ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ,
ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Պետրոսյան Գևորգ Արմենի

ՏԵՔՍՏԻ ԻՆՔՆԱՏԻՊՈՒԹՅԱՆ ԱՍՏԻՃԱՆԻ ՈՐՈՇՄԱՆ ԲԱԶՄԱԼԵԶՈՒ ՀԱՄԱԿԱՐԳԻ ՄՇԱԿՈՒՄԸ

Ե.13.02 «Ավտոմատացման համակարգեր» մասնագիտությամբ
տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի
հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2026

МИНИСТЕРСТВО ОБРАЗОВАНИЯ, НАУКИ, КУЛЬТУРЫ И СПОРТА
РЕСПУБЛИКИ АРМЕНИЯ

НАЦИОНАЛЬНЫЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ АРМЕНИИ

Петросян Геворг Арменович

РАЗРАБОТКА МНОГОЯЗЫЧНОЙ СИСТЕМЫ ОПРЕДЕЛЕНИЯ СТЕПЕНИ УНИКАЛЬНОСТИ ТЕКСТА

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата
технических наук по специальности 05.13.02-
“Системы автоматизации”

Ереван 2026

Ատենախոսության թեման հաստատվել է Հայաստանի ազգային պոլիտեխնիկական համալսարանում (ՀԱՊՀ):

Գիտական ղեկավար՝	տ.գ.դ. Ռուստամ Ռաֆիկի Սահակյան
Պաշտոնական ընդդիմախոսներ՝	տ.գ.դ. Աշոտ Գևորգի Հարությունյան ֆ-մ.գ.թ. Համլետ Ցոլակի Հակոբյան
Առաջատար կազմակերպություն՝	ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ

Ատենախոսության պաշտպանությունը կայանալու է 2026թ. հուլիսի 10-ին, ժամը 11^{00}-ին, ՀԱՊՀ-ում գործող «Կառավարման և ավտոմատացման» 032 մասնագիտական խորհրդի նիստում (հասցեն՝ 0009, Երևան, Տերյան փ., 105, 17 մասնաշենք):

Ատենախոսությանը կարելի է ծանոթանալ ՀԱՊՀ–ի գրադարանում:

Սեղմագիրն առաքված է 2026թ. հունիսի 9-ին:

032 Մասնագիտական խորհրդի գիտական քարտուղար, տ.գ.թ. Անուշ Վազգենի Մելիքյան

Тема диссертации утверждена в Национальном политехническом университете Армении (НПУА)

Научный руководитель:	д.т.н. Рустам Рафикович Саакян
Официальные оппоненты:	д.т.н. Ашот Геворкович Арутюнян к.ф-м.н. Гамлет Цолакович Акопян
Ведущая организация:	Институт проблем информатики и автоматизации НАН РА

Защита диссертации состоится 10-го июля 2026г. в 11^{00} ч. на заседании Специализированного совета 032 — "Управления и автоматизации", действующего при Национальном политехническом университете Армении, по адресу: 0009, г. Ереван, ул. Теряна, 105, корпус 17.

С диссертацией можно ознакомиться в библиотеке НПУА.

Автореферат разослан 9-го июня 2026 г.

Ученый секретарь Специализированного совета 032 к.т.н. Ануш Вазгеновна Меликян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Современный уровень развития информационных технологий и огромный объем цифровой информации, доступной в интернет-среде, обеспечивают современному исследователю практически неограниченный доступ к источникам на любом языке. Однако это создает риски, связанные с несанкционированными заимствованиями и обеспечением академической добросовестности.

В указанных условиях определение степени уникальности текста является крайне важной задачей. Проблема определения степени уникальности текста напрямую связана с поиском в нем заимствований на языке оригинала (одноязычных) и на иностранных языках (межъязыковых). Особенно для малоресурсных (инструменты обработки естественного языка, модели, текстовые корпуса, электронные словари и т.п.) языков, таких как армянский, при определении степени уникальности текста задача выявления межъязыковых заимствований порой имеет более важное значение, чем обнаружение одноязычных заимствований. Это связано с тем, что для малоресурсных языков заимствование из иностранных источников (в частности, из англоязычных и русскоязычных источников в случае армянского языка) является более распространенным явлением.

Судя по количеству статей, опубликованных в последние годы по задаче выявления заимствований в армяноязычных текстах и смежным задачам (морфологический анализ, векторные представления слов, стилометрический анализ и др.), можно сделать вывод, что данная область активно развивается, в том числе с применением методов машинного и глубокого обучения.

Следует отметить, что в большинстве научно-исследовательских и образовательных организаций Республики Армения отсутствуют специализированные автоматизированные программы для определения степени уникальности научных и научно-квалификационных работ, написанных на армянском языке. В таких случаях они вынуждены либо проверять работы вручную, что является трудоемким и неэффективным занятием, учитывая огромный объем доступной информации, либо использовать системы, разработанные для других языков (например, Turnitin, Антиплагиат), что также неэффективно, так как такие системы не учитывают морфологические и синтаксические особенности армянского языка, либо из-за отсутствия специализированных программ научные работы в основном пишутся и публикуются не на армянском, а на языках, для которых существуют автоматизированные системы для определения степени уникальности текста (например, на русском или английском).

С учетом вышеизложенного разработка автоматизированной системы определения степени уникальности армяноязычного текста и выявления содержащихся в нем одноязычных и межъязыковых текстовых заимствований из англоязычных и русскоязычных источников является актуальной и значимой научно-практической задачей.

Цель и задачи исследования. Целью данной работы является разработка автоматизированной системы определения степени уникальности армяноязычного текста и выявления содержащихся в нем одноязычных и межъязыковых заимствований из англоязычных и русскоязычных источников.

Для достижения указанной цели были поставлены и решены следующие задачи:

- Проведены исследование и сравнительный анализ существующих методов и алгоритмов поиска одноязычных и межъязыковых заимствований, а также возможных вариантов применения их модифицированных версий.
- Разработан и протестирован метод поиска источников межъязыковых заимствований (поиск кандидатов), содержащихся в армяноязычных текстах.

- Разработан и протестирован метод поиска межъязыковых заимствований (детальный анализ), содержащихся в армяноязычных текстах.
- Разработан метод структурного анализа армяноязычных текстов.
- Спроектированы двухуровневая архитектура и функциональная модель предлагаемой системы.
- Выполнена программная реализация предлагаемой системы.

Объект исследования. Разработка автоматизированной системы определения степени уникальности армяноязычного текста в многоязычной среде.

Предмет исследования. Методы, модели и алгоритмы поиска одноязычных и межъязыковых заимствований.

Методы исследования. В работе были использованы методы обработки естественного языка, статистические методы ранжирования, современные архитектуры глубокого обучения, в частности модели на основе трансформеров, экспериментальные методы оценки качества, методы IDEF0-моделирования и средства программной реализации веб-систем на основе Django.

Научная новизна диссертационной работы заключается в следующем:

- Разработан метод поиска кандидатов межъязыковых текстовых заимствований, содержащихся в армяноязычных текстах, на основе разметки частей речи.
- Разработан метод поиска одноязычных заимствований, содержащихся в армяноязычных текстах, на основе марковских цепей.
- Разработан метод детального анализа межъязыковых заимствований, содержащихся в армяноязычных текстах, на основе многоязычной трансформерной модели.
- Создан корпус параллельных и перефразированных пар предложений для армяно–английской и армяно–русской языковых пар, включающий трудные негативные примеры (hard negatives), с целью дообучения моделей в задаче межъязыкового семантического сопоставления.
- Разработан способ перефразирования армяноязычных предложений с использованием многоуровневой системы фильтрации, обеспечивающий семантическое сходство и лексическое разнообразие пар. Перефразированные предложения на армянском языке использовались с целью формирования наборов пар перефразирований для армяно-английской и армяно-русской языковых пар.
- Выполнено дообучение многоязычной трансформерной модели LaBSE с использованием разработанного корпуса параллельных и перефразированных пар, что позволило улучшить качество межъязыкового семантического сопоставления предложений на армянском, русском и английском языках.
- Адаптирован набор тестовых данных для армяно-английской и армяно-русской языковых пар путем перевода существующего набора данных, предназначенного для оценки методов решения обеих подзадач обнаружения межъязыковых текстовых заимствований.

Практическая значимость работы. В разработанной автоматизированной системе реализована двухуровневая архитектура поиска одноязычных и межъязыковых заимствований, включающая в себя этапы быстрого извлечения кандидатов и детального анализа с использованием современных моделей на основе трансформеров. Это обеспечивает производительность и высокую точность в выявлении прямых заимствований, перефразирований и переводов. Централизация хранения данных, поддержка асинхронной обработки ресурсоемких задач и интеграция научных работ из внешних источников значительно расширяют область применения системы по сравнению с локальными системами. Благодаря удобному веб-интерфейсу разработанная система может быть использована в

научно-исследовательских и образовательных организациях Республики Армения для автоматизации проверки степени уникальности научных работ.

Основные положения, выносимые на защиту:

- Метод поиска кандидатов межъязыковых текстовых заимствований.
- Многоязычная трансформерная модель, дообученная на корпусе параллельных и перефразированных предложений для армяно-английской и армяно-русской языковых пар.
- Метод поиска одноязычных заимствований на основе марковских цепей.
- Двухуровневая архитектура для определения степени уникальности армяноязычных текстов в многоязычной среде.
- Результаты экспериментальных исследований, демонстрирующие эффективность разработанных методов по сравнению с существующими подходами.
- Автоматизированная веб-система определения степени уникальности армяноязычных текстов в многоязычной среде.

Достоверность научных положений. Достоверность научных положений подтверждается результатами программной реализации, представленной в диссертации, комплексными экспериментальными исследованиями и оценками, полученными с помощью международно признанных метрик, а также математическими обоснованиями.

Внедрение. Результаты диссертации, в частности разработанные методы и система определения степени уникальности армяноязычного текста, внедрены в теоретические и практические занятия курса «Основы искусственного интеллекта» кафедры «Математика и информатика» Ванадзорского государственного университета имени О. Туманяна.

Апробация результатов работы. Основные теоретические и практические результаты исследования были представлены на:

- ежегодной конференции НПУА (Ереван, Армения, 2025 г.);
- научных семинарах кафедры «Математика и информатика» ВГУ (Ванадзор, Армения, 2025, 2026 гг.);
- научном семинаре кафедры «Математика, физика и информационные технологии» ШГУ (Гюмри, Армения, 2026 г.);
- научных семинарах кафедры «Информационные технологии и автоматизация» НПУА (Ереван, Армения, 2023-2025 гг.).

Публикации. Основные положения, представленные в диссертации, обобщены в девяти (9) научных статьях, две из которых без соавторов, а две - в научной базе данных “SCOPUS” (“СКОПУС”). Список статей приведен в конце автореферата.

Структура и объем работы. Диссертация состоит из введения, пяти глав, основных выводов, списка литературы, включающего 135 наименований, и одного приложения, в котором представлен акт внедрения. Основной объем диссертации составляет 148 страниц, а вместе с приложением - 149 страниц. Диссертация написана на армянском языке.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цель и задачи работы, представлены научная новизна, практическая значимость работы, а также основные положения, выносимые на защиту. Кратко представлено содержание диссертации.

В первой главе проведен анализ предметной области обнаружения текстовых заимствований. Определены ключевые понятия - уникальность, заимствование и плагиат. Уникальность - это доля текста, не совпадающая с документами коллекции источников. Эта метрика поддается автоматическому вычислению и не оценивает намерения автора. Заимствование - это использование чужого текста в своей работе. Корректное заимствование (цитирование) является допустимой и необходимой практикой академического и научного

дискурса. Большинство научных работ обязательно опирается на предшествующие исследования. Несанкционированное заимствование (плагиат) - это классификация заимствования, требующая человеческого суждения о намерении автора и степени несоблюдения правил цитирования. Это разграничение обосновывает позиционирование разрабатываемой автоматизированной системы как инструмента поддержки принятия решений, тогда как классификация конкретного случая заимствования как плагиат требует участия эксперта.

Проведены обзор различных видов плагиата и анализ существующих коммерческих систем обнаружения заимствований, ориентированных преимущественно на другие языки. Несмотря на существование международных систем, поддерживающих несколько языков, эти системы, как правило, не учитывают морфологические и синтаксические особенности каждого из поддерживаемых языков.

Представлен систематический обзор методов обнаружения заимствований, особое внимание уделено анализу методов обнаружения межъязыковых заимствований. Отдельно рассмотрены работы, посвященные обнаружению текстовых заимствований для армянского языка и смежным задачам.

Отдельный раздел главы посвящен архитектурам глубокого обучения для семантического анализа текстов. Представлены современные нейросетевые подходы к построению семантических представлений текста, обеспечивающие выявление смысловой близости фрагментов на разных языках без перевода. Эти модели существенно улучшают эффективность обнаружения перефразирований.

В главе также проведен анализ типологических особенностей армянского языка применительно к задаче обнаружения межъязыковых заимствований, рассмотрены ограничения, такие как дефицит размеченных корпусов, для обучения и тестирования методов и моделей.

Во второй главе представлены концептуальные и структурные решения предлагаемой автоматизированной системы.

В первом разделе представлена двухуровневая архитектура предлагаемой системы (рис. 1).

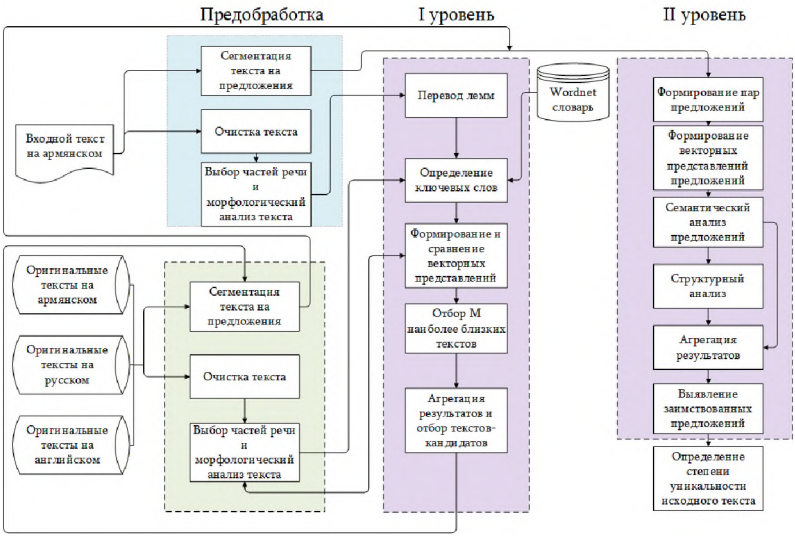

Рис. 1. Двухуровневая архитектура предлагаемой системы определения степени уникальности армяноязычного текста в многоязычной среде

Предварительная обработка организована в виде двух параллельных потоков. Первый поток обрабатывает исходный текст на армянском языке и включает токенизацию и морфологический анализ. Второй поток работает с многоязычными корпусами на армянском, русском и английском языках и проводит тот же лингвистический анализ.

Первый уровень реализует первичный отбор кандидатов. На этом уровне последовательно выполняются: выбор ключевых слов (отдельных частей речи), их ранжирование, построение векторных представлений и поиск ближайших текстов по многоязычному корпусу. Второй уровень осуществляет семантический анализ текстов, отобранных на первом уровне, с целью выявления заимствованных фрагментов. Здесь используется многоязычная трансформерная модель, которая проецирует предложения на разных языках в общее семантическое векторное пространство, а степень их семантической близости определяется с помощью косинусного сходства между их векторными представлениями. Также проводится структурный анализ исключительно армяноязычных текстов:

$$Sim\left(s_i^{hy},\ s_j^{en/ru}\right) = \Phi\left(s_i^{hy},\ s_j^{src}\right) = \cos\left(emb\left(s_i^{hy}\right), emb\left(s_j^{src}\right)\right), \quad (1)$$

$$Sim\left(s_i^{hy}, s_j^{hy}\right) = \Phi\left(s_i^{hy},\ s_j^{src}\right) \cdot (1 + \lambda \cdot \Psi(W_i^{hy}, W_j^{src}). \quad (2)$$

Затем степень уникальности входного армяноязычного текста H определяется по следующей формуле:

$$U(H) = 1 - \frac{\sum_{i=1}^{n} L_i \cdot Q_i}{\sum_{i=1}^{n} L_i}, \quad (3)$$

где L_i — это количество слов в i-м предложении, n - количество предложений в тексте H, а Q_i - бинарный индикатор наличия заимствования:

$$Q_i = \begin{cases} 1, & \text{если для } i-\text{го предложения было обнаружено} \\ & \text{совпадение в исходных текстах,} \\ 0, & \text{в противном случае.} \end{cases} \quad (4)$$

Во втором разделе главы построена функциональная модель с использованием стандарта IDEF0 технологии SADT. На уровне A0 представлена система определения степени уникальности армяноязычного текста в многоязычной среде в качестве единого функционального блока (рис. 2).

Рис. 2. Контекстная диаграмма предлагаемой системы (уровень A0)

Входной поток данных включает армяноязычный проверяемый документ для рассмотрения. Управляющие воздействия включают грамматические правила, параметры метода поиска кандидатов, правила предварительной обработки текстов, требования к текстам и гиперпараметры модели на основе трансформеров.

Механизмы реализации включают специализированные инструменты обработки естественного языка для трех языков, метод поиска кандидатов, метод структурного анализа армяноязычного текста, модель на основе трансформеров, базу данных локальных файлов, базу данных внешних файлов и оператор.

Выходные потоки включают итоговый отчет о проверке с цветовой разметкой заимствованных фрагментов непосредственно в тексте документа и метрику уникальности.

Уровень A1 - это первый уровень декомпозиции контекстной диаграммы (рис. 3). Он состоит из четырех основных подсистем:

1) подготовка текста;
2) поиск кандидатов в многоязычной среде;
3) детальный анализ;
4) обобщение результатов и определение степени уникальности.

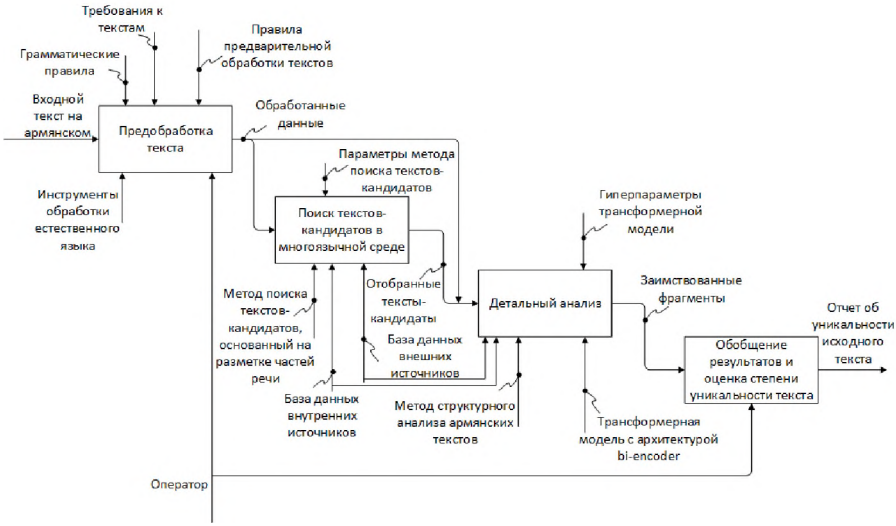

Рис. 3. Взаимодействие основных подсистем предлагаемой системы (уровень A1)

В главе также представлена диаграмма декомпозиции уровня A2.

В третьей главе представлены разработанные методы обнаружения одноязычных и межъязыковых заимствований, образующие двухуровневую архитектуру (уровень 1 - поиск кандидатов, уровень 1 - детальный анализ).

Поиск кандидатов межъязыковых текстовых заимствований на основе разметки частей речи (уровень 1)

Первичный поиск текстов или поиск кандидатов - это процесс, который помогает выделить значительно меньшее количество текстов из большого корпуса для последующего применения более детальных методов поиска заимствований, которые невозможно или нецелесообразно применять к большому количеству текстов.

Разработанный метод поиска кандидатов — CL-POSSR (Cross-Lingual Part-Of-Speech Source Retrieval) предназначен для быстрого извлечения небольшого набора кандидатов из

большой многоязычной коллекции. В основе метода лежит гипотеза о том, что основное смысловое содержание предложения несут существительные, которые при переводе заменяются своими эквивалентами, сохраняя семантику текста. За ними следуют глаголы и прилагательные. Метод включает следующие этапы обработки текстов: предварительная обработка, удаление стоп-слов, разметка частей речи, распознавание именованных сущностей и лемматизация. Для отбора топ-N слов представлены три варианта метрики: средняя частота, TF-IDF и дисперсия частотности («Term-frequency variance»). На основе отобранных слов тексты преобразуются в числовые векторы, близость которых определяется с помощью косинусного сходства.

Расширенный метод поиска кандидатов (уровень 1)

Для улучшения разработанного метода поиска кандидатов предложен его расширенный вариант - CL-POSSWR (Cross-Lingual Part-of-Speech Sentence-to-Window Retrieval). Ключевым отличием от исходного метода является переход от сравнения текстов целиком к локализованной схеме сопоставления. Проверяемый текст разбивается на отдельные предложения, а оригинальные тексты — на перекрывающиеся окна (последовательность слов) фиксированной длины. Это позволяет обнаруживать тексты, содержащие частичные и фрагментарные заимствования.

Для каждого предложения проверяемого текста извлекаются лингвистические признаки: существительные, глаголы, прилагательные, именованные сущности и числа. Армянские леммы переводятся на язык индексируемой коллекции, после чего формируется взвешенный поисковый запрос, в котором каждая группа признаков получает собственный вес. Наибольшим весом обладают именованные сущности, существительные, числа, далее следуют прилагательные и глаголы.

Поиск релевантных окон осуществляется с использованием инвертированного индекса и функции ранжирования BM25. Итоговая оценка документа-кандидата D_j вычисляется путем двухуровневой агрегации: сначала на уровне окон для каждого предложения s_i, затем на уровне всего проверяемого текста S:

$$\mu_{D_j}^{s_i} = \sum_{r=1}^{M} score_{(r)}(s_i, D_j), \tag{5}$$

$$\mu_{D_j}^{S} = \sum_{i=1}^{p} \mu_{D_j}^{s_i}. \tag{6}$$

Здесь $score_{(r)}(s_i, D_j)$ - r-е наибольшее значение BM25 среди окон текста D_j для запроса по предложению s_i, M - число учитываемых лучших окон, p - число предложений в проверяемом тексте S. Тексты-кандидаты ранжируются по убыванию $\mu_{D_j}^{S}$.

Нахождение одноязычных заимствований на основе марковских цепей

Разработано представление текста в виде вероятностного графа переходов между ключевыми леммами. Каждая уникальная лемма — вершина графа, каждая биграмма (упорядоченная пара последовательных лемм) — направленное ребро с весом, равным вероятности перехода:

$$p_{ij} = n_{ij} / \sum_k n_{ik}, \tag{7}$$

где

$$\sum_j p_{ij} = 1. \tag{8}$$

Далее выполняется оценка сходства графов путем вычисления значений PageRank для каждого узла и определения косинусного сходства между PageRank-векторами графов.

Поиск заимствований с использованием многоязычной трансформерной модели LaBSE (уровень 2)

Для семантического анализа текста на уровне предложений была использована предобученная многоязычная модель - Language-agnostic BERT Sentence Embedding (LaBSE). Важным элементом модели является механизм самовнимания (Self-Attention Mechanism), который позволяет каждому слову в последовательности учитывать контекст всех остальных слов. В общем виде механизм масштабированного скалярного внимания описывается следующим образом:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\cdot \text{V}, \tag{9}$$

где $Q \in \mathbb{R}^{n\times d_k}$ — матрица запросов; $K \in \mathbb{R}^{n\times d_k}$ — матрица ключей, а $V \in \mathbb{R}^{n\times d_v}$ — матрица значений (часто $d_k = d_v$). Трансформерная модель LaBSE состоит из 12 последовательных слоев, каждый из которых содержит 12 голов внимания. Каждое предложение, независимо от его длины и языка, с помощью модели LaBSE представляется в виде вектора размерности 768. Модель LaBSE использует механизм многоголового самовнимания. Механизмы самовнимания реализуются параллельно несколькими головками, каждая из которых может фокусироваться на различных семантических зависимостях между словами. Результаты всех головок объединяются и проецируются в выходное пространство.

Для дообучения модели был создан корпус параллельных и перефразированных пар предложений для армяно-английской и армяно-русской языковых пар.

Параллельные предложения. В качестве источника параллельных предложений были использованы коллекции из Opus (NLLB, JW300, ParaCrawl-Bonus, MultiCCAligned, QED, TED2020, NeuLab-TedTalks, Wikimedia). Поскольку параллельные предложения для выбранных языковых пар в рассматриваемых наборах данных не всегда являются точными переводами, после загрузки всех предложений они были подвергнуты многоступенчатой фильтрации для отбора только «высококачественных» пар. После фильтрации были получены два набора, содержащие 1,8 млн армяно-английской (далее — Θ_{en}) и 1 млн армяно-русской (далее — Θ_{ru}) пар предложений соответственно.

Перефразирования. Набор перефразированных пар формируется из двух источников.

а) Используются пары предложений из Θ_{en} и Θ_{ru} для создания перефразированных версий армянских предложений. Каждое армянское предложение переводится на английский, после чего с помощью модели Pegasus генерируется несколько перефразирований. Кандидаты фильтруются по нескольким критериям: семантическая близость к переведенному предложению проверяется моделью all-MiniLM-L6-v2, исключаются кандидаты с чрезмерным отклонением по длине и т.д. Итоговое перефразирование p_j^* выбирается по взвешенной формуле, балансирующей семантическую близость ($\text{sim}(e_i,p_j)$) и лексическую дивергенцию ($d_{lex}(e_i,p_j)$):

$$p_j^* = argmax_{p_j\in E_i'}\Big(\alpha \cdot d_{lex}(e_i,p_j) + (1-\alpha)\cdot \text{sim}(e_i,p_j)\Big). \tag{10}$$

б) Используются адаптированные версии готовых англоязычных наборов перефразирований — Quora Question Pairs, PAWS-X и STS-B. Для армяно–английской языковой пары одно из каждой пары предложений переводится на армянский, для армяно–русской языковой пары второе предложение переводится на русский.

Извлечение трудных негативов. Для извлечения трудных негативных примеров применяется алгоритм на основе LaBSE и FAISS. Для каждого армянского предложения извлекается 100 ближайших предложений на английском (русском) языке с помощью индекса FlatIP. Правильный перевод, как правило, находится среди них. Кандидаты трудных негативов отбираются по пороговому значению семантической близости и затем проходят

многоуровневую фильтрацию - исключаются совпадения по числам и датам, применяются метрики Жаккара по токенам и биграммам, вычисляется пересечение символьных n-грамм, нормализованное расстояние Левенштейна, BertScore, модель stsb-roberta-base в режиме cross-encoder, а также модель RoBERTa-large, дообученная на задаче NLI, - для исключения парафраз в пограничных случаях. Прошедшие фильтрацию кандидаты сортируются по убыванию семантической близости, среди которых отбираются 4 лучших.

Таким образом, извлеченные трудные негативы семантически близки к исходному армянскому предложению, но не являются ни точным, ни перефразированным переводом.

Дообучение модели. Дообучение LaBSE осуществляется с использованием функции потерь Additive Margin Softmax в стиле ArcFace. Однако вместо мега-батчевого майнинга используются внутрибатчевые негативы и заранее извлеченные трудные негативы.

Для каждой L2-нормализованной пары эмбеддингов положительного примера вводится фиксированный угловой штраф m, добавляемый к косинусному углу между эмбеддингами армянского и английского (русского) предложений. Все $\{\mathbf{e_j}\}_{j\neq i}$ внутрибатчевые негативы и $\{\mathbf{n}_i^k\}_{k=1}^4$ трудные негативы масштабируются коэффициентом s, а трудные негативы дополнительно получают коэффициент γ. Целевая функция имеет следующий вид:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{p}{p+n+\gamma\times q}\right), \tag{11}$$

где $p = e^{s\times\cos(\varphi(\mathbf{h_i},\mathbf{e_i})+m))}$, $n = \sum_{j\neq i} e^{s\times\cos(\varphi(\mathbf{h_i},\mathbf{e_j}))}$, $q = \sum_k e^{s\times\cos(\varphi(\mathbf{h_i},\mathbf{n_i^k}))}$.

Аналогичный процесс повторяется в обратном направлении - английские (русские) предложения сопоставляются с армянскими (только с внутрибатчевыми негативами), после чего градиенты обоих направлений усредняются.

Параметры дообучения. Модель была дообучена на корпусе из 100 000 пар предложений, сформированном из 60 000 параллельных и 40 000 перефразированных пар. Обучение проходило в два этапа, каждый из которых занимал одну эпоху. В обоих этапах обучения для валидации использовался набор из 3 000 пар предложений (1 000 параллельных и 2 000 перефразированных). Обучение проводилось отдельно для армяно–английской и армяно–русской языковых пар. В качестве основных метрик были использованы: Rec@k (k=1,5,10), F1 мера бинарной классификации пар, а также средние значения косинусного сходства для позитивных и негативных пар и величина разрыва между ними (margin).

Результаты дообучения для армяно-английской языковой пары. На рис. 4 представлены кривые потерь для армяно–английской языковой пары. Валидационные потери устойчиво снизились с 0,932 до 0,211 в течение двух этапов.

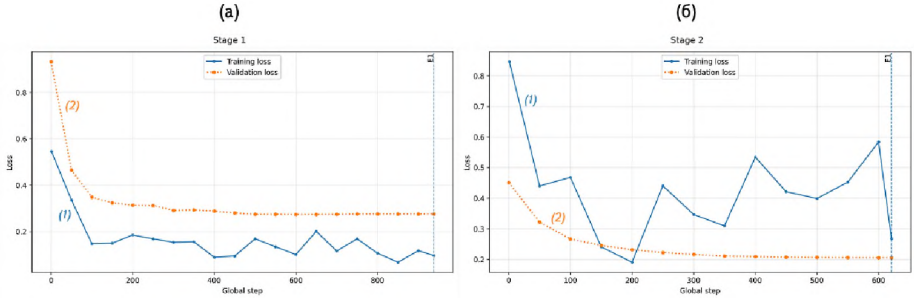

Рис. 4. Динамика потерь при обучении и валидации для армяно-английской языковой пары: 1 - кривая обучения, 2 - кривая валидации; а - этап 1, б - этап 2

На рис. 5 представлены распределения косинусного сходства для позитивных и негативных пар для трех моделей - базовая LaBSE, лучшая модель после первого этапа и лучшая модель после второго этапа. Распределение позитивных пар после второго этапа формирует выраженный пик в диапазоне 0,80…0,90, тогда как распределение негативных пар смещается в область 0,30…0,50. Разрыв между распределениями вырос с 0,190 до 0,345, что подтверждает эффективность процесса двухэтапного дообучения.

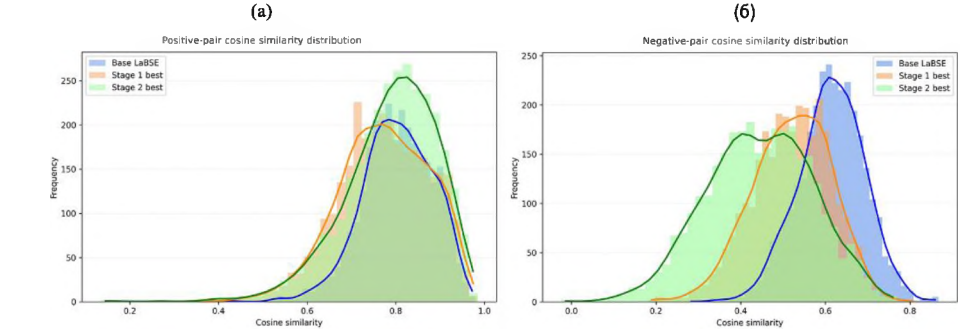

Рис. 5. Распределения косинусного сходства положительных и отрицательных пар для армяно-английской языковой пары (для базовой модели (синий) и моделей, полученных в лучших промежуточных точках первого (оранжевый) и второго (зеленый) этапов обучения): а - распределение положительных пар, б - распределение отрицательных пар

Результаты дообучения для армяно-русской языковой пары. На рис. 6 представлены кривые обучающих и валидационных потерь для двух этапов дообучения модели LaBSE для армяно–русской языковой пары. На обоих этапах валидационные потери монотонно убывают без признаков переобучения. По итогам двух этапов валидационные потери снизились с 0,885 до 0,275.

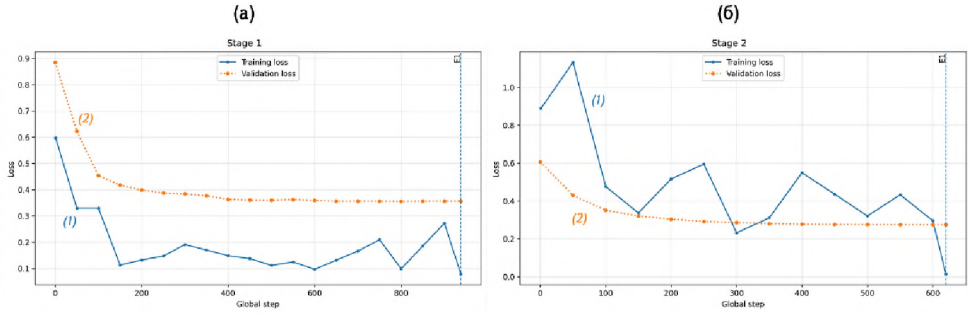

Рис. 6. Динамика потерь при обучении и валидации для армяно-русской языковой пары: 1 - кривая обучения, 2 - кривая валидации; а - этап 1, б - этап 2

На рис. 7 представлены распределения косинусного сходства для армяно-русской языковой пары. Для позитивных пар наблюдается прогрессивное смещение распределения вправо и его сужение. Лучшая модель этапа 2 формирует компактный пик в диапазоне 0,85…0,90. Для негативных пар распределение этапа 2 значительно смещено влево — в диапазон 0,30…0,55. В результате разрыв между распределениями вырос с 0,188 до 0,322.

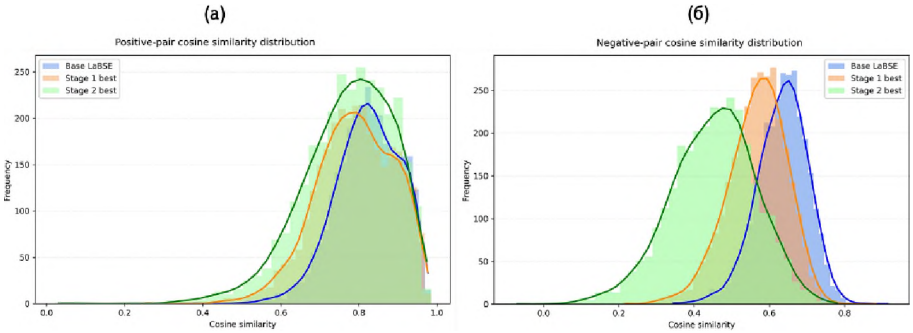

Рис. 7. Распределения косинусного сходства положительных и отрицательных пар для армяно-русской языковой пары (для базовой модели (синий) и моделей, полученных в лучших промежуточных точках первого (оранжевый) и второго (зеленый) этапов обучения): а - распределение положительных пар, б - распределение отрицательных пар

В четвертой главе представлены результаты экспериментальных исследований.

Экспериментальные исследования методов поиска кандидатов.

Экспериментальные исследования метода поиска кандидатов проводились на нескольких уровнях. Прежде всего была верифицирована гипотеза о ключевой роли частей речи на корпусе из 1 000 армяноязычных текстов «Википедии» и их английских аналогов. Метод исследовался для существительных, глаголов и прилагательных при варьировании размера вектора и метода отбора слов. Исследования подтвердили, что наилучшие результаты для задачи поиска кандидатов обеспечивают существительные (табл. 1).

Таблица 1

Результаты применения метода CL-POSSR при отборе топ-N слов по значениям показателя «Term-frequency variance» для различных частей речи и при разных размерах вектора

Размер вектора	Части речи			
	Существительное	Прилагательное	Глагол	Все слова
50	74 %; 6,49	42,3 %; 3,37	9 %; 0,59	40 %; 3,27
100	78,8 %; 7,01	46,1 %; 3,87	15,4 %; 0,99	54,5 %; 4,45
200	83,6 %; 7,49	48,6 %; 4,14	21,9 %; 1,55	65,6 %; 5,55
400	85,8 %; 7,81	48,9 %; 4,19	23,5 %; 1,73	77,4 %; 6,66

Первый показатель в каждой ячейке таблицы отражает долю текстов, для которых соответствующий исходный текст был найден среди топ-10 результатов, а второй - среднее значение рангового балла.

Оценка полноты метода проводилась на адаптированном корпусе текстов для обнаружения межъязыковых заимствований для армяно–английской и армяно–русской языковых пар. Адаптированный корпус включает два типа наборов. Набор Essays1 в основном состоит из текстов, в которых использовалось значительное количество отдельных приемов перефразирования. Essays2 - немного более сложный набор с большим количеством сильно перефразированных фрагментов. Фоновая коллекция содержала 10 000 случайно выбранных статей из англоязычной «Википедии». В данном эксперименте топ-N слов отбирались с

использованием показателя «Term-frequency variance», а сравнение текстов проводилось по существительным при размере вектора 500 (табл. 2).

Таблица 2

Оценка полноты разработанного метода поиска кандидатов CL-POSSR

Набор текстов	Языковая пара	Rec@1	Rec@5	Rec@10	Rec@20	Rec@50
Essays1	Арм. – англ.	0,821	0,707	0,76	0,816	0,873
	Арм. – русс.	0,869	0,7	0,761	0,84	0,903
Essays2	Арм. – англ.	0,813	0,685	0,736	0,808	0,875
	Арм. – русс.	0,867	0,719	0,762	0,804	0,879

Сравнение с результатами других работ по полноте показало, что предложенный метод является конкурентоспособным по отношению к существующим подходам.

С целью проверки переносимости метод был дополнительно применен к четырем малоресурсным языкам - грузинскому, финскому, румынскому и греческому - в паре с английским. Эксперимент, в котором дополнительно применялось распознавание синонимов на основе WordNet, подтвердил применимость подхода за пределами армянского языка и выявил положительное влияние учета синонимов на полноту поиска.

Расширенный метод поиска кандидатов исследовался на тех же адаптированных наборах, увеличив фоновую коллекцию до 100 000 текстов из англоязычной «Википедии». Полученные результаты подтвердили высокую полноту метода в условиях большой фоновой коллекции (табл. 3).

Таблица 3

Оценка полноты разработанного метода поиска кандидатов CL-POSSWR

Набор текстов	Языковая пара	Rec@5	Rec@10	Rec@20
Essays1	Арм. – англ.	0,800	0,848	0,917
	Арм. – русс.	0,729	0,869	0,913
Essays2	Арм. – англ.	0,798	0,822	0,893
	Арм. – русс.	0,720	0,831	0,894

Экспериментальные исследования метода детального анализа текстов.

Метод детального анализа применялся не ко всей фоновой коллекции, а к 100 текстам-кандидатам, предварительно отобранным расширенным методом поиска кандидатов.

На первом этапе сравнивались пять трансформерных моделей - LaBSE, all-MiniLM, distiluse-base-multilingual-cased-v2, snowflake-arctic-embed-l-v2.0 и paraphrase-xlm-r-multilingual-v1 - на адаптированных наборах Essays1 и Essays2 по метрикам F1, Recall, Precision, MCC (Matthews Correlation Coefficient) и AUC-ROC. Наилучшие результаты по F1-мере продемонстрировала модель LaBSE, что определило ее выбор для дальнейшего дообучения. Далее исходная (LaBSE, default) и дообученная в рамках работы на 100 000 парах (LaBSE, tuned_100k, представленная в предыдущем разделе) модели LaBSE были протестированы на переведенном наборе данных MRPC с целью сопоставления полученных результатов с результатами дообученной модели XLM-RoBERTa из работы [T. Ter-Hovhannisyan и др. – 2022], в которой сравнивались несколько моделей на основе трансформеров для нескольких языковых пар, включая армяно-английскую (табл. 4).

Таблица 4

Сравнение результатов моделей LaBSE с результатами модели XLM-RoBERTa

Модель	**Macro-F1**
XLM-RoBERTa	0,65…0,66
LaBSE, default	0,6718
LaBSE, tuned_100k	0,6909

Исходная модель LaBSE (default), модель LaBSE, обученная на 15 000 положительных парах (5 000 параллельных и 10 000 перефразированных) в течение трех эпох (epoch) (tuned_15k), и модель LaBSE, обученная на 100 000 положительных парах (tuned_100k), применялись для армяно–английской и армяно–русской языковых пар над адаптированными наборами Essays1 и Essays2. Дообученные в рамках работы модели были дополнительно протестированы на общедоступном наборе текстов, включающем 400 армянских текстов и коллекцию из 120 000 английских текстов. Обе конфигурации дообучения обеспечили прирост результата F1 по сравнению с базовой моделью (табл. 5).

Таблица 5

Сравнение результатов базовой модели LaBSE с дообученными версиями на общедоступном наборе данных для армяно–английской языковой пары

Модель	**Точность**	**Полнота**	**F1**
Default	0,810	0,752	0,78
tuned_15k	**0,812**	**0,758**	**0,784**
tuned_100k	**0,799**	**0,823**	**0,81**

Результаты дообученной на 100 000 парах модели были сопоставлены с результатами работы, в которой дообученная модель XLM-RoBERTa тестировалась на том же наборе текстов, а также результатами двух других работ, в которых методы тестировались на наборе данных сопоставимого размера для русско–английской языковой пары, с использованием метрик F1, granularity и PlagDet (табл. 6).

Таблица 6

Сравнение разработанного метода с использованием модели, дообученной на 100 000 парах, с другими методами

Метод	**Языковая пара**	**Полнота**	**Точность**	**F1**	**Gran.**	**PlagDet**
[O. Bakhteev и др. – 2019]	Русс. – англ.	0,79	0,83	0,8	-	-
[D. Zubarev и др. – 2022]	Русс. – англ.	0,924	0,824	0,871	1,08	0,825
[K. Avetisyan и др. – 2023]	Арм. – англ.	0,73	0,72	0,73	-	-
tuned_100k	Арм. – англ.	0,823	0,799	0,81	1,001	0,81

В пятой главе описана программная реализация системы CrossUnique в виде полнофункционального веб-приложения, обоснован выбор основного технологического стека, описаны специализированные инструменты для реализации алгоритмических компонентов и архитектурные решения с подробным описанием схемы базы данных.

Языком реализации выбран Python в качестве доминирующего языка в области обработки естественного языка и машинного обучения. В качестве серверного фреймворка выбран Django, предоставляющий полный набор необходимых компонентов - ORM (Object Relational

Mapping), система аутентификации, административный интерфейс, шаблонизатор, обработка форм, миграция схемы базы данных. Основной системой управления базой данных (СУБД) выбрана PostgreSQL - объектно-реляционная СУБД с открытым исходным кодом, отличающаяся высокой надежностью и поддержкой сложных запросов.

Для выполнения ресурсоемких вычислительных задач в фоновом режиме применяется связка Celery и Redis. Celery выступает в роли распределенной очереди задач, а Redis - в качестве брокера сообщений. Пользовательский интерфейс реализован на основе шаблонов Django с применением HTML, CSS, JavaScript и библиотеки HTMX, которая позволяет реализовать динамическое обновление отдельных элементов страницы без полной перезагрузки.

Структура базы данных представлена в виде ER-диаграммы, отражающей связи между основными сущностями системы: документами, пользователями и результатами проверки и т.д. (рис. 8).

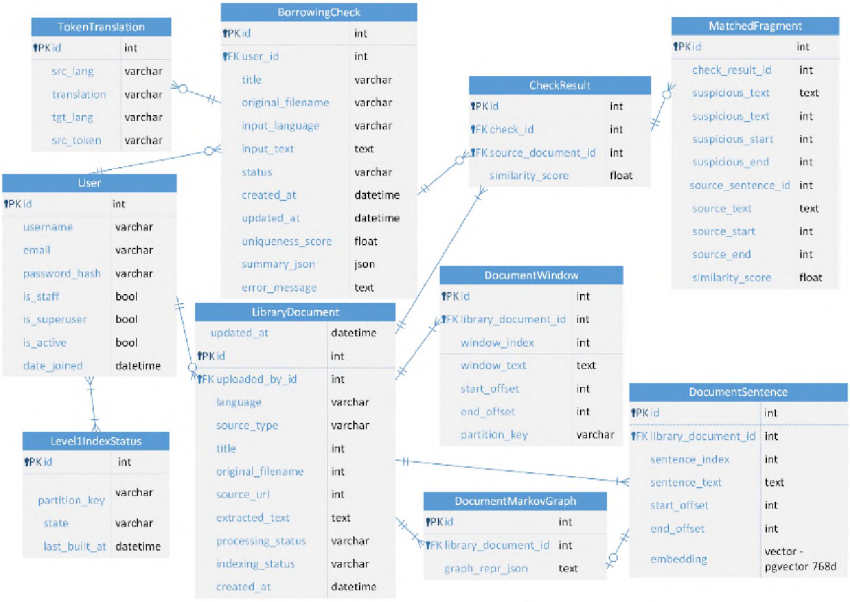

Рис. 8. ER-диаграмма модели данных системы

Разграничение функциональных возможностей реализовано по ролевому принципу. Авторизованный пользователь имеет доступ исключительно к функции проверки текста: он загружает документ и получает автоматически сформированный отчет с указанием выявленных совпадений и их характеристик. Администратор обладает расширенными правами: помимо проверки текста, он управляет базой эталонных документов — загружает новые материалы, редактирует и удаляет существующие записи, а также контролирует внешние источники, используемые при индексировании корпуса. На рис. 9 представлен интерфейс результатов проверки. На странице отображается проверяемый текст, в котором заимствованные фрагменты выделены красным цветом, что позволяет визуально локализовать совпадения. В верхней части страницы представлен показатель уникальности текста в процентах. Подобное представление результатов обеспечивает наглядность и удобство интерпретации для пользователя.

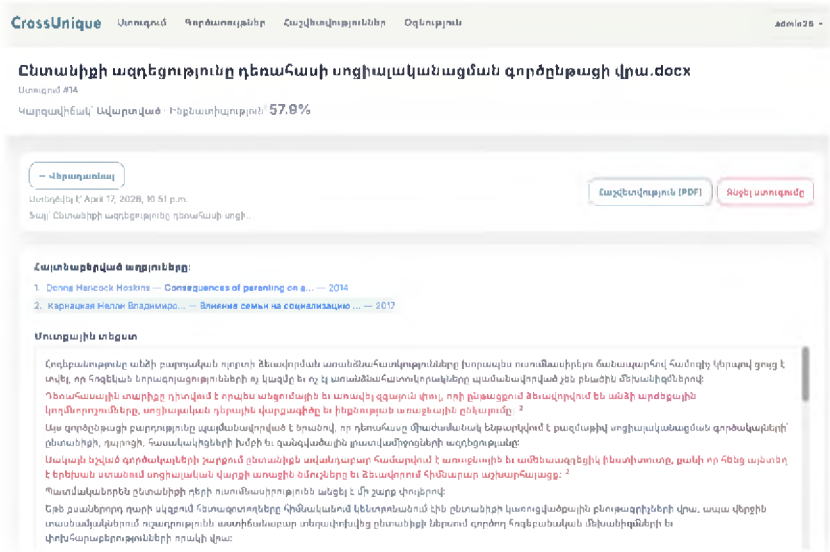

Рис. 9. Страница отчета о проверке документа с показателем уникальности и выделением заимствованных фрагментов

Проведен сравнительный анализ разработанной автоматизированной системы CrossUnique с пятью коммерческими системами, ориентированными преимущественно на английский и русский. Для тестирования были созданы и использованы два армяноязычных текста, содержащие в совокупности восемь дословных и перефразированных фрагментов из реальных источников на армянском, русском и английском языках. Разработанная система идентифицировала 7 из 8 заимствований, тогда как ни один из сравниваемых инструментов не обнаружил ни одного заимствованного фрагмента, что подтверждает практическую значимость специализированных инструментов для определения степени уникальности армяноязычного текста и выявления одноязычных и межъязыковых заимствований.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ

1. Спроектирована двухуровневая архитектура автоматизированной системы определения степени уникальности армяноязычного текста в многоязычной среде [1, 5, 8].
2. Для поиска кандидатов межъязыковых текстовых заимствований разработан метод на основе разметки частей речи [2, 6, 7, 9].
3. Разработан метод структурного анализа армяноязычных текстов на основе марковских цепей. Метод представляет текст в виде графа вероятностных переходов между ключевыми понятиями и позволяет вычислять статистику связей между ними, что обеспечивает дополнительный структурный признак при сопоставлении армяноязычных текстов [3].
4. Для поиска межъязыковых заимствований (детального анализа) разработан метод семантического сопоставления на основе предобученной многоязычной трансформерной модели LaBSE. Предложены подходы к формированию обучающих данных: стратегии отбора «трудных негативных» примеров, а также способ автоматического перефразирования армяноязычных предложений с обеспечением лексического разнообразия и семантического сходства [9].

5. Созданы наборы параллельных и перефразированных пар предложений для армяно–английской и армяно–русской языковых пар, предназначенные для дообучения модели в задаче межъязыкового семантического сопоставления. Каждая обучающая пара дополнена четырьмя трудными негативными примерами. Выполнено дообучение модели LaBSE для армяно–английской и армяно–русской языковых пар с использованием функции потерь Additive Margin Softmax в стиле ArcFace [9].
6. Для армяно-английской и армяно-русской языковых пар адаптирован набор тестовых данных путем перевода существующего набора данных, предназначенного для оценки методов решения двух подзадач обнаружения межъязыковых текстовых заимствований [7, 9].
7. Проведены экспериментальные исследования и сравнительный анализ с существующими подходами, подтвердившие эффективность разработанных методов на обоих уровнях поиска межъязыковых заимствований [6, 7, 9].
8. Разработана и реализована автоматизированная веб-система определения степени уникальности армяноязычного текста в многоязычной среде. Веб-интерфейс обеспечивает удобный пользовательский доступ к функциям проверки уникальности и детальную визуализацию заимствованных фрагментов. Система поддерживает экспорт результатов проверки в формате PDF-отчета, а также предоставляет средства управления базой данных документов. Интерфейс не требует от пользователя специальных технических знаний и пригоден для практического применения в академической среде [4, 8].

Основные результаты диссертации опубликованы в следующих работах:

1. **Սահակյան Ռ.Ռ., Պետրոսյան Գ.Ա.** Հետազոտական աշխատանքների ինքնատիպության աստիճանի գնահատման համակարգի նախագծում // Հայաստանի ճարտարագիտական ակադեմիայի լրաբեր գիտատեխնիկական հոդվածների ժողովածու. - 2022. - Հատոր 19, No. 1. - էջ 98-103.
2. **Պետրոսյան Գ.Ա., Սահակյան Ռ.Ռ.** Տեքստի ինքնատիպության աստիճանի գնահատման նախապատրաստում օգտակար բառերի ուղիղ ձևի զանգվածի ձևավորման միջոցով // Վանաձորի պետական համալսարանի գիտական տեղեկագիր, Բնական և ճշգրիտ գիտություններ. - 2022. – No. 2. – էջ 77-86.
3. **Саакян Р.Р., Шпехт И.А., Петросян Г.А.** Нахождение наличия заимствований в научных работах на основе марковских цепей // Вестник Санкт-Петербургского университета: Прикладная математика. Информатика. Процессы управления. – 2023. – Том 19, No. 1. – С. 43–50. doi: 10.21638/11701/spbu10.2023.104. **(SCOPUS)**
4. **Петросян Г.А., Саакян Р.Р., Шпехт И.А.** Разработка прототипа информационной системы определения степени уникальности выпускных квалификационных работ вуза // Научные ведомости Ванадзорского государственного университета: Естественные и точные науки. – 2023. - № 1. – С. 66-76.
5. **Պետրոսյան Գ.Ա.** Լեզվական մոդելների կիրառման հեռանկարները տեքստի ինքնատիպության աստիճանի գնահատման բանական համակարգերում // «Եվրոպական համալսարան» գիտական հոդվածների ժողովածու. – 2025. - No 16(01). - էջ 214-223.
6. **Петросян Г.А., Саакян Р.Р.** Применение метода поиска кандидатов межъязыковых заимствований на типологически разных малоресурсных языках // Вестник НПУА: Информационные технологии, электроника, радиотехника. – 2025. – No. 1. – С. 70-78.
7. **Петросян Г.А., Саакян Р.Р., Саакян В.Р.** Разработка метода поиска кандидатов межъязыковых текстовых заимствований на основе разметки частей речи // Вестник

Санкт-Петербургского университета: Прикладная математика. Информатика. Процессы управления. – 2025. – Том 21, No. 4. – С. 516–530. doi: 10.21638/spbu10.2025.405. **(SCOPUS)**

8. **Պետրոսյան Գ.Ա.** Բազմալեզու միջավայրում հայերեն տեքստի ինքնատիպության աստիճանի որոշման ինտելեկտուալ համակարգի ճարտարապետությունը // Վանաձորի պետական համալսարանի գիտական տեղեկագիր. Բնական և ճշգրիտ գիտություններ. - 2025. – No. 2. – էջ 66-71:
9. **Petrosyan G.A., Sahakyan R.R.** The Sentence-level Cross-lingual Plagiarism Detection Method for Armenian-English and Armenian-Russian Language Pairs // Proceedings of NPUA: Information Technologies, Electronics, Radio engineering. - 2025. – No. 2. - P. 78-87.

ՊԵՏՐՈՍՅԱՆ ԳԵՎՈՐԳ ԱՐՄԵՆԻ
ՏԵՔՍՏԻ ԻՆՔՆԱՏԻՊՈՒԹՅԱՆ ԱՍՏԻՃԱՆԻ ՈՐՈՇՄԱՆ ԲԱԶՄԱԼԵԶՈՒ ՀԱՄԱԿԱՐԳԻ ՄՇԱԿՈՒՄԸ
ԱՄՓՈՓԱԳԻՐ

Ատենախոսությունը նվիրված է բազմալեզու միջավայրում հայերեն տեքստի ինքնատիպության աստիճանի որոշման ավտոմատացված համակարգի մշակմանը: Ատենախոսության շրջանակներում կատարվել է տեքստային փոխառությունների որոնման մեթոդների և տեքստի ինքնատիպության աստիճանի որոշման համակարգերի վերլուծություն, մշակվել են մեթոդներ, նախագծվել և ներդրվել է ավտոմատացված համակարգ, կատարվել են փորձարարական հետազոտություններ:

1. Մշակվել է բազմալեզու միջավայրում հայերեն տեքստի ինքնատիպության աստիճանի որոշման ավտոմատացված համակարգի երկմակարդակ ճարտարապետությունը:
2. Մշակվել է հայերեն տեքստում պարունակվող միջլեզվական փոխառությունների սկզբնաղբյուրների՝ թեկնածու տեքստերի որոնման մեթոդ, որը ցուցաղրում է արդյունավետության 0.917 և 0.913 արժեքներ՝ համապատասխանաբար հայերեն-անգլերեն և հայերեն-ռուսերեն լեզվական զույգերի դեպքում ըստ Rec@20 գնահատման չափանիշի: Մեթոդը հարմարեցվել և կիրառվել է նաև հայերեն թեկնածու տեքստերի որոնման նպատակով:
3. Մշակվել է հայերեն տեքստերի կառուցվածքային վերլուծության՝ մարկովյան շղթաների վրա հիմնված մեթոդ, որը տեքստը ներկայացնում է որպես հիմնական հասկացությունների միջև հավանականական անցումների գրաֆ՝ դրանց միջև կապերի վիճակագրությունը պարզելու նպատակով:
4. Մշակվել է հայերեն տեքստում պարունակվող միջլեզվական տեքստային փոխառությունների որոնման՝ մանրամասն վերլուծության մեթոդ՝ նախապես ուսուցանված LaBSE բազմալեզու տրանսֆորմերային մոդելի միջոցով: Մեթոդը կիրառվել է նաև հայերեն միալեզու փոխառությունների որոնման նպատակով: Մոդելի վերապատրաստման համար մշակվել են «դժվար բացասական» նախադասությունների ընտրության, ինչպես նաև հայերեն նախադասությունների վերաձևակերպման մոտեցումներ՝ օգտագործելով բազմամակարդակ ֆիլտրման համակարգ, որն ապահովում է զույգերի իմաստային նմանությունը և բառապաշարային բազմազանությունը:
5. Ստեղծվել են հայերեն-անգլերեն և հայերեն-ռուսերեն զուգահեռ ու վերաձևակերպված նախադասությունների զույգերի հավաքածուներ՝ ներառյալ «դժվար բացասական» օրինակները:

6. Բազմալեզու տրանսֆորմերային LaBSE մոդելը վերապատրաստվել է 100,000 (60,000 զուգահեռ և 40,000 վերաձևակերպված) հայերեն-անգլերեն և հայերեն-ռուսերեն նախադասությունների զույգերի վրա՝ կիրառելով ArcFace ոճի Additive Margin Softmax կորստի ֆունկցիան: Հայերեն-անգլերեն լեզվական զույգի համար LaBSE մոդելի երկվույլ վերապատրաստման արդյունքում վավերացման կորուստը նվազել է 0.932-ից մինչև 0.211, լավագույն F1 արժեքը 0.894-ից աճել է մինչև 0.931, իսկ դրական և բացասական զույգերի միջին կոսինուսային նմանության արժեքների տարբերությունն աճել է 0.190-ից մինչև 0.345: Հայերեն-ռուսերեն լեզվական զույգի համար երկվույլ վերապատրաստման արդյունքում վավերացման կորուստը նվազել է 0.885-ից մինչև 0.275, լավագույն F1 արժեքն աճել է 0.906-ից մինչև 0.924, իսկ դրական և բացասական զույգերի միջին կոսինուսային նմանության արժեքների տարբերությունն աճել է՝ 0.188-ից հասնելով մինչև 0.322: Կորստի ֆունկցիայի արժեքների նվազումը հաստատում է վերապատրաստման արդյունավետությունը:
7. Հայերեն-անգլերեն և հայերեն-ռուսերեն լեզվական զույգերի համար հարմարեցվել է թեստային տվյալների հավաքածու՝ գոյություն ունեցող տվյալների հավաքածուի թարգմանության միջոցով, որը նախատեսված է միջլեզվական տեքստային փոխառությունների հայտնաբերման երկու ենթախնդիրների մեթոդների գնահատման համար:
8. Կատարվել են մշակված մեթոդների փորձարարական հետազոտություններ և համեմատություններ առկա մոտեցումների հետ, որոնք հաստատում են դրանց արդյունավետությունը միջլեզվական փոխառությունների որոնման երկու մակարդակներում: Մշակված մեթոդների համախումբը, բաց հասանելիություն ունեցող հայերեն-անգլերեն տեքստերի հավաքածուի վրա վերապատրաստված մոդելի միջոցով կիրառելիս, ըստ F1 գնահատման չափանիշի դրսևորել է 0.81 արժեք՝ սկզբնական մոդելի 0.78 արժեքի համեմատ:
9. Մշակվել է բազմալեզու միջավայրում հայերեն տեքստի ինքնատիպության աստիճանի որոշման երկմակարդակ ճարտարապետությամբ ավտոմատացված վեբ-համակարգ: Վեբ ինտերֆեյսը օգտատիրոջը հնարավորություն է տալիս օգտվելու ինքնատիպության աստիճանի որոշման և փոխառված հատվածների հայտնաբերման գործառույթից: Համակարգը աջակցում է ստուգման արդյունքների՝ PDF ձևաչափով հաշվետվության ներբեռնմանը և սրամադրում է գործիքներ՝ փաստաթղթերի տվյալների բազայի կառավարման համար: Ինտերֆեյսը օգտատիրոջից չի պահանջում մասնագիտացված տեխնիկական գիտելիքներ և հարմար է ակադեմիական միջավայրում գործնական կիրառման համար:
10. Կատարվել է մշակված CrossUnique ավտոմատացված համակարգի և տեքստի ինքնատիպության աստիճանի որոշման և գրագողության հայտնաբերման գոյություն ունեցող առևտրային համակարգերի համեմատական վերլուծություն: Արդյունքները փաստում են մշակված համակարգի արդյունավետությունը և հաստատում, որ գոյություն ունեցող բազմալեզու համակարգերը հիմնականում անարդյունավետ են հայերեն տեքստերում առկա փոխառությունների հայտնաբերման խնդրում:

PETROSYAN GEVORG ARMEN
DEVELOPMENT OF A MULTILINGUAL SYSTEM FOR DETERMINING THE DEGREE OF TEXT UNIQUENESS
SUMMARY

The dissertation focuses on the development of an automated system for determining the degree of uniqueness of Armenian texts within a multilingual environment. In this work, the existing text-borrowing detection methods and text uniqueness determination systems are analyzed, original methods are developed, an automated system is designed and implemented, and experimental investigations are conducted.

1. A two-level architecture of an automated system for determining the degree of uniqueness of an Armenian text in a multilingual environment has been designed.
2. A method for cross-lingual text borrowing source retrieval based on part-of-speech tagging has been developed, which demonstrates the efficiency values of 0.917 and 0.913 for the Armenian-English and Armenian-Russian language pairs, respectively, according to the Rec@20 evaluation criterion. The method is also adapted and applied to the retrieval of Armenian text candidates.
3. A method for the structural analysis of Armenian texts based on Markov chains has been developed. The method presents a text as a graph of probabilistic transitions between key concepts and allows for the calculation of statistics of relationships between them.
4. For the detailed analysis of cross-lingual borrowings (text alignment), a semantic matching method based on a pre-trained multilingual transformer model LaBSE has been developed. The method has also been used to search for monolingual borrowings in Armenian. To fine-tune the model, approaches have been developed for selecting "hard" negative sentences, as well as for paraphrasing Armenian sentences using a multi-level filtering system that ensures the semantic similarity and lexical diversity of the pairs.
5. Collections of Armenian-English and Armenian-Russian parallel and paraphrased sentence pairs have been created, including “hard negative” examples.
6. A two-stage fine-tuning of LaBSE has been performed on a corpus of 100,000 sentence pairs (60,000 parallel and 40,000 paraphrased) using the ArcFace style Additive Margin Softmax loss function. For the Armenian–English language pair, the validation loss has decreased from 0.932 to 0.211, the best F1 value has increased from 0.894 to 0.931, and the gap between the average cosine similarity values of positive and negative pairs has increased from 0.190 to 0.345. For the Armenian–Russian pair, the validation loss has decreased from 0.885 to 0.275, F1 improved from 0.906 to 0.924, and the margin has increased from 0.188 to 0.322. The decrease in the loss function values confirms the effectiveness of the training.
7. A test dataset is adapted for the Armenian-English and Armenian-Russian language pairs by translating the existing dataset, which is intended to evaluate the methods for two subproblems of detecting cross-lingual text borrowings.
8. Experimental studies and comparisons with the existing approaches are conducted, confirming the effectiveness of the developed methods at both levels of cross-lingual borrowing detection. The developed set of methods, when applied to a collection of open-access Armenian-English texts using the fine-tuned model, shows a result of 0.81 according to the F1 evaluation criterion compared to the result of 0.78 for the original model.
9. An automated web-based system for determining the degree of uniqueness of an Armenian text in a multilingual environment is developed and implemented. The web interface provides convenient user access to the uniqueness checking functions and detailed visualization of the borrowed fragments. The system supports export of check results in PDF report format, and also provides tools for managing a document database. The interface

does not require special technical knowledge from the user and is suitable for practical application in the academic environment.

10. A comparative analysis of the developed automated system CrossUnique has been conducted with the existing commercial systems for determining the degree of text uniqueness and detecting plagiarism. The results demonstrate the efficiency of developed system and confirm that the existing multilingual systems are largely ineffective for detecting borrowings contained in Armenian texts.