Отзыв официального оппонента
о диссертации
Смирнова Сергея Владимировича
«Технология и система автоматической корректировки результатов при распознавании архивных документов»,
представленной на соискание ученой степени кандидата технических наук по специальности
05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

1. Актуальность исследования

Несмотря на то, что в области развития современных методов корректировки ошибок оптического распознавания достигнут определенный прогресс, однако, во многих случаях системы, построенные с использованием словарей и статистических моделей языка требуют предварительного ручного обучения, предназначены для обработки современных, а не архаичных исторических текстов, прямым образом зависят от качества изображения и не адаптированы для работы с текстами на русском языке. В тоже время деятельность современных архивов настоятельно требует отказа от практики работы с сопроводительным текстовым описанием документа, его эффективной безошибочной оцифровки, что предъявляет новые требования к качеству распознавания и автоматизации корректировки неизбежно возникающих ошибок.

Решению этой важной и актуальной задачи по совершенствованию уже имеющихся и разработке новых алгоритмов эффективной корректировки ошибок, учитывающих особенности русского языка и позволяющих автоматизировать обработку корпусов текстов больших объемов и посвятил свое диссертационное исследование С.В. Смирнов.

2. Общая характеристика работы

Цель исследования заключается в разработке технологии и реализации системы оптимизации потокового распознавания архивных документов с применением методов автоматического обнаружения и корректировки допущенных ошибок.

Диссертационная работа содержит введение, четыре главы, заключение, список литературы, содержащий 122 источника и три приложения.

Структура работы логична и убедительна. Материал диссертационной работы изложен в грамотном научном стиле, содержит информативные графические иллюстрации. Постановка задачи, выводы и рекомендации в должной форме аргументированы. Диссертацию отличает высокий научный уровень, а автора — методичность, скрупулезность, хорошая теоретическая и практическая подготовка программиста, грамотная работа с лингвистическим материалом, умение анализировать и обобщать полученные наблюдения.

3. Научная новизна и основные результаты исследования

К результатам, определяющим новизну и значимость представленной диссертационной работы можно отнести следующие основные научные результаты исследования:

1) с опорой на существующую научную литературу и самостоятельно собранный корпус документов, автором разработан новый комбинированный метод автоматической корректировки ошибок распознавания архивных документов на основе рейтинго-ранговой модели текста, на основе автоматического формирования тезаурусов без предварительного обучения;

2) на основе n-грамм анализа тематических текстов и корпуса результатов распознавания выработаны правила ранжирования и выбора наилучших корректировок, учитывающие статистическую вероятность сочетаемости с предшествующими словами;

3) в целях повышения качества распознавания архивных документов автором предложен новый инструментарий, позволяющий эксперту ограничивать пространство конфигураций процесса обработки документов;

4) разработана и программно реализована новая технология распознавания архивных документов и автоматической корректировки результатов, позволяющая производить потоковую обработку больших наборов документов с учетом лексикона и специфики их предметной области.

4. Практическая ценность результатов исследования

Практическая ценность работы подтверждается актами о государственной регистрации разработанных программ и лингвистического обеспечения, а также актами внедрения результатов диссертационного исследования в Архивном комитете, Информационно-аналитическом центре и Центральном государственном архиве документов по личному составу ликвидированных предприятий, учреждений и организаций г. Санкт-Петербурга. Перечисленные документы приведены в Приложениях и свидетельствуют о высокой практической значимости работы,

результаты которой, безусловно, будут востребованы при исследовании новых аспектов распознавания архаичных машинописных документов и автоматизации коррекции ошибок. Считаем, что полученные результаты могут быть использованы также в процессе чтения вузовских курсов по компьютерному моделированию, прикладной, математической и корпусной лингвистике.

Экспериментальная проверка убедительно показала эффективность разработанного программного комплекса: количество ошибочных слов после автоматической корректировки сократилось на 46%, было исправлено 16 497 948 ошибочных слов, значение словарной точности в среднем по всем архивам увеличилось на 18%.

5. Обоснованность и достоверность полученных результатов

Обоснованность научных результатов диссертации обеспечивается:

- последовательным применением фундаментальных концепций и принципов системного подхода при проведении исследования;
- добротным аналитическим обзором научной литературы по исследуемой проблематике и преемственностью основных научных положений, сформулированных автором;
- корректностью применения аппарата теории множеств, теории вероятности, статистического анализа, корпусной и компьютерной лингвистики;
- достаточной полнотой учета факторов, влияющих на качество распознавания архаичных архивных документов и мероприятий по оптимизации автоматической корректировки полученных результатов;
- эффективной программной реализацией предложенной комбинированной системы, состоящей из Java веб-приложений, набора прикладных программ и базы данных.

Достоверность полученных результатов подтверждается:

- согласованностью результатов теоретических исследований с данными, полученными в результате практического проектирования, внедрения и эксплуатации разработанного автором исследования человеко-машинного комплекса, внушительного массива аутентичных архивных документов, содержащем около 35 тысяч документов научно-справочного аппарата пяти архивов, объемом более миллиона изображений.
- положительными экспертными оценками результатов диссертационного исследования в ходе их обсуждения на 7 – ми международных и всероссийских конференциях.

6. Основные результаты исследования, опубликованные в научной печати

По материалам диссертационного исследования опубликовано 13 печатных работы, в том числе 6 работ в рецензируемых научных изданиях из перечня ВАК РФ, также получено 2 свидетельства о государственной регистрации программы для ЭВМ. Результаты докладывались на семи научных конференциях.

Диссертационная работа имеет в целом законченный характер, написана четким языком, подробно иллюстрирована. Автореферат в целом отражает ее содержание.

7. Замечания по диссертации и автореферату

Однако, в процессе ознакомления с работой, возник ряд вопросов, не влияющих на общую высокую оценку исследования.

1. В работе практически все вводимые термины дефинированы, однако остается неясным как автор исследования трактует термин *корпус.*

2. Хотя эмпирическая база исследования очень объемна и убедительна, остается не совсем понятной информация, приведенная на страницах 90-91 диссертационного исследования. На лицо расхождение в числовых данных по объемам. Автор утверждает, что им были собраны и обработаны более 35 тысяч документов (приводится ссылка на таблицу, в которой приведена уточненная цифра 32 608 документов), научно-справочного аппарата пяти архивов, объемом более миллиона изображений (данные таблицы показывают, что корпус состоит из 708 663 изображений).

3. Какой объем эмпирической базы для распознавания позволил бы сделать полученные автором выводы; повлияло ли ее сокращение на полученный результат? Еще большее количество исследуемых текстов позволило бы обнаружить иные механизмы оптимизации процесса распознавания и коррекции ошибок?

4. Судя по тексту работы процедура нормализации является процедурой лично реализованной автором исследования. Хотелось бы прояснить, как одна из основных частей процедуры нормализации - стеммер, соотносится с такими, наиболее известными стеммерами, используемыми для обработки русского языка как SnowBall, MyStem, Stemka?

Несмотря на возникшие в процессе чтения работы вопросы, диссертация С. В. Смирнова является актуальным, новым, законченным, самостоятельным научным исследованием, обладающим высокой практической значимостью.

8. Заключение

На основании вышесказанного очевидно, что диссертационное исследование С. В. Смирнова «Технология и система автоматической корректировки результатов при распознавании архивных документов», представленное на соискание ученой степени кандидата технических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей, является научно-квалификационной работой, удовлетворяющей нормам и критериям, изложенным в п. 9 «Положения о присуждении ученых степеней», утвержденного Постановлением Правительства РФ № 842 от 24 сентября 2013 г., а ее автор, Сергей Владимирович Смирнов, безусловно, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

29. 04. 2015г.

к. техн. наук, профессор кафедры
методики обучения
математике и информатике
РГПУ им. А.И. Герцена
К.Р. Пиотровская

Сведения об оппоненте:
Пиотровская Ксения Раймондовна
кандидат технических наук,
доктор педагогических наук,
профессор кафедры методики обучения
математике и информатике
Федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Российский государственный педагогический университет им. А.И. Герцена»
Почтовый адрес: 191186, Санкт-Петербург, набережная реки Мойки, д.48
Телефон: (812)314-49-96, добавочный 20-93
e-mail: krp62@mail.ru

РГПУ им. А.И. Герцена
подпись К.Р. Пиотровской
удостоверяю «29» апреля 2015г.
Отдел персонала
управления кадров и социальной работы

Ведущий документовед
отдела персонала
И.В. Горбатовская