В.Б. Мордухович (Минск, МГЛУ)

ЛОГИКО-СЕМАНТИЧЕСКАЯ ОРГАНИЗАЦИЯ ТЕКСТА ГЛАВНОЙ СТРАНИЦЫ АНГЛОЯЗЫЧНОГО ВЕБ-САЙТА

Общеизвестно, что лексический состав текста (слова и сочетания слов) отражает его тематику. Если несколько изменить формулировку Е. Агриколы [1, с. 13], под темой текста будем по-нимать его понятийное ядро, передающее содержание текста в сжатом и абстрактном виде, в форме смыслового комплекса, выраженного словесными средствами и состоящего из имен героев (или предметов) и предикатов. Другими словами, тема — это совокупность определенных слов и словосочетаний, отражающая наиболее существенные составляющие описываемой в тексте ситуации и наиболее важные отношения между ними. Таким образом, в качестве основного носителя семантики текста можно рассматривать слово. Слова, выражающие наиболее важные для текста понятия, как правило, употребляются в тексте неоднократно. Они несут основную семантическую нагрузку, обеспечивают становление и развитие темы, связывают отдельные текстовые структуры, создавая тем самым целостность текста. Поэтому их называют ключевыми или опорными словами. Для выбора из текста ключевых слов чаще всего используются статистические, позиционные и лингвосемантические методы [2].

Исходя из цели данного исследования, рассмотрим более подробно статистические методы. Они основаны на использовании статистических параметров для оценки информативности различных элементов текста и учитывают, прежде всего, показатель частоты встречаемости слов в тексте. В результате ранжирования лексики того или иного документа определяются слова с высоким рангом и их сочетаемость в различных фразах [3, с. 144]. Именно статистические методы используются в наши дни в промышленных компьютерных системах для решения таких задач, как аннотирование и реферирование текстов, их тематическая классификация и кластеризация, порождение и понимание, смысловой поиск и т. д., которые можно отнести к проблеме тематического анализа. Статистическая информация об отдельных лексических единицах легко извлекается компьютером из текста, и есть все основания полагать, что она адекватно отражает его тему (основное содержание).

В настоящее время существует целая группа статистических методов. Так, в одном из них [4, с. 125] процедура выделения ключевых слов текста сводится к следующему. Сначала компьютер разбивает текст на абзацы и составляет по каждому абзацу алфавитно-частотный словарь словоформ. Далее алфавитно-частотные словари абзацев объединяются в распределительный алфавитно-частотный словарь словоформ всего текста. В нем указывается общая частота употребления словоформы в тексте и число абзацев, в которых она встретилась. Затем производится сокращение распределительного алфавитно-частотного словаря словоформ текста до словаря потенциальных ключевых (опорных) слов. Эта процедура включает в себя следующие операции:

а) из распределительного словаря удаляется вся служебная и общеупотребительная лексика;

б) в оставшейся части словаря суммируются частоты всех грамматических форм одного и того же слова;

в) в этом же словаре суммируются частоты синонимов (в том числе и контекстуальных);

г) из оставшейся части распределительного словаря удаляются слова, которые встретились только в одном абзаце.

Наконец, на последнем этапе словарь потенциальных опорных слов текста делится на две части. В первую часть входят главные опорные слова (ГОС), а во вторую — второстепенные опорные слова (ВОС). Данная процедура осуществляется с учетом коэффициента важности слова, который вычисляется по формуле $\mathbf{K_в=F*m/N*n}$, где **F** — абсолютная частота употребления слова; **m** — количество абзацев, в которых встретилось слово; **N** — общее число словоупотреблений в тексте; **n** — общее число абзацев в тексте. Предварительно, в зависимости от длины текста (в словах и абзацах) по специальной формуле определяется средняя (пороговая) величина коэффициента важности [5, с. 45]. Коэффициент важности каждого слова словаря сравнивается с этой величиной. Если коэффициент важности слова выше пороговой величины, то оно относится к числу главных опорных слов, если меньше пороговой величины — к числу второсте-пенных опорных слов текста. Слова, входящие в каждую из выделенных групп, не однородны по своему содержанию. Они обозначают основные компоненты описанной в тексте ситуации: адресата и адресанта (субъекты), совершаемые ими действия в определенное время и в определенном месте, а также объекты действительности, о которых идет речь в тексте.

Для адекватного формализованного представления темы текста необходимо оперировать не отдельными словами, а определенным набором ключевых слов, который целесообразно представить в виде таблицы основного статического содержания текста (ТОСС). В данном исследовании на основе статистического метода были получены ТОСС двадцати текстов главных страниц англоязычных веб-сайтов спортивной тематики. В качестве примера приведем наполнение таблицы основного статического содержания текста главной страницы веб-сайта *prodirectsoccer.com*: главные опорные слова — *football, boots, adidas, nike, range, puma, clothing, pro-direct, include, worn, new, latest*; второсте-

пенные опорные слова — *umbro, brands, soccer, service, shirt, gloves, kits, predator, footballs, balls, shorts, available, goalkeeper*.

При моделировании процесса порождения текста, нужно учитывать два аспекта темы текста: статический и динамический. Статический аспект темы текста формально представлен с помощью главных и второстепенных опорных слов, которые обеспечивают становление и развитие темы текста. Поскольку порождение текста носит процессуальный характер, тема текста имеет и ярко выраженный динамический аспект.

Рассматривая тему текста в ее динамическом развитии, необходимо обратиться к такому понятию, как микротема. В работах А.И. Новикова и Н.А. Купиной отмечается, что тема текста есть продукт синтеза микротем, каждая из которых представляет собой тему абзаца [6; 7]. По мнению О.М. Скирдач, «процесс раскрытия темы текста может быть представлен в виде иерархии микротем, последовательное появление которых в тексте дополняет, насыщает тему текста, позволяет более достоверно ее раскрыть… Сложное взаимодействие микротем текста между собой, последовательный переход от одной микротемы к другой создает динамическое развитие текста и обеспечивает его целостное единство» [8, с. 125]. Таким образом, динамическое развитие темы текста является «его важным признаком, то есть динамику микротемы текста следует рассматривать как способ существования самого текста» [8, с. 126].

Поскольку порождение текста носит процессуальный характер, динамический аспект его содержания можно описать в виде логико-семантических формул, отражающих порядок следования представленных в тексте микротем абзацев. В данной работе были выделены конкретные микротемы всех абзацев двадцати текстов главных страниц англоязычных веб-сайтов спортивной тематики. Каждой из них был присвоен свой уникальный код. В табл. 1 представлено 14 типов абзацев текста главной страницы веб-сайта *prodirectsoccer.com*.

Таблица 1

Типы абзацев текста главной страницы веб-сайта *prodirectsoccer.com*

№ п/п	Код микротемы	Обобщенное предметно-логическое содержание абзаца
1.	M001	*Items available*
2.	M002	*Football boots*
3.	M009	*Football boots personalization*
4.	M005	*Football shirts*
5.	M003	*Goalkeeper accessories*
6.	M010	*Referee accessories*
7.	M004	*Footballs*
8.	M008	*Football kits*
9.	M017	*Football shirts personalization*
10.	M007	*Training wear*
11.	M012	*Discounts*
12.	M013	*Level of service*
13.	M014	*Visit of the site*
14.	M015	*Site updates*

Логико-семантическая формула данного текста может быть записана следующим образом: T= M001 & M002 & M009 & M005 & M005 & M005 & M005 & M003 & M003 & M010 & M004 & M008 & M017 & M007 & M012 & M013 & M013 & M013 & M014 & M015.

Логико-семантические формулы двадцати текстов главных страниц англоязычных веб-сайтов спортивной тематики вошли в лингвистическую базу знаний системы автоматического порождения текста главной страницы сайта.

ЛИТЕРАТУРА

1. Agricola, E. Vom Text zum Thema. — Berlin, 1976.

2. Горняк, Л.В. Тематический фильтр текстов / Искусственный интеллект. — Донецк, 2004. — № 4. — С. 580–586.

3. Блюменау, Д.И. Информационный анализ и синтез для формирования вторичных документов / учеб.-практич. пособ. — СПб., 2002.

4. Зубов, А.В. Основы лингвистической информатики: Компью-терная лингвистика. В 3-х частях. — Часть 2. — Минск, 1992.

5. Зубов, А.В. Вероятностно-алгоритмическая модель порождения текста / Проблемы порождения текста. В 4-х частях. — Часть 2.– Минск, 1989.

6. Новиков, А.И. Семантика текста и ее формализация. — М., 1983.

7. Купина, Н.А. Опыт системно-синтаксического анализа семантики связного текста / Семантика и структура предложения. Лексическая и синтаксическая семантика. — Уфа, 1978. — С. 137–143.

8. Скирдач, О.М. Тема текста и ключевые слова / Коммуникативно-функциональная типология текстов: сб. науч. тр. Моск. ордена Дружбы народов гос. лингв. ун-та. Вып. 381. — М., 1991. — С. 124–129.