УДК 004.853

СЕГМЕНТАЦИЯ СОДЕРЖАНИЯ ОБУЧАЮЩЕЙ СИСТЕМЫ ДЛЯ ГЕНЕРАЦИИ СЕМАНТИЧЕСКОЙ СЕТИ

Н.И. Гурин, Я.А. Жук

Статья посвящена проблеме сегментации содержания обучающей системы по технической дисциплине, возникающей вследствие наличия в ней многочисленных нетекстовых элементов. В качестве решения данной проблемы предлагается использование формата HTML для представления содержания информационной системы с последующим анализом групп тегов, соответствующих нетекстовым элементам, как неделимых информационных единиц, для включения их в базу знаний.

Введение

При построении диалоговой информационной системы важнейшим этапом является построение для нее базы знаний, содержащей понятия предметной области и связи между ними для построения семантической сети [1]. От объема построенной базы знаний во многом и зависит полезность диалога с информационной системой. Однако наполнение базы знаний является весьма трудоемким процессом, если его выполнять вручную. Автоматизировать данный процесс возможно путем автоматического извлечения знаний из текстов по заданной предметной области. Одним из начальных этапов семантического анализа исходного текста является морфологический анализ слов, т. е. определение части речи слова и его формы в предложении [2]. Данная задача возникает как при разбиении сложных предложений на простые, так и при семантическом анализе простых предложений. В целом, морфологический анализ слов исходного текста является важным этапом для безошибочного наполнения базы знаний и создания на ее основе семантической сети ключевых понятий диалоговой информационной системы.

Наполнение семантической сети знаниями по предметной области образовательной системы в значительной мере затрудняется наличием в содержании нетекстовых элементов: формул, ссылок, схем, иллюстраций, таблиц, интерактивных тренажеров и др. Начальным этапом автоматической обработки содержания информационных систем является его сегментация, т. е. формирование списка слов [3]. Уже на данном этапе возникает необходимость корректного анализа перечисленных нетекстовых элементов во избежание их разрыва на несколько слов.

Представление и сегментация содержания информационной системы

В качестве решения данной проблемы предлагается использование формата HTML для представления содержания информационной системы с последующим анализом групп тегов, соответствующих нетекстовым элементам. Язык разметки HTML широко используется для реализации обучающих информационных систем. Он позволяет интегрировать в страницы таких систем различные виды содержания. Кроме того, следует отметить, что

многие другие форматы хранения информационных систем могут быть преобразованы в формат HTML без значительных временных затрат и потерь качества.

Таким образом, входными данными для генератора семантической сети являются файлы в формате HTML. Для выявления названий тегов и применяемых классов в генераторе семантической сети реализован алгоритм, добавляющий пробелы перед и после знаков препинания, включая кавычки и скобки, используемые в HTML-тегах и их атрибутах. Затем генератор семантической сети выполняет создание списка подстрок, разделенных пробелами, при помощи функции splice. Полученный список выступает в роли исходных данных для следующего этапа, на котором необходимые HTML-теги объединяются в неделимые слова, а лишние удаляются.

По результатам изучения ряда обучающих информационных систем был составлен перечень несущих смысл нетекстовых элементов содержания и соответствующих им HTML-тегов, который представлен в таблице 1.

Таблица 1

Нетекстовые элементы и соответствующие им HTML-теги

Элемент	Тег
Изображения (в т. ч. формулы)	img
Ссылки	a
Таблицы	table
Видео	video, object, embed
Аудио	audio, object, embed
Нижние и верхние индексы	sub, sup, span
Интерактивные Flash-элементы	object, embed

Перечисленные теги и стили занесены в специальный список генератора семантической сети под названием *goodtags*. При обнаружении любого из перечисленных тегов или стилей выполняется поиск соответствующего закрывающего тега, и фрагмент между открывающим и закрывающим тегами заносится в результирующий список псевдослов как одно псевдослово. Описанный алгоритм выполняется фрагментом кода генератора семантической сети на языке программирования Python, который представлен на рисунке 1.

```
if words[w_index+1] in goodtags:    #найден тег с нетекстовой информацией
    foundtag=words[w_index+1]
    start_index=w_index
    w_index=w_index+2
    # поиск закрывающего тега
    while (not (words[w_index]==">" and words[w_index-1]==foundtag and
                words[w_index-2]=="/" and words[w_index-3]=="<"))
           and (not (words[w_index]==">" and words[w_index-1]=="/")):
        w_index=w_index+1
    # запись всей строки от открывающего до закрывающего тега как
    # одного псевдослова
    words = words[0:start_index]+[" ".join(words[start_index:w_index])]+
            words[w_index+1:]
    w_index=start_index
```

Рис.1. Листинг сегментации HTML-кода

Для корректной обработки полученного списка на следующем этапе морфологического анализа в таблицу неизменяемых слов были внесены регулярные выражения, содержащие начала перечисленных HTML-тегов и оканчивающиеся символом произвольной комбинации знаков. Для данных записей был создан специальный разряд слов.

Заключение

Для автоматической обработки содержания информационных систем технической направленности целесообразно конвертировать их в формат HTML. Этот формат позволяет проводить сегментацию содержания информационных систем по HTML-коду (теги, стили, идентификаторы) и затем адаптировать ее элементы к построению базы знаний обучающей системы.

Список литературы

1. Гурин Н. И., Жук Я. А. Генератор семантической сети информационной системы в таблицу реляционной базы данных // Труды БГТУ. 2015. №6: Физ.-мат. науки и информатика. С. 181–185.
2. Гурин Н. И., Жук Я. А. Морфологический анализ текста для генерации базы знаний диалоговой информационной системы // Труды БГТУ. 2016. №6: Физ.-мат. науки и информатика (в печати).
3. Бочаров В. В. и др. Сегментация текста в проекте «Открытый корпус» // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 1: Основная программа конференции. М.: – Изд-во РГГУ, 2012. – С. 51–60.

Гурин Николай Иванович, доцент кафедры информационных систем и технологий Белорусского государственного технологического университета, кандидат физико-математических наук, доцент, ngourine@mail.ru

Жук Ярослав Александрович, аспирант кафедры информационных систем и технологий Белорусского государственного технологического университета, root@belstu.by