ОТЗЫВ

официального оппонента на диссертацию

Бутенко Юлии Ивановны

«Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе»,

представленную на соискание ученой степени доктора технических наук по специальности 2.3.8 Информатика и информационные процессы

Актуальность темы. Актуальность диссертационного исследования обусловлена современными тенденциями развития технологий обработки естественного языка, где наблюдается устойчивый рост объемов научно-технической информации, представленной в цифровой форме и, в значительной степени, в виде переводных текстов. Активное развитие систем машинного перевода и интеллектуального анализа текстов формируют устойчивый запрос на наличие качественных, репрезентативных и структурно размеченных параллельных корпусов научно-технических текстов.

Несмотря на наличие значительного количества параллельных корпусов для различных языковых пар, большинство из них либо ориентировано на художественные и медийные тексты, либо обладает ограниченными объемами и низким уровнем автоматизации разметки. Особенно остро данная проблема проявляется в отношении англо- и русскоязычных научно-технических текстов, характеризующихся сложной композиционной структурой, высокой насыщенностью терминами и наличием специфических обозначений. Отсутствие формализованных и автоматизированных подходов к их обработке существенно сдерживает развитие прикладных систем лингвистической аналитики и интеллектуальных информационных технологий.

В этих условиях разработка моделей и методов автоматической обработки научно-технических текстов, ориентированных на создание параллельных корпусов с многоуровневой разметкой, является своевременной и важной научной задачей. Таким образом, тема диссертационной работы является актуальной, соответствует приоритетным направлениям развития науки и техники и имеет существенное научное и прикладное значение.

Научная новизна. В диссертационной работе получены следующие наиболее важные научные результаты.

1. Развиты и уточнены модели представления иерархически организованных научно-технических текстов, предусматривающие учет межуровневых связей и функциональной значимости структурных элементов при автоматизированной обработке в параллельном корпусе.

2. Предложены усовершенствованные модели и методы автоматической разметки и выравнивания англо- и русскоязычных терминологических единиц, обеспечивающие корректную обработку многокомпонентных терминов с правосторонними определениями, что расширяет область применимости существующих подходов.

3. Разработаны оригинальные модели и метод автоматической разметки номенклатурных наименований в научно-технических текстах, учитывающие наличие произвольных буквенно-числовых последовательностей и смешение алфавитов, что ранее системно не рассматривалось.

4. Впервые предложены методы выявления машинно-сгенерированных и машинно-переведенных русскоязычных научно-технических текстов на основе анализа семантико-синтаксических характеристик и актуального членения предложения.

5. Создан прототип системы управления корпусными данными, обеспечивающий комплексную поддержку всех этапов обработки параллельных научно-технических текстов и формирование специализированных наборов данных для задач машинного обучения.

Практическая значимость диссертационной работы заключается в возможности использования разработанных моделей, методов и программных средств при создании и сопровождении параллельных корпусов научно-технических текстов, а также в системах автоматической обработки естественного языка для специальных целей. Результаты исследования могут быть применены при разработке интеллектуальных информационно-поисковых систем, терминологических баз данных, средств поддержки машинного перевода, а также в лингводидактических и прикладных инженерных задачах. Практическая ценность работы подтверждена внедрением полученных результатов в прикладные системы текстовой аналитики.

Достоверность и обоснованность научных результатов. Достоверность полученных результатов обеспечивается корректным использованием современных методов компьютерной лингвистики, машинного обучения, математической статистики и программной инженерии. Теоретические положения согласуются с известными научными результатами в области обработки естественного языка, а выводы подтверждены экспериментальными исследованиями на реальных коллекциях научно-технических текстов.

Основные результаты диссертации обсуждались на международных и всероссийских научных конференциях. Содержание исследования достаточно полно изложено в 60 печатных работах, 27 из которых опубликованы в изданиях, входящих в перечень ВАК, 20 – в изданиях, входящих в международные базы Web of Science и Scopus.

Содержание работы.

Диссертация включает введение, 6 глав, заключение и список литературы.

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследования, показаны научная новизна и практическая значимость работы.

В первой главе диссертации проведен детальный анализ существующих параллельных корпусов текстов и методов их автоматизированной обработки. Рассмотрены этапы создания параллельного корпуса научно-технических текстов, проанализированы современные средства разметки и выравнивания, выявлены их ограничения и обоснована необходимость разработки новых подходов, ориентированных на научно-технический дискурс.

Во второй главе предложены модели композиционной структуры различных типов научно-технических текстов, включая научные статьи, учебно-научные тексты и нормативные документы. Автор последовательно раскрывает особенности их иерархической организации и демонстрирует, каким образом структурные характеристики могут быть формализованы для автоматической обработки в параллельном корпусе.

Третья глава посвящена разработке моделей и методов разметки и выравнивания специальной лексики. Подробно рассмотрены структурные модели терминологических единиц и номенклатурных наименований, а также предложены алгоритмы их автоматической обработки в англо- и русскоязычных текстах, что является важным вкладом в развитие терминологической обработки.

В четвертой главе исследуются методы выявления машинно-сгенерированных и машинно-переведенных текстов. Автор обосновывает выбор семантико-синтаксических признаков и актуального членения предложения как маркеров машинного происхождения текста и подтверждает эффективность предложенных методов экспериментальными данными.

В пятой главе описана концепция и архитектура системы управления корпусными данными, а также реализованные программные средства автоматической разметки и выравнивания. Показаны возможности

использования системы в лингвистических и информационных исследованиях.

Шестая глава демонстрирует применение разработанных моделей и методов для решения практических задач, включая лингводидактику, анализ научных тенденций, информационный поиск и обработку специализированных баз знаний, что подтверждает, в том числе, прикладную направленность диссертационного исследования.

В заключении представлены основные результаты работы.

Замечания по диссертации.

1. В ряде разделов диссертации представляется целесообразным более широко обсудить сопоставление предложенных методов с современными подходами, основанными на машинном обучении и нейросетевых моделях. В частности, в главе 3 при оценке методов извлечения и выравнивания многокомпонентных терминов основное внимание уделено сравнению подходов на основе правил и шаблонов, тогда как включение результатов сравнения с моделями на базе машинного обучения, в том числе, с нейросетевыми моделями, позволило бы более полно охарактеризовать место предложенного подхода среди существующих решений. Аналогично, в главе 4 при рассмотрении методов выявления машинно-переведённых и машинно-сгенерированных текстов сопоставление с распространёнными классификационными моделями затронуто лишь в ограниченной степени, и его более развёрнутое представление могло бы дополнительно усилить интерпретацию полученных результатов.

2. Оценка качества предложенного метода извлечения терминов выполнена на выборке 20 статей по космонавтике из одного журнала за 2018–2019 гг. (см. стр. 125), которая не является достаточно репрезентативной.

3. В главах 3 и 4, где используются результаты экспертной разметки и оценка качества автоматической аннотации, в тексте диссертации не зафиксированы ключевые параметры процесса разметки, такие как количество аннотаторов, принципы формирования эталонных данных и показатели согласованности разметки. Отсутствие этих сведений затрудняет воспроизводимость экспериментов и оценку достоверности полученных количественных результатов.

4. В параграфе 5.3 показано, что обеспечение корректности корпуса в значительной степени опирается на последующую ручную проверку результатов автоматической обработки. При этом для такой важной категории научно-технической лексики, как аббревиатуры и сокращённые обозначения, в системе не описан отдельный механизм автоматического распознавания и

нормализации, вследствие чего их корректная обработка фактически выносится на этап экспертной правки, что снижает общий уровень автоматизации предложенного решения.

Указанные замечания не влияют на общую положительную оценку диссертации.

Заключение. В целом, считаю, что диссертационная работа Бутенко Юлии Ивановны соответствует требованиям пунктов 9–14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24 сентября 2013 года № 842, предъявляемым к докторским диссертациям, а соискатель достоин присуждения учёной степени доктора технических наук по специальности 2.3.8 Информатика и информационные процессы.

Котельников Евгений Вячеславович
доктор технических наук, доцент,
профессор Автономной некоммерческой
образовательной организации высшего образования
«Европейский университет в Санкт-Петербурге»
e.kotelnikov@eu.spb.ru

30.01.2026 г.

Подпись Котельникова Е.В. заверяю
Специалист отдела кадров Аржанова И.В.