АЛГОРИТМ СЕМАНТИЧЕСКОГО АНАЛИЗА ДОКУМЕНТОВ С ЦЕЛЬЮ СОЗДАНИЯ СЕМАНТИЧЕСКИХ МЕТАОПИСАНИЙ

Губин М. Ю.
Научный руководитель: Тузовский А.Ф. д.т.н., профессор
Томский политехнический университет
e-mail: gubin.m.u@gmail.com

Введение

Основным препятствием для повсеместного распространения технологий Semantic Web на сегодняшний день служит недостаточное количество семантических метаданных, пригодных для обработки с использованием семантических технологий и подходов. В данной работе описывается постановка задачи создания семантических метаданных на основе текстов на русском языке и предлагается алгоритм решения этой задачи.

Постановка задачи

Под **документом** D_i будем понимать фрагмент текста на естественном языке.

Семантическое метаописание документа строится согласно **онтологии предметной области** O, представляющей собой набор **понятий** C_i, связанных между собой **отношениями** R_i. Также в онтологию предметной области входят **экземпляры объектов** E_i. Понятия, отношения и экземпляры имеют одну или более текстовых меток T_i. **Текстовая метка** T_i элемента онтологии – слово либо словосочетание естественного русского языка, соответствующее некоторому элементу онтологии.

Для построения базового семантического метаописания на основе текста документа для каждого его предложения L_i формируется сеть элементов документа, представляющая собой граф, состоящий из множества вершин W_i и соединяющих их рёбер L_i. Элементарная сеть представляет результат синтаксического анализа и дополнительных семантических трансформаций дерева синтаксических зависимостей между словами в отдельном предложении. **Вершинами** W_i сети элементов являются сущности, встречающиеся в предложении, а **рёбра** L_i представляют собой семантические отношения между сущностями.

Сети элементов предполагается получать из результатов синтаксического разбора текстов на естественных языках.

Задача синтаксического разбора текстов на данный момент в различной степени решена для русского [6,7] и английского [3,4,5] языка. Также существуют работы по синтаксическому разбору текстов на французском, норвежском, корейском и греческом [4], а также испанском и японском [4,5] языках. В данной работе рассматривается частный случай с русским языком.

Семантическое метаописание – это набор извлечённых из предложений документа RDF-триплетов T_i, представляющих собой кортежи вида $<S_i,P_i,O_i>$, где S_i включен в объединение C_i и E_i, P_i включен в R_i, а O_i включен в объединение C_i и E_i.

Поскольку для анализа документа необходимо использовать онтологию соответствующей предметной области, возникает необходимость определения предметной области документа. Для определения предметной области предлагается использовать наивный байесовский классификатор. Классом Ci в данном применении классификатора будет считаться предметная область документа, а классифицирующим признаком Wi — вхождение в данный текст данного термина. Тогда вероятность принадлежности документа к предметной области можно вычислить по формуле

$$p(C_i|W_1, W_2, \dots, W_n) = \left(\frac{1}{Z}\right) p(C_i) \prod_{j=1}^{n} p(W_j|C_i)$$

где Z = 1/n, а p(Ci) и p(Wi|Ci) получаются путём анализа обучающего набора документов на этапе предварительной настройки классификатора.

Решаемой задачей является получение на основе текста на русском языке его семантического метаописания. процесс построения семантического метаописания

Для построения метаописания предполагается использовать редактор метаописаний, который позволяет пользователю создавать метаописания в полу-ручном режиме с использованием онтологического инструментария. Построение метаописания производится с применением онтологии предметной области, а также базы знаний.

Для этого, вначале производится предварительный анализ текста.

Производится семантический анализ текста, преобразующий текст в слова с номером их начальных символов, смысловые связи между словами, обнаруженные и преобразованные в RDF триплеты.

Подсчитывается количество вхождений слов в текст. При этом не учитываются так называемые «стоп-слова» (предлоги, союзы и частицы). Остальные слова нормализуются и подсчитывается количество вхождений нормы слова.

Составляется ранговое распределение слов в документе. Слова с одинаковым количеством

вхождений объединяются в классы, которые затем нумеруются в порядке убывания количества вхождений слов-членов класса в тексте, начиная с 1. [8]

Производится сопоставление важных для документа слов с терминами из онтологии предметной области и базы зныний.

Производится поиск класса, слова в котором являются значимыми для текста, с наибольшим номером. Все классы, идущие после него, отсеиваются. [8]

Выставляется первичное значение «веса» слов в документе. Оно равняется N_{max}/N_i, где N_{max} — количество вхождений слов первого ранга, а N_i — количество вхождений слова t_i. [8]

Производятся корректировки значений весов для упорядоченных пар слов, входящих в одни и те же триплеты либо предложения.

Из множества выделенных из текста RDF-триплетов выбираются триплеты-кандидаты, каждая из позиций которых (субъект, предикат и объект) заняты в естественно-языковом представлении вхождением метки (соответственно, субъект и объект – метками понятия либо экземпляра, а предикат – меткой свойства).

После предварительного анализа текста, список триплетов-кандидатов, а также список ключевых слов документа и объектов онтологии и базы знаний, текстовые метки которых соответствуют ключевым словам, предоставляются пользователю для корректировки и формирования основы метаописания документа. Данный подход позволяет существенно ускорить формирование метаданных документа пользователем по сравнению с полностью ручным формированием, при этом сохраняя высокую точность и полноту создаваемых метаданных.

Разработанная методика позволяет создавать пригодные для использования в целях семантического поиска метаданные документов. Дальнейшее развитие системы предварительного разбора документа позволит формировать метаданные в полностью автоматическом режиме.

Литература

1. Люгер Д. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. 4-е издание. – М.: Вильямс, 2003 – 864 с.
2. Хорошилов А. А. Белоногов Г. Г. Калинин Ю. П. Компьютерная лингвистика и перспективные информационные технологии: теория и практика. // НТИ. Сер. 2. Информ. процессы и системы / ВИНИТИ. - 2004. - N 8. - С.30-43.
3. Poon H., Domingos P. Unsupervised semantic parsing. ACL Anthology. A Digital Archive of Research Papers in Computational Linguistics / [Электронный ресурс]. Режим доступа: www.aclweb.org/anthology/D/D09/D09-1001.pdf, свободный (дата обращения: 02.10.2010).
4. Deep linguistic processing with hpsg. [Электронный ресурс]. – Режим доступа: http://www.delph-in.net, свободный (дата обращения: 02.10.2010).
5. Сайт лаборатории speech technology копрорации microsoft. [Электронный ресурс]. – Режим доступа: http://research.microsoft.com/en-us/groups/srg/default.aspx, свободный (дата обращения: 02.10.2010).
6. Сайт рабочей группы «Автоматическая обработка текстов». [Электронный ресурс] / Режим доступа: http://aot.ru/, свободный (дата обращения: 02.10.2010).
7. Сайт компании RCO [Электронный ресурс] / Режим доступа: http://www.rco.ru, свободный (дата обращения: 02.10.2010).
8. Thomas Roelleke , Jun Wang, TF-IDF uncovered: a study of theories and probabilities // Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore