УДК 004.415

Лесько О.М., Рогушина Ю.В.

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ АНАЛИЗА СЕМАНТИКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Предлагается использовать онтологии для автоматизированной семантической разметки естественноязыковых текстов с учетом как морфологических и синтаксических свойств естественного языка (в частности, украинского), так и структуры ПрО, а также знаний пользователя об этой ПрО. Разработан алгоритм, который осуществляет выделение слов и поименованных сущностей ЕЯ-текста, связанных с определенными понятиями выбранной пользователем ПрО (например, с терминами онтологии). Чтобы полученная семантическая разметка была пригодна для обработки различными распределенными системами, представляется целесообразным использовать технологии и стандарты, разработанные в рамках проекта Semantic Web.

Введение

Сегодня пользователи Интернет получают доступ к огромному количеству информационных ресурсов, значительная часть которых представлена на естественном языке (ЕЯ). Возрастание их объема приводит ко многим проблемам – проанализировать эту информацию вручную за удовлетворительное время человек не способен, а полностью формализовать содержание ЕЯ-текстов невозможно даже теоретически. Решение проблемы связано с переходом от хранения и обработки данных к накоплению и обработке знаний, в частности, с переходом от традиционного Web к Semantic Web [1], базирующемуся на использовании метаданных для описания семантики информационных ресурсов (ИР) и средств обработки этих метаописаний. Для такого перехода необходимо использовать знания о предметной области (ПрО), т. е. нужно связывать фрагменты текста с какими-то понятиями ПрО. Одним из способов такого связывания является семантическая разметка текста, или семантическая аннотация. Примером системы, в которой используется семантическая разметка, является Semantic Wiki [2].

Аннотация – это метаданные, которые описывают документ или его часть. Она может быть вставлена в тот же документ или сохранена отдельно. Семантическая аннотация – аннотация, которая написана на формальном языке с хорошо определенной семантикой, и базирующаяся на онтологии.

При формировании семантической разметки нужно использовать не только знания ПрО (или хотя бы ее терминологическую базу), но и правила того конкретного естественного языка, на котором написан текст. К сожалению, создание такой разметки является нетривиальной и довольно трудоемкой задачей. Семантическая разметка зависит и от того, какие именно средства используются для описания ПрО.

Постановка задачи

Для семантической разметки ЕЯ-текстов необходимо разработать алгоритм, который обеспечит выделение фрагментов (слов) ЕЯ-текста, связанных с определенными понятиями выбранной пользователем ПрО (например, с терминами онтологии). Для этого предлагается анализировать ЕЯ-тексты определенной ПрО с учетом как морфологических и синтаксических свойств естественного языка (в частности, украинского), так и структуры ПрО и знаний пользователя об этой ПрО. Для полученной разметки нужно разработать средства и методы, позволяющие с ее помощью осуществлять поиск информации, релевантной персональным информационным потребностям конкретного пользователя.

Чтобы семантическая разметка была пригодна для обработки различными распределенными системами, целесообразно использовать технологии и стандарты, разработанные в рамках проекта Semantic Web.

Лингвистический анализ

Различные типы систем извлечения знаний из текстов основываются на метаданных; на лингвистическом анализе текста; на анализе структуры документа; на анализе формальных свойств документа.

Лингвистические методологии применяются для ЕЯ-текстов произвольной, четко не выделенной структуры. Они в значительной мере зависят от языка, на котором написан текст, требуют больших вычислительных мощностей и также не всегда позволяют однозначно идентифицировать семантику текста.

Лингвистически методы позволяют выделить в тексте слова, связанные с понятиями (классами) ПрО (например, «стол» связан с понятием «мебель»), и слова, являющиеся именами, т. е. связанные с экземплярами понятий (классов) онтологии (например, «Лада» является экземпляром класса «собака»). Рассмотрим детальнее методы лингвистического анализа и те сведения, которые можно получить из текста этими методами

Традиционно лингвистический анализ включает этапы морфологического, синтаксического и семантического анализа [3].

Для выделения лексем в ЕЯ-тексте применяют морфологический анализ. Слово (лексема) с грамматической точки зрения определяется как система словоформ, основы которых тождественны по значению, а одноименные морфы основ, также тождественные по значению, фонематически близки или тождественны друг другу. В одну лексему объединяются разные словоформы одного слова (например, «словарь, словарём, словарю» и т. п.).

На этапе морфологического анализа возникают две задачи: определение того, какой частью речи является слово в предложении; определение морфологических характеристик слова (числа, рода, падежа, времени и т. п.). При решении первой задачи особую трудность представляют омоформы и омографы. Омоформы – слова, совпадающие в одной, реже – в нескольких грамматических формах, например, три – тереть (глагол) и три (числительное). Омонимия может также возникать на уровне форм слова одной и той же части речи. Для снятия омонимии используется контекст, в котором встретилось слово.

Синтаксический анализ заключает в распознавании синтаксической структуры предложений на основе морфологической информации и синтаксических правил объединений слов и словосочетаний данного языка. Синтаксическая структура – это связь между словами предложения. Для единообразного описания синтаксических правил языка используются формальные грамматики.

Семантический анализ направлен на распознавание смысла текста. Способы описания семантики текста и предложения, также алгоритмы построения такого описания определяются целями анализа. Назначение семантического анализа – извлечь из ЕЯ-текста содержащиеся в нем знания, заложенные в него автором, и предоставить в форме, пригодной для автоматизации их обработки.

Рассмотрим различные этапы лингвистического анализа ЕЯ-текстов на примере задачи распознавания поименованных сущностей (ПС). В тексте ПС обозначаются собственными именами, которые пишутся с большой буквы. Поименованными сущностями (Named entities, NE) считаются люди, организации, города, страны, реки, имеющие свое уникальное имя. ПС в онтологии соответствуют экземпляры классов. Более широкая интерпретация позволяет считать поименованными сущностями также некоторые скалярные величины (дату, время, валюту, цену и т. п.) и адреса.

Алгоритмы классификации в анализе ЕЯ-текстов применяются в первую очередь для классификации поименованных сущностей.

На этапе морфологического анализа в ЕЯ-тексте выделяются слова и словосочетания, начинающиеся с большой буквы и

не стоящие в начале предложения, определяется (если удается) их род и число.

На этапе синтаксического анализа для ПС уточняются их род и число, место в предложении и связь с другими членами предложения.

В пределах более широкой интерпретации ПС могут считаться также некоторые скалярные величины (числа, суммы денег, даты) и адреса.

На этапе семантического анализа делаются попытки связать ПС с каким-либо классом из соответствующей онтологии ПрО (например, «Сидоров» – человек, потому что он «трудолюбивый», «читает» и «работает», «Киев» – город, потому что он «находится» и «является столицей», «очень большой»).

В результате лингвистического анализа текста получаем два множества – слов текста с соответствующей морфологической информацией (часть речи и т. д.) и синтаксических связей между словами и словосочетаниями в предложениях текста (члены предложения)

На основании этих сведений можно получить информацию о связях между словами и словосочетаниями. Но этого недостаточно, чтобы понять смысл текста. Понимание смысла (семантики) требует использования знаний о значениях слов для установления семантических связей между словами и понятиями предметной области.

Онтологии как источник знаний о ПрО

В инженерии знаний под онтологией понимается детальное описание некоторой ПрО, которое используется для формального и декларативного определения ее концептуализации. *Онтология* – это явная спецификация концептуализации на уровне знаний [4, 5]. Онтология обязательно включает словарь понятий ПрО и указания о связях между ними, что задает структуру ПрО область и ограничивает возможные интерпретации терминов. *Формальная модель* онтологии – тройка O = < P, R, F>, где P – множество понятий ПрО, R – множество связей между понятиями ПрО, F – множество аксиом и правил вывода ПрО

Для использования онтологий в задачах понимания смысла ЕЯ-текстов необходимы алгоритмы отображения синтаксических отношений, присутствующих в ЕЯ-текстах, на отношения, имеющиеся в онтологиях. При этом возможен как перевод исходного текста на язык формальной грамматики в категориях род, число, падеж, так и непосредственное получение семантических отношений из морфологической формы слов [3, 6]. Примеры использования онтологий для извлечения фактов из ЕЯ-текстов определенной ПрО приведены в [7, 8].

Следующий алгоритм предлагается для поиска и классификации ПС в ЕЯ-текстах .

Алгоритм семантической разметки текстов

Семантическая разметка ЕЯ-текстов для определенной ПрО создается в два этапа. На первом этапе производится обучение. На первом этапе используется алгоритм накопления лингвистических сведений о ПрО (АНЛС).

На этапе обучения необходимо сформировать следующие множества:

- P_w – словоформы, связанные с понятиями онтологии ПрО. Эта информация может быль извлечена из различных слоарей синонимов, лингвистических баз данных, а также явным образом вручную из корпуса текстов;
- R_w – словоформы, связанные с отношениями онтологии ПрО (аналогично);
- I – отношения именования (ОИ), связывающие а. ПС и классы , б. классы и подклассы;
- I_w – словоформы, связанные с ОИ;
- шаблоны, связывающие ПС и имена их классов (в общем случае слабо зависящее или вообще не зависящие от предметной области). Эта операция может быть выполнена один раз, но в дальнейшем множество шаблонов может расширяться для учета специфики ПрО. Каждый шаб-

лон представляет собой строку символов, состоящую из имени предиката и модели управления, например «называется кто как».

Это осуществляется следующим образом. В процессе обучения в корпусе текстов находятся слова, написанные с большой буквы и не входящие в общий словарь; состоящие из больших букв; слова, взятые в кавычки. Для таких слов выделяются синтаксические шаблоны, определяющие указание на принадлежность ПС к определенному классу. Затем в предложении с такими ПС обнаруживаются имена классов , к которым принадлежат эти ПС. Если такое имя класса присутствует, то осуществляется попытка выделить слова, связывающие синтксически ПС и имя ее класса. Если это удается, то для этих слов – ОИ – строится шаблон.

На вход АНЛС подаются: онтология О, характеризующая знания пользователя об интересующей его ПрО; словарь лексем естественного языка; обучающая выборка ЕЯ-текстов, по которой при участи пользователя формируется набор словоформ, связанных с терминами онтологии О (корпус текстов).

После того, как формируется набор множеств слов ЕЯ-текстов, каждый из которых соответствует определенному термину онтологии $\forall t_i \in T, t = \overline{1,n}, \exists S_i = \{s_{i_1},...,s_{i_m}\}$, эти множества могут быть преобразованы для более эффективной обработки текста. Часть элементов множества $S_i = \{s_{i_1},...,s_{i_m}\}$ могут быть опознаны пользователем как одна словоформа и заменены элементом , являющимся общей частью этих элементов,

$$\exists l, l \subseteq s_{i_k}, k = \overline{1,p}, p \geq 2, s_{,i_k} \in S_i .$$

В результате обучения системы каждому термину онтологии *O* приписывается 0, 1 или несколько словоформ, соответствующих в ЕЯ данному понятию. Словоформы извлекаются из обучающего множества ЕЯ-текстов, отнесенных пользователем к определенной ПрО, описанной в *O* с точки зрения информационных интересов пользователя. Полученная информация заносится в таблицу *C*.

Алгоритм построения шаблонов

Работа алгоритма начинается с процесса обучения распознаванию ПС. Предполагаем, что указание на класс ПС находится в том предложении, где имя ПС встречается впервые, либо в следующем предложении. Он состоит из следующих шагов:

1) сформировать множество предложений, содержащих ПС, $-Q$. Для этого нужно найти в текстах предложения, в которых встречаются слова, которые могут являться именами ПС – слова, не найденные ни в словаре пользователя, ни в общем словаре, и взятые в кавычки либо состоящие из больших букв или начинающиеся с большой буквы;

2) сформировать множество предложений, содержащих указание на класс ПС, $-Q_{Class}$. Для этого нужно вручную проанализировать предложения, содержащиеся в множестве Q, и исключить из него те, в которых нет указания на класс ПС (при первом появлении имени ПС в тексте предложение может не содержать указания на класс ПС, и тогда нужно анализировать следующие предложения с данной ПС, например, «Я купил мебель в АВС. (АВС – это магазин)»;

3) по множеству Q_{Class} сформировать множество шаблонов Т. Каждый шаблон включает слово из I_w и морфологическую информацию для связанных с ним слов в соответствии с моделью управления [9]. Например, «называется (кто/что, nm (кто/что, кем/чем)» - > «называется (класс, имя)».

На этом процесс обучения заканчивается.

Алгоритм автоматической семантической разметки

На вход алгоритма автоматической семантической разметки (ААСР) подается:

- P_w – словоформы, связанные с понятиями онтологии ПрО;

- R_w – словоформы, связанные с отношениями онтологии ПрО;

- I_w – словоформы, связанные с ОИ шаблоны, связывающие ПС и имена их классов;

- ЕЯ – тексты, для которых надо создать семантическую разметку.

На этапе анализа нового ЕЯ-текста нужно выделить в тексте:

- словоформы, связанные с понятиями онтологии ПрО;

- словоформы, связанные с отношениями онтологии ПрО;

- слова, которые могут быть именами ПС;

- ОИ.

Работа алгоритма включает нахождение в ЕЯ-текстах словоформ из P_w и I_w, приписывании в их начале и конце тэгов, которые соответствуют понятиям множества Р онтологии ПрО.

Размеченный таким образом текст в дальнейшем анализируется для определения класса ПС следующим образом.

Вначале в ЕЯ-текстах обнаруживаются слова и словосочетания, которые могут являться именами ПС.

Затем к тексту нужно применить шаблоны, описывающие правила, связывающие имена ПС с именами их классов. Например, «называется (кто/что, nm (кто/что, кем/чем))» -> «называется (класс, имя)». Эти шаблоны позволяют определить класс найденных в тексте ПС.

Для этого надо выполнить следующие действия:

1) проверить, есть ли в предложении ПС;

2) проверить, есть ли в предложении шаблоны отношений (имена классов из онтологии или их лингвистические словоформы);

3) если 1 и 2 – да, то разобрать синтаксическую структуру предложения.

Если ПС, имя понятия и имя отношения именования ОИ занимают место в предложении, соответстующие шаблону места (подлежащее, сказуемое, дополнение, обстоятельство….), то считать ПС относящейся к соответствующему классу. Например, в шаблоне «называется (кто/что, nm (кто/что, кем/чем)» кто/что заменяется на «Эта улица», а nm (кто/что, кем/чем) – на Зеленая.

В процессе работы алгоритма формируется словарь, в котором каждому имени ПС ставится в соответствие ее класс.

Процесс анализа, в котором используются результаты такого обучения, состоит из таких шагов:

- найти предложение, в котором впервые встречается имя некоторой ПС;

- осуществить его синтаксический анализ;

- вычислить номер шаблона по номерам составляющих предложения;

- по полученному шаблону определить название класса и имя ПС;

- добавить в текст тэги языка XML [10].

Результат работы этого алгоритма – множество семантически размеченных по правилам языка XML ЕЯ-текстов, пригодных для автоматического анализа, например, для поиска интересующих пользователя сведений, связанных с определенными понятиями ПрО, описанной в онтологии О.

Перспективы использования семантической разметки

Рассмотрим работу алгоритма на примере задачи подбора эксперта для рецензирования научной статьи.

Если тематика статьи соответствует направлению работы журнала в целом, то необходимо из множества специалистов-экспертов, известных редакции журнала, выбрать одного (или нескольких), чья специализация ближе всего к направлению работы. При этом человек, который осуществляет такой выбор, не является экспертом в данной ПрО. Если автор статьи отнес свою работу к четко определенному подразделу, а один или несколько специалистов явно классифицированы как эксперты именно в этой области – это сделать довольно просто.

Но часто на практике каждый из специалистов декларирует себя как эксперта в нескольких ПрО, хотя его компетентность распространяется только на некото-

рые подобласти этих ПрО и по состоянию несколько лет назад. Так, если специалист 10 лет назад занимался проблемами логического вывода или представлением знаний, это не значит, что сегодня он может объективно оценить исследования в области Semantic Web. Кроме того, одну и ту же научную работу можно отнести сразу к нескольким разделам.

В результате статья часто попадает на рецензию к специалисту, не способному объективно оценить ее научную ценность и новизну.

Использование семантической разметки ЕЯ-текстов позволяет в той или иной мере решить эту проблему (или по крайней мере добавить новый инструмент для ее решения). Будем считать, что объективным отражением знаний специалиста (научного сотрудника) являются его публикации, которые в основном представляют собой ЕЯ-текст.

ПрО, в которой специализируется научный журнал (и его подразделы), может быть формализована при помощи онтологии. Затем, воспользовавшись вышепредложенным алгоритмом семантической разметки, необходимо разметить публикации потенциальных экспертов понятиями этой онтологии. Такая разметка позволит охарактеризовать специализацию каждого из них (причем с учетом динамики). Так, можно определить, что в подразделе «экспертные системы» Иванов был специалистом на 70 %, а 5 лет назад – на 55 %.

Затем производится семантическая разметка вновь поступившей статьи и осуществляется поиск эксперта, который в своих публикациях использует те же понятия (и те же связи). При равных показателях предпочтение отдается тому эксперту, чьи знания более актуальны.

Решаемая задача близка к поиску текста, похожего на образец (подобную функцию предлагают сегодня многие ИПС, но результаты такого поиска крайне непредсказуемы). В данном случае цель – найти множество текстов, использующих не просто похожие слова или словосочетания, а одинаковые понятия и близкие наборы понятий. Если оказывается, что ни один из экспертов своей компетентностью не покрывает полностью тематику статьи, то можно предложить отрецензировать статью нескольким экспертам независимо, предложив обратить внимание на определенные аспекты работы.

При наличии семантической разметки статьи и публикаций различных экспертов можно сравнивать их предметные области с точки зрения онтологии конкретного издания и подбирать набор экспертов, чьи компетенции охватывают все значимые (с точки зрения этой онтологии) аспекты статьи. Кроме того, выделение ПС позволяет определить индекс цитируемости различных экспертов и таким образом оценить их относительную квалификацию в той или иной ПрО.

Выводы

Предложенный в работе подход ориентирован на явное выделение семантики ЕЯ-текстов с помощью понятий определенной ПрО. Способная к обучению система позволяет накапливать лингвистическую информацию (синонимы, словоформы, правила определения понятий и классов, способы выделения поименованных сущностей и т.д.), связанные с определенным комплексом понятий и связей между ними, формально представленным в виде онтологии. В дальнейшем эти сведения позволят относительно легко обнаруживать в новых ЕЯ-текстах фрагменты, интересные пользователю и связанные с определенными понятиями интересующей его ПрО. Кроме того, такая семантическая разметка может стать основой для автоматического создания метаописаний ИР с точки зрения фиксированной ПрО, которые будут интероперабельны и могут использоваться приложениями Semantic Web.

1. *Semantic Web*. - http://www.w3.org/2001/sw/.
2. *Semantic wiki* – http: // en. wikipedia. org/wiki/Semantic_wiki.
3. *Палагін О. В., .Світла С. Ю, . Петренко М. Г, .Величко В.Ю.* Про один підхід до аналізу та розуміння природномовних об’єктів // Комп’ютерні засоби, мережі та системи. – 2008, № 7. – С. 128 – 137.

4. *Gruber T.R.* A Translation Approach to Portable Ontology Specifications // Knowledge Acquisition. – 1993. – N 5, P. 199 – 220.
5. *Гладун А.Я,. Рогушина Ю.В* Онтологии в корпоративных системах. Часть II // Корпоративные системы. – 2006. – № 1 http: // www. management. com. ua/ims/ims 116. html
6. *Gladun V., Velichko V., Svyatogor L.* Hierarchical Three-level Ontology for Text Processing. International Book Series "INFORMATION SCIENCE & COMPUTING", N. 7 – FOI ITHEA Sofia, Bulgaria. – 2008. – P. 11 – 17.
7. *Добров Б.В., Лукашевич Н.В.* Онтологии для автоматической обработки текстов: описание понятий и лексических значений. //http://www.dialog21.ru/dialog2006/materials/html/Dobrov_files/editdata.mso.
8. *Невзорова О.А.* Онтологическая поддержка методов решения задач семантико-синтаксического анализа текстов. – //http://www.raai.org/cai-08/files/cai-08_paper_234.doc
9. *Апресян Ю. Д.*. Типы коммуникативной информации для толкового словаря. – http://www.philology.ru/linguistics2/apresyan-88.htm
10. *Extensible* Markup Language (XML). – http://www.w3.org/XML/

Получено 24.04.2009

Об авторах:

Лесько Ольга Николаевна,
инженер-программист,

Рогушина Юлия Витальевна,
кандидат физико-математических наук, старший научный сотрудник.

Место работы авторов:

Институт программных систем
НАН Украины,
03187, Киев 187,
Проспект Академика Глушкова, 40.