Михаил Концевой
Брестский государственный университет имени А.С. Пушкина

СЕМАНТИЧЕСКАЯ РАЗМЕТКА ЭЛЕКТРОННОГО ЖУРНАЛИСТСКОГО ТЕКСТА

В контексте бурного развития интернет-журналистики значение эффективного семантического анализа текста электронных публикаций многократно возросло, став одной из основ функционирования информационной инфраструктуры современного социума. Это обусловлено как резким ростом числа самих электронных публикаций, так и востребованностью такого анализа все большим числом информационных служб и сервисов, и прежде всего поисковых систем.

Традиционный филологический и лингвистический структурно-смысловой анализ текстов, осуществляющийся непосредственно человеком и опирающийся на его знания, жизненный опыт, интеллектуальную интуицию, творческое вдохновение, уже не адекватен информационным вызовам эпохи. Это объясняется рядом причин. Во-первых, качество и количество требующей анализа текстовой информации несоизмеримы с возможностями отдельного человека и неприемлемы в социальном плане. Во-вторых, недостижима для человека и заданная технологической необходимостью скорость данного анализа. В-третьих, достаточный уровень эффективности такого анализа для подавляющего большинства задач не требует участия человеческого разума. Значительная часть текстовой информации в компьютерных сетях автоматически генериру-

ется программами и предназначена для программного считывания, что предполагает использование программных алгоритмов для их анализа. Полем деятельности человеческого разума, освобожденного от решения алгоритмизируемых задач, остается недоступная компьютерам художественная и научная аналитика шедевров культурного наследия в области словесности и актуального вербального творчества журналиста.

Интеллектуальным ответом на новые вызовы в области структурно-семантического анализа стала концепция Big Data (серия подходов, инструментов и методов обработки неструктурированных данных огромных объёмов и значительного многообразия для получения человекочитаемых результатов, эффективных в условиях их непрерывного прироста и распределения по узлам вычислительной сети) [1].

В журналистике появилось новое направление – журналистика данных (datajournalism), в которой журналисты призваны эффективно анализировать и использовать открывшиеся грандиозные информационные ресурсы в своей работе по донесению до широкой общественности содержащихся в них знаний. С точки зрения журналистики данных, вся информация Интернета может быть представлена как машиночитаемые данные, которые, как правило, являются открытыми (общедоступными), но требуют отбора, фильрации, структурирования в базы по заданным признакам и автоматического анализа. Популярный источник данных – это социальные сети, которые собирают информацию о своих многочисленных пользователях, и эти данные в анонимизированном виде вполне доступны при использовании специальных скриптов, позволяющих собирать их с сайтов (web-scraping). Данные могут быть самыми разными – начиная с того, сколько пользователей у сети в той или иной стране, и заканчивая частотой употребления каких-нибудь слов [7]. Журналистика данных основана на убежденности в том, что если к данным правильно применить технологии, то из них можно извлечь и донести до пользователя новое жизненно важное для него знание, получить которое нельзя никакими иными способами. Для работы с большими массивами данных журналисты должны располагать определенными компетенциями и владеть современным информационно-технологическим инструментарием, одним из которых является семантическая разметка текста.

Для компьютерного анализа текста особое значение имеет собственно лингвистическая разметка, которая заключается в приписывании текстам (их компонентам) специальных меток, обеспечивающих возможность автоматически идентифицировать тексты по различным параметрам, осуществлять их синтаксический и семантический анализы.

Полная лингвистическая разметка включает в себя следующие типы: морфологическую (выделение аффиксов, сложных слов и т.п.); лемматизацию (указание для каждой словоформы из текста ее исходной формы); морфолого-синтаксическую (выделение основ, определение части речи и признаков грамматических категорий); синтаксическую (синтаксические связи, типы и члены предложений и т.п.); семантическую (снятие семантической омонимии, разрешение анафоры и кореферентности, фиксирование информационной структуры и т.п.); дискурсивную (реплики, коммуникативные акты и т.п.).

Комплексный подход к лингвистической XML-разметки реализуется в рамках проекта TEI (Text Encoding Initiative) [4]. Другой подход, основан на понятии микроформатов и позволяет создавать свои собственные спецификации к семантической разметке. В 2011 году создатели крупнейших поисковых систем объединились в проекте Schema.org [3] – инициативе по разработке единой схемы для семантической разметки на основе эффективной структуризации поставляемых информационных ресурсов и их семантической разметки микроформатами. Метаданные на ресурсах, использующие предлагаемые Schema.org схемы, представляют собой семантическую разметку, предназначенную для поисковых роботов, и могут быть непосредственно проанализированы ими с целью извлечения и обработки информации о содержимом веб-ресурсов. Таким образом, Schema.org открывает новое направление в контексте становления Semantic Web. В качестве основного формата разметки веб-страницы метаданными Schema.org предлагаются microdata (микроданные) – теги и атрибуты для разметки структурированной информации на веб-страницах. Помимо этого формата, существует возможность применения онтологии schema.org, выраженной в формате RDFS при разметке RDF-данных.

Микроформаты – это сущности поверх HTML, с помощью которых можно описывать любую информацию на Web-страницах. Спецификация микроформатов представляет собой способ разметки содержания для определения таких специальных типов информации, как отзывы, информация о человеке, мероприятии. Стандарт представляет собой набор классов, описывающих всевозможные сущности и их свойства. Сейчас их уже несколько сотен [6].

Наиболее обобщенный тип сущности – это Thing (нечто), у которого есть четыре свойства: name (название), description (описание), url (ссылка) и image. Более специализированные, частные типы имеют общие свойства с более универсальными. Например, Place (место) — частный

случай Thing, а LocalBusiness (местная компания) — частный случай Place. Частные типы наследуют свойства родительских типов, которых может быть несколько. В качестве примеров популярных типов сущностей отметим: CreativeWork (творческое произведение), Book (книга), Movie (фильм), MusicRecording (музыкальная запись), Recipe (рецепт), TVSeries (телесериал), AudioObject (аудио), ImageObject (изображение), VideoObject (видео), Event (событие), Organization (организация), Person (человек), Place (место), LocalBusiness (местная фирма), Product (продукт), Offer (предложение), Review (отзыв), AggregateRating (сводный рейтинг). В настоящее время поисковые системы корректно поддерживают микроформатную разметку веб-страниц в результатах поиска людей, событий, обзоров, товаров, кулинарных рецептов и элементов навигации. Вместе с тем стандарт schema.org предусматривает возможность добавлять свойства и дочерние типы для имеющихся типов сущностей. Разметка микроформатами не требует создания отдельных экспортных файлов и происходит непосредственно в HTML-коде страниц оборачиванием описания определенного типа в контейнер и указанием схемы разметки отдельных свойств с помощью специальных атрибутов. Каждый тип информации описывает определенный тип элемента (субъект, событие, отзыв). Например, человек имеет такие свойства, как имя, место жительства, место работы, занимаемая должность и т.д.

Существенно, что разметку Schema.org можно использовать на веб-страницах на любом языке. Код микроформатов прост для написания в любом текстовом редакторе, но лучше воспользоваться специальными программами, которые позволяют добавлять микроформатированный контент в создаваемые с их помощью ресурсы. Существует несколько специализированных сервисов, с помощью которых можно проверить корректность разметки и выявить возможные ошибки. Для проверки корректности формата данных, размеченных с помощью схем, полезно использовать инструменты Google Rich Snippets Validator [2] и валидатор от Яндекса [5], которые позволяют не только выяснить, есть ли в коде разметки ошибки, которые могут помешать корректной обработке данных, но и проверить, как поисковые роботы данных систем видят и обрабатывают предложенную семантическую разметку страницы.

Микроформаты – полностью открытый формат. Следовательно, данные, размеченные по стандарту семантической разметки schema.org, становятся общедоступными и могут быть извлечены и использованы любыми сервисами. Успех новой поисковой технологии зависит от того, насколько широкое она получит распространение, но уже сейчас применение микроформатов создает качественно новую среду для автоматического анализа текстовых ресурсов Интернета на основе их семантической разметки.

Литература

1. Big data [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/Big_data. – Дата доступа: 23.02.2014.

2. Google Rich Snippets Validator [Электронный ресурс]. – Режим доступа: http://www.google.com/webmasters/tools/richsnippets. – Дата доступа: 10.01.2014.

3. Schema.org [Электронный ресурс]. – Режим доступа: http://schema.org/. – Дата доступа: 12.01.2013.

4. Text Encoding Initiative [Электронный ресурс]. – Режим доступа: http:// www.tei-c.org – Дата доступа: 25.02.2014.

5. Валидатор от Яндекса [Электронный ресурс]. – Режим доступа: http:// webmaster.yandex.ru/microtest.xml – Дата доступа: 09.01.2014.

6. Расширенные описания веб-страниц (микроданные, микроформаты RDFa) [Электронный ресурс]. – Режим доступа: http://support.google.com/webmasters/bin/topic.py?hl=ru&topic=21997. – Дата доступа: 12.01.2014.

7. Саакян, А. Данные для журналистов / А. Саакян [Электронный ресурс]. – http://polit.ru/article/2013/05/04/data_journalism/. – Дата доступа: 09.01.2014.