УТВЕРЖДАЮ

Проректор по научной деятельности
федерального государственного автономного
образовательного учреждения высшего
образования «Пензенский государственный
университет»

д.т.н., проф. Д.К. Нургалиев

«03» декабря 2021 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию Гукасяна Цолака Гукасовича «Методы и программные средства для выявления заимствований в текстах на армянском языке», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Актуальность темы диссертации. Диссертация Ц.Г. Гукасяна посвящена методам и программным инструментам для выявления заимствований в текстах на армянском языке. Тематика диссертации относится к области обработки естественного языка, и ее актуальность несомненна: нераскрытые заимствованные работы негативно влияют на научный, учебный процессы, и для армянского языка необходима программная система, решающая эту задачу, учитывая специфику языка. Такие программные системы выполняют важную роль в образовательных, научных и других организациях, помогая контролировать качество выполняемых работ и поддерживая академическую доброчестность.

Структура и содержание диссертации. Диссертация состоит из введения, пяти глав, заключения, списка литературы и двух приложений. Объем диссертации составляет 188 страниц. Список литературы содержит 186 наименований.

Во введении обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена описанию изучаемых форм заимствований.

Вторая глава посвящена исследованию и разработке внутренних методов обнаружения заимствований, в частности стилометрического анализа и выявления технических методов маскировки. Для стилометрического анализа реализованы модели (Nath et al. 2019), (Karaś et al. 2017), и метод обнаружения границ нарушений стиля на основе иерархической кластеризации. Для признакового описания стиля написания текста создан набор лингвистических ресурсов (списки аббревиатур, редких, жаргонных слов).

Третья глава посвящена исследованию внешних методов обнаружения заимствований. Изучены глобальные и локальные методы обнаружения полных и нечетких дубликатов текстов. Описан метод шинглов, используемый для поиска нечетких дубликатов. Также

изучены подходы к поиску источников заимствований в Интернете, в частности, и реализован существующий алгоритм. Для детального анализа текстов изучены модели обнаружения парафраза, предложен полуавтоматический подход к генерации парафразов предложений на основе обратного перевода. Используя этот подход, для армянского языка впервые разработан набор парафразов с высоким уровнем лексического разнообразия. Путем дообучения нейронной сети M-BERT, для армянского языка впервые создан программный инструмент обнаружения парафраза.

В четвертой главе приведено описание вспомогательных методов обработки текстов, в частности описано исследование и разработка инструментов лемматизации, предложены методы векторного представления слов для языков с богатой морфологией, описано исследование методов по исправлению ошибок автоматического распознавания текста, а также методов распознавания именованных сущностей. Для последней задачи, предложена модификация существующего метода создания размеченного набора данных, позволяющая полностью автоматизировать этот процесс.

В пятой главе приведено описание программной системы для обнаружения текстовых заимствований. Приводится краткий обзор существующих систем, описывается общая архитектура реализованной системы, его модули и компоненты. Описаны программные средства для полнотекстового поиска, технологии поиска источников заимствований в Интернете, инструменты извлечения текста из документов, реализация асинхронности для вычисления трудоемких задач.

Таким образом, **основные результаты** диссертационной работы состоят в следующем:

- Предложены методы модификации модели векторного представления fastText, которые при вычислении вектора слова используют исключительно векторы частиц этого слова.
- Предложены новый подход генерации парафразов, сокращающий роль экспертов в создании наборов данных, и новый подход к генерации размеченных текстов для задачи распознавания и классификации именованных сущностей, полностью автоматизирующий процесс создания наборов данных.
- Впервые для армянского языка созданы размеченные текстовые наборы для задач распознавания именованных сущностей, обнаружения парафраза, векторного представления слов, стилометрического анализа, и исправления ошибок автоматического распознавания текстов. Созданы программные инструменты для соответствующих задач, превосходящие существующие аналоги.
- Разработана и внедрена программная система обнаружения текстовых заимствований для армянского языка, которая позволяет обнаружить прямое и частичное копирование, парафраз, техническую маскировку, выполняет поиск заимствований в проверочной базе документов и в Интернете.

Обоснованность и достоверность результатов диссертации. Результаты диссертации являются новыми, их достоверность не вызывает сомнений. Ошибок в доказательствах, выводах и постановках экспериментов не обнаружено. Все полученные результаты подтверждаются экспериментами.

Научная новизна работы. Предложен новый метод генерации парафразов предложений на основе обратного машинного перевода, где этап ручного изменения

результатов перевода заменяется увеличением количества итераций и ручной проверкой корректности результатов. Предложена модификация полуавтоматического метода генерации размеченных данных на основе Википедии для задачи распознавания именованных сущностей, позволяющая с помощью элементов Викиданных полностью автоматизировать процесс генерации размеченных примеров.

Предложены модификации модели векторов fastText на основе подслов, которые решают проблему разреженности данных для языков с богатой морфологией и существенно сокращают размер этих моделей без серьезной потери точности в задачах лемматизации и морфологического анализа.

Соответствие содержания диссертации специальности 05.13.11. Содержание и результаты работы соответствуют паспорту специальности 05.13.11 — «математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» по следующим областям исследований: 1. Модели, методы и алгоритмы проектирования и анализа программ и программных систем, их эквивалентных преобразований, верификации и тестирования.

Теоретическая и практическая значимость работы. Основная практическая значимость диссертации заключается в разработанной системе оценки уникальности текстов, которая может быть применена в работе высших учебных заведений и других похожих организаций. Для армянского языка впервые разработаны программные инструменты, позволяющие выполнять внутренний стилометрический анализ текстов на наличие заимствований, обнаруживать парафраз, исправлять ошибки в результатах оптического распознавания текстов.

Впервые для армянского языка созданы тестовые наборы данных с ручной разметкой для задач распознавания именованных сущностей, определения парафраза, а также оценки качества векторных представлений слов. Созданные размеченные наборы текстов могут быть использованы в будущих исследованиях для разработки и оценки качества инструментов обработки армянских текстов.

Предложенные автоматические методы генерации размеченных данных позволят сократить использование человеческих и других ресурсов при создании обучающих и тестовых данных для соответствующих задач, могут быть применены для создания размеченных наборов для других языков.

Рекомендации по использованию результатов диссертации. Разработанные инструменты могут быть использованы в других программных системах для обработки армянских текстов. Представленные в диссертационной работе подходы к генерации размеченных данных могут быть использованы для построения аналогичных наборов для других языков. На основе созданных размеченных наборов данных могут быть обучены машинного обучения для задач обнаружения парафраза, исправления ошибок автоматического распознавания текстов, распознавания и классификации именованных сущностей, стилометрического анализа, для применения в прикладных программных инструментах.

Оформление текстов диссертации и автореферата. Оформление диссертации соответствует требованиям, установленным Минобрнауки России. Автореферат в полной мере отражает содержание диссертации и позволяет составить достаточно полное представление о ней.

Апробация и публикация результатов диссертации. Результаты диссертационной работы докладывались на международных и локальных научных конференциях; материалы диссертации достаточно полно представлены в 7 статьях, опубликованных соискателем, в том числе в 4 статьях в журналах, входящих в список изданий, рекомендованных ВАК, и 3 статьях в изданиях, индексируемых в международных базах данных Scopus. Количество публикаций в рецензируемых научных изданиях соответствует требованиям Положения о порядке присуждения ученых степеней.

По диссертации имеются следующие **замечания**:

1. Для тестового набора данных с ручной разметкой для задачи распознавания именованных сущностей, не указан коэффициент согласия аннотаторов.
2. В нескольких местах встречается непоследовательное использование терминологии, используются разные термины для обозначения одного и того же понятия (например, «корпус», «эталонный корпус», «набор данных», «датасет»).
3. В результатах экспериментов по обнаружению границ нарушений стиля методы сравниваются по метрике точность. Оценки качества методов по метрикам F1 и полнота вынесены в Приложение, хотя было бы полезно описать эти результаты в основной части текста.

Тем не менее, указанные замечания не ставят под сомнение ценность основных результатов работы. Диссертация является законченной научно-квалификационной работой, выполненной автором самостоятельно на высоком научном уровне. Основные этапы работы, её выводы и результаты полностью отражены в автореферате.

Заключение. Диссертационная работа Гукасяна Цолака Гукасовича «Методы и программные средства для выявления заимствований в текстах на армянском языке» является законченным научным исследованием по актуальной теме. В работе представлены результаты, имеющие важное научное и практическое значение для специальности 05.13.11 - «теоретические основы информатики». Результаты исследований, представленные в диссертации, делают существенный вклад в решение актуальной проблемы автоматической обработки текста на естественном языке.

Считаем, что диссертация Ц.Г. Гукасяна соответствует требованиям к кандидатским диссертациям, включая пункт 9 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, предъявляемым к диссертациям на соискание ученой степени кандидата наук, и является самостоятельным и завершенным научным исследованием, содержащим решение задачи поиска заимствований в текстах на армянском языке, имеющей важное значение в области методов разработки программного обеспечения, а Ц.Г. Гукасян заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертация и отзыв обсуждены на заседании научно-исследовательской лаборатории «Медицинская информатика» Федерального государственного автономного образовательного учреждения высшего образования «Казанский (Приволжский) федеральный университет» (Протокол № 6 от 3 апреля 2021 г.).

Сведения о ведущей организации: Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет».

Адрес: 420008, г. Казань, ул. Кремлевская, 18
Тел.: (843) 233-71-09
Электронная почта: public.mail@kpfu.ru
Сайт: https://kpfu.ru

Д.ф.-м.н., профессор
КФУ, Институт филологии и межкультурной коммуникации,
Высшая школа русской и зарубежной филологии им. Льва Толстого,
кафедра прикладной и экспериментальной лингвистики

КФУ, Институт информационных технологий и интеллектуальных систем,
Кафедра Интеллектуальные технологии поиска

Соловьев Валерий Дмитриевич

К.т.н.
Заместитель директора по научной деятельности,
КФУ, Институт информационных технологий и интеллектуальных систем

Заведующий кафедрой, к.н.,
КФУ, Институт информационных технологий и интеллектуальных систем,
Кафедра Интеллектуальные технологии поиска (внутренний совместитель)

Зуев Денис Сергеевич