ОТЗЫВ

официального оппонента на диссертацию

Гукасяна Цолака Гукасовича

на тему «Методы и программные средства для выявления заимствований в текстах на армянском языке»

по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

на соискание ученой степени кандидата технических наук

Актуальность темы

Наличие неправомерных заимствований в научных и выпускных студенческих работах является серьезной проблемой, снижающей качество научных исследований и качество образования. Доступность информации в Интернете стала фактором, который значительно облегчил возможность неправомерных заимствований. Для контроля заимствований в научных и студенческих работах создаются специализированные программы борьбы с плагиатом, для которых важна адаптация к конкретному языку, на котором написаны анализируемые научные работы.

Диссертационная работа Гукасяна Цолака Гукасовича посвящена задаче создания программных средств для установления степени уникальности текстов на армянском языке, включающих реализацию стилометрических методов нахождения заимствований для армянского языка, методов борьбы с техническими методами маскировки заимствований для армянского языка, нахождения нечетких дубликатов для армянского языка и др. Гукасяном Цолаком Гукасовичем выполнен большой объем работы по разработке ресурсов и инструментов для армянского языка, которые необходимы для создания системы обнаружения заимствований в текстах на армянском языке.

Достоверность и степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

Диссертация основана на применении методов машинного обучения, математической статистики, информационного поиска нечетких дубликатов,

математического анализа и линейной алгебры. Все полученные результаты подтверждаются экспериментами, проведенными в соответствии с общепринятыми стандартами.

Новизна исследования, полученных результатов, выводов и рекомендаций, сформулированных в диссертации

Новизна исследования состоит в том, что

- Представлен новый метод генерации парафразов предложения с помощью обратного машинного перевода в несколько итераций и ручной проверки корректности результатов перевода;
- Разработан метод автоматической генерации обучающих и тестовых примеров для задач нахождения и исправления ошибок оптического распознавания текстов;
- Представлен новый подход к использованию Викиданных и статей Википедии для полной автоматизации процесса генерации обучающих примеров для задачи распознавания именованных сущностей,
- 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.

Значимость для науки и практики полученных автором результатов

Практическая значимость исследования заключается в том, что разработана системе оценки уникальности текстов для армянского языка, которая может быть применена в работе высших учебных заведений, для анализа качества диссертационных работ. 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, внутреннего стилометрического анализа текстов, а также оценки качества векторных представлений слов.

Оценка содержания диссертации, ее завершенность.

Диссертация написана ясным и четким языком и состоит из введения, пяти глав и заключения. Библиография включает 186 наименований.

В **первой главе** рассматривается используемая типология заимствований. В данной работе рассматриваются методы обнаружения следующих форм заимствований: буквальные заимствования, техническая маскировка, замена синонимов и парафраз.

Вторая глава представляет исследование внутренних методов обнаружения заимствований, в частности стилометрического анализа и выявления технических методов маскировки. Для стилометрического анализа были проведено исследование подходов для разных постановок задач внутреннего анализа: обнаружение изменения стиля в документе, обнаружение границ нарушений , кластеризация по авторству. Для признакового описания стиля написания текста были скомпилированы лингвистические ресурсы (списки аббревиатур, редких, жаргонных слов), которые в дальнейшем также могут быть использованы в разработке других инструментов обработки армянских текстов. Для тестирования моделей автоматически были сгенерированы примеры из учебников, диссертаций, энциклопедий, новостных и художественных текстов.

В **третьей главе** рассматриваются внешние методы обнаружения заимствований путем сопоставления с существующими коллекциями научных текстов, публикациями в Интернете. Предложен полуавтоматический подход к генерации парафразов предложений на основе обратного перевода. Используя этот подход, для армянского языка впервые был создан набор парафразов с высоким уровнем лексического разнообразия.

В **четвертой главе** приведено описание вспомогательных методов автоматической обработки текстов, в частности описаны:

- Методы лемматизации, которые используются для нормализации текста во время индексации, стилометрического анализа и на этапе детального анализа. Дополнительно для языков с богатой морфологией предлагаются

модификации стандартных подходов, использующие только символы и морфемы слов для вычисления вектора слова;

- Методы исправления ошибок автоматического распознавания текстов;
- Методы распознавания именованных сущностей, используемые для признакового описания текстов в стилометрическом анализе;
- Способы обучения и оценки моделей векторного представления слов армянского языка, используемых в качестве признаков в задачах обработки текстов.

В пятой главе приведено описание программной системы для обнаружения текстовых заимствований.

Недостатки в содержании и оформлении диссертации

В качестве **замечаний** к тексту диссертации следует отметить следующие:

- примеры на армянском языке не переведены на русский язык (например, стр. 64-66),
- странное высказывание: векторные представления слов содержат порядка 108 параметров (стр. 98)?
- стр. 102 "модель fasttext на основе суффиксов" непонятно, сколько таких суффиксов обычно выделяется для конкретного слова,
- в тексте диссертации и автореферате встречаются предложения с нарушениями правил русского языка и опечатками (в автореферате стр. 3 фраза с нарушениями правил русского языка: расследования уникальности работа часто требуют сотен рабочих часов от затронутых учреждений, автореферат стр. 7 разрабатан, представен новый подход и др.).

Заключение о соответствии диссертации критериям, установленным Положением о порядке присуждения ученых степеней

Результаты диссертационной работы опубликованы в 7 печатных работах, из которых в том числе три статьи в изданиях и сборниках научных конференций, индексируемых в Scopus, две статьи в рецензируемых научных журналах из перечня ВАК РФ по специальности 05.13.11.

Материал диссертации изложен последовательно и логично. Структурные составляющие диссертационной работы (введение, главы, заключение, библиографический список, приложения) позволяют получить полное представление о проделанных исследованиях и полученных результатах.

Автореферат соответствует диссертации, отражает её содержание и дает представление об актуальности темы, целях, задачах, объекте и методах исследования, научной новизне, практической ценности, реализации, апробации, объеме, кратком содержании и результатах работы.

Таким образом, диссертация Гукасяна Цолака Гукасовича является научно-квалификационной работой, в которой сделан существенный вклад в разработку ресурсов и программных инструментов для автоматической обработки текстов на армянском языке, что соответствует всем критериям, предъявляемым к диссертациям на соискание ученой степени кандидата наук ВАК РФ, а ее автор заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей".

Официальный оппонент,
Лукашевич Н.В., д.т.н.,
ведущий научный сотрудник
НИВЦ МГУ имени М.В. Ломоносова
119899, Москва, Ленинские горы 1, стр. 4,

30.04.2021