УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«МИНСКИЙ ГОСУДАРСТВЕННЫЙ ЛИНГВИСТИЧЕСКИЙ
УНИВЕРСИТЕТ»

УДК 811.111’322.2 (043.3)

ЧЕРНЫШЕВИЧ
Марина Валерьевна

СИСТЕМА АВТОМАТИЧЕСКОГО СЕНТИМЕНТ-АНАЛИЗА ТЕКСТОВ НА АНГЛИЙСКОМ ЯЗЫКЕ

Автореферат
диссертации на соискание ученой степени
кандидата филологических наук

по специальности 10.02.21 – прикладная и математическая лингвистика

Минск, 2019

Работа выполнена в учреждении образования «Минский государственный лингвистический университет»

Научный руководитель: **Совпель Игорь Васильевич**, доктор технических наук, профессор, заместитель генерального директора по информационным технологиям ООО «АйЭйчЭс Глобал»

Официальные оппоненты: **Беляева Лариса Николаевна**, доктор филологических наук, профессор, академик РАЕН, ГОУ ВПО «Российский государственный педагогический университет им. А. И. Герцена», кафедра образовательных технологий в филологии

Головня Анастасия Ивановна, кандидат филологических наук, доцент, Белорусский государственный университет, кафедра компьютерной лингвистики и лингводидактики

Оппонирующая организация: УО «Гродненский государственный университет имени Янки Купалы»

Защита состоится 24 декабря 2019 года в 14.00 на заседании совета по защите диссертаций Д 02.22.01 в учреждении образования «Минский государственный лингвистический университет» по адресу: 220034, г. Минск, ул. Захарова, 21, ауд. Б-202; e-mail: info@mslu.by; тел. ученого секретаря: (017) 284-47-48.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Минский государственный лингвистический университет».

Автореферат разослан « » ноября 2019 года.

Ученый секретарь
совета по защите диссертаций
кандидат филологических наук, доцент Р. В. Детскина

ВВЕДЕНИЕ

Одной из характерных черт настоящего времени является стремительное развитие социальных интернет-платформ, таких как блоги, форумы и сети. Пользователи генерируют огромное количество текстовых сообщений и это, несомненно, ценный источник знаний, а также данных об их отношении к различным событиям, конкретным лицам, товарам, услугам и т. п. Очевидно, что такие объемы информации фактически невозможно обработать вручную, поэтому активно разрабатываются системы автоматизации их обработки и анализа с целью решения многих актуальных прикладных задач, в том числе и автоматического сентимент-анализ текста – АСАТ. Он предполагает распознавание эмоционально окрашенных фрагментов текста, также распознавание объектов (предметов, фактов, процессов, событий и т. п., их атрибутов и свойств), в отношении которых в анализируемом тексте высказано мнение, и формирование по определенным критериям оценки (тональности) этого мнения.

Большинство существующих решений задачи АСАТ ориентировано на определенные, достаточно узкие предметные области, и на бинарную шкалу тональности (положительная и отрицательная), к которой иногда добавляется третья – нейтральная. Но массовым пользователем очень востребованы более гибкие шкалы, ориентированные, в частности, на решение задачи на уровне объектов и их аспектов, а не предложений и тем более документов в целом, на использование наряду с общими и частных оценок, содержащих в том числе и дескриптивные значения, указывающие на основание оценки, а также сравнительных оценок. Предлагаемые при этом решения задачи должны носить промышленный характер и осуществляться с ориентацией на современные технологии обработки больших данных, тексты произвольной предметной области и произвольных источников (особенно социальных сетей) со значительно более высокими по сравнению с существующими качественными показателями. А это требует разработки процедур определенной фильтрации текстовых документов, распознавания в них и корректировки лексических единиц, не соответствующих нормам естественных языков (ЕЯ), разработки и использования развитых лингвистических баз знаний, ориентированных на глубокий лингвистический анализ текстов.

Настоящая диссертационная работа посвящена исследованию и решению перечисленных задач, разработке на основе полученных результатов прототипа промышленной системы автоматического сентимент-анализа текстов на английском языке.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь с крупными научными программами и темами. Диссертационное исследование выполнялось на кафедре информатики и прикладной лингвистики УО «Минский государственный лингвистический университет» в рамках госбюджетной НИР «Создание автоматического англо-белорусского терминологического словаря по информационным технологиям и разработка методики его использования для перевода текстов» (задание 4.1.07 подпрограммы «Белорусский язык и литература» Государственной программы научных исследований «Экономика и гуманитарное развитие белорусского общества» на 2016–2020 гг.), а также программы научных исследований отдела разработки средств интеллектуализации информационных систем ООО «АйЭйчЭс Глобал».

Цель и задачи исследования. Целью диссертационной работы является разработка методов, алгоритмов и лингвистического обеспечения автоматического сентимент-анализа текстов на английском языке и их реализация в виде промышленного прототипа.

Для достижения поставленной цели необходимо решить следующие основные **задачи**:

1) сформулировать концепцию системы ACAT, ориентированную на промышленный характер решаемой задачи и высокие показатели эффективности;

2) определить необходимый для исследования текстовый материал и построить, опираясь на его анализ и актуальность решения целевой задачи для массового пользователя, развитую шкалу тональности мнений, включающую как абсолютные, так и сравнительные оценки;

3) разработать структурно-функциональную схему системы ACAT, сформулировать основные требования к ее базовому лингвистическому обеспечению и определить общий метод решения задачи, а также основные принципы реализации системы и ее качественные показатели;

4) разработать методы, алгоритмы и в дополнение к базовому – собственное лингвистическое обеспечение для автоматического сентимент-анализа текстов на английском языке;

5) построить прототип системы ACAT для английского языка, определить на основе тестирования его качественные характеристики и внедрить в промышленную эксплуатацию.

Объектом исследования являются смысловые сентимент-ориентированные связи в текстах на английском языке.

Предмет исследования составляют методы, алгоритмы и лингвистические ресурсы автоматического сентимент-анализа текстовых документов на английском языке.

Материалом исследования являются три корпуса текстов:

- тексты на английском языке, представляющие собой выборку сообщений интернет-пользователей в социальных сетях (Twitter.com, Facebook.com), интернет-магазине (Amazon.com) и форумах (Fixia.com, Tripadvisor.com), объемом 21 740 предложений;
- выборка из англоязычных новостных статей, опубликованных международным агентством Reuters, объемом 3 640 предложений;
- тексты на английском языке из научных статей международного научного журнала IEEE и патентов патентного фонда США общим объемом 3 238 предложений.

Всего для исследования был образован корпус из 28 618 предложений, содержащих 8 637 сентиментов.

Научная новизна

1. Сформулирована и обоснована концепция промышленной системы ACAT с высокими показателями эффективности. Новизна заключается в ее ориентации на глубокий, вплоть до уровня семантики, базовый лингвистический анализ текстовых документов и одновременное использование технологии машинного обучения, развитой шкалы тональности мнений и широких лингвистических ресурсов, на универсальность по отношению к языкам текстовых документов и их соответствию языковым нормам, по отношению к их источникам и предметной области.
2. Построена гибкая шкала тональности мнений и дано соответствующее ей формальное определение понятия «мнения». Новизна заключается в том, что в отличие от применяемых бинарных шкал (положительная и отрицательная оценка), к которым иногда добавляется нейтральная тональность, данная шкала включает как абсолютные (общие и частные), так и сравнительные оценки, всего девять типов тональности.
3. Разработана структурно-функциональная схема системы ACAT, определены ее базовый лингвистический процессор, общий метод решения задачи, а также основные принципы работы и качественные показатели. Новизна заключается в самой функциональности этой системы (базовый лингвистический анализ, фильтрация спам-сообщений, предобработка, т. е. его лексическая нормализация и фильтрация слов и предложений, собственно сентимент-анализ текста), ориентированной на сформулированную ее концепцию, разработанную шкалу тональности мнений, а также в сформулированных для такого класса си-

стем принципах их реализации и обязательного тестирования как системы в целом, так и ее отдельных модулей.

4. Разработаны методы, алгоритмы и лингвистическое обеспечение для фильтрации спам-сообщений, предобработки англоязычного текста и его сентимент-анализа. Впервые в основу этого процесса положено комбинирование методов машинного обучения с оригинальными лингвистическими ресурсами типа обучающих корпусов текстов, признаковых пространств, специальных словарей и множеств лингвистических паттернов. Это в совокупности определило универсальную по отношению к языкам текстовых документов, их источникам и предметной области технологию решения целевой задачи, существенно оптимизировало это решение и обеспечило его эффективность.

5. Разработан прототип оригинальной системы автоматического сентимент-анализа текстов на английском языке. Осуществлено ее внедрение в промышленную эксплуатацию, которая показала высокие качественные характеристики в силу использования разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов.

Решение поставленных задач определило содержание работы и позволило сформулировать **положения диссертации, выносимые на защиту**:

1. Концепция системы АСАТ, универсальной по отношению к языку текста и ориентированной на промышленный характер решаемой задачи и высокие показатели эффективности.

2. Формальное определение мнения и шкала тональности мнений, которая в отличие от существующих включает как абсолютные (общие и частные), так и сравнительные оценки, всего девять типов тональности.

3. Структурно-функциональная схема системы АСАТ, разработанная в соответствии с предложенной концепцией, ее базовый лингвистический процессор, общий метод решения задачи, а также основные принципы реализации и качественные показатели.

4. Методы, алгоритмы и лингвистическое обеспечение для фильтрации спам-сообщений, предобработки англоязычного текста и его сентимент-анализа, в основу которого впервые положено комбинирование методов машинного обучения с оригинальными лингвистическими ресурсами в виде обучающих корпусов текстов, признаковых пространств, специальных словарей и множеств лингвистических паттернов.

5. Прототип оригинальной системы автоматического сентимент-анализа текстов на английском языке, его качественные характеристики и результаты внедрения в промышленную эксплуатацию.

Личный вклад соискателя ученой степени. Все основные результаты и положения, выносимые на защиту, получены автором самостоятельно и составляют его личный вклад в исследование темы диссертации.

Апробация диссертации и информация об использовании ее результатов. Основные результаты диссертационной работы докладывались и обсуждались на заседаниях кафедры информатики и прикладной лингвистики МГЛУ в 2015–2018 гг., ежегодных конференциях преподавателей и аспирантов МГЛУ (Минск, Беларусь, 2017 г.; Минск, Беларусь, 2018 г.), международном конгрессе по информатике (CSIST) (Минск, Беларусь, 2016 г.), международных конференциях по оценке семантических систем SemEval (Дублин, Ирландия, 2014 г.; Сан-Диего, США, 2016 г.).

Разработанная система автоматического сентимент-анализа внедрена в состав промышленного многоязычного лингвистического процессора системы инженерии и управления знаниями Goldfire Innovator, используемой для решения инновационных задач крупнейшими компаниями мира.

Опубликованность результатов исследования. По теме диссертации опубликовано 10 научных статей, среди которых 5 – в рецензируемых периодических изданиях (2,28 авторского листа) и 4 – в сборниках материалов научных конференций (1,4 авторского листа). Общий объем опубликованных материалов составляет 3,68 авторского листа. Все публикации выполнены без соавторства.

Структура и объем диссертации. Диссертация состоит из введения, общей характеристики работы, трех глав с выводами по каждой из них, заключения, библиографического списка, списка публикаций автора и шести приложений. Полный объем диссертации составляет 147 страниц машинописного текста, из них 89 страниц занимает основной текст, 3 рисунка и 16 таблиц; библиография из 153 источников, включая 10 публикаций соискателя, на русском и английском языках, изложена на 13 страницах, 6 приложений – на 30 страницах.

Во **введении** обоснована актуальность темы диссертационной работы, определены цели и основные задачи исследования.

Общая характеристика работы содержит обоснование актуальности, новизны диссертационного исследования и его значимость. Здесь же определяются цели и задачи, объект и предмет исследования, излагаются основные положения, выносимые на защиту.

В **первой главе** дано общее описание комплексной задачи ACAT и определены её основные подзадачи, осуществлён анализ методов и результатов их решения, а также существующих промышленных систем ACAT.

Во **второй главе** рассмотрен круг вопросов, касающихся формирования необходимого для исследования текстового материала (корпуса текстов), разработки шкалы тональности мнений, метода и принципиальной схемы решения АСАТ и ее базового лингвистического процессора, а также принципов реализации и качественных показателей.

В **третьей главе** представленные в предыдущих главах диссертации результаты получили своё дальнейшее развитие с целью разработки системы автоматического сентимент-анализа текстов на английском языке.

В **заключении** подводятся итоги проведенного исследования и даются рекомендации по практическому использованию его результатов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В **первой главе** дано общее описание комплексной задачи АСАТ и определены её основные подзадачи, осуществлён анализ методов и результатов их решения, а также представлен обзор существующих промышленных систем АСАТ. Отмечено, что задача АСАТ является относительно новой в сфере автоматической обработки ЕЯ и в связанной с ней терминологии все еще наблюдаются разночтения. При исследовании и решении этой задачи целесообразно исходить из того, что понятия мнения и сентимента, а также оценки и тональности являются эквивалентными. Основная цель задачи АСАТ – распознавание эмоционально окрашенных фрагментов текста и, далее, распознавание объектов (предметов, фактов, процессов, событий и т. п., их атрибутов и свойств), в отношении которых в анализируемом тексте высказано мнение, и формирование по определенным критериям оценки этого мнения. Оценка как абстрактная категория выражается в языке с помощью оценочных предложений, являющихся целостным лингвистическим явлением, и содержащих оценку человеком того, что он считает ценным, что плохим, а к чему относится безразлично. Эффективное решение актуальной на настоящее время задачи АСАТ должно быть ориентировано на распознавание и анализ оценочных предложений и высказываний, являющихся проекцией логической категории оценки как структурной организации. Основными элементами последней являются: субъект оценки, объект оценки и аспектные характеристики объекта, по отношению к которому высказывается мнение, а также собственно оценка, формализованная на основе предварительной классификации тональности мнений. Важное место в выборе типа оценки занимают такие аспекты, как учет аксиологического значения, чисто оценочного и дескриптивного значений, абсолютного и сравнительного характера оценки, влияния контекста на реализацию семантики оценочных единиц. В этом плане, как показал анализ, наибольший интерес представляют те

типы оценок, которые, во-первых, обладают формальными признаками для автоматического анализа, в первую очередь ингерентные оценки, во-вторых, оценки, наиболее востребованные массовым пользователем. В данном случае это частные оценки, содержащие наряду с оценочным и дескриптивное значение.

Основными подзадачами комплексной задачи АСАТ являются:

- классификация тональности мнений, иначе – разработка шкалы тональности;
- распознавание эмоционально окрашенных текстовых фрагментов (оценочных предложений);
- распознавание и категоризация объектов и их характеристик (аспектов), о которых высказывается мнение;
- определение тональности мнений по отношению к объектам и их аспектам в соответствии с разработанной шкалой.

В связи с этим подчёркнуто, что большинство существующих решений задачи АСАТ ориентировано на бинарную шкалу тональности (положительная и отрицательная), к которой иногда добавляется третья – нейтральная, а также количественные показатели для любой из используемых категорий. Но массовым пользователем сейчас востребованы более гибкие шкалы, ориентированные, в частности, на анализ объектов и аспектов, а не предложений и тем более документов в целом, на использование наряду с общими и частных оценок, содержащих в том числе и дескриптивные значения, указывающие на основание оценки, а также сравнительных оценок.

Решение подзадач автоматического распознавания эмоционально окрашенных фрагментов, объектов и их аспектов, определения тональности мнений должно осуществляться с ориентацией на тексты произвольной предметной области и произвольных источников (особенно – социальных сетей) со значительно более высокими по сравнению с существующими показателями полноты и точности. А это требует разработки процедур фильтрации недостоверных мнений, распознавания и корректировки в обрабатываемых текстах лексических единиц, не соответствующих нормам ЕЯ, разработки и использования лингвистических баз знаний, ориентированных на глубокий, вплоть до уровня семантики, лингвистический анализ текстовых документов.

Таким образом, качественное, особенно с точки зрения промышленных систем автоматической обработки текста, решение комплексной задачи АСАТ может быть построено на основе комбинированного метода, объединяющего технологии машинного обучения и экспертных лингвистических правил с опо-

рой на функциональность развитого лингвистического процессора и его лингвистической базы знаний (ЛБЗ).

Во **второй главе** рассмотрен круг вопросов, касающихся формирования необходимого для исследования текстового материала (корпуса текстов), разработки шкалы тональности мнений, метода и принципиальной схемы решения АСАТ и его базового лингвистического процессора, а также принципов реализации системы и ее качественных показателей.

Необходимый для исследования текстовый материал, обусловлен поставленной задачей, предполагаемым приложением полученных результатов и известных принципов построения репрезентативных корпусов текстов, включил в себя сообщения пользователей (21 740 предложений, 8 612 мнений), новостные статьи (3 640 предложений, 25 мнений) и технические тексты (3 238 предложений, 0 мнений), а имеющиеся в них мнения были аннотированы. Такое многообразие текстового материала очень важно с точки зрения универсальности разрабатываемой системы АСАТ и её ориентации на обработку произвольных текстов ЕЯ. Экспертный анализ текстового материала позволил выявить ряд особенностей, существенно влияющих на качество решения задачи АСАТ, в частности, на этапе автоматического лингвистического анализа текста. Это, прежде всего, наличие спам-сообщений, а также сленговых слов и выражений, которые были подвергнуты классификации. Указанные особенности в значительной степени предопределили принципиальную схему решения АСАТ и состав его базового лингвистического процессора.

Опираясь на анализ исследуемого текстового материала, актуальность получаемых по ходу исследования практических результатов для массового пользователя и выдвигаемые им требования к решению задачи АСАТ, разработана наиболее приемлемая шкала тональности мнений, включающая как абсолютные (общие и частные), так и сравнительные оценки. Для общих оценок предложена тональность типа Positive, Negative и Desire, для частных – Appreciation, Deficiency и Wish, для сравнительных – ComparisonPositive, ComparisonNegative и ComparisonEqual (таблица 1).

Мнение (C) пользователя можно формально представить в виде кортежа из четырех компонентов:

$$C = (O_1, O_2, A, V) \quad (1)$$

где O_1 – один или более объектов оценки, иначе фокус (Focus) оценки;

O_2 – один или более объектов, с которым сравнивается фокус оценки (данный компонент является непустым только в случае сравнительных оценок);

A – основание оценки (этот компонент является непустым только в случае частных оценок);

V– значение тональности в соотвествии с построенной шкалой тональности.

Таблица 1. – Шкала тональности мнений

Абсолютные оценки		**Сравнительные оценки**
Общие	**Частные**	
Positive – положительное мнение по отношению к объекту	Appreciation – положительное мнение по отношению к объекту с указанием основания оценки (какой конкретный аспект объекта нравится)	ComparisonPositive – положительное сравнительное мнение
Negative – отрицательное мнение по отношению к объекту	Deficiency – отрицательное мнение по отношению к объекту с указанием основания оценки (какой конкретный аспект объекта не нравится)	ComparisonNegative – отрицательное сравнительное мнение
Desire – желание приобрести какой-либо объект	Wish – условно отрицательное отношение к объекту с указанием основания оценки (желание изменить что-либо в объекте)	ComparisonEqual – нейтральное сравнительное мнение

Разработана также принципиальная схема решения задачи АСАТ, которая включает в качестве основных следующие этапы обработки входного текста: его базовый лингвистический анализ, фильтрацию спам-сообщений, предобработку (лексическую нормализацию) текста и его собственно сентимент-анализ (рисунок 1).

Рисунок 1. – Принципиальная схема решения задачи АСАТ

Исходя из указанной выше принципиальной схемы и построенной шкалы тональности мнений, в качестве базового лингвистического процессора разрабатываемой системы АСАТ определен известный многоязычный БЛП IHS Goldfire. Полученное в результате базового лингвистического анализа текста (подзадача 1), включающего лексический, лексико-грамматический, синтаксический и семантический анализы, представление входного текста (тегированное лексико-грамматическими классами (ЛГК) слов предложение с распознанными в них именными и глагольными группами и субъект-акция-объект (САО) отношениями) поступает далее на вход модуля фильтрации спам-сообщений (подзадача 2). Затем отфильтрованный текст поступает на вход модуля предобработки текста, где осуществляется его лексическая нормализация, а также удаляются те фрагменты текста, которые несущественны при дальнейшей обработке, например, ссылки, хэш-теги (подзадача 3). Следует отметить, что если в результате этих процедур произошла лексическая замена или удаление каких-либо слов и выражений текста, он повторно поступает на вход модуля базового лингвистического анализа и далее – на вход модуля сентимент-анализа текста (подзадача 4), гдеосуществляется распознавание в этом тексте предложений, содержащих мнения, а также всех их компонентов в соответствии с формулой (1). Распознанные компоненты помечаются в тексте тегами.

Учитывая достоинства и недостатки, присущие при решении задачи АСАТ как методам машинного обучения, так и методам, основанным на лингвистических паттернах, в настоящей работе предложено их комбинирование. Метода машинного обучения с учителем дополнен процедурами оперирования, при необходимости, результатами синтаксического, а также семантического анализа текста. Все это позволяет максимально обобщать признаковое описание прецедентов и тем самым обеспечивает высокое качество решения задачи АСАТ. Метод машинного обучения предполагает наличие обучающей выборки (корпуса), представляющей собой совокупность описаний прецедентов с использованием измеряемых у них признаков. Учитывая, что этот метод используется на трех этапах решения задачи АСАТ, для лексической нормализации взят один из известных корпусов коротких сообщений (твитов) из социальной сети Твиттер, каждое из которых (их всего 2 954) содержит лексические единицы, подлежащие нормализации, и их эталонные варианты. Для задачи фильтрации спам-сообщений на основе исследуемого в работе текстового материала создан также аннотированный корпус, включающий 25 583 прецедента. На основе этого же материала создан аннотированный корпус для задачи собственно сентимент-анализа текста. Всего же аннотировано 28 618 предложений, которые содержат 8 637 мнений.

Учитывая высокую востребованность именно промышленных систем АСАТ, в работе сформулированы наиболее важные принципы, которые должны быть положены в основу разрабатываемой системы и максимально соответствовать их параметрам. Исходя из того, что сентимент-анализ текста является одной из задач информационного поиска, для оценки эффективности разрабатываемых решений предложено использовать такие классические показатели, как точность, полнота и *F*-мера. Дано их определение для задачи АСАТ.

В **третьей главе** результаты, представленные в предыдущих главах диссертации, получили своё дальнейшее развитие с целью разработки системы автоматического сентимент-анализа текстов на английском языке.

Эта работа осуществлялась в соответствии с полученной ранее принципиальной схемой решения задачи АСАТ. Функциональность ее базового лингвистического анализа текста обеспечивается БЛП IHS Goldfire. Что касается остальных трех задач (фильтрация спам-сообщений, предобработка текста и его собственно сентимент-анализ), в основу их решения положены методы машинного обучения с некоторыми дополнениями в виде разработанных множеств лингвистических паттернов. Система ориентирована на обработку произвольных текстов на английском языке, в том числе и текстов из социальных сетей.

Поскольку создание обучающего корпуса для фильтрации спам-сообщений является очень трудоемким, для автоматической выборки текстов (сообщений пользователей), принадлежащих классу «спам», были сформулированы соответствующие критерии, а в их рамках разработаны 26 лингвистических паттернов. Всего обучающий корпус включил 25 583 сообщения (10 214 для класса «спам» и 15 369 для класса «не спам»). В основу классифицирующего алгоритма решения этой задачи положен метод опорных векторов, а построенное пространство включило 10 типов статистических признаков, ориентированных на лексический, лексико-грамматический и семантико-синтаксический уровни языка. Тестирование показало высокое качество решения задачи (точность 98,76 %, полнота 97,68 %).

Задача предварительной обработки текста включает две подзадачи: его лексической нормализации и фильтрации слов и предложений. Первая подзадача была решена в два этапа: распознавания во входном тексте лексических единиц-кандидатов для процедуры нормализации и их собственно нормализации. Первый этап реализован на основе методов машинного обучения. В качестве обучающего корпуса использовался один из известных, уже размеченных, корпусов, включающий 2 954 сообщения. Построенное пространство включило 11 типов признаков, оперирующих теми же, что и при решении предыдущей задачи, уровнями языка. Эффективность решения задачи на данном этапе также

оказалась высокой (точность 91,08 %, полнота 81,21 %). Собственно нормализация лексических единиц-кандидатов реализована на основе разработанных четырех обобщенных лингвистических паттернов и списка замен сленговых слов и сокращений на слова и выражения, соответствующие нормам ЕЯ. Такой список построен с использованием известных онлайн-ресурсов и включил 745 пар (например, *4ever – forever, becoz – because, fab – fabulous*). Общая эффективность решения первой подзадачи подтверждена достаточно высокими качественными показателями (точность 85,03 %, полнота 80,17 %). Анализ влияния этих результатов на модули лингвистической обработки текста, например, на корпусе твитов (около 16 000 предложений) показал, что 55,32 % его предложений были подвергнуты процессу нормализации, 39,42 % CAO-отношений – откорректированы и 9 % новых CAO-отношений – распознаны. Осуществлено решение второй подзадачи предварительной обработки текста – удаление определенных слов и конструкций внутри предложений обрабатываемого текста (имён пользователей, ссылок и электронных адресов, некоторых междометий и т.п.) и даже целых предложений, явно не содержащих мнения. В его основу положены 57 разработанных лингвистических паттернов, что в конечном счете существенно повысило точность решения целевой задачи.

Для решения задачи собственно сентимент-анализа на основе исследуемого текстового материала создан обучающий корпус. Он включил 28 618 предложений, содержащих 8 637 мнений, тональность которых по отношению к помеченным в предложениях объектам экспертно классифицирована в соответствии с построенной шкалой, которая имеет девять классов тональности. Поскольку объекты оценки обычно выражаются именными группами, для автоматического распознавания объектов-кандидатов во входном тексте функциональность используемого БЛП была дополнена возможностью распознавания вложенных именных групп и последующего их анализа и редактирования. С этой целью разработано 18 лингвистических паттернов. В итоге каждое предложение как входного текста, так и обучающего корпуса трансформировалось во столько прецедентов, сколько объектов-кандидатов в нем присутствовало. Так, корпус, производный от исходного обучающего, включил в итоге 54 665 отдельных прецедентов (предложение и один из его объектов-кандидатов).

Согласно методам машинного обучения каждому прецеденту должна соответствовать совокупность признаков заданного множества. В результате проведенного экспертного анализа наше множество включало такие признаки, как лексическая единица (ЛЕ), лексико-грамматический класс, лексико-семантическая группа (ЛСГ, одна или несколько), числовое значение тональной ориентации, принадлежность ЛЕ определенному компоненту CAO-отношений.

Распознавание в разработанной системе АСАТ таких признаков, как собственно ЛЕ, ЛГК ЛЕ и принадлежность ЛЕ к определенному компоненту САО-отношений осуществляется с помощью БЛП. Что касается признака ЛСГ, он введен с целью повышения обобщающей способности нейронной сети и эффективного обучения классификатора. Всего введено 60 ЛСГ (слова, специфические для социальных сетей; именнованные сущности; определённого класса существительные, глаголы, прилагательные и т.п.) Их распознавание осуществляется на основе двух разработанных лингвистических паттернов и словарей прилагательных, наречий, глаголов и существительных с положительной и отрицательной экспрессивной коннотацией общим объемом 9 870 слов. Для получения числовых значений тональной ориентации (ТО) ЛЕ, на основе исследуемого материала и формулы поточечной взаимной информации PMI, был разработан лингвистический ресурс, включающий 89 380 слов со значением тональности для каждого из них в пределах от –∞ до +∞. В соответствии с вышеизложенным, например, для предложения *My sister doesn't actually love this Audi!!!* система автоматически построит следующее признаковое описание:

1. (**ЛЕ:** *My*, **ЛГК:** PP$, **ЛСГ:** –, **ТО:** +0,1)
2. (**ЛЕ:** *sister*, **ЛГК:** NN, **ЛСГ:** %nn_Person, **ТО:** 0,0)
3. (**ЛЕ:** *does*, **ЛГК:** DOZ, **ЛСГ:** –, **ТО:** 0,0)
4. (**ЛЕ:** *n't*, **ЛГК:** XNOT, **ЛСГ:** %XNOT %gr_Dislike, **ТО:** –0,6)
5. (**ЛЕ:** *actually*, **ЛГК:** RB, **ЛСГ:** %gr_Dislike, **ТО:** –0,3)
6. (**ЛЕ:** *love*, **ЛГК:** VB, **ЛСГ:** %vb_Like %gr_Dislike, **ТО:** +1,9)
7. (**ЛЕ:** *this*, **ЛГК:** DT, **ЛСГ:** –, **ТО:** 0,0)
8. (**ЛЕ:** *Audi*, **ЛГК:** NP, **ЛСГ:** –, **ТО:** 0,0)
9. (**ЛЕ:** *!!!*, **ЛГК:** !, **ЛСГ:** –, **ТО:** +0,4)

САО:

Субъект: *My sister*	Предлог: –,
Акция: *doesn't love*	Непрямой объект: –
Объект: *this Audi*	Атрибут наречие: *actually*
Атрибут прилагательное: –	

В этом предложении система распознает два объекта-кандидата (*My sister* и *this Audi*), поэтому для них автоматически будут сформированы два прецедента, каждый из которых включит одно и то же представленное выше признаковое описание. И именно эти два прецедента будут представлять данное предложение в том случае, если оно окажется во входном тексте, подлежащем решению задачи АСАТ (классификатор назначит одну из возможных меток). Если же это предложение включается в состав обучающей выборки, в одном прецеденте объект-кандидат *My sister* получит метку «нейтральное мнение/не мне-

ние», а во втором – *this Audi* – метку «фокус негативной оценки». И, таким образом, обученный на фактических данных классификатор при дальнейшем решении задачи определит, что в данном предложении присутствует мнение и его фокусом является объект *this Audi*.

При реализации системы ACAT для физического представления текстовой информации используется многомерное векторное пространство, называемое семантическим пространством. С целью повышения эффективности решения целевой задачи, данное пространство было трансформировано с помощью разработанных путем экспертного анализа словарей синонимов (3 560 пар) и антонимов (4 670 пар).

Разработанная система автоматического сентимент-анализа текстов на английском языке является оригинальной, она протестирована на контрольной выборке из 6 041 прецедента. Ее качественные показатели оказались достаточно высокими и составили: точность – 86,09 %, полнота – 81,80 %. Ниже представлены фрагменты результатов, полученных системой ACAT при обработке конкретных входных текстов. Эти результаты обозначены компонентами представленного ранее формального описания мнения в виде формулы (1).

Входной текст: *The one drawback of the LG camera is its startup speed.*

Мнение 1	**Мнение 2**
O: *LG camera*	O: *startup speed*
V: Deficiency	V: Negative
A: *startup speed*	

Входной текст: *Google has destroyed our world.*

Мнение 1
O: *Google*
V: Negative

Входной текст: *The password identification for Safari finally got better.*

Мнение 1	**Мнение 2**
O: *Safari*	O: *password identification*
V: Apprecition	V: Positive
A: *password identification*	

Входной текст: *T-Mobile is definetly better carrier than AT&T.*

Мнение 1	**Мнение 2**
O1: *T-Mobile*	O1: *AT&T*
O2: *AT&T*	O2: *T-Mobile*

V: ComparisonPositive V: ComparisonNegative

Входной текст: *Pulp looks beautiful. I wish it was available on android :(.*

Мнение 1	**Мнение 2**
O: *Pulp*	O: *Pulp*
V: Positive	V: Wish
	A: *available on android*

Входной текст: *Gatorade is the only thing I need.*

Мнение 1
O: *Gatorade*
V: Desire

Входной текст: *BTW my bugaboo stroller seems miniature compared to the duallie T!*

Мнение 1
O1: *bugaboo stroller*
O2: *duallie T*
V: ComparisonEqual

Система автоматического сентимент-анализа внедрена в состав промышленного многоязычного лингвистического процессора известной системы автоматизации инженерии знаний и управления знаниями IHS Goldfire, используемой для решения инновационных задач такими крупнейшими компаниями мира, как Boeing, Sony, Renault, LG Electronics, Samsung, Nestle, Northrop Grumman, Johnson&Johnson, Henkel, Hewlett-Packard, Daimler Chrysler, Shell.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Сформулирована и обоснована концепция промышленной системы АСАТ с высокими показателями эффективности, ориентированная на глубокий, вплоть до уровня семантики базовый лингвистический анализ текста и комбинирование методов машинного обучения с развитыми шкалой тональности мнений и лингвистическими ресурсами, на универсальность по отношению к языку текста и соответственно его языковым нормам по отношению к его источнику и предметной области [2; 3; 4; 5].

2. Создан необходимый для исследования корпус текстов общим объемом 28 618 предложений, содержащих 8 637 сентиментов, осуществлено их аннотирование. Проведен экспертный анализ указанного корпуса, который поз-

волил выявить ряд особенностей входящих в него текстов. Это прежде всего наличие спам-сообщений, а также сленговых слов и выражений. Произведена их классификация. Учет этих особенностей в значительной степени предопределил принципиальную схему решения задачи АСАТ и состав его базового лингвистического процессора. Дано формальное определение мнения и построена гибкая шкала тональности мнений, включающая как абсолютные (общие и частные), так и сравнительные оценки, всего девять типов тональности [3; 4].

3. Разработана структурно-функциональная схема системы АСАТ, которая включает в качестве основных следующие этапы обработки входного текста: его базовый лингвистический анализ, фильтрацию спам-сообщений, предобработку (лексическую нормализацию) и собственно сентимент-анализ. Созданы три аннотированные обучающие корпуса текстов, необходимые для решения всех тех подзадач, которые осуществляются методом машинного обучения, всего три таких корпуса. Оперирование при этом глубоким уровнем лингвистического анализа текста позволило максимально обобщить признаковое описание прецедентов и тем самым обеспечило высокое качество решения задачи АСАТ. Сформулированы основные принципы реализации системы и ее качественные показатели [4].

4. Разработаны методы, алгоритмы и лингвистическое обеспечение для фильтрации спам-сообщений, предобработки англоязычного текста и его собственно сентимент-анализа. В основу этого положено комбинирование методов машинного обучения с оригинальными лингвистическими ресурсами типа обучающих корпусов текстов, признаковых пространств, специальных словарей и множеств лингвистических паттернов. Это в совокупности определило универсальную по отношению к языкам текстовых документов, их источникам и предметной области технологию решения целевой задачи, существенно оптимизировало это решение и обеспечило его эффективность [1; 8; 9].

5. Разработан прототип оригинальной системы автоматического сентимент-анализа текстов на английском языке. Осуществлено тестирование, которое позволило определить качественные показатели решения как отдельных задач, так и задачи АСАТ в целом: точность фильтрации спам-сообщений составила 98,76 %, полнота – 97,68 %, нормализация текста – соответственно 85,03 % и 80,17 %, сентимент-анализ в целом – 86,09 % и 81,80 %. Система внедрена в промышленную эксплуатацию и в силу разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов, имеет высокие показатели эффективности, сопоставимые, а по отдельным алгоритмическим решениям превосходящими существующие [6; 7].

Рекомендации по практическому использованию результатов

Разработанные методы, алгоритмы и лингвистические ресурсы рекомендуются к использованию при создании лингвистических процессоров различных систем автоматической обработки текстовых документов (сентимент-анализа, информационного поиска, экспертных и вопросно-ответных систем и т. д.), а также в учебном процессе учреждений высшего образования, осуществляющих подготовку специалистов в области интеллектуальных информационных систем и компьютерной лингвистики.

Построенная система АСАТ внедрена в состав промышленного многоязычного лингвистического процессора известной системы автоматизации инженерии и управления знаниями Goldfire Innovator, используемой для решения инновационных задач крупнейшими компаниями мира.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых изданиях

1. Чернышевич, М. В. Автоматическая нормализация англоязычных сообщений пользователей социальных сетей для задачи их сентимент-анализа / М. В. Чернышевич // Вестник МГЛУ. Сер. 1, Филология. – 2017. – № 5 (90). – С. 66–73.

2. Чернышевич, М. В. Обзор существующих систем автоматического сентимент-анализа текста / М. В. Чернышевич // Вестник МГЛУ. Сер. 1, Филология. – 2017. – № 6 (91). – С. 111–117.

3. Чернышевич, М. В. Актуальные аспекты решения задачи автоматического сентимент-анализа текста / М. В. Чернышевич // Вестник МГЛУ. Сер. 1, Филология. – 2018. – № 1 (92). – С. 100–106.

4. Чернышевич, М. В. Принципиальная схема решения задачи АСАТ и его лингвистическое обеспечение / М. В. Чернышевич // Вестник МГЛУ. Сер. 1, Филология. – 2018. – № 3 (94). – С. 72–80.

5. Чернышевич, М. В. Классификация тональности мнений для задачи автоматического сентимент-анализа текста / М. В. Чернышевич // Ученые записки УО "ВГУ им. П. М. Машерова": сб. науч. трудов. – Витебск. ВГУ имени П. М. Машерова, 2018. – Т. 28. – С. 136–140.

Материалы и тезисы докладов научных конференций

6. Chernyshevich, M. IHS R&D Belarus: Cross-domain Extraction of Product Features using Conditional Random Fields / M. Chernyshevich // SemEval-2014: Proceedings of SemEval-2014. – Dublin, Ireland, 2014. – P. 309–313.

7. Chernyshevich, M. IHS-RD-Belarus at SemEval-2016 Task 5: Detecting Sentiment Polarity Using the Heatmap of Sentence / M. Chernyshevich // SemEval-2016: Proceedings of SemEval-2016. – San Diego, California, USA, 2016. – P. 296–300.

8. Чернышевич, М. В. Лексическая нормализация коротких сообщений пользователей социальных сетей / М. В. Чернышевич // Материалы ежегодной научной конференции преподавателей и аспирантов университета, Минск, 5–6 мая 2017 г.: в 4 ч. / отв. ред. А. М. Горлатов. – Минск: МГЛУ, 2017. – Ч. 2. – С.153–155.

9. Чернышевич, М. В. Автоматическое распознавание недостоверных мнений в сообщениях пользователей на английском языке / М. В. Чернышевич // Материалы ежегодной научной конференции преподавателей и аспирантов университета, Минск, 19–20 апреля 2018 г.: в 4 ч. / отв. ред. Л. А. Тарасевич. – Минск: МГЛУ, 2018. – Ч. 2. – С.138–141.

РЕЗЮМЕ

Чернышевич Марина Валерьевна

СИСТЕМА АВТОМАТИЧЕСКОГО СЕНТИМЕНТ-АНАЛИЗА ТЕКСТОВ НА АНГЛИЙСКОМ ЯЗЫКЕ

Ключевые слова: автоматический сентимент-анализ текста, анализ тональности, анализ мнений.

Цель исследования: разработка методов, алгоритмов и лингвистического обеспечения решения задачи автоматического сентимент-анализа текстов на английском языке и их реализация в виде промышленного прототипа.

Методы исследования: методы компьютерной лингвистики и машинного обучения, экспертное тестирование.

Полученные результаты и их новизна. В работе сформулирована и обоснована концепция системы ACAT, ориентированная на гибкую шкалу тональности мнений, универсальность по отношению к языкам текстовых документов и их соответствию языковым нормам, к их источникам и предметной области; разработана структурно-функциональная схема системы ACAT, предполагающая базовый лингвистический анализ текста, фильтрацию в нём спам-сообщений, лексическую нормализацию и сентимент-анализ; разработаны методы, алгоритмы и лингвистическое обеспечение для решения указанных задач, в основу которых положено комбинирование методов машинного обучения с лингвистическими ресурсами типа обучающих корпусов текстов, признаковых пространств, специальных словарей и множеств лингвистических паттернов, что в совокупности определяет универсальность технологии решения целевой задачи, существенно оптимизирует его и повышает эффективность; разработан прототип оригинальной системы автоматического сентимент-анализа текстов на английском языке, который в силу использования разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов, имеет высокие качественные покаазатели, осуществлено его внедрение.

Рекомендации к использованию и область применения. Разработанные методы, алгоритмы и лингвистические ресурсы могут быть использованы при построении лингвистических процессоров систем автоматической обработки текста, а также в учебном процессе при подготовке специалистов в области интеллектуальных информационных систем и компьютерной лингвистики. Полученные результаты внедрены в состав промышленного многоязычного лингвистического процессора известной системы автоматизации инженерии и управления знаниями Goldfire Innovator, используемой для решения инновационных задач крупнейшими компаниями мира.

РЭЗЮМЭ

Чэрнышэвіч Марына Валер’еўна

СІСТЭМА АЎТАМАТЫЧНАГА СЕНТЫМЕНТ-АНАЛІЗУ ТЭКСТАЎ НА АНГЛІЙСКАЙ МОВЕ

Ключавыя словы: аўтаматычны сентымент-аналіз тэксту, аналіз танальнасці, аналіз меркаванняў.

Мэта даследавання: распрацоўка метадаў, алгарытмаў і лінгвістычнага забеспячэння рашэння задачы аўтаматычнага сентымент-аналізу тэкстаў на англійскай мове і іх рэалізацыя ў выглядзе прамысловага прататыпа.

Метады даследавання: метады камп’ютарнай лінгвістыкі і машыннага навучання, экспертнае тэсціраванне.

Атрыманыя вынікі і іх навізна. У працы сфармулявана і абгрунтавана канцэпцыя сістэмы АСАТ, арыентаваная на гнуткую шкалу танальнасці меркаванняў, універсальнасць у дачыненні да моў тэкставых дакументаў і іх адпаведнасці моўным нормам, да іх крыніц і прадметнай вобласці; распрацавана структурна-функцыянальная схема сістэмы АСАТ, якая прадугледжвае базавы лінгвістычны аналіз тэксту, фільтрацыю ў ім спам-паведамленняў, лексічную нармалізацыю і сентымент-аналіз; распрацаваны метады, алгарытмы і лінгвістычнае забеспячэнне для вырашэння названых задач, у аснову якіх пакладзена камбінаванне метадаў машыннага навучання з лінгвістычнымі рэсурсамі тыпу навучальных корпусаў тэкстаў, прыкметавых прастор, спецыяльных слоўнікаў і шматлікіх лінгвістычных патэрнаў, што ў сукупнасці вызначае ўніверсальнасць тэхналогіі рашэння мэтавай задачы, істотна аптымізуе яго і павышае эфектыўнасць; распрацаваны прататып арыгінальнай сістэмы аўтаматычнага сентымент-аналізу тэкстаў на англійскай мове, які по сваіх канцэптуальных і алгарытмічных рашэннях, а таксама лінгвістычных рэсурсах, мае высокія якасныя характарыстыкі, ажыццёўлена яго ўкараненне.

Рэкамендацыі па выкарыстанні і гал іна прымянення. Распрацаваныя метады, алгарытмы і лінгвістычныя рэсурсы могуць быць выкарыстаны пры пабудове лінгвістычных працэсараў сістэм аўтаматычнай апрацоўкі тэксту, а таксама ў навучальным працэсе пры падрыхтоўцы спецыялістаў у галіне інтэлектуальных інфармацыйных сістэм і камп'ютарнай лінгвістыкі. Атрыманыя вынікі ўкаранены ў склад прамысловага шматмоўнага лінгвістычнага працэсара сістэмы аўтаматызацыі інжынерыі і кіравання ведамі Goldfire Innovator, якая ўжываецца для вырашэння інавацыйных задач буйнейшымі кампаніямі свету.

SUMMARY

Chernyshevich Maryna Valerievna

AUTOMATIC SENTIMENT ANALYSIS OF ENGLISH TEXTS

Key words: automatic sentiment-analysis, sentiment analysis, opinion mining, opinion sentiment analysis.

Goal of research: elaboration of methods, algorithms and linguistic support for solving the problem of automated sentiment analysis of texts in English and their deploying as an industrial prototype.

The research methods: methods of computational linguistics; machine learning; expert analysis.

The obtained results and their novelty: in this work elaborated concept of the ASAT system, oriented to flexible sentiment scale of opinions, to universality towards a language of text document, substandard vocabulary, text source and its subject domain; developed structural-functional scheme of ASAT system, based on deep linguistic analysis of text, identifying and filtering spam messages, normalization of substandard words and expressions and sentiment-analysis of text; proposed methods, algorithms and linguistic support for solving these tasks are based on the combination of machine learning methods with the linguistic resources such as training corpora, features spaces, special dictionaries and sets of linguistic patterns, that are building language and subject-domain independent technology to perform sentiment-analysis with high quality; developed and implemented a prototype of the original system of automatic sentiment analysis of English texts, that by virtue of proposed conceptual and algorithmic decisions, as well as linguistic resources, demonstrats high quality.

Recommendations on the use and field of application: developed methods, algorithms and linguistic resources are recommended to use by creating linguistic processors of various automatic text processing systems, as well as in the educational process in the fields of intelligent information systems and computational linguistics. The obtained results are implemented in the industrial multilingual linguistic processor of known engineering and knowledge management system Goldfire Innovator, used for solving innovative tasks by many companies of the world.

Научное издание

Чернышевич Марина Валерьевна

СИСТЕМА АВТОМАТИЧЕСКОГО СЕНТИМЕНТ-АНАЛИЗА ТЕКСТОВ НА АНГЛИЙСКОМ ЯЗЫКЕ

Автореферат
диссертации на соискание ученой степени
кандидата филологических наук

по специальности 10.02.21 – прикладная и математическая лингвистика

Ответственный за выпуск *М. В. Чернышевич*

Подписано в печать 18.11.2019. Формат 60x84$^{1}/_{16}$. Бумага офсетная. Гарнитура Таймс. Ризография. Усл. печ. 1,39 л. Уч.-изд. 1,20 л. Тираж 100 экз. Заказ 56.

Издатель и полиграфическое исполнение: учреждение образования «Минский государственный лингвистический университет». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий от 02.06.2014 г. № 1/337.
ЛП № 02330/458 от 23.01.2014 г. Адрес: 220034, г. Минск, ул. Захарова, 21