А. В. Скрябина

ОПРЕДЕЛЕНИЕ ОСНОВНОГО СОДЕРЖАНИЯ АНГЛОЯЗЫЧНЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

В настоящее время тексты на естественном языке (ЕЯ) являются основным способом хранения и передачи знаний. В связи с усиливающейся тенденцией к хранению текстов в цифровом виде и с быстрым ростом объёма текстовой информации актуальной является проблема автоматизации обработки подобной информации, в частности проблема машинного анализа текста [1].

Задаче компьютерного анализа текста на естественном языке посвящено множество теоретических и практических работ. Системы анализа текста решают сегодня целый ряд задач, такие, как формирование информационного портрета текста в терминах ключевых понятий, выявление смысловых связей между понятиями, автоматическое реферирование. Доступные сегодня вычислительные мощности позволили применить широкий класс математических методов анализа неструктурированных данных для обработки больших массивов документов, эффективно решая задачи поиска информации, классификации, кластерного анализа, выявления скрытых закономерностей и др. К сожалению, внедрение математических методов в обработку текста проходит на фоне некоторой неразработанности собственно лингвистической составляющей алгоритмов, что не позволяет достичь высокого качества работы прикладных систем [2].

В отличие от смысла, содержание у текста может быть только одно, оно постоянно и неизменно. Ввиду этого факта, представляется возможным определить основное содержание текста с помощью методов прикладной лингвистики. На сегодняшний день компьютер в состоянии определить основное содержание любого текста, хранящегося в его памяти. Что же касается его смысла, то тут проблема состоит в множественности смыслов одного текста.

Для определения основного содержания текста существует несколько критериев [3]:

- позиционный (критерий положения слова в тексте);
- статистический (критерий выделения «опорных слов»);
- логико-семантический.

В основе статистического метода лежит критерий выделения ключевых или опорных слов, которые пронизывают текст и влияют на понимание всего текста. В структуре «опорных слов» или «смысловых вех» существует определённая иерархия. Часть из них является «главными», «важными словами», «номенклатурными дескрипторами», «выступающими точками» всего предложения. Они определяют главный предмет сообщения и по существу явля-

ются свёрнутым замыслом текста. При этом один и тот же денотат может выражаться отдельными словами – контекстуальными синонимами – в тех пределах, при которых возможно их взаимное приравнивание. Ввиду линейности означающего текста, такие слова оказываются разнесёнными по разным абзацам текста.

Основным недостатком подобных систем является то, что они работают со смыслами отдельных слов и словосочетаний, реже – со связями между ключевыми словами, но структура предложения целиком не анализируется. Это можно определить как недостаток, потому что часто смысл слов зависит от контекста. В настоящее время не существует (или находятся в экспериментальной стадии) автоматизированных систем, анализирующих семантическую структуру текста на уровне выше предложения.

В настоящее время ни одна из теорий не может претендовать на полноту, хотя наиболее совершенные теории достигли удовлетворительных теоретических результатов. Подобные теории используют предложение как структурную единицу и имеют слабые средства представления связного текста. Кроме того, все они требуют дополнительной формализации для их реализации на компьютере.

Исходя из этого, задача компьютерного анализа и определения основного содержания текста на естественном языке остается одной из ведущих тем многочисленных теоретических и практических исследований в области компьютерной лингвистики. В нашей работе мы используем статистический метод для определения основного содержания публицистического текста.

В нашей работе материалом исследования послужили англоязычные новостные публицистические тексты по теме культура, взятые с Интернет-ресурсов www.bbc.com и www.theguardian.com. Для определения основного содержания каждого из текстов мы использовали статистический метод. В основе этого метода лежит критерий выделения ключевых или опорных слов, которые пронизывают текст и влияют на понимание всего текста.

Процесс выделения опорных слов каждого анализируемого текста включает в себя следующие этапы.

1. Построение частотно-алфавитного словаря для каждого анализируемого текста. В частотно-алфавитном словаре для каждого слова текста указывается, сколько раз оно встретилось в данном тексте, в каком количестве абзацев, в каких абзацах анализируемого текста, а также сколько раз в каждом из данных абзацев. В данной работе мы использовали программу «**DICT**», чтобы получить частотно-алфавитные словари по анализируемому материалу. Для начала, мы привели отобранные тексты к виду, соответствующему требованиям, предъявляемым данной компьютерной программой: обозначили начало каждого абзаца символом «*», элиминировали переносы, преобразовали строчные символы в прописные символы.

2. Выделение из частотно-алфавитного словаря потенциальных опорных слов. Далее они служат основой для деления опорных слов на главные и второстепенные, которые и определяют, в конечном счёте, основное содержание анализируемого текста. Для осуществления данной операции мы использовали компьютерную программу «**UNIFY**», которая выделяет из частотно-алфавитного словаря слова и словосочетания, имеющие частоту два и более.

Данная операция заключалась в выполнении следующей последовательности действий:

a) удаление из полученных частотно-алфавитных словарей всех служебных слов (предлогов, артиклей, местоимений и так далее);

b) объединение всех грамматических форм одного слова в начальную, словарную форму, т.е. все личные формы глаголов сводились к инфинитиву, а существительные – к форме единственного числа именительного падежа;

c) пополнение базы данных потенциальных опорных слов словами схожими по тематике;

d) указание в базе данных возможных словоформ опорных слов.

С учетом того, что английский язык – аналитический язык, является целесообразным построить словарь словоформ для всех входящих в базу данных опорных слов.

Рассмотрим ход выполнения данных операций на примере одного англоязычного публицистического текста.

Итак, для текста «Andrew Garfield and Michael Shannon to team up for 99 Homes movie» с помощью программы «**DICT**» был построен частотно-алфавитный словарь, который показал, что текст состоит из 6 абзацев и содержит 247 слов. Для каждого слова указана частота его встречаемости в тексте и в каждом абзаце. Таким образом, в данном тексте имеются слова с частотой от 1 до 9. С помощью программы «**UNIFY**» данный частотно-алфавитный словарь был сведён к потенциальному словарю опорных слов, при этом из частотно-алфавитного словаря были удалены следующие служебные слова: артикли – *the, a, an,* местоимения – *he, you, him,* союз *and,* предлоги – *of, about, with, in, to,* вспомогательные глаголы *is, was,* имена собственные *Garfield, Michael Shannon, General Zod,* числительные *30, 2*. Все глагольные формы были приведены к форме инфинитива, существительные – к форме единственного числа именительного падежа. В итоге, полученные опорные слова были сведены в таблицы: «Главные опорные слова к тексту «Andrew Garfield and Michael Shannon to team up for 99 Homes movie» (табл. 1), «Второстепенные опорные слова к тексту Andrew Garfield and Michael Shannon to team up for 99 Homes movie» (табл. 2).

Таблица 1

Главные опорные слова к тексту «Andrew Garfield and Michael Shannon to team up for 99 Homes movie»

Потенциальное опорное слово	F-Частота слова в тексте	m-число абзацев, в которых встретилось слово	0 АБЗАЦ	1 АБЗАЦ	2 АБЗАЦ	3 АБЗАЦ	4 АБЗАЦ	5 АБЗАЦ	6 АБЗАЦ
DRAMA	4	4	1	1	1	0	0	1	0
FILM	4	2	0	0	1	0	3	0	0
PLAY	3	3	1	0	1	0	1	0	0
STAR	3	3	1	1	1	0	0	0	0
TAKE	3	3	0	0	1	1	0	1	0
WORK	2	2	1	1	0	0	0	0	0

Таблица 2

Второстепенные опорные слова к тексту «Andrew Garfield and Michael Shannon to team up for 99 Homes movie»

Потенциальное опорное слово	F-Частота слова в тексте	m-число абзацев, в которых встретилось слово	0 А Б З А Ц	1 А Б З А Ц	2 А Б З А Ц	3 А Б З А Ц	4 А Б З А Ц	5 А Б З А Ц	6 А Б З А Ц
CONTRACTOR	2	2	1	1	0	0	0	0	0
MAN	2	2	1	0	1	0	0	0	0
EVICT	2	2	1	1	0	0	0	0	0
DEADLINE	2	2	0	1	0	1	0	0	0
FINANCIAL	2	2	1	0	1	0	0	0	0
CRISIS	2	2	1	0	1	0	0	0	0
OPPORTUNITY	2	2	0	0	0	1	1	0	0
AMAZING	2	2	0	0	0	0	1	1	0
HOME	2	2	0	1	0	1	0	0	0

ЛИТЕРАТУРА

1. *Захаров, В. П.* Информационные системы (документальный поиск) / В. П. Захаров. – Санкт-Петербург, 2002. – 71 с.

2. *Шемакин, Ю. И.* Тезаурус в автоматизированных системах управления и обработки информации / Ю. И. Шемакин. – М., 2006. – 266 с.

3. *Зубов, А. В.* Статистический аспект содержания текста и его формальное представление / А. В. Зубов // Квантитативная лингвистика и автоматический анализ текстов: уч. зап. Тартуск. гос. ун-та. – Тарту, 1986. – № 745. – С. 75–91.