ОТЗЫВ

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

Актуальность темы

Диссертация Гукасяна Ц.Г. посвящена методам и программным инструментам текстовых заимствований для армянского языка. Направление исследования находится на стыке компьютерных наук и вычислительной лингвистики, и имеет большую актуальность в связи с увеличением количества открытых источников и распространением заимствований. Раскрытие заимствованных работ выполняет важную роль в образовательных и научных организациях, так как помогает контролировать качество работ учащихся, статей и других публикаций. Для армянского языка актуальность разработки такой программной системы обусловлена отсутствием аналогов, учитывающих морфологические, синтаксические и другие особенности языка.

Теоретическая и практическая значимость работы

Основная практическая значимость диссертации заключается в разработанной системе «полного цикла» оценки уникальности текстов, которая содержит все ключевые модули, необходимые для сопоставления текстов. Для армянского языка впервые разработаны программные инструменты, позволяющие выполнять внутренний стилометрический анализ текстов на наличие заимствований, обнаруживать парафраз, исправлять ошибки в результатах автоматического распознавания текстов. Разработанная системы может применяться в работе высших учебных заведений и других образовательных и научных организаций.

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

Методы автоматической генерации размеченных данных, предложенные в диссертации, позволят сократить использование ресурсов при создании обучающих и тестовых данных для соответствующих задач, могут быть применены для создания размеченных наборов для других языков.

Научная новизна работы

Научная новизна работы заключается в разработанных методах векторного представления и алгоритмах автоматической генерации размеченных коллекций текстов. Предложены модификации модели векторов fastText на основе внутренних N-грамм и морфем, 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.

Также разработаны несколько подходов автоматизации процесса создания размеченных текстов, в частности, для задачи обнаружения парафраза предложен метод генерации парафразов на основе обратного машинного перевода, где этап ручного изменения результатов перевода заменяется увеличением количества итераций и ручной проверкой корректности результатов. Для задачи распознавания именованных сущностей предложен универсальный метод генерации размеченных данных на основе Википедии, использующий атрибуты Викиданных.

Обоснованность и достоверность результатов диссертации

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

Структура и содержание диссертации

Диссертация состоит из введения, пяти глав, заключения, списка литературы и двух приложений. Библиография включает 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, перечислены исследуемые методы преобразования оригинальных текстов. Также отмечается, что не рассматриваются политические и юридические аспекты задачи поиска заимствований.

Вторая глава посвящена исследованию и разработке внутренних методов обнаружения заимствований, в частности стилометрического анализа и выявления попыток использования технических обходов для скрытия заимствований. Для стилометрического анализа описаны реализация и адаптация методов из серии открытых научных соревнований в рамках конференций PAN, и создание синтетических тестовых наборов, других актуальных ресурсов для признакового описания стиля написания текстов.

Третья глава посвящена исследованию внешних методов обнаружения заимствований. Изучены глобальные и локальные методы обнаружения полных и частичных дубликатов текстов. Для поиска нечетких дубликатов изучен метод шинглов, реализованный с помощью алгоритма MinHash. Также изучены подходы к поиску источников заимствований в Интернете. Для анализа текстов на локальном уровне 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.

Рекомендации по использованию результатов диссертации

Разработанные инструменты могут быть использованы для обработки армянских текстов в рамках смежных программных систем. Представленные в диссертационной работе подходы к генерации размеченных данных могут быть использованы для построения аналогичных наборов для других языков. Кроме того, представленные подходы допускают масштабирование, что позволяет рассчитывать в перспективе на повышение качества решаемых задач путем перенастройки алгоритмов на расширенных наборах данных. 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.

Апробация и публикация результатов диссертации

Результаты диссертационной работы докладывались на международных и локальных научных конференциях. Также опубликованы четыре научные работах в журналах, из списка ВАК, и три статьи в изданиях, индексируемых в Scopus.

Достоинства и недостатки в содержании и оформлении диссертации

По диссертации имеются следующие **замечания**:

1. В пятой главе было бы желательно более подробно описать используемый метод разбиения текстов для полнотекстового поиска, в частности указать минимальный и максимальный размеры участков текста, для которых выполняется поиск источников заимствования и дальнейшая проверка на наличие парафраза.

2. Помимо оценок качества точности реализованных методов, также было бы полезно представить оценки их производительности и скорости обработки текстов.

Указанные замечания не умаляют достоинств работы. Диссертация является законченной научно-квалификационной работой и оформлена в соответствии с требованиями, установленными Минобрнауки России. Автореферат диссертации адекватно отражает содержание основной работы.

Заключение

Таким образом, диссертация Гукасяна Ц.Г. «Методы и программные средства для выявления заимствований в текстах на армянском языке» является самостоятельным, завершенным научным исследованием и полностью соответствует требованиям п. 9 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, а Цолак Гукасович Гукасян

заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент,

Чехович Юрий Викторович,
к.ф.-м.н., заведующий отделом,
Вычислительный центр им. А.А. Дородницына РАН
Федеральный исследовательский центр «Информатика и управление» РАН

Дата « 30 » апреля 2021 г.