2. *Синицын И. Н., Новиков С. О., Ушмаев О. С.* Развитие технологий интеграции биометрической информации // Системы и средства информатики. 2004. Вып. 14. С. 5—36.

3. *Тропченко А. А., Тропченко А. Ю.* Нейросетевые методы идентификации человека по изображению лица // Изв. вузов. Приборостроение. 2012. Т. 55, № 10. С. 31—36.

4. *Ушмаев О. С.* Методы мультибиометрической идентификации. М.: Изд-во ИПИ РАН, 2009. 114 с.

5. *Dass S. C., Nandakumar K., Jain A. K.* A principled approach to score level fusion in multimodal biometric systems // Audio- and Video-Based Biometric Person Authentication. 2005. P. 1049—1058.

6. *Karam W., Bredin H., Greige H., Chollet G., Mokbel C.* Talking-face identity verification, audiovisual forgery, and robustness issues // EURASIP J. Adv. Signal Process. 2009. Vol. 4. P. 1—15.

7. *Ross A., Govindarajan R.* Feature level fusion using hand and face biometrics // Proc. of the SPIE Conf. on Biometric Technology for Human Identification. Orlando, USA, 2005. P. 196—204.

Сведения об авторе

Андрей Александрович Тропченко — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники;
E-mail: zayka_98rus@mail.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.12.13 г.

УДК 004.912

И. В. Калинин, С. В. Клименков, А. Е. Харитонова, Е. А. Цопа

ПРЕОБРАЗОВАНИЕ ЕСТЕСТВЕННОГО ЯЗЫКА В ФОРМАТ RDF С ПОМОЩЬЮ СЕМАНТИЧЕСКИХ АНАЛИЗАТОРОВ ТЕКСТОВОЙ ИНФОРМАЦИИ

Решена задача автоматического преобразования естественного языка (русского) в формат RDF средствами семантического анализа. Приведен алгоритм работы программных модулей, созданных для решения задачи.

Ключевые слова: *текст, естественный язык, RDF, семантический анализ, тезаурус, AOT, Jena.*

Введение. Консорциумом Всемирной паутины для машиночитаемого представления данных, в особенности — метаданных, была разработана модель RDF (Resource Description Framework) [1]. Одним из направлений развития сети Интернет является реализация механизмов машинной обработки информации [2].

В основе этих механизмов лежит работа с метаданными, однозначно идентифицирующими характеристики и содержание ресурсов Интернета. Обработка метаданных должна прийти на смену используемому в настоящий момент текстовому анализу документов [3].

Формирование RDF-описаний ресурсов обычно осуществляется вручную — авторами. Во многих случаях такой подход неэффективен: в частности, при необходимости формирования большого объема метаданных или при формировании содержимого ресурсов пользователями (интернет-энциклопедии, социальные сети). Таким образом, возникает потребность в автоматизации процесса формирования RDF-метаданных.

В настоящей статье рассмотрено решение задачи автоматизации преобразования текста на естественном языке в формат RDF с помощью технологий семантического анализа. Приведен пример разработки алгоритма анализа текстов технической документации на русском языке.

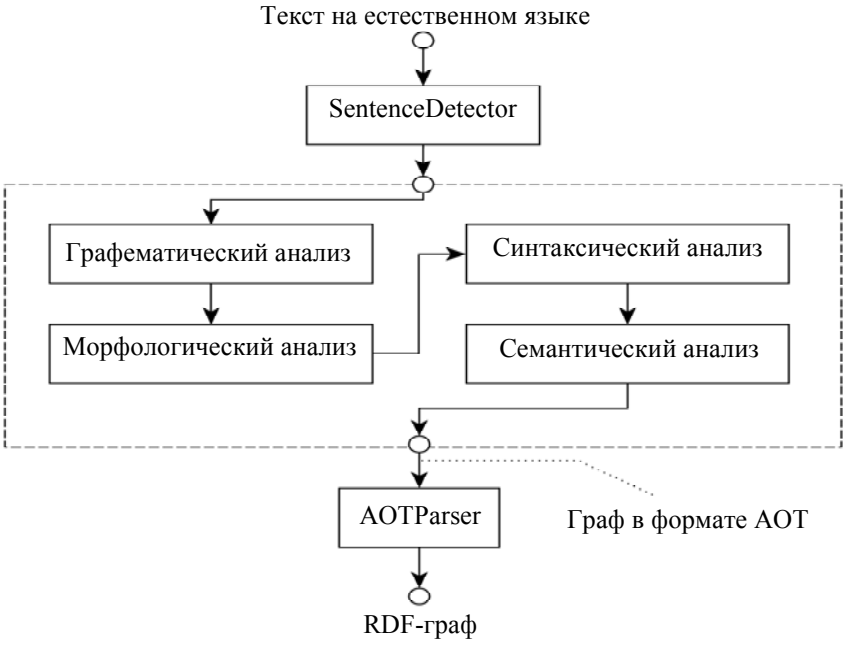
Алгоритм преобразования. Ввиду комплексности поставленной задачи в алгоритме преобразования текста в формат RDF целесообразно выделить несколько последовательных этапов.

Так как минимальной единицей, с которой работают семантические анализаторы текста, является предложение [4, 5], на первом этапе необходимо разбить массив на отдельные предложения. После этого для построения модели необходимо определить семантические отношения между словами предложения. Далее, на основании полученной модели, можно синтезировать итоговый RDF-граф.

1. Разбор текста на предложения. Наибольшую трудность на этом этапе представляет однозначное определение роли того или иного знака препинания в предложении. Так, например, точка может обозначать сокращение слов, отделять целую часть числа от дробной (в десятичных дробях), быть частью многоточия (которое не всегда обозначает конец предложения) или, например, адреса электронной почты. Особенности обработки сокращений слов и знаков препинания (определения того, какие из знаков могут обозначать конец предложения) зависят от языка анализируемого текста. Важно учитывать, что „универсального“ алгоритма разбиения текста на предложения не существует, и любой алгоритм всегда будет давать некоторый процент некорректных срабатываний [6].

Авторами статьи предложен алгоритм разбиения текста, который для определения спорных случаев использует тезаурусы с набором терминов и сокращений, характерных для конкретной предметной области. При этом реализация алгоритма позволяет динамически переключать тезаурусы в случае изменения предметной области анализируемого текста.

На основании предложенного алгоритма авторами был разработан модуль SentenceDetector (см. рисунок), который [7] показал эффективность на тестовом наборе документов (сопроводительная документация к программным проектам) — корректно определились границы примерно 95 % предложений. Такая точность определения границ является приемлемой для решения поставленной задачи.



2. Семантический анализ предложения. После того как исходный текст на естественном языке успешно разбит на отдельные предложения, необходимо выполнить его семантический анализ. Цель семантического анализа — извлечение из текста информации посредством определения и формализации смысловых связей между словами в предложении.

В настоящей работе для семантического разбора предложения была адаптирована система автоматической обработки текста (АОТ), разработанная группой „Aot.ru“ и доступная по лицензии с открытым исходным кодом. Система АОТ предназначена для анализа предложений на естественном языке (в данный момент доступны русский, английский и немецкий словари [8—10]), она обрабатывает информацию по цепочке лингвистических процессоров [11]. Вход одного процессора является выходом другого.

АОТ разбирает заданное предложение по словам, переводит слова в начальную форму и определяет семантические связи между ними [12]. Для анализа текста используются подключаемые через унифицированный интерфейс словари и тезаурусы. На базе исходных кодов АОТ авторами настоящей статьи разработан модуль семантического анализа текста. Результатом работы модуля является семантический граф, представленный в собственном формате АОТ.

3. Преобразование выдачи АОТ в формат RDF. Формат, используемый системой АОТ для представления результатов семантического анализа, несовместим с какими-либо другими инструментами обработки текста. Поэтому необходимо приводить полученные на предыдущем этапе данные к некоему стандартному виду, которым, в рамках поставленной задачи, является формат RDF.

Эту задачу позволяет решить разработанный модуль AOTParser. Преобразование выходных данных АОТ в нем реализовано по собственному алгоритму, а генерация RDF-модели осуществляется с помощью библиотеки Apache Jena [13]. Полученная RDF-модель должна быть преобразована в формат, удобный для хранения и представления семантической информации. В качестве такого формата был выбран XML.

Тестирование разработанного программного модуля. На основе предложенного алгоритма преобразования текста был разработан программный модуль, блок-схема которого приведена на рисунке. Соответствие спецификации полученного в результате преобразования формата XML было проверено с помощью W3C Validation service [14]. Проведенное тестирование показало достаточно высокое качество преобразования — более 90 % загруженного текста было корректно преобразовано в RDF-модель.

Эксперименты выявили некоторые недостатки предложенного алгоритма преобразования текста, например, частичную потерю семантической информации (выполняемое преобразование не является эквивалентным), однако применительно к поставленным задачам (формирование метаданных), это несущественно, так как выполнение обратного преобразования не предполагается.

Заключение. В настоящей статье предложен алгоритм автоматического преобразования естественного языка в формат RDF и представлена программная реализация, включающая три модуля, осуществляющих преобразование в RDF-граф технических текстов на русском языке.

Тестирование разработанных модулей показало эффективность предложенного алгоритма и доказало возможность его дальнейшего применения при решении практических задач. Важным достоинством программной реализации является возможность ее использования для автоматического преобразования в формат RDF текстов другой тематики и на других языках без необходимости модификации алгоритма — с этой целью достаточно подключить к нему дополнительные тезаурусы.

СПИСОК ЛИТЕРАТУРЫ

1. RDF Primer. W3C Recommendation [Электронный ресурс]: <http://www.w3.org/TR/rdf-primer/>.

2. *Бессмертный И. А.* Семантическая паутина и искусственный интеллект // Научно-технический вестник информационных технологий, механики и оптики. 2009. № 6 (64).

3. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American. May, 2001.

4. *Сокирко А.* Первичный семантический анализ [Электронный ресурс]: <http://aot.ru/docs/seman.html>.

5. LinkParser Grammar Tutorial [Электронный ресурс]: <http://www.link.cs.cmu.edu/link/dict/index.html>.

6. *O'Neil J.* Things with Words, Part Two: Sentence Boundary Detection [Электронный ресурс]: <http://www.attivio.com/attivio/blog/263-doing-things-with-words-part-two-sentence-boundary-detection.html>.

7. Тезаурусы ДИАЛИНГ [Электронный ресурс]: <http://www.aot.ru/docs/thes.html>.

8. *Сокирко А.* Технологии автоматической обработки текста [Электронный ресурс]: <http://www.aot.ru/technology.html>.

9. *Сокирко А.* Семантические отношения, используемые в модуле поверхностно-семантического анализа „Диалинг“ [Электронный ресурс]: <http://www.aot.ru/docs/SemRels.htm>.

10. Русский общесемантический словарь [Электронный ресурс]: <http://www.aot.ru/docs/ross.html>.

11. *Рубинштейн М. Л.* Описание английского общесемантического словаря [Электронный ресурс]: <http://www.aot.ru/docs/aoss.html>.

12. Немецкий общесемантический словарь [Электронный ресурс]: <http://sourceforge.net/p/seman/svn/HEAD/tree/trunk/Dicts/>.

13. Apache Jena RDF API Documentation [Электронный ресурс]: <http://jena.apache.org/documentation/rdf/index.html>.

14. World Wide Web Consortium. RDF Validator [Электронный ресурс]: <http://www.w3.org/RDF/Validator/>.

Сведения об авторах

Игорь Владимирович Калинин — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; E-mail: kalinin1989@mail.ru

Сергей Викторович Клименков — Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; ассистент; E-mail: Serge.Klimenkov@tune-it.ru

Анастасия Евгеньевна Харитонова — Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; ассистент; E-mail: Anastassia.Kharitonova@Elcom.SPb.ru

Евгений Алексеевич Цопа — Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; ассистент; E-mail: Evgenij.Tsopa@Elcom.SPb.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.12.13 г.