

Исследование списков цитат экспертом для фиксации новых способов выражения в тексте отношений из онтологии – новых лексико-грамматических конструкций, используемых впоследствии для настройки лингвистического обеспечения системы автоматического извлечения знаний из текста.

Заключение

Экспериментальная проработка и успешная апробация методики проводилась специалистами компании “ЭР СИ О” в ходе построения онтологии для предметной области “Автомобили”. Онтология предназначена для оценки конкретных марок автомобилей с точки зрения характеристик (*положительная/отрицательная*) их потребительских свойств, высказываемых в отзывах потребителей, размещенных в Интернете. При составлении онтологии использовался реальный языковой материал, полученный из автомобильных сообществ блога “Живой Журнал” (http://www.livejournal.ru/auto) – около 30 Мбайт текстовых сообщений. Результирующая онтология содержит более 1200 терминов (не считая конкретных марок автомобилей), из которых 211 представляют собой наименования узлов автомобиля (*движок*, *коробка*

передач, ходовая часть); 71 - наименование их свойств (*ходовые качества*, *комфорт*, *надежность*, *стоимость содержания*); 882 - возможные наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой*, *поломка*, *глючить*, *отстойно*), 37 эмоциональных характеристик (*любить*, *жалоба*, *плеваться*). Возможные связи в предложении между классами терминов из онтологии описываются 150 лексико-грамматическими шаблонами. В результате для каждой модели автомобиля в блогах удается "выловить" положительные и отрицательные отзывы, классифицировав их по темам "за что хвалят/ругают".

Литература

1. Gruber T. R. A translation approach to portable ontologies // Knowledge Acquisition, 1993, V. 5(2), P.199-220.

2. Овдей О.М., Проскудина Г.Ю. Обзор инструментов инженерии онтологий // Электронные библиотеки – Москва: Институт развития информационного общества, т.7 вып.4, 2004. – Электронный журнал, посвященный созданию и использованию электронных библиотек. – (http://www.elbib/).

3. Гладун А., Рогушина Ю. Онтологии в корпоративных системах // Корпоративные системы, – 2006. – № 1. – с. 41-47.

4. Гаврилова Т.А. Использование онтологий в системах управления знаниями // Труды международного конгресса «Искусственный интеллект в XXI веке», Дивноморское, Россия, М., Физматлит. 2001 - с. 21-33.

5. Гаврилова Т.А. Извлечение знаний: лингвистический аспект // Корпоративные системы (Enterprise Partner), 2001. - № 10 (25). - с. 24-285.

6. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва, Наука, 2004. – С. 282-285.

7. Ермаков А.Е. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза. // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. – Москва, Наука, 2003. - С. 136-140.