ПРИМЕНЕНИЕ СЕМАНТИЧЕСКИХ СЕТЕЙ И ЧАСТОТНЫХ ХАРАКТЕРИСТИК ТЕКСТОВ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ ДЛЯ СОЗДАНИЯ СЕМАНТИЧЕСКИХ МЕТАОПИСАНИЙ

М. Ю. Губин, В. В. Разин, А. Ф. Тузовский

Институт кибернетики Национального исследовательского
Томского политехнического университета, 34034, Томск, Россия

УДК 004.89:004.93

Представлен подход к созданию семантических метаданных с использованием семантических сетей и частотных характеристик для анализа документов на естественных языках.

Ключевые слова: онтология, семантическая сеть, фрейм, RDF.

The article describes an approach to creation of semantic metadata using semantic networks and frequency characteristics to analyze natural language text documents.

Key words: ontology, frame, semantic network, RDF.

Введение. Естественные человеческие языки обладают большой выразительностью и сложностью, существенное влияние на смысл текста в них оказывают контекст и эмоциональная составляющая. Понимание естественного языка не только включает разбор предложений по частям речи и поиск значений слов в словаре, оно основано на обширном фоновом знании о предмете, идиомах, используемых в этой области, а также на способности применять общее контекстуальное знание для понимания недомолвок и неясностей, присущих естественной человеческой речи. Поэтому системы, использующие натуральные языки с характерными для человеческой речи гибкостью и общностью, находятся за пределами существующих методологий [1]. Однако в определенных условиях (когда документ имеет достаточно строгую грамматическую структуру, а следовательно, содержит достаточно информативную формальную составляющую) данная задача решаема с достаточно высоким качеством распознавания смысла [2]. В настоящей работе описаны условия, выполнение которых необходимо для успешного распознавания, и предложен соответствующий алгоритм.

Постановка задачи. Предлагаемый алгоритм позволяет решать задачу создания метаописаний документов для последующего семантического поиска по ним на данном множестве документов D_i, относящихся к одной предметной области. Под документом D_i в рамках данного исследования будем понимать фрагмент текста на естественном языке.

Для реализации семантического поиска по документам необходимо создать достаточно полные семантические метаописания документов T_i.

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 эк-

земпляры 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 семантическая сеть, представляющая собой граф, состоящий из множества вершин W_i и соединяющих их ребер L_i. Элементарная сеть является результатом синтаксического анализа и дополнительных семантических трансформаций дерева синтаксических зависимостей между словами в отдельном предложении. Вершинами W_i семантической сети являются сущности, встречающиеся в предложении, а ребра L_i представляют собой семантические отношения между сущностями. Семантические сети предполагается получать на основе результатов синтаксического разбора текстов на естественных языках. В настоящее время задача синтаксического разбора текстов в различной степени решена для русского (сайт рабочей группы "Автоматическая обработка текстов": http://aot.ru/; сайт компании RCO: http://www.rco.ru) и английского [3, 4] языков (сайт лаборатории speech technology корпорации Microsoft: http://research.microsoft.com/en-us/groups/srg/default.aspx). Также существуют работы по синтаксическому разбору текстов на французском, норвежском, корейском и греческом [4], а также испанском и японском [4] языках (сайт лаборатории speech technology корпорации Microsoft: http://research.microsoft.com/en-us/groups/srg/default.aspx). В данной работе рассматривается частный случай на примере русского языка.

Программный интерфейс большинства существующих семантических анализаторов позволяет получить для каждой сущности набор направленных связей, исходящих от нее к другим сущностям. Обычно направление связи соответствует направлению синтаксического подчинения (для равноправных однородных членов предложения создается пара одинаковых связей, одна из которой направлена от первого члена ко второму, а другая – от второго к первому). Семантические сети, соответствующие описанным выше критериям, могут быть использованы в разрабатываемом алгоритме с незначительными преобразованиями.

Семантическое метаописание – набор извлеченных 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 – в объединение R_i; O_i – в объединение C_i и E_i.

Также для ускорения актуализации метаданных алгоритмом генерируются частотные характеристики слов в документе – TF- и IDF-терминов [5].

Алгоритм формирования метаданных отдельного документа. На вход алгоритма поступает исходный текст файла, а также набор текстовых меток элементов онтологии.

Шаги алгоритма:

1. Проводится семантический анализ текста. Выходом этого шага является программная структура, содержащая всю требуемую информацию о тексте – слова с номером их начальных символов, смысловые связи между словами, обнаруженные и преобразованные в RDF-триплеты (части предложений, соответствующие одному из описанных выше фреймов). Эта программная структура приводится к семантической сети, пригодной для обработки алгоритмом.

2. Определяется количество вхождений слов в текст. При этом не учитываются так называемые стоп-слова. Стоп-словами являются предлоги, союзы и частицы. Остальные слова нормализуются, и количество вхождений определяется именно для нормы слова.

3. 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 единицы [5].

4. Производится поиск класса с наибольшим номером, слова в котором являются значимыми для текста. Все классы, следующие за ним, отсеиваются и в дальнейшей работе алгоритма не участвуют [5].

5. Выставляется первичное значение веса слов в документе, равное N_{max}/N_i, где N_{max} – количество вхождений слов первого ранга; N_i – количество вхождений слова t_i [5].

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

7. Из множества выделенных из текста RDF-триплетов выбираются:

– триплеты, каждая из позиций которых (субъект, предикат и объект) занята в естественно-языковом представлении вхождением метки (субъект и объект – метками понятия либо экземпляра, а предикат – меткой свойства);

– триплеты, одна из позиций которых занята вхождением ключевого слова, а две другие – вхождением метки (так называемые триплеты-кандидаты).

Выход алгоритма – метаописание документа, в которое входит набор записей вида $<E_i,S_i>$, где E_i – идентификатор элемента онтологии (так называемый universal resource identifier (URI)); S_i – индекс значимости этого элемента для документа. При этом S_i имеет вид $S_i = <S_iTF, S_iIDF,S_iC>$, где S_iTF – коэффициент значимости элемента с точки зрения документа (модифицированный коэффициент TF); S_iIDF – коэффициент значимости элемента с точки зрения набора документов (коэффициент IDF); S_iC – итоговое значение коэффициента значимости термина. В метаописание входят также все обнаруженные в тексте триплеты, все позиции которых заняты вхождениями меток элементов онтологии.

Кроме того, по завершении работы алгоритм генерирует набор вспомогательных записей, уменьшающих время возможной последующей повторной обработки документа.

Результаты работы алгоритма – семантические метаописания, которые позволяют реализовать семантический поиск и семантическую навигацию по обработанному множеству текстов. В зависимости от полноты онтологии предметной области и глубины анализа текста качество распознавания составляет приблизительно 60 % качества распознавания человеком.

Обнаружение RDF-триплетов в тексте. Для выявления RDF-триплетов в анализируемом тексте осуществляется преобразование документа D_i в семантическую сеть с использованием утилит семантического анализа. После преобразования к семантической сети применяется набор фреймов.

В традиционной терминологии искусственного интеллекта фреймом называется логическая схема некоторой ситуации. Фрейм имеет имя, которое идентифицирует класс описываемых им ситуаций, а также содержит слоты, которые имеют свои имена, идентифицирующие роли участников ситуации. Для конкретной ситуации, описанной в тексте, часть слотов может быть заполнена именами ее конкретных участников, упомянутых в тексте (например, "покупатель=Иванов, продавец=?, эмитент акций=Лукойл, количество акций=10 %, сумма сделки=?, дата=2010").

Модель фрейма задается множеством семантических шаблонов, каждый из которых описывает множество изоморфных семантических сетей, соответствующих некоторому типовому способу описания ситуации в тексте.

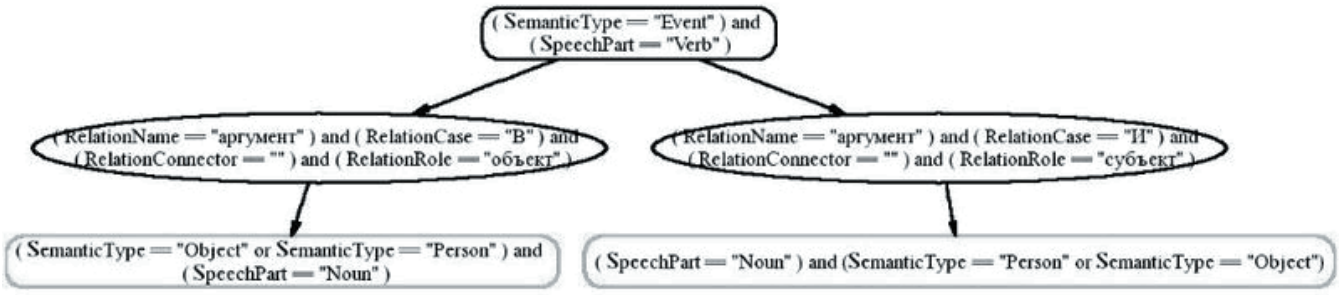

Рис. 1. Фрейм "Существительное – глагол – существительное"

Таким образом, задача непосредственного выделения RDF-триплетов из текста сводится к задаче поиска такого набора фреймов, которые описывали бы все возможные ситуации вида <субъект, предикат, объект>. При этом следует учитывать, что предикат может быть как явным (представленным, например, глаголом, явно применяемым в тексте: "Университет проводит набор"), так и неявным (например, предикат эквивалентности либо предикат вида "is_a", т. е. задающий отношение подкласс – суперкласс, может быть представлен с помощью знака препинания: "Сервер – логический или физический узел сети, обслуживающий запросы к одному адресу и (или) доменному имени").

Предлагаемый подход для текста общего вида заключается в описании минимального количества фреймов, с использованием которых можно извлечь достаточную долю триплетов. Для текста общего вида применяются три основных фрейма.

1. Существительное – глагол – существительное – естественно-языковая интерпретация "классической" формы RDF-триплета субъект – предикат – объект. Данный фрейм представлен на рис. 1. В данном случае предикатом является глагол, представляющий собой текстовую метку некоторого отношения в онтологии. Шаблон является основным в системе. Графическое представление данного фрейма показано на рис. 2.

2. Существительное – прилагательное – вспомогательный атрибутивный шаблон. Предполагается, что одна сущность является атрибутом другой сущности. Атрибут охватывает значительную часть литеральных отношений (например: "инвертер стабилен") и при этом, как правило, является экземпляром концепта-перечисления ("цвет", "надежность" и т. д.).

3. Существительное – существительное – вспомогательный шаблон, описывающий отношения эквивалентности либо генерализации. Данный фрейм представлен на рис. 3. Конструкции вида "сервер – компьютер", "Николаев – директор" подразумевают либо принадлежность некоторого объекта к классу, либо отношение подкласс – суперкласс. Этот шаблон целесообразно использовать в тех случаях, когда основной шаблон, вычленяющий значительную

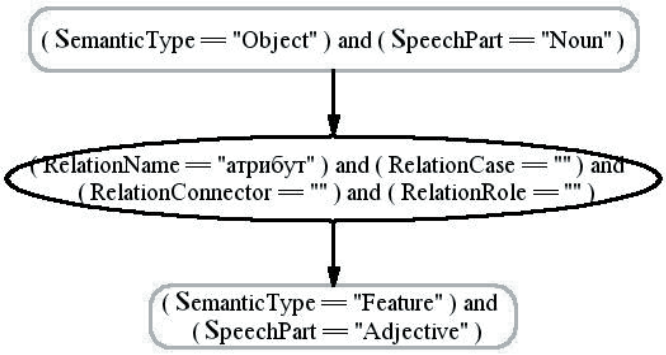

Рис. 2. Фрейм "Существительное-прилагательное"

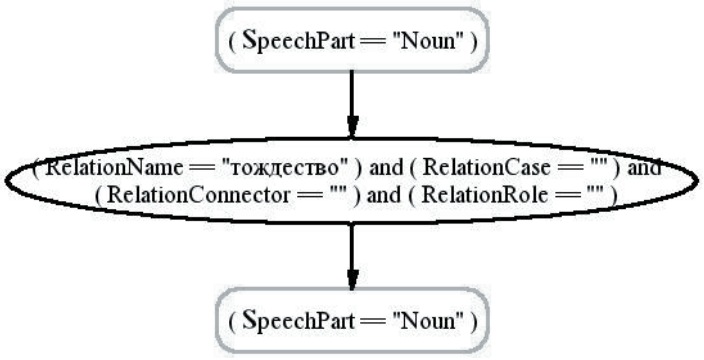

Рис. 1. Фрейм "Существительное – существительное"

часть найденных в тексте отношений эквивалентности (по меткам "быть", "являться", "представлять собой" и т. д.), не может корректно обработать отношение генерализации или эквивалентности, выраженное в тексте с помощью знака препинания.

Существует возможность сделать распознавание текста более глубоким путем введения в систему дополнительных фреймов, соответствующих стилевым особенностям текстов, характерным для конкретной предметной области. Однако вследствие существенного снижения скорости обработки текста при увеличении количества фреймов увеличение количества универсальных фреймов может быть нецелесообразным. Более того, в случае, когда производительность критична, а глубины распознавания заведомо более чем достаточно, может иметь смысл отказ от фрейма эквивалентности.

Список литературы

1. Люгер Д. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. 4-е изд. М.: Вильямс, 2003.

2. Хорошилов А. А., Белоногов Г. Г., Калинин Ю. П. Компьютерная лингвистика и перспективные информационные технологии: теория и практика // Науч.-техн. информ. Сер. 2. Информ. процессы и системы. 2004. № 8. С. 30–43.

3. Poon H., Domingos P. Unsupervised semantic parsing. ACL Anthology. A digital archive of research papers in computational linguistics. [Электрон. ресурс]. www.aclweb.org/anthology/D/D09/D09-1001.pdf.

4. Deep linguistic processing with hpsg. [Электрон. ресурс]. http://www.delph-in.net.

5. Roelleke T., Jun Wang. TF-IDF uncovered: a study of theories and probabilities // Proc. of the 31st Annual intern. ACM SIGIR conf. on research and development in information retrieval, Singapore, July 20–24, 2008. New York: SIGIR'08. ACM, 2008. P. 435–442.

Губин Максим Юрьевич – асп. Института кибернетики Томского политехнического университета; тел.: (382-2) 44-46-53, 8-906-950-73-14; e-mail: gubin.m.u@gmail.com;
Разин Владимир Викторович – асп. Института кибернетики Томского политехнического университета; тел.: (382-2) 46-52-23; e-mail: mooonbreeze@sibmail.com;
Тузовский Анатолий Федорович – д-р техн. наук, проф. Института кибернетики Томского политехнического университета; тел.: (382-2) 42-14-85; e-mail: tomo@tpu.ru

Дата поступления – 31.10.11