АННОТАЦИЯ
диссертационной работы Әйтім Әйгерім Қайратқызы «Модели и методы автоматической обработки неструктурированной информации», представленной на соискание степени доктора философии (PhD) по ОП: 8D06101 – «Интеллектуальные системы»

Актуальность исследования определяется острой необходимостью в разработке интеллектуальных языковых технологий для казахского языка, что соответствует стратегическим задачам цифровой трансформации, научных инноваций и модернизации культуры в Республике Казахстан. Работа находится на пересечении искусственного интеллекта, обработки естественного языка (NLP) и национальной языковой политики, решая как научные, так и социотехнические задачи интеграции казахского языка в современные цифровые системы.

Одним из основных стимулов для проведения данного исследования является реализация государственной программы «Цифровой Казахстан», запущенной в 2017 году, целью которой является модернизация экономики, государственных услуг и инфраструктуры за счет внедрения цифровых технологий. Важнейшим компонентом данной программы является создание инклюзивных цифровых сервисов на казахском языке. Однако отсутствие надежных инструментов автоматической обработки казахскоязычных текстов - таких как определители частей речи, морфологические анализаторы и системы семантического анализа - представляет собой серьезный барьер на пути к достижению этих целей. Настоящая диссертация вносит вклад в преодоление данного пробела путем создания вычислительных моделей, способных обрабатывать неструктурированные тексты на казахском языке с высокой лингвистической точностью.

Кроме того, исследование поддерживает цели программы «Рухани жаңғыру» («Духовная модернизация»), направленной на сохранение, развитие и цифровую популяризацию казахского языка и культуры. Разрабатываемые в рамках работы инструменты обработки естественного языка - такие как морфологический анализатор, модели теггирования на основе KazBERT и аннотированные корпуса - способствуют использованию казахского языка в современных коммуникационных технологиях, образовательных платформах и культурных архивах. Обеспечивая автоматизированный лингвистический анализ и понимание крупномасштабного цифрового контента, данное исследование вносит вклад в укрепление языкового суверенитета и цифровой идентичности Казахстана.

Работа также соответствует приоритетам, изложенным в Стратегии развития науки Казахстана до 2025 года, где подчеркивается важность развития технологий искусственного интеллекта, методов обработки

больших данных и цифровых лингвистических ресурсов для казахского языка. В диссертации предлагается научно обоснованная методология построения языковых ресурсов и моделей глубокого обучения, адаптированных к особенностям казахской языковой системы, что способствует технологическому развитию обработки малоресурсных языков в стране.

Дополнительно исследование поддерживает инициативу «Білімді ұлт» («Образованная нация»), направленную на цифровизацию образования и разработку инновационных средств персонализированного обучения. Разработанные в рамках диссертации компоненты NLP могут быть интегрированы в интеллектуальные обучающие системы, приложения для изучения казахского языка и цифровые учебники, обеспечивая высококачественный лингвистически обоснованный образовательный контент на государственном языке.

Таким образом, все перечисленные национальные программы формируют стратегический запрос на разработку передовых вычислительных моделей, способных к автоматической обработке и пониманию неструктурированной информации на казахском языке. Результаты данного исследования, включая программную библиотеку QNLP, лингвистическую онтологию и аннотированные датасеты, предлагают практические решения для интеграции казахского языка в цифровую инфраструктуру Казахстана. Это гарантирует, что будущие технологические системы страны будут не только поддерживать пользователей казахского языка, но и укреплять его роль в информационную эпоху.

Объект исследования - неструктурированная текстовая информация на казахском языке, представленная в реальных доменах, включая цифровые новости, социальные сети, художественную литературу и официальные документы.

Предмет исследования - особенности и принципы построения вычислительных моделей и методов автоматической обработки, анализа и лингвистической аннотации неструктурированных казахских текстов с фокусом на фундаментальные задачи обработки естественного языка, такие как токенизация, определение частей речи, морфологический анализ, синтаксический разбор и семантическая разметка.

Цель исследования заключается в разработке интегрированной системы и набора лингвистических инструментов, объединяющих методы машинного обучения и языковые правила, для автоматической и точной обработки неструктурированной текстовой информации на казахском языке. Особое внимание уделяется анализу реальных казахскоязычных текстов из новостных источников, где структура предложений варьируется и не подчиняется формализованным шаблонам, что требует гибких и адаптивных решений в области компьютерной лингвистики.

Задачи исследования:

– провести обзор литературы по методам и моделям автоматической обработки неструктурированных текстовых данных, с особым акцентом на агглютинативные и тюркские языки;

– выявить ограничения и вызовы при применении существующих архитектур NLP к казахскому языку;

– собрать корпус неструктурированных казахских текстов из онлайн-источников;

– разработать и реализовать автоматизированные пайплайны для очистки текстов, удаления дубликатов, фильтрации по языку и предварительной аннотации;

– разработать и оценить модели теггирования частей речи с использованием как правил, так и методов машинного и глубокого обучения;

– построить модели для выделения именованных сущностей и синтаксического анализа с учетом специфических особенностей казахского языка и использования предварительно обученных языковых моделей;

– реализовать гибридную архитектуру моделей для задач последовательной разметки;

– спроектировать и разработать библиотеку QNLP как модульный программный инструмент, включающий все разработанные модели и ресурсы.

Научные вопросы исследования:

– Как можно создать модели обработки естественного языка (NLP), способные учитывать агглютинативные и богатые морфологические особенности казахского языка?

– Какие методы наиболее подходят для построения аннотированных корпусов в условиях ограниченности ресурсов?

– В какой степени возможно адаптировать технологии трансферного обучения и мультиязычные модели к казахскому языку?

– Как эффективно сочетать правила и методы машинного обучения для решения морфологических и синтаксических задач?

Научная гипотеза постулируется, что интеграция морфологических моделей на основе правил с предварительно обученными трансформерами казахского языка в сочетании с глубокими нейронными архитектурами приведет к значительному повышению точности и устойчивости систем автоматической обработки текстов на казахском языке. Также предполагается, что создание тщательно сконструированного корпуса и лингвистической онтологии обеспечит эффективное моделирование уникальных лингвистических структур казахского языка, способствуя обобщению моделей на различных задачах NLP в условиях ограниченности ресурсов.

Научная новизна исследования:

В ходе диссертационной работы получены следующие основные научные результаты:

-Собран, очищен, аннотирован оригинальный корпус неструктурированных текстов на казахском языке для воспроизводимых экспериментов.

– Разработана новая библиотека на Python - QNLP для лингвистической обработки, предлагающая полный набор инструментов для морфологического анализа, POS-теггинга, синтаксического разбора зависимостей и распознавания именованных сущностей (NER), адаптированных к особенностям казахского языка.

– Впервые реализована гибридная модель на основе правил и модели KazBERT + BiLSTM + CRF, обученной на аннотированном казахском корпусе для задач последовательной разметки.

Научные положения, выносимые на защиту:

– Корпус неструктурированных казахских текстов для воспроизводимых экспериментов.

– Библиотека QNLP на Python для морфологического анализа, POS-теггирования, синтаксического анализа и распознавания именованных сущностей казахского языка.

– Гибридная модель на основе правил и KazBERT + BiLSTM + CRF для задач последовательной разметки казахских текстов.

Теоретическая значимость заключается в развитии теории и методологии обработки естественного языка для морфологически богатых и агглютинативных языков, в частности в определении вычислительных моделей, адаптированных к тюркской типологии.

Практическая значимость состоит в разработке готовых к использованию инструментов NLP и аннотированных корпусов, применимых в задачах машинного перевода, поисковых систем, цифровых помощников и образовательных платформ на казахском языке.

Практическая ценность исследования заключается в создании открытой библиотеки QNLP для казахского языка с модулями для морфологического анализа, определения частей речи и выделения именованных сущностей, разработке протоколов лингвистической аннотации и цепочек аннотирования, а также в содействии сохранению казахского языка через цифровизацию текстовых ресурсов. Инструменты и ресурсы, разработанные в рамках исследования, могут быть интегрированы в национальную цифровую инфраструктуру и продукты языковых технологий.

Методы исследования включают сочетание качественных и количественных методов: сбор и предобработку крупномасштабных текстовых корпусов на казахском языке, морфологическое моделирование на основе правил, построение моделей определения частей речи и

синтаксической грамматики зависимостей, применение статистических и нейронных моделей для задач разметки последовательностей, использование метрик точности и F1-меры для оценки качества моделей, а также применение гибридных методов ручной и автоматической аннотации.

Апробация работы - основные положения и научные результаты докладывались на семинарах кафедры «Информационные системы» Международного университета информационных технологий и на международных конференциях:

1. The 6th International Conference on Engineering & MIS 2020, ICEMIS'20 (DTESI'20), September 14–16, 2020, ACM International Conference Proceeding Series, 2020.
2. The 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks/ EUSPN, Procedia Computer Science, 2024.
3. The 9th International Conference Digital Technologies in Education, Science, and Industry, 2024.

Публикации. Основные результаты, полученные в ходе выполнения диссертационной работы, были опубликованы в тринадцать печатных работах, включая 4 статьи в изданиях, рекомендованных Комитетом по контролю в сфере образования и науки Министерства образования и науки Республики Казахстан, а также 2 статьи в журнале Eastern-European Journal of Enterprise Technologies (Q3), индексируемом в базе Scopus, с cite score 2.0 и перцентилем 46. Кроме того, опубликовано 4 статей в материалах международных конференций, среди которых одна научная статья с cite score 4.5 и перцентилем 69 (Q2).

Результаты, полученные по теме диссертации, представлены в следующих публикациях:

1. Aitim A.K., Satybaldiyeva R.Zh., Linguistic ontology as means of modeling of a coherent text, Bulletin of Abai KazNPU. Series of Physical and mathematical sciences. 3 (Sep. 2022), https://doi.org/10.51889/2022-3.1728-7901.18.

2. Aitim A.K., Developing methods for automatic processing systems of Kazakh language, Bulletin of KazATC 133 (4), 2024, ISSN 1609-1817, ISSN Online 2790-5802, https://doi.org/10.52167/1609-1817.

3. Aitim A.K., Satybaldiyeva R.Zh., A systematic review of existing tools to automated processing systems for Kazakh language. Bulletin of Abai KazNPU. Series of Physical and mathematical sciences. 87, 3 (Sep. 2024), 106–122, https://doi.org/10.51889/2959-5894.2024.87.3.009.

4. Aitim A.K., Satybaldiyeva R.Zh., Building methods and models for automatic processing systems of Kazakh language, Bulletin of KazATC №2-137-2025, https://doi.org/10.52167/1609-1817-2025-137-2-346-356

5. Aitim A.K., Satybaldiyeva R.Zh. A comparison of Kazakh language processing models for improving semantic search results Eastern-European

Journal of Enterprise Technologies, 1(2 (133), 66–75, 2025. https://doi.org/10.15587/1729-4061.2025.315954

6. Aitim, A., Sattarkhuzhayeva, D., & Khairullayeva, A. (2025). Development of a hybrid CNN-RNN model for enhanced recognition of dynamic gestures in Kazakh Sign Language. Eastern-European Journal of Enterprise Technologies, 2025, 2(2 (134), 58–67. https://doi.org/10.15587/1729-4061.2025.315834

7. Aitim A.K., Satybaldiyeva R.Zh., Wojcik W.The construction of the Kazakh language thesauri in automatic word processing system the 6th International Conference on Engineering & MIS 2020, ICEMIS'20 (DTESI'20), September 14–16, 2020, ACM International Conference Proceeding Series, 2020, https://doi.org/10.1145/3410352.341078

8. Aitim A.K., Abdulla M.A. Data processing and analyzing techniques in UX research15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks/ EUSPN, Procedia Computer Science, volume 251, 2024, Pages 591-596, https://doi.org/10.1016/j.procs.2024.11.154

9. Aitim A.K., Abdulla M.A., Altayeva A.B. Sentiment Analysis using Natural Language Processing 9th International Conference Digital Technologies in Education, Science and Industry, October 16-17, 2024, Almaty.

10. Aitim A.K., Abdulla M.A., Altayeva A.B. Human-Centric AI: Improving User Experience with Natural Language Interfaces, 9th International Conference Digital Technologies in Education, Science and Industry, october 16-17, 2024, Almaty.

11. Aitim A.K., I. Khlevna. Models of natural language processing for improving semantic search results // International Journal of Information and Communication Technologies. 2022. Vol. 3. Is. 2. Number 10. Pp. 82–91. https://doi.org/10.54309/IJICT.2022.10.2.008.

12. Aitim A.K., Satybaldiyeva R.Zh. Analysis of methods and models for automatic processing systems of speech synthesis. International Journal of Information and Communication Technologies, 1(2), 2020. https://doi.org/10.54309/IJICT.2020.2.2.019

13. Aitim A.K., Wojcik W. Satybaldiyeva R.Zh. Methods of applying linguistic ontologies in text processing. Journal, News of the scientific and technical society KAKHAK, 2021, Volume-1, Issue - 72, Pages - 132-137.

Структура диссертации. Работа состоит из введения, четырёх разделов, заключения, списка использованной литературы и приложений.

Основное содержание диссертации.

Диссертация включает четыре основных главы.

Первая глава посвящена подробному обзору литературы и концептуальной основе обработки казахского языка. В ней проанализирована история развития и современное состояние инструментов NLP для казахского языка, подчёркнута сложность

агглютинативной морфологии и представлены современные нейросетевые архитектуры, такие как KazBERT, CRF и BiLSTM. Также рассматриваются ранее использованные лингвистические методы, первые системы обработки и трудности адаптации универсальных моделей к спецификации казахского языка.

Вторая глава представляет QCrawler - специально разработанную систему для автоматического сбора казахскоязычных текстов из интернета. Описываются архитектура, алгоритмы и математическая модель краулера, а также проводится сравнение его эффективности и охвата с базовыми методами парсинга. В главе также приведено первичное применение задачи распознавания именованных сущностей (NER) с использованием модели KazBERT+CRF на собранном корпусе.

Третья глава сосредоточена на лингвистической основе платформы QNLP, включая разработку морфологического анализатора, тезауруса казахского языка и онтологии казахской морфологии. Подробно описывается обработка словесной структуры на основе модели «корень + суффиксы», обеспечивающая высокоточную сегментацию и генерацию словоформ. Эти лексические и правил-ориентированные компоненты способствуют более глубокому лингвистическому анализу и повышают интерпретируемость моделей.

Четвёртая глава описывает полноценный инструментарий QNLP - полнофункциональную открытую платформу для обработки казахского языка. Излагается модульная архитектура, взаимодействие компонентов и внутренние математические формулировки. Глава включает результаты экспериментов по POS-разметке, NER и морфологии, исследование влияния компонентов KazBERT, BiLSTM и CRF, а также сравнение с существующими инструментами. Заканчивается глава визуализациями, графиками точности и сравнительными характеристиками, демонстрирующими передовые результаты QNLP.