

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию Юлии Ивановны Бутенко «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 исследования

В настоящее время наблюдается стремительный рост объемов новой научно-технической информации, что в свою очередь требует создания эффективных инструментов ее обработки, прежде всего, автоматических или, как минимум, автоматизированных. Такие же инструменты необходимы для обработки и структурирования уже накопленной информации.

Научно-техническая информация является значительной частью современных «больших данных», само понятие которых подразумевает работу с огромными потоками информации, которая поступает из разных источников и регулярно обновляется. Для обработки и анализа больших объемов информации, в частности, научно-технической, сегодня эффективно применяют различные методы искусственного интеллекта, в том числе нейросетевые модели и машинное обучение как одни из подклассов таких методов. Для построения названных методов имеется большой арсенал математических средств (математической статистики, численных методов, методов оптимизации, теории графов и др.), предназначенных для работы с цифровыми данными. Вместе с тем, создание и эффективность использования методов искусственного интеллекта применительно к массивам научно-технических текстов напрямую зависят от наличия больших размеченных корпусов таких текстов, имеющих структурированное представление. Примером структурированного представления больших научно-технических данных являются параллельные корпуса, представленные в виде множества текстов-оригиналов, написанных на каком-либо исходном языке, и соответствующих текстов-переводов этих исходных текстов на один или несколько других языков.

Параллельные корпуса научно-технических текстов, существующие в настоящее время, имеют незначительные объемы размеченных данных, а отдельные подкорпуса по ряду узких предметных областей вообще отсутствуют. При этом имеющиеся средства автоматической обработки текстов в этих корпу-

сах чаще всего ограничены лишь лемматизацией, морфологической разметкой и выравниванием на уровне предложений. К препятствиям использования таких текстовых корпусов для решения различных задач машинного обучения можно также отнести невозможность получения доступа к результатам разметок, реализованных в этих корпусах.

Таким образом, отсутствие структурированных и упорядоченных коллекций научно-технических текстов, а также созданных на их основе терминологических баз данных и знаний существенно тормозит развитие и совершенствование средств искусственного интеллекта применительно к автоматической обработке научно-технической информации. Поэтому разработка комплекса моделей и методов автоматической разметки и выравнивания научно-технических текстов в параллельных корпусах является актуальной задачей. Ее решению и посвящена представленная диссертация.

Содержание работы

Диссертация состоит из введения, 6 глав, заключения и списка использованных источников. Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи проведенного исследования, охарактеризованы научная новизна и практическая значимость работы.

В первой главе дан обзор имеющихся параллельных и специальных корпусов научно-технических текстов с позиций возможности автоматизации их разметки и выравнивания, а также проанализированы современные методы и средства автоматической обработки научно-технических текстов в целях их применения для создания параллельных корпусов.

Во второй главе предложены модели композиционной структуры научно-технических текстов, являющихся источниками для наполнения параллельного корпуса. Для каждого типа научно-технических текстов выделены уровни выравнивания лингвистических единиц, обязательные и факультативные элементы текста, а также рассмотрены проблемы асимметрии параллельных текстов и предложены способы их отражения в параллельном корпусе.

В третьей главе представлены разработанные модели многокомпонентных терминов и номенклатурных наименований, а также методы их разметки и выравнивания в параллельном корпусе. Предложена обобщенная модель наследования морфологических признаков многокомпонентных терминов, а на

ее основе – способ приведения таких терминов к естественной форме. Выделены структурные модели англо- и русскоязычных многокомпонентных терминов и номенклатурных наименований, которые учитывают не только разнообразие их структур от двух до шести компонентов, но и возможности их обработки морфологическими анализаторами. Разработаны методы разметки англо- и русскоязычных многокомпонентных терминов и номенклатурных наименований из параллельных научно-технических текстов, а также предложен метод выравнивания англо- и русскоязычных многокомпонентных терминов в параллельных научно-технических текстах с использованием пакета SimAlign.

В четвертой главе предложены методы выявления машинно-сгенерированных и машинно-переведенных текстов на основе особенностей актуального членения предложений в русском языке. Представлен сравнительный анализ изменения порядка слов в машинно-сгенерированных и антропогенных русскоязычных текстах, и на его основе разработан метод выявления русскоязычных машинно-сгенерированных текстов. Предложен метод выявления машинно-переведенных русскоязычных текстов или их фрагментов, в основу которого положены данные о синтаксической структуре предложения и наличия в нем переводческих трансформаций. Описан статистический подход к выявлению машинно-сгенерированных и машинно-переведенных текстов.

Пятая глава посвящена разработке прототипа параллельного корпуса научно-технических текстов. Представлена концепция построения системы управления корпусными данными, предложена информационная технология обработки научно-технических текстов, где каждый новый уровень разметки текстов основан на результате предыдущего уровня: структурная, морфологическая, терминологическая разметка, разметка номенклатурных наименований, семантико-синтаксическая разметка, разметка машинных текстов. Разработаны система разметки многокомпонентных терминов и номенклатурных наименований из параллельных англо- и русскоязычных научно-технических текстов; система обработки семантико-синтаксических структур в научно-технических текстах; предложена система структурной разметки и выравнивания научно-технических текстов, а также показаны варианты использования параллельного корпуса при решении ряда прикладных задач в области лингвистики и информатики.

В шестой главе дано применение разработанных моделей и методов обработки научно-технических текстов при решении ряда прикладных задач учебной лексикографии на примере создания программного средства для практического овладения специальной терминологией, сертификации программного обеспечения на этапе формирования нормативного профиля объекта сертификации, выявления тенденций развития научных направлений, а также инженерии знаний при организации информационного поиска и извлечении информации из базы знаний.

В заключении сформулированы основные результаты работы.

Тема и содержание диссертации Ю.И. Бутенко «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» полностью соответствуют выбранной специальности 2.3.8 – «Информатика и информационные процессы»: в диссертации и автореферате указаны разделы паспорта специальности, которым соответствует диссертационная работа. Такое соответствие не вызывает никаких сомнений.

Обоснованность и достоверность научных положений, сделанных выводов и полученных результатов

В диссертационной работе предложен и всесторонне исследован комплекс моделей и методов, позволяющих решать широкий класс задач автоматической обработки научно-технических текстов. Обоснованность и достоверность научных положений и сделанных выводов подтверждены результатами сравнения разработанных моделей и методов с известными; практическим их применением при решении различных прикладных задач автоматической обработки научно-технических текстов, а также непротиворечивостью и согласованностью полученных результатов с фактами и исследованиями, известными, в частности, в компьютерной лингвистике. Обоснованность и достоверность результатов диссертационного исследования подкреплены также соответствующими публикациями в профильных ведущих научных изданиях.

Научная новизна и практическая значимость

В диссертационной работе подучены следующие **наиболее важные научные результаты**:

1. Разработаны модели и методы разметки и выравнивания терминологических единиц из научно-технических текстов, а также модели и метод разметки номенклатурных наименований в научно-технических текстах на русском и английском языках.

2. Предложены модели иерархически-структурированных научно-технических текстов, которые учитывают межуровневые лингвистические единицы и оценку значимости каждого структурного элемента при создании параллельного корпуса.

3. Разработаны методы выявления машинно-сгенерированных и машинно-переведенных текстов на основе семантико-синтаксических особенностей русского языка.

Практические ценность и значимость диссертационной работы состоят в создании программных средств обработки научно-технических текстов в параллельном корпусе, которые позволяют автоматизировать наиболее трудоемкие процессы обработки и структурирования этих текстов, такие как извлечение многокомпонентных терминов и номенклатурных наименований, структурная разметка многокомпонентных терминов и выявление машинных текстов и их фрагментов.

Основные результаты, содержащиеся в диссертации Ю.И. Бутенко, были доложены и обсуждены на ряде всероссийских и международных конференций, отражены в 24 печатных работах, опубликованных в рецензируемых научных журналах, входящих в перечень ВАК РФ, в 9 работах в изданиях, проиндексированных в международных наукометрических базах данных Scopus и Web of Science, а также в 27 публикациях в других изданиях.

Замечания по диссертационной работе

1. Многие из результатов, содержащихся в диссертации, находятся на стыке информатики и целого ряда других научных дисциплин (терминоведения и терминографии, грамматики, лексикологии и лексикографии, переводоведения, контрастивной лингвистики, лингвостилистики, сертификации программного обеспечения, наукометрического анализа и ряда других), что, с одной стороны, является одним из несомненных достоинств рассматриваемого диссертационного исследования. С другой стороны, чтение и понимание представленного материала во многих случаях требует значительных усилий, потому что в диссертации практически всюду, когда используется терминология из названных выше научных дисциплин, отсутствуют определения и даже минимальные разъяснения этой терминологии. С моей точки зрения, стиль диссертационного текста только выиграл бы, если бы были даны соответствующие определения, а лучше – во введении был дан перечень всех используемых понятий, не имеющих прямого отношения к специальности, по которой диссертация защищается. К таковым терминам я бы отнес: код класса слов или код частеречной принадлежности; анафорическая, метатекстовая, морфологическая, грамматическая, семантическая, структурная и терминологическая разметки; падежная грамматика; обобщенная модель многокомпонентной терминологической единицы; лемматизация; номен; актуальное членение предложения, тема и рема в нем; литеральные термины (литералы); левые и правые определения термина, квантитативная лингвистика и ряд других.

2. Как указано на с. 10 диссертации, «практическая ценность работы подтверждается внедрением результатов диссертационной работы в ряд прикладных промышленных систем текстовой аналитики, о чем имеются акты о внедрении результатов диссертационного исследования». Никакой более подробной информации на эту тему, как и самих упомянутых актов, в диссертации нет. С другой стороны, с некоторыми большими подробностями эти документы упомянуты на с. 7 автореферата. Возможно, названные акты, оценивающие результаты практического использования достижений диссертации, будут представлены в процессе защиты, что представляется мне необходимым.

3. При описании современного состояния корпусной лингвистики, а также аспектов создания, использования и обработки параллельных корпусов текстов на сс. 6 и 7 диссертации приведена масса фамилий разных исследователей без единой ссылки на их публикации, что не только не в полной мере подтверждает выводы, сделанные соискателем в этой части диссертации, но даже не позволяет посмотреть конкретные публикации, о которых говорится в такой форме. Думаю, можно было найти и указать ряд обзорных статей и монографий, в полной мере отражающих информацию, представленную в диссертации. Аналогичное замечание относится еще к нескольким разделам диссертации. Так, например, на с. 58 говорится, что «к настоящему времени разработано множество методов автоматического извлечения терминов, и число работ новых работ в последние годы только увеличивается», и снова нет ни одной ссылки.

Выскажу также общее замечание ко всему тексту диссертации, когда цитируются определенные публикации со словами, что они отражают современное состояние исследований по тому или иному вопросу: достаточно часто некоторые из этих публикаций представляются устаревшими – они опубликованы более 10–15 лет назад.

4. Достаточно часто в диссертации в разделах, в которых в действительности отражены существенные достижения автора по анализу разных вопросов, отсутствуют необходимые подробности (например, не сказано, что и как анализировалось, какой инструментарий был использован и пр.), а просто сделаны ссылки на приведенные таблицы и рисунки со ссылками на собственные публикации. Считаю, что соответствующие подробности послужили бы дополнительным подтверждением сделанных выводов. Сказанное относится, например, к рисункам 1.1 (с. 22) и 1.2 (с. 51), таблицам 1.7 (с. 55) и 1.8 (с. 61).

5. Как правильно указано на с. 26, одним из ключевых аспектов проектирования параллельного корпуса научно-технических текстов являются их метаописания. В диссертации нет ни слова об использовании каких-то стандартов метаописаний.

6. На с. 130 впервые появилось упоминание пакета SimAlign, предназначенного для высококачественного выравнивания слов (элементов терминов) в параллельных текстах без использования обучающих данных, ссылок на авторов и другой информации. Судя по публикациям, диссертант активно его ис-

пользовала, но не является автором. Почему нет никаких ссылок на его использование кем-то еще и вообще ссылок, где этот пакет размещен? Вообще в диссертации содержится много указаний на использованные ресурсы (машинный переводчик, морфологический анализатор и др. – DeepL, Wordybot.ai, Rymorphy2, VOSviewer) без ссылок, минимального описания и указания путей доступа к ним. Одновременно упомянуто (с. 188) авторское приложение, написанное на языке Python, для извлечения и последующей разметки многокомпонентных терминов из англо- и русскоязычных текстов (применению этого приложения по сути посвящена глава 5 диссертации). Никаких подробностей нет, хотя, с моей точки зрения, эта программа является существенным достижением диссертации в рамках специальности, по которой диссертация защищается, и могла бы быть зарегистрирована, например, как программа для ЭВМ. Проявленная здесь скромность диссертанта представляется мне излишней. Сказанное относится и к лексическому тренажеру, созданному для оказания содействия студентам инженерных специальностей МГТУ им. Н.Э. Баумана при изучении английского языка.

7. Имеются некоторые разделы диссертации, не имеющие прямого отношения к основной содержательной части работы, которые, с моей точки зрения, описаны поверхностно и без существенного ущерба могли бы быть удалены. В частности, это замечание относится к особенностям разметки математических выражений и поиска по формулам.

8. К сожалению, в тексте диссертации достаточно много грамматических и стилистических погрешностей (несогласование слов, неверные окончания и пр.), что несколько снижает впечатление о диссертации.

Еще имеются, например, такие погрешности, как ссылка на мнение А.С. Никонов и статью, автором которой указанный автор не является (см. с. 68). Интересным показался незнакомый мне термин «документы с суровой сквозной нумерацией» (с. 93), данный снова без каких-либо пояснений. Встретился и совсем забавный вариант (с. 216): «проблематика создания учебных словарей ... довольно широко **освящена** в научной литературе».

Было бы целесообразно пояснить и дать ссылку на использованный термин (с. 77 и далее) «нотация Бекуса–Наура», а к таблице 3.5 (с. 126) дать разъяснение, что понимается под Полнотой, Точностью и F-мерой.

Приведенные замечания не снижают высокую научную ценность результатов представленной диссертационной работы и ее положительную оценку в целом. Надеюсь, они будут способствовать дальнейшему совершенствованию полученных результатов и их представления.

Заключение

Считаю, что диссертация Юлии Ивановны Бутенко «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» представляет собой законченное исследование, в котором решена важная научная проблема разработки моделей и методов обработки научно-технических текстов в параллельном корпусе. Диссертационная работа соответствует требованиям пунктов 9–14 «Положения о порядке присуждения ученых степеней», утвержденного Постановлением Правительства Российской Федерации от 24.09.2013 г. № 842. Диссертационная работа выполнена автором самостоятельно, отвечает критериям, предъявляемым к диссертациям на соискание ученой степени доктора технических наук, а ее автор, Юлия Ивановна Бутенко, заслуживает присуждения искомой ученой степени доктора технических наук по специальности 2.3.8 – «Информатика и информационные процессы».

Официальный оппонент,
доктор физико-математических
наук, профессор, Заслуженный
деятель науки Российской Феде-
рации, Заслуженный деятель
науки Республики Татарстан,
профессор Института информа-
ционных технологий и интеллек-
туальных систем Казанского
(Приволжского) федерального
университета
Тел. 8-9376151553
e-mail: amelizarov@gmail.com

А.М. Елизаров
21.04.2026

Елизаров Александр Михайлович

