

А.А. БРИТИКОВ
**АВТОМАТИЗАЦИЯ СИНТЕЗА КОРОТКИХ ТЕКСТОВЫХ
СООБЩЕНИЙ С ЗАДАННОЙ СЕМАНТИЧЕСКОЙ И
ЭМОЦИОНАЛЬНОЙ НАПРАВЛЕННОСТЬЮ**

Бритиков А.А. Автоматизация синтеза коротких текстовых сообщений с заданной семантической и эмоциональной направленностью.

Аннотация. В статье рассматриваются вопросы, посвященные созданию автоматической системы синтеза по ключевым словам коротких текстовых сообщений с заданной семантической и эмоциональной направленностью в сети Интернет. Предложена эскизная структура прототипа такой системы, определен состав ее основных компонентов.

Ключевые слова: семантический анализ текста, автоматическая генерация текста, текстовое сообщение, информация, Интернет.

Britikov A.A. Automating the synthesis of the short text messages with a given semantic and emotional orientation.

Abstract. The questions about creation of automatic synthesis system for keyword shot text messages with a given semantic and emotional orientation in the Internet are discussed in this paper. The draft structure of prototype of this system is proposed, the composition of its basic components is determined.

Keywords: semantic analysis of text, text automatic generation, text message, information, Internet.

1. Введение. В настоящее время сеть Интернет является неотъемлемой частью жизни любого современного человека. Именно поэтому практически все компании стремятся реализовать свой бизнес в виде Интернет-ресурса, на котором будет отражена деятельность компании, а также будет осуществляться продажа товаров и услуг, предоставляемых указанной компанией. Как и в реальности, в сети Интернет присутствует рыночная конкуренция, которая вынуждает участников конкурентной гонки искать все новые пути привлечения покупателей. Следовательно, выделяются средства для проведения кампаний по продвижению сайта/товара/услуги, проводятся технические работы по поисковой оптимизации сайтов, покупается трафик. Все вышеуказанное остается бесполезным до тех пор, пока у пользователей не появится доверие к компании, к продаваемым ею товарам/услугам.

Практически на всех сайтах разрешено оставлять комментарии под размещенными новостями/товарами. Чем больше положительных отзывов, тем больше доверия у пользователя к предлагаемому его вниманию товару. Соответственно, на пользователя оказывает непосредственное влияние мнение других пользователей, которые оставили свой комментарий. Отсюда возникает потребность компаний форми-

ровать у пользователя «правильное» мнение в отношении предмета его интереса. Существует возможность «вручную» оператору рекламной службы зарегистрировать несколько аккаунтов под различными именами и размещать необходимую информацию от их лица, но такой путь верен только в случае малого количества необходимых комментариев. Если же необходимо ежедневно размещать несколько тысяч подобных комментариев, то содержать штат таких операторов становится экономически невыгодным. Дополнительной сложностью размещения большого количества комментариев является обязательное отсутствие в них шаблонности. Пользователь не должен определить, что формирование комментария происходило на основе минимальных изменений в заранее подготовленном тексте. Решением указанной проблемы может являться создание системы автоматической генерации комментариев (САГК).

2. Состав и функции системы автоматической генерации комментариев. САГК — совокупность компьютерных программ, основной целью которой является создание по ключевым словам комментариев с их внутренним разделением по тональности, стилистике и объему размещенного текста. В дальнейшем сформированный пул комментариев будет автоматически размещен на необходимых сайтах.

При разработке САГК необходимо учитывать человеческий фактор и возможность распознавания человеком "неестественности" написанного комментария. Таким образом, каждый комментарий должен быть индивидуальным, иметь свою тональную окраску (быть положительным, негативным, нейтральным), иметь свою стилистику, которую легко можно сопоставить с определенной возрастной, социальной, образовательной группой. Необходимо учитывать, что каждый сформированный комментарий должен адекватно соотноситься с комментируемым товаром/услугой, а также в общей массе они должны формировать «правильное» мнение.

Исходя из изложенного выше, САГК должна включать в себя следующие программные модули:

1. Модуль морфологического и лексического анализа, полученных ключевых слов или текста.
2. Модуль синтаксического анализа.
3. Модуль семантического анализа.
4. Модуль генерации комментария.
5. Модуль тональной окраски сформированного комментария.
6. Модуль стилистической окраски сформированного комментария.

В настоящее время вопросы анализа текста проработаны глубоко и объемно [1,2]. Существует множество работающих систем по анализу и разбору представленного на вход в систему текста, но большинство разработок имеет целью сформировать реферат текста [3,4] или выделить основную тему текста[5]. В нашем случае первоначальный анализ поданного на вход текста не является ключевым моментом функционирования системы.

Вопрос оценки тональности сообщений также серьезно проработан и существуют как научные разработки, так и практическая реализация систем тональной оценки сообщений [6, 7, 8, 9]. Наиболее широко указанные системы используются в области оценки удовлетворенности пользователей той или иной услугой [6,8], либо для оценки новостных материалов [6,8]. Таким образом, научно-методическая база для разработки программных средств генерации коротких текстовых сообщений, существует и в целом обеспечивает решение поставленных задач.

Основными задачами САГК является генерация комментария и придание ему стилистической окраски с учетом особенностей заданной социальной группы.

При генерации комментария возникает три основных проблемы:

1. Определение основной темы генерации комментария.
2. Определение множества возможностей выбора.
3. Построение связного текста.

При задании ключевых слов оператором их набор должен быть достаточным для того, чтобы САГК смогла определить основную тему комментария. Таким образом, первоочередной задачей является формирование «понимания» системой области используемых слов. Также необходимо принять во внимание, что используемые слова могут иметь несколько значений (омонимы) или несколько слов могут означать одно и то же (синонимы).

Возможности выбора определяются областью основной темы комментария. САГК должна быть в состоянии решить, какую информацию сообщить, когда и что написать, и какие слова или синтаксические структуры наилучшим образом могут выразить заданную тему. При выражении основной темы необходимо решать задачи синтеза связанных слов в предложении.

На заключительном этапе необходимо связать сформированные предложения в связный текст. Текст должен отражать ту информацию, которая существенна и актуальна для решаемой задачи. Например, если заданной темой будет создание серии положительных коммента-

риев о маршрутизаторах фирмы Cisco, то будет неверным выдать комментарий, повествующий о маршрутизаторах в целом, как о классе. Определив, какая информация актуальна в настоящее время и для текущей задачи, САГК также должна определить с чего необходимо начать и чем завершить построение комментария. Причем порядок информации в тексте может иметь решающее значение для понимания текста читающим.

Стилистическая окраска также крайне важна для большего доверия к комментарию, потому что на один и тот же товар пишут отзывы разные люди из разных групп — возрастных, социальных, образовательных и т.д. Необходимо внести в комментарий такие поправки, которые однозначно заставят читающего поверить, что это было написано реальным человеком.

3. Заключение. Таким образом, создание программных средств генерации коротких тестовых сообщений представляет собой сложную задачу, решение которой должно опираться на современные достижения в различных областях компьютерной лингвистики.

Литература

1. *Ермаков А.Е.* Полнотекстовый поиск: проблемы и их решение. // Мир ПК, №5. 2001.
2. *Ермаков А.Е.* Неполный синтаксический анализ текста в информационно-поисковых системах. // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2002. В двух томах. Т.2. "Прикладные проблемы". М., Наука, 2002. С. 180–185.
3. *Белоголов Г.Г., Калинин Ю.П., Хорошилов А.А.* Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2004. 246 с.
4. *Алыгулиев Р.М.* Автоматическое реферирование документов с извлечением информативных предложений // Вычислительные технологии. 2007. Т. 12. № 5. С. 5–15.
5. *Большакова Е.И., Баева Н.В., Васильева Н.Э.* Структурирование и извлечение знаний, представленных в научных текстах // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Т. 2. М: Физматлит, 2004. С. 480–488.
6. Компания «Ай-Текко». Система «Аналитический курьер». [Электронный ресурс]. — Доступ: <http://www.i-teco.ru/ac.html>
7. *Меньшиков И.Л., Кудрявцев А.Г.* Обзор систем анализа тональности текста на русском языке // Молодой ученый. 2012. №12. С. 140–143.
8. Проект ВААЛ. [Электронный ресурс]. — Доступ: <http://www.vaal.ru/>
9. *Ермаков А.Е., Киселев С.Л.* Лингвистическая модель для компьютерного анализа тональности публикаций СМИ. // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. М., Наука, 2005.

Бритиков Алексей Алексеевич — военнослужащий в/ч 64829. Область научных интересов: инфокоммуникационные системы, компьютерная лингвистика, обработка текстовых материалов. chil86@rambler.ru; в/ч 64829, ул. Мясницкая, Москва, 101000, РФ; р.т. +7(926)379-52-50.

Britikov Alexey Alexeevich — military man v/ch 64829. Research interest: infocommunication systems, computing linguistics, dataminig systems. chil86@rambler.ru; v/ch 64829, Myasnitskaya st., Moscow, 101000, Russia; office phone +7(926)379-52-50.

Рекомендовано лабораторией автоматизации научных исследований СПИИРАН.
Статья поступила в редакцию 14.03.2013.

РЕФЕРАТ

Бритиков А.А. **Автоматизация синтеза коротких текстовых сообщений с заданной семантической и эмоциональной направленностью.**

Сеть Интернет является неотъемлемой частью жизни любого современного человека. Именно поэтому практически все компании стремятся реализовать свой бизнес в виде Интернет-ресурса, на котором будет отражена деятельность компании, а также будет осуществляться продажа товаров и услуг, предоставляемых указанной компанией. Возникает потребность компаний формировать у пользователя «правильное» мнение в отношении предмета его интереса путем размещения большого количества положительных комментариев о предлагаемых товарах и услугах на различных электронных ресурсах, причем комментарии должны быть индивидуальными, т.е. иметь различную стилистику, семантику и эмоциональную окраску. Решением такой проблемы может являться создание системы автоматической генерации комментариев (САГК) — совокупности компьютерных программ, основной целью которой является создание по ключевым словам комментариев с их внутренним разделением по тональности, стилистике и объему размещенного текста. В дальнейшем сформированный пул комментариев будет автоматически размещен на необходимых сайтах.

САГК должна включать в себя следующие программные модули:

1. Модуль морфологического и лексического анализа, полученных ключевых слов или текста.
2. Модуль синтаксического анализа.
3. Модуль семантического анализа.
4. Модуль генерации комментария.
5. Модуль тональной окраски сформированного комментария.
6. Модуль стилистической окраски сформированного комментария.

Основными задачами САГК является генерация комментария и придание ему стилистической окраски с учетом особенностей заданной социальной группы.

При генерации комментария возникает три основных проблемы:

1. Определение основной темы генерации комментария.
2. Определение множества возможностей выбора.
3. Построение связного текста.

Сформированный текст должен отражать ту информацию, которая существенна и актуальна для решаемой задачи. Стилистическая окраска также важна для большего доверия к комментарию.

SUMMARY

Britikov A.A. Automating the synthesis of the short text messages with a given semantic and emotional orientation.

The internet is an essential part of modern human being. Hence this almost all companies are trying to realize their business as an internet resource which reflects company's activity and commences trading operations and services. The need to form the "right" opinion about product in consumers mind emerges. This is being achieved by placing of a big number of positive feedbacks about product on various internet resources. Besides all feedback comments must be individual and differ by various semantics, style and emotional color.

The development of automated comment generation system (ACGS) can solve this problem. ACGS — is a set of computer programs aimed at generation of comments varied by tone, style and size of textual information using keywords. The comments from formed pool are automatically placed on specific internet resources by ACGS.

ACGS should include following program modules:

1. The morphology and lexical analysis of received keywords or text.
2. The syntactic analysis module.
3. The semantic analysis module.
4. The comment generation module.
5. The module of tonality (emotional coloration) of formed comment.
6. The stylistic coloration module.

The main goal of ACGS is in comment generation and applying stylistic coloration accounting for common features of targeted social group.

There are three main issues in comment generation:

1. Determination of the main topic for comment generation.
2. Determination of variety of choice possibilities.
3. The coherent text formation.

The generated text should reflect relevant and important for the problem being solved information. The stylistic coloration is important for comment to be trusted.