«УТВЕРЖДАЮ»
Проректор по научной работе
НИУ «МЭИ»
_____ Драгунов В.К.
«01» 09 _____ 2015 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет "МЭИ"» на диссертацию Швеца Александра Валерьевича «Взаимодействие информационных и лингвистических методов в задачах анализа качества научных текстов», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики

Актуальность темы. В последнее время регулярно появляются публикации, содержащие в тексте различные некорректности, такие как: нарушение требований к лексике научного текста, нарушение структуры научного текста, правил согласования, синтаксической и семантической связности, последовательности изложения и другие. Оценивание публикаций в большинстве случаев выполняется лишь на основе косвенных наукометрических показателей, главным образом при помощи анализа цитирований. Недостатком такого подхода является возможность оценивания только публикаций, попавших в цитатные базы, в которые, как правило, включено небольшое число научных журналов. К тому же требуется длительное время на изучение новой статьи научным сообществом, написание и издание работ с цитированием и индексацией этих работ цитатными базами. Такие ограничения наукометрических методов приводят к необходимости создания методов, позволяющих непосредственно на основе текста публикации автоматически выявлять нарушения и извлекать признаки, которые характеризуют его качество. Представленная работа посвящена решению именно этих задач, что свидетельствует о ее актуальности.

Основной целью диссертационного исследования является автоматизация процесса определения качества научных текстов.

Структура и содержание диссертации. Диссертация состоит из введения, трех глав, заключения, списка использованных источников и

приложения. Диссертация содержит 120 страниц, 21 таблицу, 24 рисунка, 94 источника в списке использованной литературы.

Во введении обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, научная новизна, теоретическая и практическая значимость полученных результатов.

В первой главе рассмотрена типология нарушений в научных публикациях. Приведены примеры, показывающие, что нарушения влекут неоднозначность толкования текста и снижение ясности изложения. Исследовано, каким образом нарушения проявляются в тексте, и установлено, что они могут быть обнаружены автоматически посредством анализа лексики, синтаксических и семантических структур. Во второй части главы выполнен обзор методов, предназначенных для анализа научных текстов, исследована возможность их применения к определению качества текстов, выявлены недостатки этих методов. В заключительной части главы приведены основные выводы и сформулированы задачи исследования.

Вторая глава содержит описание предложенных в работе методов извлечения признаков, характеризующих качество текстов научной сферы. Предложен метод формирования общенаучного словаря устойчивых словосочетаний, описано применение соответствующего алгоритма и рассмотрены результаты использования полученного словаря для определения значений признака «относительное количество общенаучных словосочетаний в тексте». Показано, что с помощью данного признака научные тексты могут быть автоматически отделены от научно-популярных и ненаучных текстов. Затем предложен метод извлечения маркеров структурных разделов и выявления структуры научной публикации. Проведенные эксперименты, показывают эффективность применения метода для определения типовых разделов «Постановка проблемы», «Методы», «Результаты» и «Выводы». Далее предложен метод формирования правил для обнаружения нарушений согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения. Показано, что правила, полученные согласно методу, применимы для автоматического выявления указанных нарушений.

Третья глава посвящена решению задачи автоматического обнаружения псевдонаучных текстов с использованием описанных во второй главе признаков. Приводится разработанный автором метод автоматического определения псевдонаучных фрагментов, эффективность метода подтверждается экспериментально на большой выборке публикаций. Представлено сформированное множество признаков с возможными дискретными значениями, которое используется при построении правил для обнаружения псевдонаучных текстов посредством индуктивного ДСМ-метода порождения гипотез. Рассмотрено применение полученных правил и проведено исследование различных методов машинного обучения, показавшее, что используемое сформированное пространство признаков

позволяет решать задачу распределения текстов по классам «научные» и «псевдонаучные» с достаточно высокой точностью и полнотой.

В заключении приведены основные результаты, полученные в работе.

Приложение содержит описание разработанных автором программных модулей, реализующих предложенные в работе методы и алгоритмы.

Научная новизна исследований. Сильной стороной диссертационной работы является удачное сочетание лингвистических и информационных методов, которое лежит в основе полученных автором результатов. Это позволяет использовать достоинства как синтактико-семантического анализа, выделяющего в тексте конструкции, которые передают смысл текста, так и статистических методов, методов индуктивного порождения гипотез и машинного обучения, применяемых для формирования правил, классификации текстов, взвешивания признаков и других операций, необходимых для достижения поставленной в работе цели.

В работе получен ряд новых результатов, среди которых отметим следующие:

1) метод автоматического формирования общенаучного словаря устойчивых словосочетаний;

2) метод автоматического выявления структуры научной публикации;

3) метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения, а также метод автоматического выявления псевдонаучных фрагментов текстов научной сферы;

4) метод получения признаковых описаний научных текстов, учитывающий соответствие текста работы устоявшимся научным нормам лексики и структуры работы, на основе которого сформировано множество признаков, характеризующих качество текстов научной сферы.

Практическая значимость полученных результатов. Результаты исследований по теме диссертационной работы использованы при выполнении ряда научно-исследовательских работ по проектам Министерства образования и науки РФ, программам РАН, грантам РФФИ. Это свидетельствует о том, что полученные в работе результаты обладают не только теоретической значимостью, но и практической ценностью. Практическая значимость и востребованность результатов работы подтверждена также внедрением реализованных программных продуктов в информационную систему «ЭКБСОН» Государственной публичной научно-технической библиотеки, электронно-библиотечную систему «Руконт» ООО «Национальный цифровой ресурс «Руконт», электронно-библиотечную систему «Znanium.com» ООО «Научно-издательский центр ИНФРА-М», систему интеллектуального поиска и анализа научных публикаций «Exactus Expert» ЗАО «РосИнтернет технологии».

Рекомендации по использованию результатов и выводов диссертации. Полученные в диссертации результаты могут быть использованы в ФИЦ ИУ РАН, ИСП РАН, ИПС РАН, НИУ ВШЭ, ВМиК МГУ и в других ведущих научных учреждениях Российской Федерации и коллективах, где проводятся исследования в области интеллектуального анализа текстов.

Достоверность полученных результатов подтверждена проведенными вычислительными экспериментальными исследованиями.

Апробация полученных результатов. Результаты диссертационной работы докладывались автором на 8 российских и международных конференциях. По теме исследования опубликовано 9 работ, в том числе: 4 – в рецензируемых изданиях из Перечня ВАК и приравненных к ним, 3 – в трудах международных и российских конференций, 2 - зарегистрированные программные системы.

Автореферат достаточно полно отражает основное содержание работы.

Замечания по работе.

1. В работе фактически отсутствуют сравнения с ранее известными методами решения аналогичных задач. Следовало бы включить в сравнение методов классификации также методы, основанные на известных классических методах признакового описания текстов.
2. Автор в недостаточной степени исследует детали предлагаемых им алгоритмов и методов. Например, автор предлагает способ вычисления численной оценки соответствия текстового фрагмента структурному разделу, но не обосновывает конкретный вид формулы, не исследует другие возможные способы вычисления указанной оценки и не проводит эмпирического сравнения этих способов.
3. Не обоснован выбор типа нейронной сети при сравнении различных методов классификации.
4. Из рисунка 10 следует, что статьи разных отраслей науки отличаются уровнем использования общенаучной лексики. Однако, статьи также могут иметь отличия в структуре и степени выраженности различных нарушений. Целесообразно было бы проанализировать особенности статей разных отраслей науки и установить для них разные (наиболее соответствующие данной отрасли) критерии оценки качества текстов.
5. Имеются ошибки и неточности при описании алгоритмов. Так, во всех описаниях алгоритмов формально отсутствуют критерии окончания алгоритмов. Алгоритмы 2.4 и 2.5 содержат недетерминированные шаги (шаги с недетерминированным выбором), а именно: «**Шаг 1**. Разделить текст T на фрагменты Fi равной длины. **Шаг 2.** Выбрать один из фрагментов F…» (алгоритм 2.4); «**Шаг 1.** Выбрать одно из правил русского языка r′.»

(алгоритм 2.5). Кроме того, для алгоритма 2.4 указана сложность $O(N)$, где N – число семантических и синтаксических конструкций в тексте, но какова сложность выявления этих конструкций не указывается.

Приведенные замечания не снижают общей положительной оценки работы.

Диссертация А.В. Швеца соответствует критериям, установленным Положением о присуждении ученых степеней: в ней содержится решение задачи, имеющей существенное значение для теоретических основ информатики, она написана автором самостоятельно, обладает внутренним единством и содержит новые научные результаты, полученные автором лично. Таким образом, представленная работа полностью соответствует требованиям ВАК, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а ее автор заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики.

Отзыв обсужден и одобрен на заседании кафедры прикладной математики ФГБОУ ВО «Национальный исследовательский университет "МЭИ"» 31 августа 2015 г., протокол № 1.

Зав. кафедрой прикладной математики НИУ «МЭИ»
лауреат Премии Президента РФ в области образования,
д.т.н., профессор А.П. Еремеев

Ученый секретарь
кафедры прикладной математики НИУ «МЭИ»
к.т.н., доцент К.Г. Меньшикова