АННОТАЦИЯ

диссертации на соискание степени доктора философии (PhD)
по специальности «6D075100 – Информатика, вычислительная техника и управление»

Еркебұлан Гүлнұр Тұратайқызы

«Система идентификации паттернов полиязычных текстов»

Актуальность темы исследования. В списке языков по количеству носителей языка английский язык занимает 3 место (379 млн. человек), русский – 7 место (154 млн.человек), а казахский – 76 место (12,9 млн.человек) [1]. Согласно анализу многоязычности количество человек, свободно разговаривающих на трёх языках, составляет 13% от населения мира, а на двух языках – 42% [2]. В Республике Казахстан знание трёх языков является почти обязательным условием карьерного роста и хорошей оплаты труда. Государственным языком Республики Казахстан является казахский язык, наравне с казахским официально употребляется русский язык. Английский язык начали изучать с 1 класса с 2014 года.

Знание нескольких языков открывает окно в большой глобальный мир с его колоссальным потоком информации и инноваций. Полиязычность – это веление времени и государство уделяет большое внимание данному направлению развития.

На настоящий момент занчивается реализация Дорожной карты развития трёхъязычного образования на 2015-2020 годы [3], по результатам которой подготовлен график перехода на трёхъязычное образование в школах РК с 2023 года [4]. В 2022-2023 годах намечается переход на англоязычное образование в полиязычных школах, а в 2023-2024 годах – переход всех общеобразовательных средних учебных заведений. Внедрение трёхъязычного образования станет реализоваться по выбору на базе коллегиального заключения педсовета организации образования и комитета родителей. Переход к обучению на трёх языках осуществляется в рамках реализации 79 шага Плана Нации «100 шагов» [5] и Государственной программы развития образования и науки РК на 2016-2019 [6].

Языковая индустрия - это большой бизнес. Аналитики TechNavio в своём исследовании «Global Language Services Market 2020-2024» прогнозируют рост рынка на 9,72 млрд долларов США в течение 2020-2024 годов, с достижением показателя CAGR 4% [7].

Сегодня существует множество онлайн-переводчиков и расширений для браузеров, помогающих перевести незнакомое слово на иностранном языке [8]. У всех разные алгоритмы перевода, разные базы данных слов и словосочетаний, поэтому и результаты перевода могут отличаться. Как же найти наилучший перевод? В интернете найдено только одно расширение для браузера, предлагающее одновременно несколько вариантов перевода от разных онлайн-компаний, его название MyTranslator. Приложение предлагает переводы от Google Translate, Microsoft Bing Translator, Яндекс Переводчик и DeepL

Translator, для просмотра вариантов нужно переключаться по вкладкам «G», «M», «T» и «Y». Расширение было запущено в январе 2020 год и имеет более 10000 пользователей [9]. Вместе с тем, для перевода текста по предложениям необходимо вызывать расширение для каждого предложения заново, нет возможности просмотра всех вариантов от переводчиков без переключения вкладок, нет возможности каким-то образом зафиксировать выбранный вариант перевода, чтобы в последующем корректировать переведённый текст самостоятельно (кроме стандартного способа «Выделить» - «Скопировать» - «Вставить»). Первое направление диссертационного исследования, «Определение наиболее достоверного перевода в работе с полиязычными текстами», не содержит перечисленных недостатков. Кроме того, данное направление предоставляет оценку каждого варианта перевода, подсказывая пользователю наилучший выбор.

Интернет-ресурсы с автоматической генерацией текста [10-12] вместе с сайтами-переводчиками [13-15] могут создавать множество интерпретаций одного и того же текста. Чтобы отловить оригинал, используют возможности нейронных сетей [16-18].

На сегодняшний день в задачах прогнозирования, классификации и управления широко используются нейросети. К их сильным сторонам можно отнести «решение задач при неизвестных закономерностях, устойчивость к шумам во входных данных, потенциальное сверхвысокое быстродействие и отказоустойчивость при аппаратной реализации нейронной сети» [19]. Вместе с тем, для их обучения и работы необходимы огромные вычислительные ресурсы [20].

Во втором направлении диссертационного исследования предложен альтернативный способ обнаружения перевода текста на основе энтропии Реньи, который не требует значительных компьютерных мощностей и большого времени на поиск. В качестве ядра разработки энтропия Реньи выбрана не случайно. Энтропийный подход уже исследуют и применяют при работе с текстами, но в других контекстах. Так, в работе [21] 2019 года приведены результаты исследований извлечения понятий для структурированного текста с использованием метода энтропийного веса, в работе [22] 2020 года приведены результаты исследований энтропийной связи между длиной текста и лексическим богатством, в работе [23] 2019 года исследуется энтропийный анализ сомнительных текстовых источников на примере рукописи Войнича, в работе [24] 2019 года, связывающей энтропийную оценку с машинным переводом, предлагаются методы решения недостаточного перевода, путём двухфазового разбиения процесса. Подробнее об этих и других работах можно узнать в пункте 2.3.2 диссертационного исследования.

Поскольку интернет-технологии продолжают расти и расширяться, становясь все более доступными и широко распространенными, ценность сайта также растет. Сайт является очень ценным активом для любой компании в современном глобализированном мире с его способностью привлекать новых

посетителей, информировать их о продуктах или услугах, и проводить продажи, независимо от типа бизнеса.

На основе анализа ответов на вопрос «Сделали бы вы покупку на веб-сайте, на котором есть контент на вашем родном языке, если бы качество этого контента было низким?» выборочной группы (3000 человек, 50% мужчин / 50% женщин в 8 странах, примерно на 3 континентах, в возрасте от 25 до 65 лет и с различным уровнем владения английским языком) было выяснено, что 45,3% не согласились бы сделать покупку на сайте с плохим переводом контента на их родном языке [25]. Значит, вопрос локализации сайта становится очень актуальным, поэтому появляются новые дисциплины в университетах, посвящённые данной тематике, с изучением лингвистических и маркетинговых особенностей, а также различных технических аспектов [26, 27].

Каждое учебное заведение заинтересовано в привлечении зарубежных студентов и инвесторов, поэтому сайт университета должен является соответствующим представительством для международного сотрудничества, как и сайты государственных и частных организаций. В системе идентификации паттернов полиязычных текстов рассматривается ещё одно направление, определяющее недостающие переводы полиязычного сайта на всех его языковых версиях с возможностью генерирования машинного перевода недостающей информации.

Новостные агрегаторы также являются полиязычными сайтами, но в диссертационном исследовании рассматриваются не в третьем, а в четвёртом направлении. Это связано с тем, что алгоритм распространяется на генерацию информации в период постпарсинга и препубликации, а не на опубликованные материалы.

Последнее направление, посвящённое созданию тестов и учебных материалов для нескольких языков, рассматривает важность аутентичности информации на разных языках. Аутентичность информации – свойство, гарантирующее, что перевод и оригинал идентичны. Существуют научные эксперименты, результаты которых показали, что материалы на языке перевода усваиваются лучше, если материал аутентичен оригиналу [28,29]. Трудно найти программу, позволяющую проверить правильность перевода, в автоматическом режиме. До сих пор онлайн-переводчики содержат ошибки или отсутствие вариантов перевода некоторых слов и выражений на казахском языке, поэтому для проверки переводов иногда применяют двойную проверку экспертной группой.

В методических рекомендациях [30, 31] по составлению тестовых заданий в пункте 6.5 шестой главы прописано: «Перевод тестовых заданий – при разработке, актуализации тестовых заданий на государственном, русском и других языках вопросы и ответы должны быть аутентичными и придерживаться терминологических словарей, утверждённых Государственной терминологической комиссией при Правительстве РК». В диссертационной работе предлагается рассмотреть возможность автоматической проверки и подсказки при формировании переводов тестов и учебных материалов на основе

рядов синонимов и терминологических словарей, а также общего анализа информации в целом, с целью выявления недостающих или ошибочно переведённых частей оригинального текста.

В итоге, в диссертации «Система идентификации паттернов полиязычных текстов» (далее – СИППТ) были выделены пять основных направлений исследования:

1) определение наиболее достоверного перевода в работе с полиязычными текстами среди онлайн-переводчиков;

2) определение разницы между настоящим и поддельным переводом исходного текста на основе энтропийного подхода;

3) определение недостающих переводов полиязычного сайта на всех его языковых версиях;

4) определение недостающих переводов в работе полиязычных новостных агрегаторов;

5) создание тестовых и учебных материалов для нескольких языков.

Все перечисленные факты говорят о том, что идентификация перевода и оригинального текста в работе с полиязычными текстами является актуальной.

Цель исследования: разработать модель управления и алгоритмы Системы идентификации паттернов полиязычных текстов по направлениям, связанным с полиязычными текстами, с использованием паттернов «предложение» и «параграф», в рамках реализации 79 шага Плана Нации «100 шагов».

Для достижения цели были поставлены следующие задачи исследования:

1. Анализ систем обнаружения плагиата и сервисов онлайн-переводчиков для использования в СИППТ, систематизация данных для выявления основных направлений СИППТ.

2. Выполнение расчётов для выбора наиболее точного алгоритма нечёткого сравнения строк для сравнения текста и его перевода (от онлайн-переводчиков) на основе программной реализации и экспертных оценок. Создание корпуса параллельных текстов на русском, казахском, английском языках для расчётов.

3. Разработка алгоритма для определения наиболее достоверного перевода в работе с полиязычными текстами среди онлайн-переводчиков и алгоритма создания тестов и учебных материалов для нескольких языков для предотвращения ошибок перевода с использованием выбранного алгоритма нечёткого сравнения строк.

4. Выполнение расчётов для проверки работоспособности энтропийного подхода (формирование ключевых рядов высокочастотных слов текстов, расчёт энтропийных координат для паттернов «предложение» и «параграф», расчёт расстояний между множествами энтропий текстов в соответствии с метрикой Минковского) для определения близости текстов на разных языках с программной реализацией расчёта координат.

5. Разработка алгоритма определения разницы между настоящим и поддельным переводом исходного текста, алгоритма определения недостающих переводов полиязычного сайта на всех его языковых версиях и алгоритма

определения недостающих переводов в работе полиязычных новостных агрегаторов с использованием энтропийного подхода.

6. Разработка модели управления Системы идентификации паттернов полиязычных текстов с опорой на результаты вышеперечисленных исследований.

Объектом исследования является информационное поле, содержащее в себе существенные информационные блоки текста на различных языках.

Предметом исследования являются модель управления и алгоритмы идентификации паттернов полиязычных текстов.

Методы исследования: алгоритм нечёткого сравнения строк Оливера, алгоритм нечёткого сравнения строк FuzzyWuzzy, стеммеры Портера, нормализация, энтропия Шеннона, энтропия Реньи, метрика Минковского, расстояние Хэмминга, декартовое расстояние, расстояние между центрами масс, расстояние между геометрическими центрами, расстояние между центрами параметрических средних.

Научная новизна: научная новизна заключается в обосновании предложенной автором СИППТ как результата синтеза методов идентификации паттернов текстовых материалов с учётом особенностей полиязычности, параметризуемой энтропии и общеизвестных php-решений.

Новшества полученных результатов:

- разработан корпус параллельных текстов на русском, казахском, английском языках;
- разработаны алгоритм для определения наиболее достоверного перевода в работе с полиязычными текстами среди онлайн-переводчиков и алгоритм создания тестов и учебных материалов для нескольких языков перевода с использованием алгоритма нечёткого сравнения строк;
- разработан энтропийный подход к обнаружению близости полиязычных текстов;
- разработаны алгоритм определения разницы между настоящим и поддельным переводом исходного текста, алгоритм определения недостающих переводов полиязычного сайта на всех его языковых версиях и алгоритм определения недостающих переводов в работе полиязычных новостных агрегаторов с использованием энтропийного подхода;
- разработана модель управления СИППТ в рамках реализации 79 шага Плана Нации «100 шагов».

Теоретическая значимость: основными теоретическими открытиями стали обнаружение способности авторского энтропийного подхода к определению близости текстов на разных языках, а также экспериментальное выявление более точного алгоритма нечёткого сравнения строк при работе с полиязычными текстами. Рекомендации по построению корпуса текстов будут полезны исследователям в области полиязычных текстов, с использованием онлайн-переводчиков. Разработанная модель управления и алгоритмы Системы идентификации паттернов полиязычных текстов представляют собой описание

оригинальных решений существующих проблем, связанных с многоязычностью населения.

Практическая значимость: Практическая значимость работы заключается в применимости разработанной системы как в организациях, взаимодействующих с информационным полем с учётом полиязычности, так и любым заинтересованным пользователем, поскольку программные элементы этой системы размещены в открытом доступе в интернете. Разработанные алгоритмы с возможностью использования заданных паттернов можно применять в части выявления наиболее достоверного перевода в работе с полиязычными текстами в интересах переводчиков, аналитиков и других заинтересованных пользователей, путём размещения необходимых скриптов в открытом доступе в интернете (с генерацией переводов на основе вариантов перевода нескольких онлайн-переводчиков, как правило, требующих оплаты при программном использовании).

В части определения обнаружения перевода, адекватного исходному тексту, разработанные алгоритмы с параметризуемой энтропией можно предложить:

- аналитическим компаниям (поиск первоисточника статьи/новости, разбиение статьи/новости на части заимствований и авторской работы и т.д.);
- организациям из области информационной безопасности (поиск дубликатов, поиск первоисточников материалов на других языках, представляющих собой угрозу национальной безопасности и т.п.);
- в высших учебных заведениях (поиск переводного плагиата в студенческих работах) и др.

В направлениях определения недостающих переводов в работе полиязычных новостных агрегаторов и определения недостающих переводов полиязычного сайта на всех его языковых версиях разработанные алгоритмы можно применить прежде всего для владельцев информагентств, а также для государственных и частных полиязычных интернет-ресурсов.

В части создания тестов и учебных материалов на разных языках предложенные алгоритмы можно применять широкому кругу заинтересованных лиц, имеющим дело с обработкой объёмных текстовых материалов.

Положения диссертации, выносимые на защиту (научные результаты):

- рекомендации по построению корпуса текстов как результат исследования способов поиска паттернов полиязычных текстов с помощью сервисов онлайн-переводчиков;
- энтропийный подход к обнаружению близости полиязычных текстов;
- модель управления и алгоритмы Системы идентификации паттернов полиязычных текстов;
- программная реализация выполнения расчётов сравнения паттернов текста и его перевода с помощью нечёткого сравнения строк, а также программная реализация расчёта энтропийных координат.

Личный вклад автора заключается в проведении исследований, обосновывающих основные выносимые на защиту положения, а также значимая роль при обобщении и анализе полученных результатов.

Структура и объем диссертации. Диссертация имеет классическую структуру: вводная часть, основная часть (три главы), заключение, список использованных источников и приложения. Работа включает 65 рисунков, 14 таблиц и 113 наименований использованных источников.

Во введении обоснован выбор темы исследования, раскрыта актуальность пяти направлений СИППТ, сформулирована цель исследования, определяющие её задачи, представлены объект и предмет исследования, раскрыты научная новизна, практическая значимость работы.

В первой главе проведено сравнение систем обнаружения плагиата в Рунете. Выполнен подробный анализ работы модуля поиска переводных заимствований системы «Антиплагиат. Вуз». Раскрыты способы поиска паттернов полиязычных текстов с помощью сервисов Google и Яндекс. Выполнена постановка задач диссертационного исследования.

Во второй главе предложены модели и методы СИППТ. Исследовано направление определения наиболее достоверного перевода с помощью php-кода. Подробно описана классификация текстов с пятью этапами. Рассмотрено применение энтропийного подхода в качестве оценки степени связности текстов. Экспериментально доказаны работоспособность энтропийного подхода для задач диссертационного исследования и выбор алгоритма нечёткого сравнения строк для направлений СИППТ. В результате экспериментов созданы скрипты, на которые получены два свидетельства о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом (Приложения Д, Е Диссертации).

Лучший алгоритм нечёткого сравнения строк, выявленный в проведённых экспериментах, был применён в первом и пятом направлениях СИППТ, а разработанный энтропийный подход – во втором, третьем и четвёртом направлениях СИППТ.

Третья глава посвящена модели управления и алгоритмам пяти направлений СИППТ с применением методов и методик из второй главы. В третьей главе подробно описаны программная реализация выполнения расчётов сравнения паттернов текста и его перевода с помощью нечёткого сравнения строк, а также программная реализация расчёта энтропийных координат (скрипты, на которые получены авторские права).

В заключении представлены результаты исследований, включающие основные выводы по итогам диссертационного исследования.

Апробация работы. Результаты диссертационного исследования докладывались и обсуждались на научных конференциях:

- VII Международная научно-практическая конференция «GLOBAL SCIENCE AND INNOVATIONS 2019: CENTRAL ASIA» в г. Нур-Султан;
- Международная конференция в г. Варшава в рамках издания MODERN SCIENTIFIC CHALLENGES AND TRENDS (2019 г.);
- Международная научно-методическая конференция «Современный университет как пространство цифрового мышления» в г. Новосибирск (2020 г.).

Проведена научная стажировка. Получено 2 свидетельства о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом.

Публикации и авторские свидетельства. Основные результаты исследования опубликованы в 8 научных работах, в том числе, в 2 статьях, напечатанных в международных рецензируемых научных журналах (Scopus), в 3 статьях в научных изданиях, включённых в Перечень научных изданий, рекомендуемых для публикации основных результатов научной деятельности, утверждаемый уполномоченным органом, в 3 работах - в трудах международных научных конференций. Получено 2 авторских свидетельства.

Список научных публикаций:

1. Еркебұлан Г.Т., Куликова В.П. Особенности реализации поиска переводных заимствований в системе «Антиплагиат»: сильные и слабые стороны // Матер. междунар. науч. конф. «Modern scientific challenges and trends» Polish Science Journal. – Варшава, 2019. – № 9(20). – С. 42-46.

2. Еркебұлан Г.Т., Куликова В.П. Поисковые системы в качестве обнаружения заимствований в полиязычных текстах // Global Science and Innovations 2019: Central Asia. – Нур-Султан, 2019. – № 2(3). – С. 171-173.

3. Еркебұлан Г.Т., Куликова В.П. Сравнительный анализ систем обнаружения кросс-языкового (переводного) плагиата // Вестник КазНИТУ им. К. Сатпаева. – Алматы, 2019. – № 6(136). – С. 178-183.

4. Еркебұлан Г.Т., Куликова В.П., Куликов В.П. О применении Google Custom Search и Google Translate API в обнаружении кросс-языкового плагиата // Вестник АУЭС. Серия «Информационные технологии». – Алматы, 2019. – № 4(47). – С. 109-116.

5. Еркебұлан Г.Т., Куликова В.П., Куликов В.П. О применении Яндекс.XML и API Яндекс.Переводчика в системе идентификации паттернов полиязычных текстов // Вестник АУЭС. Серия «Информационные технологии». – Алматы, 2020. – № 1(48).– С. 110-117.

6. Еркебұлан Г.Т., Куликова В.П., Куликов В.П., Крылова Е.М. Модели и методы классификации текстовых запросов в Системе идентификации паттернов полиязычных текстов // Матер. междунар. науч. конф. «Современный университет как пространство цифрового мышления». – Новосибирск: СГУГиТ, 2020. – С. 130-134.

7. G. Yerkebulan, V. Kulikova, V. Kulikov. Google/Yandex Translation Detection in the Patterns Identifying System of Multilingual Texts // Research Institute for Intelligent Computer Systems, West Ukrainian National University, журнал «International Journal of Computing», ISSN 1727-6209 (print). – 2021. – Vol. 20, Issue 1. – P. 72-77. *// https://doi.org/10.47839/ijc.20.1.2094*

8. G. Yerkebulan, V. Kulikova, V. Kulikov, Z. Kulsharipova. Devising an entropy based approach for identifying patterns in multilingual texts // PC TECHNOLOGY CENTER, Kharkiv, Ukraine, журнал «Eastern-European Journal of Enterprise Technologies», ISSN 1729-3774 (print). – 2021. – №2/2 (110). – P. 16-22. // *https://doi.org/10.15587/1729-4061.2021.228695*

Свидетельства о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом:

1. Свидетельство о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом №19775 от «17» августа 2021 года «Программа для определения наиболее достоверного перевода в работе с полиязычными текстами» (программа для ЭВМ), авторы: Еркебұлан Г.Т., Куликова В.П., Куликов В.П.

2. Свидетельство о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом №19562 от «30» июля 2021 года «Программа расчёта энтропийных координат (по паттернам «параграф» и «предложение») для определения близости полиязычных текстов» (программа для ЭВМ), авторы: Еркебұлан Г.Т., Куликова В.П., Куликов В.П.