АННОТАЦИЯ

диссертации на соискание степени доктора философии (PhD)
по специальности «6D070400 – Вычислительная техника и программное обеспечение»

Мухсиной Куралай Женисбековны
«Разработка системы анализа многоязычной текстовой информации на основе машинного обучения»

Актуальность работы. В современном полиязычном и мультикультурном мире, как никогда актуальна проблема сопряженности языков, поиск эффективных и жизнеспособных программ в области языков по консолидации обществ. Инициированный Главой государства и реализуемый в республике Казахстан проект триединства языков служит базой для информационного обмена как внутри страны, так и за ее пределами. Данный проект, рассматривая казахский язык, как государственный язык, русский язык, как язык межнационального общения и английский язык, как язык успешной интеграции в глобальную экономику, содействует активному развитию информационного сообщество внутри страны и интеграции Казахстана в мировое глобальное информационное сообщество. Для осуществления международного взаимодействия, продвижения продукции, знаний и коммуникации Республики Казахстан в мировом информационном пространстве необходимо развивать как специальные информационные приложения по обработке казахского языка, так и возможности его включения в приложения мультиязычной обработки.

Сегодня существует достаточное количество приложений, обрабатывающих казахский язык. В тоже время, остается достаточно много проблем, связанных с необходимостью повышения качества автоматической обработки казахского языка.

К настоящему времени в решение задачи автоматической обработки казахского языка существенный вклад внесли такие видные ученые как Калдыбай Бектаев, Шарипбаев Алтынбек Амирович, Амиргалиев Едилхан Несипханович, Тукеев Уалшер Ануарбекович, Хусейн Атакан Варол, Рахимова Диана Рамазановна, Мусабаев Рустам Рафикович Мансурова Мадина Есимхановна, Мусиралиева Шынар Женисбековна и многие другие.

Однако, основная часть существующих исследований направлена на анализ автоматизацию морфологии и синтаксиса когда как задача его семантического анализа по-прежнему не решена. Анализ научных исследований в области компьютерной лингвистики, интеллектуального анализа и искусственного интеллекта, связанных с развитием знаний и включением казахского языка в многоязычные проекты, показывает, что существующих в данном научном направлении решений не достаточно для удовлетворения на качественном уровне существующих потребностей по разработке систем автоматической обработки текстов казахского, русского и английского языков.

Существующие многоязычные приложения NLP, в своем большинстве, используют только грамматический этап обработки текста, тогда как, семантический анализ текста или анализ смысла естественного языка, по-прежнему остается одной из ключевых проблем, как теории искусственного интеллекта, так и компьютерной лингвистики.

Поскольку методы, базирующиеся на правилах, требуют высоких интеллектуальных затрата, а компьютерные процессоры обладают все большей мощью, все большее распространение приобретают методы машинного обучения. Однако, для использования методов машинного обучения при семантической и грамматической обработки многоязычной информации необходимы заранее созданные грамматически и семантически размеченные корпуса каждого естественного языка.

Все вышеперечисленное и обуславливает актуальность диссертационной работы, посвященной комплексному исследованию и решению проблемы анализа многоязычной текстовой информации на основе машинного обучения.

Целью диссертационной работы является повышение качества работы систем автоматической обработки текстовой информации казахского, русского и английского языков за счет использования моделей интеллектуального анализа и методов машинного обучения. В рамках поставленной цели решается научная задача моделирования процессов интеллектуальной обработки многоязычных текстов, разработке моделей, методов, алгоритмов, осуществляющих анализа многоязычной текстовой информации, с целью определения основных характеристик текстов для построения моделей машинного обучения. Полученные алгоритмы должны получить свою практическую реализацию в виде экспериментальных систем обработки текстов, с возможностью оценки их работы на базе созданных размеченных корпусов.

Для достижения поставленной цели в диссертации решаются нижеследующие **задачи**.

- Разработка модели извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов и ее адаптация для казахского, русского и английского языков;
- Модификация метода вероятностного POS-тегинга, использующего скрытую Марковскую модель;
- Разработка метода определения семантической близости многоязычных текстовых документов, базирующегося на использовании VSM;
- Формирование методики экспертной оценки качества работы системы анализа семантической близости текстов;
- Создание программного приложения, имплементирующего разработанные модели, методы и алгоритмы.

Научная новизна диссертационной работы.

- Модифицирован гибридный метод автоматической морфологической и семантической разметки текстовых корпусов казахского языка, отличительной особенностью которого является одновременное

использование HMM и правил, представленных регулярными выражениями; что позволило снять часть морфологической многозначности и повысить полноту и точность разметки;

- Разработана логико-лингвистическая модель семантического анализа, идентифицирующая факты в многоязыковых текстах, что позволило извлекать из текстов казахского, русского и английского языков знания, явным образом представленные в виде RDF-триплетов и формировать семантически размеченные обучающие корпуса
- Усовершенствован метод определения семантической близости многоязычных текстовых документов на базе VSM, который отличается использованием весовой функции PPMI для определения принадлежность текста к узкоспециализированной предметной области;
- Создана информационная технология определения семантической близости текстов к заданной узкоспециализированной тематике, базирующаяся на предложенном методе вычислении среднего значения косинусного сходства векторов документов обучающего корпуса.

Методы исследований базируются на комплексном использовании теории интеллекта, общей теории систем, системного анализа, алгебры конечных предикатов и методов машинного обучения. Алгебра конечных предикатов используется для формализации семантической информации, передаваемой предложениями естественного, с последующим формированием обученного корпуса. Методы машинного обучения используются для построения моделей определения принадлежности текстов узкой предметной области и алгоритма семантической разметки многоязычных корпусов.

Объект исследования. Системы автоматической обработки текстовой информации на казахском, русском и английском языках.

Предметом исследования являются модели и алгоритмы интеллектуального семантического анализа многоязычной текстовой информации.

Практическая значимость работы заключается в разработке на основе положений выносимых на защиту программного приложения, позволяющего осуществлять автоматическую семантическую разметку мультиязычных корпусов текстов казахского, русского и английского языков, и приложения, позволяющего определить возможную криминальную составляющую анализируемого текста. А также в разработке семантически размеченных корпусов криминально окрашенных текстов казахского, русского и английского языков.

Прикладная ценность результатов работы заключается в возможности выявления криминально-окрашенных текстов в компьютерных сетях любыми заинтересованными государственными органами.

Положения, выносимые на защиту. По результатам исследования были решены нижеследующие задачи:

- Разработана модель извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов, которая адаптирована для казахского, русского и английского языков. Обоснован выбор

математического аппарата алгебры конечных предикатов для моделирования семантики предложений естественного языка.

- Модифицирован метод вероятностного POS-тегинга, использующий скрытую Марковскую модель.
- Разработан метода определения семантической близости многоязычных текстовых документов, базирующейся на использовании VSM;
- Сформирована методика экспертной оценки качества работы системы анализа семантической близости текстов;
-Создан, программный комплекс, позволяющий определить наличие криминального смысла в поступающих текстах и осуществляющее семантическую разметку.

Связь темы с планами научно-исследовательских программ

Диссертационная работа выполнялась в соответствии с календарным планом научно-исследовательских грантовых работ: «Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах» Института информационных и вычислительных технологий Комитета науки и МОН РК.

Публикации. Основные результаты, проведенных исследований по теме диссертации, представлены в 17 публикациях, из которых 4 – в научных изданиях, рекомендуемых КН МОН РК, 6 – в международных научных изданиях, входящих в базу данных Scopus и Web of Science, 7 – в материалах международных научно практических конференций.

Структура диссертации включает введение, 4 раздела , заключение, список использованных источников и пяти приложений. Общий объем диссертации составляет 121 страниц, 26 рисунков, 6 приложении. Список литературы состоит из 145 источников

Во введении дано обоснование актуальности выбранной темы диссертационной работы. Сформулированы цель, объект, предмет и задачи научно-исследовательской работы. Описаны результаты проведенных исследований, показаны их научная новизна и практическая значимость. Приведены данные об апробации основных результатов диссертационной работы.

В первом разделе диссертационной работы осуществлен обзор современных лингвистических ресурсов и систем автоматической обработки текстов казахского языка, проведен анализ существующих проблем формализации и алгоритмизации автоматической обработки текстов казахского языка. В разделе осуществлен анализ современных методов машинного обучения, используемых при обработки текстовой информации и выделены подходы Open IE, позволяющие получать информацию из многоязычных неструктурированных текстов. На основе проведенного анализа осуществлена постановка задачи исследования.

В втором разделе обоснован выбор математического аппарата алгебры конечных предикатов для моделирования процессов интеллектуальной обработки многоязычной текстовой информации. Рассмотрены основы инструментария алгебры предикатов и предикатных операций применительно

к его использованию для формализации конструкций естественных языков, идентифицирующих отношения между участниками действия в предложении. В разделе приведена разработанная логико-лингвистическая модель Open IE, описывающая семантические функции партиципантов предложения через отношения грамматических и семантических характеристик слов предложений заданного языка. Показана адаптация данной модели для автоматической генерации структурированной машиночитаемой информации из текстов казахского, русского и английского языков. В разделе приводится алгоритм перефразирования факта побуждения к действию в предложениях английского языка, полученный на базе разработанной математической модели извлечения фактов из многоязычной текстовой информации

Третий раздел посвящен разработке методов и алгоритмов морфологического и семантического анализа многоязычных текстов, на базе моделей машинного обучения. В разделе показан метод вероятностной морфологической и семантической разметки, использующий скрытую Марковскую модель (HMM). Используемая в методе функция оценки вероятности цепочки тегов зависит от двух вероятностей: условной вероятности последовательности тегов и условной вероятности обозначения токена данным тегом. Описан алгоритм семантической разметки текстов казахского языка. Первичная разметка корпуса казахского языка, на базе которого, осуществляется обучение, базируется на использовании списка суффиксов и лингвистических правил. В третьем разделе также показан метод, определения семантической близости многоязычных текстовых документов, базирующийся на вычислении косинусного сходство между двумя векторами документов VSM, который использует в качестве координат векторов меру PPMI, представляющую весовую функцию.

В четвертом разделе приведена практическая реализация полученных результатов. В разделе обосновано использования метрики числовых оценок, использующей в качестве объективно измеряемых показателей эффективности разработанных моделей кортеж, включающий коэффициенты полноты, точности и меру Ван Ризбергена. Также показаны практические результаты реализации разработанной модели Open IE на трех корпусах русского, казахского и английского текстов. Общий объем построенных корпусов: 6000 текстов из 700 000 слов. Точность извлечения триплета факта для английского языка представляет 87,2%, для русского языка 82,4% и для казахского 71,0%. Кроме того, в разделе описывается разработанная информационная технология определения семантической близости текстов к заданной узкоспециализированной тематике, базирующаяся на предложенных в диссертационном исследовании методах машинного обучения, и приводится модель оценки качества данной технологии .

В заключении изложены основные результаты и выводы данной диссертационной работы.

Обоснованность выносимых на защиту научных положений, выводов и рекомендаций подтверждается корректностью использования математического аппарата; корректной постановкой экспериментов;

качественным и количественным соответствием результатов теоретических исследований и экспериментальных данных; практическим применением результатов исследования.

Апробация работы. Результаты диссертационной работы докладывались на международных научных конференциях, ежегоных научных конференциях Института вычислительных и информационных технологий, научных конференциях молодых ученых и специалистов Казахского национального университета, а также научных семинарах кафедры «Информатика» КазНУ имени аль-Фараби.

Получены свидетельства о государственной регистрации прав на объект авторского права.

Научные публикации:

1. Мамырбаев О.Ж., Мухсина К.Ж. Мәтін үндесітілігін анықтауға арналған қолданыстағы жүйелерді талдау//«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы», 2017. - №5 (315). – Б.149-155.

2. Мамырбаев О.Ж., Мухсина К.Ж. Анализ текстовых сообщений с применением векторной формы//"Международная научно-практическая конференции «Математический методы и информационные технологии макроэкономического анализа и экономической политики», посвященной празднования 80-летнего юбилея академика НАН РК Абдыкаппара Ашимовича Ашимова", Алматы, 11.04.2017-12.04.2017.-С.136-144.

3.Хайрова Н.Ф., Избасаров Е.Ж., Мамырбаев О.Ж., Мухсина К.Ж. Формальная модель оценивания качества экстракции и идентификации знаний из слабоструктурированной текстовой информации// Материалы научной конференции ИИВТ МОН РК «Современные проблемы информатики и Вычислительных технологий». – 2018. - С.306 – 310.

4.Мамырбаев О.Ж., Хайрова Н. Ф., Мухсина К.Ж. Моделирование грамматических способов выражения семантики факта в английском предложении // III Международной научной конференции «Информатика и прикладная математика», посвященная 80-летнему юбилею профессора Бияшева Р.Г.и 70-летию профессора Айдарханова М.Б. 26-29 сентября 2018 года, Алматы. -С.136-144.

5.Petrasova S., Khairova, N., Lewoniewski W., Mamyrbaev O., Mukhsina K. Similar text fragments extraction for identifying common wikipedia communities// MDPI № 66 от 10.12.2018 https://doi.org/10.3390/data3040066.

6.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Automatic extraction of synonymous collocation pairs from a text corpus // Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, 2018, DOI: 10.15439/2018F186 Номер статьи 8511195, -P. 485-488.

7.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Comparative analysis of the informativeness and encyclopedic style of the popular web information sources// Lecture Notes in Business Information Processing 320, 2018, DOI: 10.1007/978-3-319-93931-5_24 -P. 333-344.

8.Mamyrbaev O., Turdalyuly M., Mekebayev N., Mukhsina K., Keylan A., Bagher B., Nabieva G., Duisenbayeva A., Akhmetov B. Continuous Speech Recognition of Kazakh Language // AMCSE 2018 - International Conference on Applied Mathematics, Computational Science and Systems Engineering. Vol. 24 – 2019.

9.Мамырбаев О.Ж., Мухсина К.Ж., Хайрова Н. Ф., Колесник А.С. Лингвистические инструменты выявления криминально окрашенной текстовой информации веб-контента // Қазақстан-Британ техникалық университетінің Хабаршысы – 2018. - №3(46). – Б. 112-117.

10.Хайрова Н. Ф., Мамырбаев О.Ж., Мухсина К.Ж., Колесник А.С. Автоматическая генерация структурированной машинно-читаемой информации из мультиязычных текстов // Информатика и прикладная математика: Матер. IV междунар. науч. конф. – Алматы, 2019. – Ч.2. - С. 509 – 519.

11.Мамырбаев О.Ж., Хайрова Н.Ф., Мухсина К.Ж. Қазақ тіліндегі мәтіндердегі қылмыстық мәнді коллакцияларды анықтау // Вестник КазАТК. – 2019. – № 3 (110). – С. 170-175.

12.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme // 3rd International Conference on Computational Linguistics and Intelligent Systems, 2019, Volume 2362.

13.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Detecting Collocations Similarity via Logical-Linguistic Model // In Proceedings of the Workshop on meaning relations between phrases and sentences - May 23, 2019, Gothenburg, Sweden, pp. 15-22.

14. Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Influence of Various Text Characteristics on the Readability and Content Informativeness // In Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-372-8, DOI: 10.5220/0007755004620469 - pp. 462-469.

15.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Open Information Extraction as Additional Source for Kazakh Ontology Generation // ACIIDS 2020, LNAI 12033, 2020. https://doi.org/10.1007/978-3-030-41964-6_8 -P. 86–96,

16.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. Logical-linguistic model for multilingual Open Information Extraction // Cogent Engineering (2020), https://doi.org/10.1080/23311916.2020.1714829 00: 1714829.

17. Хайрова Н. Ф., Колесник А.С., Мамырбаев О.Ж., Мухсина К.Ж. Выровненный казахско-русский параллельный корпус, ориентированный на криминальную тематику// Вестник Алматинского университета энергетики и связи № 1 (48) 2020- С. 84-92.

Свидетельства о государственной регистрации прав на объект авторского права :

Свидетельство № 9180 от 8 апреля 2020 г. о внесении сведений в Государственный реестр прав на объекты, охраняемые авторским правом, авторы: Мамырбаев О.Ж., Жұмажанов Б.Ж., Мухсина К.Ж.