На правах рукописи

Аветисян Карен Ишханович

Метод обнаружения межъязыковых заимствований в текстах

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Автореферат

диссертации на соискание ученой степени
кандидата технических наук

Москва-2023

Работа выполнена в Российско-Армянском университете.

Научный руководитель:	кандидат физико-математических наук **Турдаков Денис Юрьевич**
Оппоненты:	**Котельников Евгений Вячеславович,** доктор технических наук, доцент профессор кафедры прикладной математики и информатики Федерального государственного бюджетного образовательного учреждения высшего образования «Вятский государственный университет», **Чехович Юрий Викторович,** кандидат физико-математических наук, исполнительный директор АО «Антиплагиат».
Ведущая организация:	Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет им. М. В. Ломоносова».

Защита состоится 07 декабря 2023 г. в 15 часов на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования им. В. П. Иванникова РАН по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институт системного программирования им. В. П. Иванникова РАН.

Автореферат разослан «___»_____ 2023 года.

Ученый секретарь
диссертационного совета 24.1.120.01,
кандидат физико-математических наук Зеленов С. В.

Общая характеристика работы

Актуальность темы.

В современном мире обнаружение текстовых заимствований является важной задачей в обеспечении честной и справедливой оценки научных работ. Текстовым заимствованием считается процитированный или использованный без должного цитирования фрагмент текста.

С развитием современных систем машинного перевода особую сложность для выявления стали представлять заимствования, совершенные из ресурсов других языков, такие заимствования называются межъязыковыми. Сложность выявления подобного рода заимствований и отсутствие инструментов их обнаружения для многих языков актуализируют данную задачу.

Особенно остро задача стоит для научных работ, написанных на языках, являющихся малоресурсными. *Малоресурсные языки* - это те языки, для которых существует малое количество данных в цифровом виде. Малое количество ресурсов на определенном языке приводит к совершению заимствований из ресурсов других языков.

Существующие методы обнаружения межъязыковых заимствований опираются на использовании инструментов машинного перевода, мультиязычных тезаурусов, векторных представлений слов. Также в некоторых методах используются инструменты разрешения лексической неоднозначности слов, которые являются специфичными для конкретных языков. Недостатками подобных методов являются их применимость к очень ограниченному количеству языков, обычно не являющихся малоресурсными, или, при неимении такого ограничения, низкое качество работы для малоресурсных языков. Таким образом, разработка метода обнаружения межъязыковых заимствований, применимого к большому количеству языков, в том числе малоресурсных, является актуальной проблемой.

Примером малоресурного языка может служить армянский язык. Для армянского языка не существует системы обнаружения межъязыковых

заимствований, что открывает возможности использования подобного типа заимствований и актуализирует задачу разработки подобной системы.

Объектом исследования диссертации являются текстовые документы, написанные на литературном языке, предметом — анализ оригинальности текстовых документов в условиях, когда возможно их полное или частичное заимствование из текстов, написанных на другом языке. Литературный язык - это наднациональный язык, который был приведен к общим письменным нормам для его использования в качестве официального.

Задача ставится следующим образом: имея набор с большим количество документов-источников на одном языке, в анализируемом документе на другом языке требуется найти и сопоставить те фрагменты, которые были заимствованы из фрагментов этих документов-источников. *Документы-источники* - это документы, из текстов которых могло быть произведено заимствование, *анализируемые (подозрительные) документы* - это документы, в которых потенциально возможно содержание межъязыковых заимствований.

Целью работы является разработка метода и программных средств обнаружения заимствований между текстами различных языков, в том числе применимого к малоресурсным языкам.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать существующие методы обнаружения межъязыковых текстовых заимствований.
2. Разработать и реализовать метод обнаружения межъязыковых заимствований между текстами различных языков, также применимый к текстам малоресурных языков.
3. Провести экспериментальное сравнение существующих методов с разработанным методом с использованием общепринятых метрик качества и эталонных наборов тестовых данных.
4. Проверить алгоритм на уязвимость к состязательным атакам.

5. На основе разработанного метода создать программные средства для обнаружения межъязыковых заимствований.

Научная новизна: Разработан новый метод обнаружения межъязыковых заимствований, с использованием собранного в рамках диссертационной работы словаря “межъязыковых синонимов”. Разработанный метод применим к малоресурсным языкам и, в отличие от других подобных методов, не использует инструменты машинного перевода и разрешения лексической многозначности слов. Метод показывает более высокое качество обнаружения межъязыковых заимствований, по сравнению с методами, результаты которых представлены в открытом доступе [6].

Практическая значимость заключается в разработке программных средств обнаружения межъязыковых заимствований, которые возможно использовать в работе высших учебных заведений, в том числе тех стран, в которых государственный язык является малоресурсным [1-6]. Метод, в частности, опробован на текстах армянского языка, и, исходя из полученных результатов, применим в работе с текстами армянского языка. Метод был дополнен детальным анализатором и алгоритмом склейки обнаруженных текстовых фрагментов, которые используются в программной системе “Антиплагиат.ру”. Дополнение изначального метода дало дополнительный прирост в эффективности обнаружения заимствований. Таким образом, дополненный метод может быть использован в качестве нового метода обнаружения межъязыковых заимствований с лучшей эффективностью.

Представлена методика выбора языковой модели для этапа детального анализа, учитывающая угрозу возможности осуществления состязательных атак.

Также, были сгенерированы тестовые выборки обнаружения межъязыковых заимствований в двух различных настройках: где в рамках одного анализируемого документа возможно содержание заимствований из текстов на нескольких различных языках, и, где в рамках одного анализируемого документа возможно содержание заимствований из текстов только на одном языке [6].

Основные положения выносимые на защиту:

1. Разработан новый метод обнаружения межъязыковых заимствований, превосходящий по эффективности существующие. Дополнительно, метод применим к задаче обнаружения межъязыковых заимствований в текстах малоресурсных языков.
2. Разработан новый метод генерации словаря “межъязыковых” синонимов, позволяющего достичь высоких показателей метрики полноты для этапа извлечения кандидатов в задаче обнаружения межъязыковых текстовых заимствований.
3. Разработан новый метод генерации состязательных атак “черного ящика” на языковые модели бинарной классификации, превосходящий по доле успешных атак, а также по дистанции Левенштейна и семантической близости все существующие аналоги.
4. Разработана методика выбора языковой модели для этапа детального анализа, учитывающая угрозу возможности осуществления состязательных атак.

Апробация работы. Результаты данной работы докладывались на конференциях, форумах:

1. XIV Годичная научная конференция РАУ, 2021, Ереван, РА;
2. LREC 2022 Workshop on Processing Language Variation: Digital Armenian (DigitAm), 2022, Марсель, ФР;
3. Международная конференция ”Иванниковские чтения 2022”, Казань, РФ;
4. AINL: Artificial Intelligence and Natural Language Conference, 2023, Ереван, РА;
5. DataFest Yerevan, 2023, Ереван, РА.

Публикации. По теме диссертации опубликовано 4 печатных работ, в том числе в изданиях и сборниках научных конференций индексируемых в Scopus [2-4], а также 1 свидетельство о государственной регистрации программы для ЭВМ [1].

Личный вклад. Предлагаемые в диссертации инструменты, текстовые наборы данных и исследования разработаны и выполнены автором или при его непосредственном участии.

Внедрение результатов. Результаты, полученные в рамках данной работы, внедрены в инструмент обнаружения заимствований “Sieve”, который в свою очередь внедрен в следующих учреждениях:

1. Российско-Армянский Университет
2. Высший Аттестационный Комитет Республики Армения;

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объем диссертации составляет 139 страницы текста, включая 21 рисунок и 39 таблиц. Список литературы содержит 111 наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований проводимых в рамках диссертационной работы, формулируется цель, ставятся задачи работы, перечисляются основные положения, выносимые на защиту, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена исследованию существующих методов обнаружения межъязыковых текстовых заимствований. Проводится обзор и выделяются основные подходы к решению данной задачи.

В большинстве существующих алгоритмов обнаружения межъязыковых заимствований используется двухэтапный подход. Первый этап, извлечения кандидатов, нацелен на быстродействующее уменьшение количества документов-кандидатов для конкретного анализируемого документа в процессе поиска. На втором этапе детального анализа документы-кандидаты, извлеченные на первом этапе, проходят более детальную проверку для нахождения конкретных фрагментов текстов из которых было произведено заимствование.

Раздел 1.1 посвящен обзору существующих методов, позволяющих решить задачу уменьшения количества документов-кандидатов для конкретного

анализируемого документа. Представляются различные проблемы использования того или иного подхода для большого количества языков, в том числе малоресурсных.

В рамках данной работы используется метод, основанный на мультиязычном тезаурусе, который обходит задачу определения лексической неоднозначности слов на этапе построения словаря, приводящего тексты в независимую от языка форму.

Раздел 1.2 посвящен обзору существующих методов, решающих задачу нахождения для конкретных фрагментов анализируемого документа конкретные фрагменты текстов, из которых было произведено заимствование. Представляются методы, основанные на использовании векторных представлений слов, а также метод, использующий BERT-основанные модели для классификации.

В рамках представляемого метода на этапе детального анализа был использован подход с применением BERT-основанной мультиязычной языковой модели XLM-RoBERTa, которая является межъязыковым текстовым энкодером и была обучена на 2.5 терабайтах данных для 100 языков.

Во **второй главе** описывается новый двухэтапный метод обнаружения межъязыковых текстовых заимствований решающий проблемы низкого качества или отсутствия специализированных инструментов для решения поставленной задачи для большого количества языков. В рамках решения этапа извлечения кандидатов представляется также метод генерации словаря “межъязыковых” синонимов. Дополнительно, описывается методика выбора модели для этапа детального анализа с учетом угрозы быть подверженной состязательным атакам. Также описывается алгоритм генерации состязательных атак “черного ящика” на языковые модели бинарной классификации разработанный в рамках методики

В разделе 2.1 описывается *первый этап метода обнаружения межъязыковых текстовых заимствований - извлечение кандидатов*, который представляет из себя процесс первичной фильтрации кандидатов из проверочной базы, позволяющий резко снизить количество документов или текстовых фрагментов подлежащих более дорогостоящему процессу детального анализа. В

качестве проверочной базы могут служить как заранее собранные коллекции документов, так и тексты из Интернета, при этом тексты проверочной базы написаны на отличном от анализируемого документа языке.

В рамках представляемого метода поиск релевантных кандидатов производится на уровне фрагментов текстов документов. Тексты анализируемых документов, а также документов-источников разбиваются на более мелкие фрагменты, такие как: предложения и параграфы соответственно.

Таким образом, наша задача на данном этапе состояла в том, что имея коллекцию из *N* документов, разбитую на N_m фрагментов на некотором языке L_1, в качестве коллекции источников, и некоторый фрагмент S_i из анализируемого документа S на языке L_2 $(L_1 \neq L_2)$, для фрагмента S_i нужно отфильтровать топ k релевантных ему фрагментов из имеющихся N_m, где $k << N_m$.

В качестве метода фильтрации релевантных фрагментов используется метод, основанный на построении инвертированного индекса по прошедшим предобратку фрагментам-источникам. Предобработка фрагментов проходит в несколько этапов. Для начала производится токенизация фрагментов, затем их лемматизация, все стоп-слова и пунктуационные символы удаляются, а также все заглавные буквы приводятся к строчным.

Последним этапом предобработки является представление слов фрагментов в независимой от языка форме. Анализируемые документы разбиваются на анализируемые предложения, после чего данные предложения подвергаются такому же процессу предобработки. Таким образом, поиск релевантных параграфов-источников производится на уровне анализируемых предложений методом полнотекстового поиска с использованием инвертированной индексации и функции оценки близости между двумя текстовыми фрагментами Okapi BM25, и в независимой от языка форме. Весь процесс фильтрации релевантных фрагментов описан на рис. 1.

Для осуществления этапа фильтрации релевантных параграфов-источников на языке L_1 для некоторого анализируемого предложения на языке L_2 требуется

либо привести их к одному языку, либо привести их к независимой от языка форме.

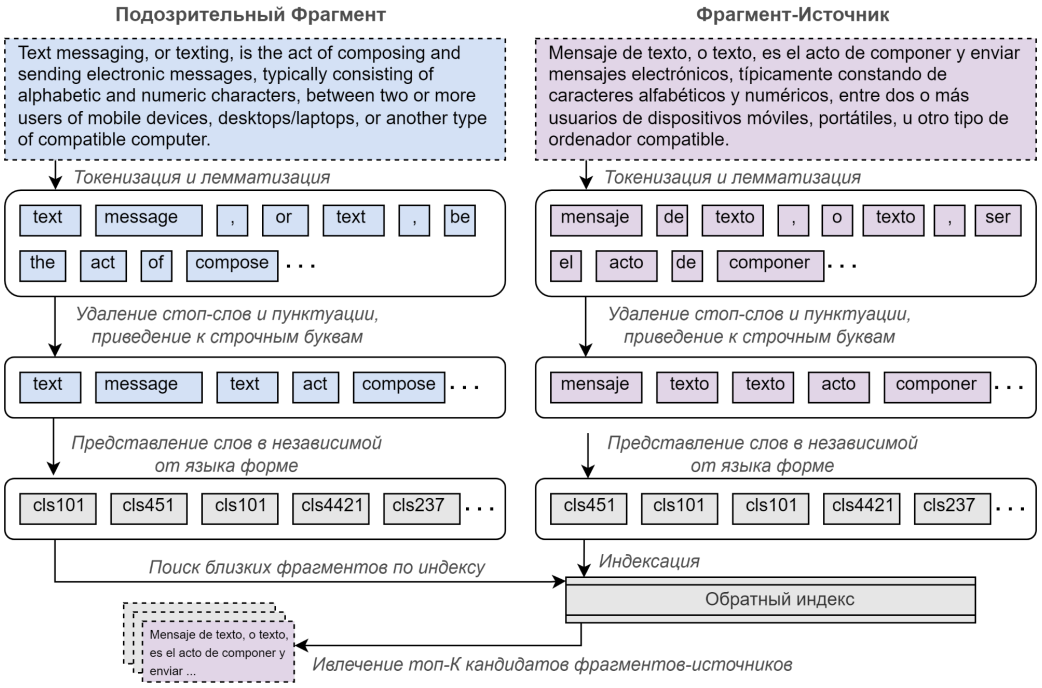

Рисунок 1 – Процесс предобработки анализируемых (подозрительных) фрагментов и фрагментов-источников, приведение их к независимой от языка форме и процесс поиска фрагментов-кандидатов.

В рамках данной работы мы рассмотрим метод приведения фрагментов к независимой от языка форме. Метод основывается на использовании *алгоритма генерации словаря “межъязыковых” синонимов*. Словарь “межъязыковых” синонимов - это словарь, в котором содержатся слова разбитые на определенные нумерованные группы. Группы должны содержать в себе слова на различных языках, при условии, что слова в рамках одной группы, независимо от языка, являются семантически близкими друг к другу.

Для сбора подобного словаря использовался тезаурус Universal WordNet (UWN), содержащий в себе более чем полтора миллиона слов на 200 языках и связи между ними. В данном тезаурусе, в рамках одного языка, слова объединены в синонимические группы, где каждой такой группе соответствует некая смысловая концепция. Смысловые концепции в свою очередь являются

межъязыковыми. Таким образом, каждой смысловой концепции соответствуют синонимические группы множества языков.

Объединяя синонимические группы смысловых концепций для нужных нам языков, получаются группы “межъязыковых” синонимов, которые и образовали базовый вариант нашего словаря. Однако, главной проблемой разбиения слов по смысловым концепциям является их смысловая многозначность. Также, одно слово может принадлежать множеству смысловых концепций, что приведет к ситуации, в которой данное слово содержится во множестве групп “межъязыковых” синонимов, что может негативным образом сказываться на процессе извлечения фрагментов-кандидатов. Во избежание данных проблем словарь был модифицирован путем использования только самых популярных смысловых концепций конкретных слов. Оценка популярности каждого из смысловых “концептов” определенного слова была извлечена из тезауруса Princeton WordNet 3.1, так как данный тезаурус, для английских слов, содержит в себе те же концепции, что представлены в UWN. Оценка смысловых концептов извлекалась только для слов английского языка, так как Princeton WordNet 3.1 является тезаурусом английского языка. Таким образом, используя самые популярные значения слов, было снижено количество “межъязыковых” синонимических групп, в которых встречается определенное слово.

Пусть L_{UWN} будет множеством языков поддерживаемых в UWN, и пусть C_{UWN} множество всех смысловых концептов во всех языках, и w некоторое слово английского языка. В качестве $S^{L}_{C_w}$ обозначим множество синонимических групп, принадлежащих всем смысловым концептам слова w (C_w- множество смысловых концепций слова w, $C_w \subset C_{UWN}$) на языке $L \in L_{UWN}$. Используя данные обозначения, определим процесс получения “межъязыковых” синонимических групп для нашего итогового словаря:

Итоговый словарь (в дальнейшем: Top1). Для каждого английского слова w, по отдельности для каждой из возможных частей речи данного слова,

выбирались только самые частоиспользуемые смысловые концепты C_w^{Top1}. В итоговые “межъязыковые” синонимические группы объединялись только синонимические группы данного смыслового концепта $S_{C_w^{Top1}}^{L}$ для всех языков $L \in L_{UWN}$.

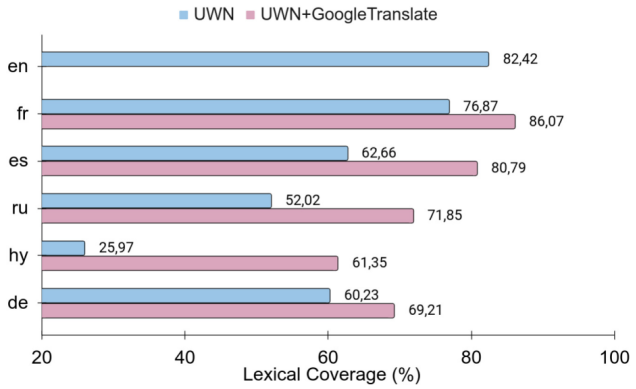

Рисунок 2 – Процент покрытия слов из 120 тысяч статей Wikipedia словарем UWN и его дополненной версией.

Для проверки качества словаря, полученного вышеописанным методом, была посчитана статистика его лексического покрытия текстов из 120 000 случайных статей из Wikipedia. Данная статистика представлена на рис. 2.

Исходя из представленной статистики, для некоторых смысловых концепций в словаре Universal WordNet не содержалось слов на некоторых из рассматриваемых языков. Во избежание данной проблемы, было решено искусственно дополнить полученный словарь.

Для тех “межъязыковых” синонимических групп, полученного ранее словаря, в которых не содержались слова на некотором языке, извлекались слова на английском языке и производился их перевод на недостающий язык. Для дополнительного обогащения словаря перевод производился и на те языки, которые присутствовали в “межъязыковых” синонимических группах.

Процесс перевода был осуществлен с использованием системы машинного перевода Google Translate[1]. Перевод каждого слова производился с учетом части речи, которой принадлежит концепция рассматриваемой “межъязыковой”

[1] https://translate.google.com/

синонимической группы и соответственно все остальные слова данной группы. Весь процесс получения словаря межъязыковых синонимов Top1 представлен на рис. 3.

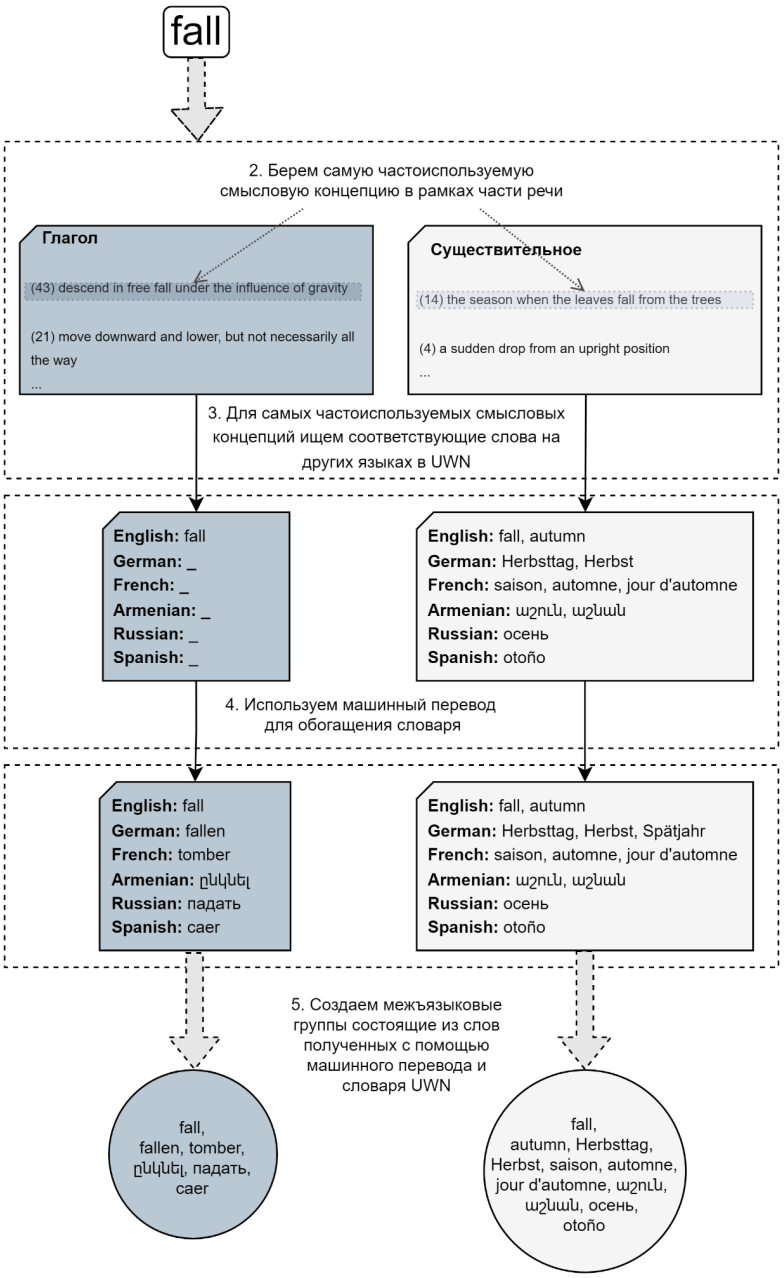

Рисунок 3 – Процесс создания первичного словаря синонимов с использованием тезауруса Universal WordNet, и процесс его дополнения с использованием машинного перевода.

Для сравнения словаря Top1 и его дополненной с помощью машинного перевода версии (табл. 1) в качестве документов-источников выступали 10 000 документов извлеченных из английской Wikipedia, используемые в рамках корпуса

CrossLang (Bakhteev et al., 2019). В качестве анализируемых документов также выступали представленные в корпусе CrossLang 316 автоматически сгенерированных документов на русском языке. В каждом из 316 анализируемых документов содержались заимствованные фрагменты из 10 000 документов-источников.

Таблица 1 – Результаты сравнения словаря, основанного только на словах из UWN, со словарем дополненным с помощью машинного перевода, на задаче извлечения фрагментов-кандидатов.

	Recall@K			
Словарь на основе	K=1	K=5	K=10	K=50
UWN	0,175	0,265	0,307	0,426
UWN + Translate	**0,747**	**0,827**	**0,853**	**0,899**

Так как в процессе извлечения топ-К фрагментов-кандидатов достаточно чтобы фрагмент из которого было произведено заимствование содержался в этих К фрагментах независимо от его позиции, то в качестве метрики для оценки работы этапа извлечения кандидатов была выбрана метрика Recall@K (1).

$$Recall@K = \frac{\text{Количество релевантных элементов среди топ}-K\text{ предсказаний}}{\text{Количество всех релевантных фрагментов}} \quad (1)$$

Исходя из результатов показанных в табл. 1, видно, что модифицированный словарь Топ1 показывает лучшие результаты.

В разделе 2.2 рассматривается *второй этап метода обнаружения межъязыковых текстовых заимствований - детальный анализ.* Данный этап предназначен для фильтрации кандидатов, полученных после окончания первого этапа, и сопоставления конкретных фрагментов текстов анализируемого документа с конкретными фрагментами документов-кандидатов.

В качестве метода для произведения детального анализа был выбран метод попарного сравнения предложений анализируемого фрагмента со всеми предложениями его топ-К фрагментов-кандидатов (анализируемые фрагменты и фрагменты кандидаты разбивались по предложениям). В рамках процесса попарного сравнения этих предложений для каждой пары производилась бинарная

классификация – являются ли предложения данной пары переводом друг друга. Процесс классификации производился с использованием многоязычной языковой модели нейронной сети. Работа модели представлена на рис. 4.

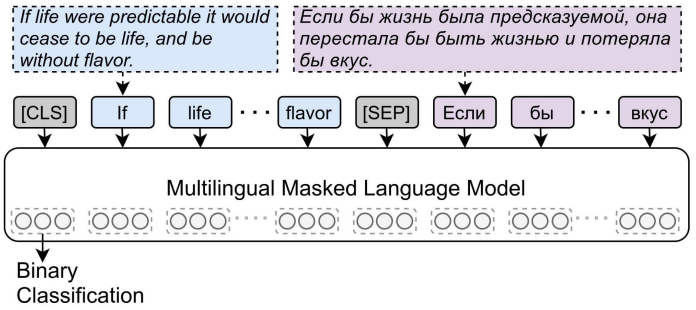

Рисунок 4 – Работа модели бинарной классификации является ли одно предложение переводом другого, которая используется на этапе детального анализа.

Если для определенного предложения из анализируемого документа процесс классификации выдает несколько возможных предложений-переводов, в таком случае переводом считается только то предложение-кандидат, в паре с которым модель выдала наибольшее значение.

Для реализации процесса бинарной классификации было решено использовать многоязычные языковые модели основанные на архитектуре “трансформер”. Выбор модели осуществлялся на основе сравнения 6 подобных моделей для пар предложений на 10 языках индо-европейской семьи: армянский, русский, английский, французский, немецкий, сербский, шведский, голландский, испанский, чешский.

Модели были протестированы на существующих для пар предложений русского и английского языках тестовых корпусах Neg-1 и Neg-2, а также на сгенерированной сложной тестовой выборке.

Задача классификации, является ли одно предложение переводом другого, близка к задаче определения перефразирований между парой предложений (перефразирование - это фактически перевод в тот же язык), то для генерации сложной тестовой выборки возникла идея использования корпуса Microsoft Research Paraphrase Corpus (MRPC). В рамках данного корпуса содержатся

примеры отмеченные меткой “*не перефразирование*”, которые, однако, семантически близки друг другу, но при этом лексически далеки. Тестовые данные были сгенерированы для 10 языков в паре с английским с использованием машинного перевода поверх корпуса MRPC.

Таблица 2 – Результаты косинусной близости пар предложений негативных примеров существующих и сгенерированных наборов данных, где векторизация предложений произведена 5 разными моделями.

Набор Данных	Языки	MUSE	MPnet	MiniLM	LaBSE	LASER
Сгенерированные на основе MRPC	EN - ES	0,71	0,76	0,74	0,75	0,79
	EN - PT	0,71	0,76	0,74	0,74	0,79
	EN - FR	0,71	0,76	0,74	0,74	0,79
	EN - RU	0,70	0,74	0,73	0,73	0,78
	EN - SR	-	0,74	0,71	0,73	0,78
	EN - CS	-	0,76	0,74	0,74	0,78
	EN - DE	0,70	0,76	0,74	0,74	0,79
	EN - NL	0,71	0,76	0,74	0,74	0,79
	EN - SV	-	0,76	0,74	0,74	0,79
	EN - HY	-	0,73	0,71	0,74	0,80
Neg-1	EN - RU	0,58	0,63	0,62	0,62	0,71
SemEval	EN - ES	0,64	0,66	0,64	0,69	0,74
	EN - AR	0,63	0,65	0,63	0,68	0,73
	EN - NL	0,64	0,67	0,64	0,69	0,74
	EN - IT	0,65	0,66	0,64	0,69	0,73
	EN - DE	0,64	0,66	0,64	0,69	0,73
	EN - FR	0,64	0,66	0,64	0,69	0,74
	EN - TR	0,63	0,67	0,65	0,69	0,74

Процесс генерации тестовых данных на основе MRPC описан на рис. 5. Для каждой пары предложений из корпуса одно из этих предложений переводится с помощью Google Translate, на один из 10 рассматриваемых языков помимо английского. Далее каждое переведенное предложение спаривается с предложением, которое являлось парой для его оригинала. Однако, прежде чем спариваться с соответствующими предложениями на английском, во избежание получения примеров, на которые влияла точность машинного перевода, переведенное предложение вместе с его оригиналом подавалось в систему оценки рисков перевода ModelFront. Данная система принимает на вход два параллельных

текста, на выходе выдает процент риска того, что перевод был произведен плохо. Те пары, для которых система ModelFront выдавала процент риска выше 50%, были отфильтрованы и не входили в конечные тестовые выборки.

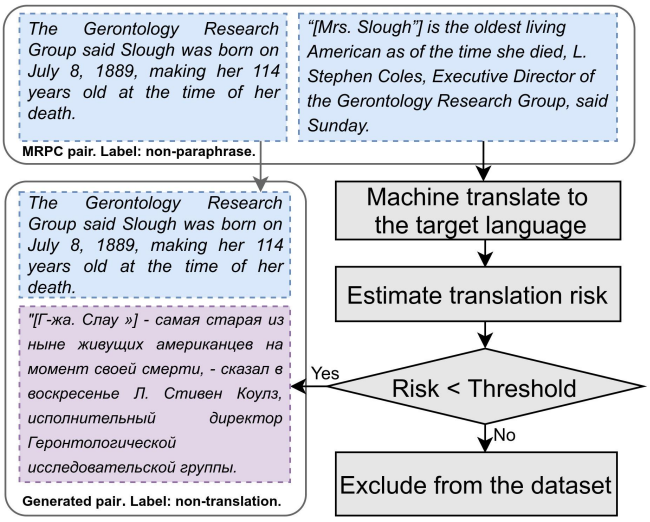

Рисунок 5 – Процесс генерации сложной тестовой выборки на основе перевода предложений из набора перефразирований MRPC.

Учитывая показанные в табл. 2 результаты, можно утверждать, что получаемые при генерации тестовых наборов данных негативные примеры являются семантически более близкими, чем негативные пары встречающиеся в других существующих наборах данных.

Таблица 3 – Количество тестовых выборок, на которых модель достигает лучших (Топ 1) и вторых лучших (Топ 2) результатов. Средний ранг обозначает среднее место которое занимали модели по всем рассматриваемым тестовым выборкам.

Модели	Топ 1	Топ 2	Средний Ранг
mBERT	0/12	8/12	2,6
mDistilBERT	0/12	0/12	4,9
XLM-RoBERTa	12/12	0/12	1,0
MiniLM	0/12	0/12	4,4
MPNet	0/12	1/12	4,2
DistilSBERT	0/12	3/12	3,8

В табл. 3 представлено, какая из сравниваемых моделей для скольких тестовых выборок показывала лучший и второй лучший результаты, а также среднее место, которое занимали модели по всем тестовым выборкам. Исходя из

данной статистики, можно сделать вывод, что лучше всех себя показала модель основанная на XLM-RoBERTa. В итоге на этапе детального анализа в рамках данной работы использовалась модель основанная на XLM-RoBERTa.

В рамках разработки системы обнаружения межъязыковых заимствований встает вопрос уязвимости моделей, используемых в алгоритме, к состязательным атакам, с помощью которых возможно “обмануть” алгоритм. Для нахождения уязвимостей моделей в рамках данного раздела также был *разработан метод генерации состязательных атак “черного ящика” на языковые модели бинарной классификации*, обходящий существующие методы по количеству успешных атак и семантической близости состязательных примеров к их оригиналам. Метод основан на произведение изменений как на уровне букв, так и на уровне слов.

Изменения на уровне букв производились с учетом расположения изменяемых букв на клавиатуре, а также с учетом разбиения на токены с помощью WordPiece токенизации. Каждое оригинальное слово, и его кандидаты, подвергались процессу WordPiece токенизации, после чего кандидаты отфильтровывались способом “минимального пересечения токенов”, в рамках которого считалось пересечение между токенами оригинального слова и токенами всех кандидатов, и в качестве кандидатов оставлялись только те измененные слова, которые имели минимальное значение пересечения с оригинальным.

Изменения на уровне слов производились с использованием в качестве генератора синонимов модели ChatGPT. Использование данной модели открывает возможности к получению синонимов практически для любого слова, в том числе и для слов отсутствующих в существующих словарях синонимов. Для сохранения максимальной естественности слов, поверх сгенерированных синонимов исправлялись их окончания тремя способами: с использованием морфологических анализаторов, с маскировкой последних частей слов и предсказании маскированного токена с помощью BERT-основанных моделей, и с изначальной настройкой ChatGPT, позволяющей генерировать синонимы с правильными окончаниями.

Итоговый метод генерации состязательных атак комбинировал в себе изменения на уровне слов и на уровне букв.

Алгоритм: Минимизация Изменений
Вход: *оригинальныйТекст*, *атакованныйТекст*, *изменения*
Выход: *атакованныйТекст* с минимальным количеством изменений
оригинальныйТекст: оригинальный текст из набора данных
атакованныйТекст: измененный текст вынуждающий модель ошибаться
изменения: информация о всех изменениях (слово и его изменение)

```
1: procedure МинимизироватьИзменения
2:     урон = [ ]
3:     временныйТекст = оригинальныйТекст
4:     for each слово in изменения do
5:         оценка = предсказание(временныйТекст)
6:         временныйТекст = заменить(временныйТекст, слово)
7:         урон ← |предсказание(временныйТекст) - оценка|
8:     Отсортировать изменения по урону
9:     новыеИзменения = [ ]
10:    while изменения ≠ новыеИзменения do
11:        новыеИзменения = изменения
12:        изменения = перевернуть(изменения)        // перевернуть - расставить
   элементы в обратном порядке
13:        for each слово in изменения do
14:            временныйТекст = вернутьОригинальноеСлово(атакованныйТекст,
   слово)
15:            if успешнаяАтака(временныйТекст) then
16:                атакованныйТекст = временныйТекст
17:                удалить слово из изменения
18:        return атакованныйТекст
```

Алгоритм 1: Псевдокод процесса минимизации изменений после совершения успешной атаки.

Для минимизации количества слов, которые подвергаются изменениям, и максимального сохранения семантической близости к оригинальным примерам был предложен алгоритм минимизации изменений (алгоритм 1).

Таблица 4 – Средние значения метрик по наборам данных IMDB, MR, YELP для представляемых методов и лучших существующих методов.

	Представляемый	**PWWS**	**TextFooler**	**Tampers**	**Bert-attack**	**TextBugger**
AR↑	**96,19%**	91,31%	94,56%	95,35%	83,06%	78,62%
PR↓	5,83%	8,79%	12,65%	**4,43%**	10,56%	22,44%
USE↑	**0,95**	0,92	0,90	0,93	0,88	0,93
Lev.↓	**18,33**	33,93	52,71	35,25	118,29	35,48

Результаты сравнения существующих методов с лучшим из представляемых на популярных наборах данных описаны в табл. 4. Для каждого из методов генерации состязательных атак были подсчитаны следующие метрики: процент успешных атак (AR), процент измененных слов (PR), дистанция Левенштейна (Lev.), семантическая близость подсчитанная с помощью Universal Sentence Encoder (USE).

Для задачи классификации перевода между двумя предложениями с помощью нового алгоритма генерации атак на уязвимость были проверены две модели: XLM-RoBERTa и DistilSBERT (табл. 5).

Таблица 5 – Значение метрик атакуемых моделей для русско-английского набора Neg-1 (изменения производились над предложением на русском языке).

Model	AR	PR	USE	Lev.
mBERT	**90,92%**	14,29%	0,91	12,09
mDistilBERT	**74,37%**	16,10%	0,90	13,59
XLM-RoBERTa	**93,89%**	13,07%	0,93	13,30
MiniLM	**65,64%**	14,17%	0,92	13,20
MPNet	**70,08%**	14,30%	0,92	12,74
DistilSBERT	**81,60%**	12,30%	0,93	9,64

Исходя из всех полученных результатов была *разработана методика выбора модели для этапа детального анализа*, состоящая из четырех этапов:

1. Генерация сложных тестовых выборок.
2. Проверка качества модели на сгенерированных выборках.
3. Проверка моделей на устойчивость к новому методу генерации состязательных атак.
4. Выбор модели в зависимости от степени угрозы подвержения состязательным атакам.

В данной главе представлены новые методы решения для каждого из двух основных этапов обнаружения межъязыковых заимствований, применимые для большого количества языков. В рамках алгоритм использующегося на этапе извлечения кандидатов был представлен новый метод генерации словаря “межъязыковых синонимов” позволяющего достичь высоких показателей метрики полноты всей системы. Дополнительно, в рамках решения задачи детального анализа, была разработана методика выбора моделей для данного этапа учитывающая угрозу возможности осуществления состязательных атак. Также представлен новый алгоритм генерации состязательных атак “черного ящика” на языковые модели бинарной классификации.

В **третьей главе** рассматриваются различные тестовые корпусы обнаружения межъязыковых заимствований для исследуемых языков. Производится оценка работы представляемого алгоритма и сравнение его точности работы с точностью других подобных алгоритмов.

В разделе 3.1 представляются различные метрики оценки качества обнаружения заимствований, такие как: микро и макро точность и полнота, granularity, plagdet.

В разделе 3.2 описываются различные существующие тестовые наборы для проверки точности методов обнаружения межъязыковых заимствований. Представляются существующие корпусы для пар языков: английский-русский, английский-французский, английский-испанский. В дополнение к этому также были сгенерированы 5 новых корпусов для следующих пар языков: английский-русский, английский-французский, английский-испанский, английский-армянский, английский-немецкий.

Генерация новых корпусов производилась с использованием алгоритма представленного в статье Bakhteev et al., 2019[2]. В каждом корпусе содержалось 400 анализируемых документов, в которых содержалось от 0% до 80% заимствований из от 1 до 10 документов. В проверяемой коллекции содержалось 120 000 статей, извлеченных из английской Wikipedia.

В разделе 3.3 производится оценка разработанного метода, а также его сравнение с другими SOTA алгоритмами обнаружения межъязыковых заимствований, с использование корпусов представляемых в разделе 3.2.

Далее представлены результаты сравнения на всех существующих корпусах: в табл. 6 сравнение для пары английский-русский, в табл. 7 для пары английский-французский, в табл. 8 для пары испанский-английский. В табл. 9 представлены результаты полученные для 5 пар языков на новосгенерированных тестовых корпусах.

---

[2] Bakhteev O. et al. CrossLang: the system of cross-lingual plagiarism detection // Workshop on Document Intelligence at NeurIPS. – 2019.

Таблица 6 – Сравнение представляемого алгоритма с алгоритмом, получившим лучшие результаты на англйско-русском корпусе CrossLang.

	CrossLang		
	Полнота	**Точность**	**F1**
Представляемый алгоритм	0,77	**0,86**	**0,81**
Bakhteev et al., 2019	**0,79**	0,83	0,80

Таблица 7 – Сравнение F1-мер представляемого алгоритма с алгоритмом достигающим лучших результатов на корпусах представленных в Ferrero et al. для пары языков английский-французский.

	JRC-Aquis	**APR**	**TALN**
Представляемый алгоритм	71,80 $\pm$ 0,444	**96,67 $\pm$ 0,387**	**89,72 $\pm$ 0,474**
Ferrero et al., 2017	72,70 $\pm$ 1,446	78,91 $\pm$ 1,005	80,89 $\pm$ 0,944

Таблица 8 – Сравнение представляемого алгортима с алгоритмом, получившим лучшие результаты на испанско-английской части корпуса PAN-PC-11.

	PAN-PC-11 ES-EN	
	Полнота	**Точность**
Представляемый алгоритм	**0,79**	**0,85**
Roostaee et al., 2020	0,75	0,79

Таблица 9 – Результаты показанные представляемым алгоритмом на сгенерированном наборе данных для 5 пар языков.

	Сгенерированный набор данных		
Языки	**Полнота**	**Точность**	**F1**
EN - HY	0,72	0,73	0,73
EN - RU	0,81	0,82	0,81
EN - ES	0,90	0,86	0,88
EN - FR	0,88	0,81	0,84
EN - DE	0,71	0,64	0,67

В данной главе описаны различные тестовые корпусы обнаружения межъязыковых заимствований для исследуемых языков. На основе результатов полученных на описанных корпусах можно утверждать, что представленный в рамках второй главы метод обнаружения межъязыковых заимствований достигает

сравнимых, а для некоторых тестовых корпусов лучших, результатов относительно других существующих методов.

В рамках **четвертой главы** проводится сравнительный анализ представляемого в работе метода с методом, представленным компанией “Антиплагиат.ру” в виде, описанном в Bakhteev et al., 2019, а также рассматриваются различные способы их слияния для нивелирования слабых мест друг друга, тем самым увеличив точность нахождения межъязыковых заимствований.

В разделе 4.1 описываются пять этапов работы алгоритма представленного компанией “Антиплагиат.ру”. Алгоритм основан на использовании машинного перевода и последующем моноязычном поиске заимствований. Дополнительно, учитывается неоднозначность переводов с использованием синонимических групп и векторной модели представления текстовых фрагментов.

В разделе 4.2 представляется процесс генерации нового тестового набора, в котором содержались тексты на 4 языках: Армянский, Русский, Английский и Испанский. Генерация производилась на основе алгоритма, представленного в статье Bakhteev et al., 2019, однако в данном случае тексты на одном языке могли содержать в себе заимствования из нескольких языков. На данном наборе и были произведены эксперименты по сравнению представляемого метода с методом “Антиплагиат.ру”, а также были произведены тесты по слиянию двух методов.

В разделе 4.3 представлены два варианта слияния двух методов: комбинированное слияние и последовательное слияние. Исходя из схожести структур двух методов, был предложен вариант комбинированного слияния, замены различных этапов одного метода теми же этапами второго метода, получая таким образом различные комбинации. Таким образом, было протестировано 8 комбинаций слияния. В конце каждой из комбинаций использовался алгоритм постобработки результатов, который сшивал стоящие рядом заимствованные фрагменты. Комбинированное слияние не дало улучшения по сравнению с работой изначальных методов.

Исходя из того, что алгоритм “Антиплагиат.ру” на сгенерированной выборке достигал высоких значений точности, а представляемый в работе алгоритм достигал высоких значений полноты (табл. 10) было решено последовательно слить методы дополнив представляемый метод этапом детального анализа из алгоритма “Антиплагиат.ру”, добавив его сразу после прохождения всех этапов представляемого алгоритма. В конце работы последовательного слияния также использовался алгоритм постобработки. Полученные в рамках последовательного слияния, усредненные по языкам фрагментов-источников, результаты представлены в табл. 11. Исходя из данных результатов, можно увидеть, что алгоритм слияния двух методов привел к значительному приросту метрики Plag Score для всех пар языков.

Таблица 10 – Результаты представляемого алгоритма и алгоритма “Антиплагиат.ру” на новосгенерированном тестовом наборе.

Комбинация	**Язык**	**Точн.**	**Полн.**	**Gran.**	**F1**	**Plag Score**
“Антиплагиат.ру”	Ru	**0.88**	0.30	1.0	0.45	**0.45**
	Es	**0.93**	0.42	1.0	0.58	**0.58**
	En	**0.95**	0.30	1.0	0.46	0.46
	Hy	**0.77**	0.05	1.0	0.09	0.09
Представляемый алгоритм	Ru	0.45	**0.82**	1.54	**0.58**	0.43
	Es	0.46	**0.83**	1.63	**0.59**	0.42
	En	0.61	**0.87**	1.60	**0.72**	**0.52**
	Hy	0.37	**0.66**	1.41	**0.48**	**0.37**

Таблица 11 – Результаты последовательного слияния двух методов.

Слияние	**Язык**	**Точн.**	**Полн.**	**Gran.**	**F1**	**Plag Score**
Последовательное слияние	Ru	0.97	0.66	1.0	0.78	**0.78**
	Es	0.98	0.71	1.0	0.83	**0.83**
	En	0.99	0.69	1.0	0.81	**0.81**
	Hy	0.90	0.45	1.0	0.60	**0.60**

На основе результатов, полученных на сгенерированном тестовом корпусе, можно утверждать, что последовательное слияние метода представляемого в рамках работы и метода представленного компанией “Антиплагиат.ру” (Bakhteev et al., 2019) повышает точность обнаружения межъязыковых заимствований.

В **заключении** сформулированы основные результаты работы:

1. Разработан новый метод обнаружения межъязыковых заимствований, превосходящий по эффективности существующие. Дополнительно, метод применим к задаче обнаружения межъязыковых заимствований в текстах малоресурсных языков.
2. Разработан новый метод генерации словаря “межъязыковых синонимов”, позволяющего достичь высоких показателей метрики полноты для этапа извлечения кандидатов в задаче обнаружения межъязыковых текстовых заимствований.
3. Разработан новый метод генерации состязательных атак “черного ящика” на языковые модели бинарной классификации, превосходящий по доле успешных атак, а также по дистанции Левенштейна и семантической близости все существующие аналоги.
4. Разработана методика выбора языковой модели для этапа детального анализа учитывающая угрозу возможности осуществления состязательных атак.

Публикации автора по теме диссертации

1. Аветисян К.И., Асатрян А.А., Гукасян Ц.Г., Ешилбашян Е.М., Маладжян А.А., Недумов Я.Р., Скорняков К.А., Тигранян Ш.Т., Турдаков Д.Ю. «Sieve» / Свидетельство о государственной регистрации программы для ЭВМ, рег. №2021668213 от 11.11.2021 - Российская Федерация, 2021.
2. Avetisyan K., Gritsay G., Grabovoy A. Cross-Lingual Plagiarism Detection: Two Are Better Than One //Programming and Computer Software. – 2023. – Т. 49. – №. 4. – С. 346-354. - Scopus.
3. Ghukasyan T., Yeshilbashyan Y., Avetisyan K. Subwords-only alternatives to fastText for morphologically rich languages //Programming and Computer Software. – 2021. – Т. 47. – С. 56-66. - Scopus.
4. Ter-Hovhannisyan T., Avetisyan K. Transformer-Based Multilingual Language Models in Cross-Lingual Plagiarism Detection //2022 Ivannikov Memorial Workshop (IVMEM). – IEEE, 2022. – С. 72-80. - Scopus.
5. Avetisyan K. Dialects Identification of Armenian Language //Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference. – 2022. – С. 8-12.
6. Avetisyan K., Malajyan A., Ghukasyan T. A Simple and Effective Method of Cross-Lingual Plagiarism Detection //arXiv preprint arXiv:2304.01352. – 2023.