"УТВЕРЖДАЮ"
Проректор
МГУ имени М.В. Ломоносова,
профессор
_____ Федянин А.А.
17 ноября 2023 г.

Отзыв ведущей организации

федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет им. М. В. Ломоносова» на диссертационную работу Аветисяна Карена Ишхановича «Метод обнаружения межъязыковых заимствований в текстах», представленную на соискание ученой степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Актуальность темы диссертации

Диссертационная работа Аветисяна К.И. посвящена исследованию и разработке методов обнаружения межъязыковых текстовых заимствований. Современные методы обнаружения межъязыковых заимствований основываются на использовании специализированных для конкретных языков инструментов. Эти методы показывают низкое качество при их применении для большого количества различных языков, в частности малоресурсных. В рамках данной работы представляется метод применимый для большого количества языков, в том числе малоресурных.

Структура и содержание диссертации

Диссертация состоит из введения, 4 глав, заключения, списка литературных источников из 111 наименований, и 2 приложений. Общий объем работы 139 страниц текста, включая 21 рисунок и 39 таблиц.

Во введении обоснована актуальность темы диссертации, сформулированы цель и задачи работы, показаны научная новизна, практическая ценность полученных результатов, приведены основные положения, выносимые на защиту.

В первой главе представляется обзор существующих решений. По отдельности рассматриваются методы, решающие две основных задачи обнаружения межъязыковых заимствований: извлечение кандидатов и детальный анализ.

Вторая глава посвящена представлению нового алгоритма обнаружения межъязыковых заимствований, который решает проблему низкого качества существующих методов или использования ими специализированных для конкретных языков инструментов. Представляются по отдельности алгоритмы решения двух основных задач обнаружения межъязыковых заимствований, а также результаты тестирования данных алгоритмов. В главе также представляется новый метод генерации состязательных атак на модели бинарной классификации, обходящий по эффективности все существующие аналоги. Также, представляется методика выбора модели классификации для этапа детального анализа, с учетом риска системы быть подверженной искусственным атакам.

В третьей главе описываются итоговые результаты, полученные представляемым алгоритмом на различных тестовых выборках. Представляются существующие тестовые выборки, а также процесс генерации и детали новых выборок, для, например армянского языка. Представляются различные метрики качества. В данной главе производится сравнение разработанного метода с другими существующими в открытом доступе.

Четвертая глава посвящена улучшению разработанного метода путем его слияния с методом представленным в Bakhteev et al., 2019. Улучшение производится путем слияния двух методов: поверх детального анализа разработанного метода производится детальный анализ метода Bakhteev et al., 2019, а также этап постобработки. Показывается, что слияние двух методов приводит к резкому улучшению метрик обнаружения межъязыковых заимствований.

В заключении подведены итоги и обобщены результаты проведенных исследований.

Основные результаты диссертации

В диссертационной работе Аветисяна К.И. получены следующие результаты:

1. Разработан новый метод обнаружения межъязыковых заимствований, превосходящий по эффективности существующие в открытом доступе методы, и также применимый к большому количеству языков, в том числе малоресурсных.

2. Разработан метод генерации словаря межъязыковых синонимов. Подобные словари могут быть применимы в различных системах содержащих поиск между текстами различных языков.

3. Разработан новый метод генерации состязательных атак на модели бинарной классификации обходящий по доле успешных атак, а также по

семантической близости и расстоянию Левенштейна все существующие аналоги.

4. Разработана методика выбора модели классификации перевода между парой предложений, с учетом риска данной модели быть подвергнутой состязательным атакам.

5. Сгенерированы новые тестовые корпуса для задачи обнаружения межъязыковых заимствований, которые в дальнейшем могут быть применимы для сравнения различных систем обнаружения подобных заимствований.

Достоверность полученных результатов

Достоверность полученных в диссертации результатов подтверждается результатами проведенных экспериментов и анализом эффективности разработанных методов, а также апробацией на конференциях и научных мероприятиях всероссийского и международного уровней и публикациями, среди которых 3 в рецензируемых журналах, в базах научных работ Scopus и Web of Science. Кроме того, получено свидетельство о государственной регистрации программы для ЭВМ.

Практическая значимость

Основная практическая значимость диссертации заключается в разработанном методе обнаружения межъязыковых заимствований применимого к большому количеству языков. Применимость данного метода к большому количеству языков и, в том числе, малоресурсных позволяет использовать данный метод в системах обнаружения плагиата в различных странах мира. Разработанный метод генерации словаря межъязыковых синонимов применим в различных поисковых системах, где производится поиск между текстами различных языков.

Представленный метод генерации состязательных атак может быть использован в различных системах для проверки их уязвимости к состязательным атакам и последующей борьбы с ними.

Новые сгенерированные тестовые выборки для задачи обнаружения межъязыковых заимствований могут быть использованы для сравнения и тестирования различных подобных систем.

Результаты диссертационной работы могут быть использованы для создания систем антиплагиата для поиска межязыковых некорректных заимствований.

Замечания

По диссертации имеются следующие замечания:

1. В работе недостаточно четко указаны границы применимости представленных методов с точки зрения тематики документов, в рамках которых производится поиск межъязыковых заимствований.

2. В рамках выбора модели детального анализа установлено ограничение в 300 миллионов параметров. Выбор данного конкретного числа недостаточно четко обоснован.

3. В тексте можно заметить несистемное использование терминов, где одни и те же понятия обозначаются различными словами, как, например, "корпус", "выборка", "набор данных".

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, ее выводы и результаты полностью отражены в автореферате.

Заключение

Диссертационная работа Аветисяна Карена Ишхановича «Метод обнаружения межъязыковых заимствований в текстах» является законченным научным исследованием по актуальной теме. В работе представлены результаты, имеющие важное научное и практическое значение для специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей». Результаты исследования вносят существенный вклад в методы обнаружения межъязыкового плагиата и методы автоматической обработки текстов на естественных языках.

Диссертационная работа соответствует требованиям ВАК РФ, представляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей», а ее автор Аветисян К. И. заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Отзыв был обсуждении и одобрен на научно-методологическом семинаре Научно-исследовательского вычислительного центра Московского государственного университета имени М.В. Ломоносова 17 октября 2023 года, протокол № 3/2023.

Заведующий лабораторией
вычислительного эксперимента и моделирования
Научно-исследовательского
вычислительного центра
Московского государственного
университета имени М.В. Ломоносова,
доктор физико-математических наук,
профессор Тихонравов А.В.
«14» ноября 2023г.

Подпись А.В. Тихонравова заверяю

Директор
Научно-исследовательского
вычислительного центра
Московского государственного
университета имени М.В. Ломоносова,
Воеводин В.В.
«14» ноября 2023 г.

Адрес ведущей организации:
119991, г. Москва, ул. Ленинские горы, д. 1
Тел.: (495) 939-10-00
http://www.msu.ru,
E-mail: info@rector.msu.ru