УДК 004.942

Микроформатное моделирование лингвистических метаданных

М.П. Концевой

Брестский государственный университет имени А.С. Пушкина

В моделировании метаданных для автоматического лингвистического анализа электронного текста эффективным инструментом является язык разметки и набор инструкций (управляющих дескрипторов, тегов), моделирующих в документах структуры токенизации, лемматизации, морфологической категоризации и др., и определения отношений между их элементами. Традиционный подход к лингвистической разметке реализуется в рамках проекта TEI (Text Encoding Initiative) и стандарта XCES (Corpus Encoding Standard for XML).

Преимуществами разметки в стандарте TEI являются ее полнота, легкость адаптации к конкретной семантике, независимость от платформы. Вместе с тем избыточность TEI для решения практических задач создает существенные трудности для разработки программных средств обработки, анализа и публикации текстов, размеченных в этом стандарте. Глубокую филологическую разметку, учитывающую разночтения и варианты интерпретации фрагментов текста на разных уровнях иерархии языковых структур, очень трудно совместить с использованием инструментов автоматической лингвистической разметки.

Снять эти ограничения позволяет разметка на основе микроформатов – микроданных поверх HTML, с помощью которых можно описывать любую информацию. Микроформаты применяются исключительно для обозначения неязыковых сведений о данных. Предлагается на основе микроформатов определить новую онтологическую сущность для описания собственно лингвистической составляющей электронного текста.

В качестве такого микроформата предлагается основанная на стандарте Sema.org спецификация XLD (XHTML Linguistic Data) – микроформат для пометки лингвистических метаданных. XLD можно использовать как на лингвистических интернет-ресурсах, так и для разметки любого текстового контента. Необходимо сообщить браузерам и поисковикам, что страница поддерживает XLD. Для этого в теге <head> веб-страницы надо добавить атрибут profile:

<head profile="http://gmpg.org/xld/17">.

Для каждой гиперссылки на странице нужно добавить атрибут rel.

Пример:

<a href="http://brsu.by" rel="text thesis"> ... </a>.

Значений атрибута rel может быть несколько, в таком случае они перечисляются через пробел.

Парсинг микроформатов имеет особенности в сравнении с XML-парсингом. Причина этих особенностей в том, что микроформаты рассчитаны на работу с HTML, для которого не обязательно документ должен быть корректно сформирован (допускается отсутствие закрывающих тегов и т.п.). Если использовать для извлечения микроформатов XSLT, то необходимо, прежде всего, откорректировать HTML-документ с помощью TIDY, HTMLlib или loadHTML, после чего можно использовать шаблоны XSLT, выполняющие большинство работы по извлечению микроформатных данных (доступны на hg.microformats.org).

Микроформаты – полностью открытый формат. Следовательно, лингвистические данные, размеченные по стандарту семантической разметки schema.org, становятся общедоступными и могут быть извлечены и использованы любыми сервисами.

Литература

1. *Концевой М.П.* Микроформатное моделирование лингвистических данных / М.П. Концевой // Вычислительные методы, модели и образовательные технологии : сб. материалов VI I междунар. науч.-практ. конф., Брест, 18 окт. 2019 г. / Брест. гос. ун-т им. А.С. Пушкина ; под общ. ред. А.А.Козинского. – Брест : БрГУ, 2019. – 272 c– С. 179–180.