

Сарсенбаева Асия на тему «Исследование моделей и разработка интеллектуальной вопросно-ответной системы на основе семантических подходов для государственного языка в сфере законодательства Республики Казахстан», представленной на соискание степени доктора философии (PhD) по образовательной программе 8D06101 – Информационные системы

АННОТАЦИЯ

Актуальность темы исследования.

Необходимость обеспечения доступа к законодательной информации на государственном языке Республики Казахстан обусловлена самой потребностью. В настоящее время процесс получения гражданами законодательной информации сложен, занимает много времени и зачастую труден для понимания из-за терминологической сложности правовых текстов. В этой связи актуальной научной проблемой является разработка интеллектуальных систем вопросов и ответов, анализирующих смысл и автоматически предоставляющих точные и понятные правовые ответы.

Быстрое развитие технологий искусственного интеллекта, особенно генеративных языковых моделей (LLM), открывает новые возможности для обработки и анализа законодательных текстов. Хотя такие модели, как GPT, LLaMA, Gemini, показали высокие результаты в задачах понимания и генерации текстов, их адаптация к агглютинативным и малоресурсным языкам, таким как казахский, изучена недостаточно. Это, в свою очередь, требует разработки специальных методов, учитывающих морфологическую сложность, разнообразие словообразования и особенности законодательной терминологии.

Современные системы (Adilet, eGov, LawVision) в основном используют поиск по ключевым словам или подходы, основанные на правилах, и не в полной мере понимают семантику вопроса пользователя. Кроме того, современные модели LLM страдают от проблемы «галлюцинаций» из-за отсутствия связи с реальными законодательными источниками, что негативно сказывается на точности юридических ответов.

В этой связи возникает необходимость в разработке интеллектуальной системы вопросов и ответов, учитывающей морфологические особенности казахского языка, сочетающей методы семантического анализа, метод генерации с расширенным поиском (RAG) и fine-tuning LLM модели. В частности, научно обоснованным является внедрение методов семантической предварительной обработки, сопоставляющих вопросы на естественном языке с законодательной терминологией.

Данная исследовательская работа направлена на повышение точности систем вопросов и ответов по законодательству на казахском языке, расширение доступности правовой информации и развитие интеллектуальных систем на национальном уровне.

Цель диссертационной работы

Исследование и разработка законодательной системы вопросов и ответов, учитывающей морфологические особенности казахского языка и сочетающей семантический анализ и генеративную модель.

Задачи исследования:

- Проанализировать текущее состояние систем вопросов и ответов по законодательству на казахском языке и определить исследовательские проблемы.
- Собрать документы с официальных порталов (adilet.kz, zan.gov.kz) и создать специализированную базу данных законодательства.
- Предложить двухэтапный метод семантической предварительной обработки, который определяет семантическую связь вопроса и преобразует его в законодательную форму.
- Повысить уровень понимания юридической терминологии на казахском языке путем fine-tuning больших языковых моделей (LLM) в законодательном корпусе.
- Разработать пайплайн для интеллектуальной системы вопросов и ответов по законодательству.
- Сравнить предложенную систему с базовыми моделями и экспериментально доказать ее эффективность.

Методы исследования

Обработка естественного языка (NLP), большие языковые модели (LLM) на основе архитектуры трансформера, технология генерации с расширенным поиском (RAG), семантический анализ SVD/LSA, трансформер mT5, метрики оценки и методы экспертной оценки.

Объект исследования

Интеллектуальные системы вопросов и ответов для обеспечения доступа к законодательной базе Республики Казахстан на государственном языке, а также поиска и обработки правовой информации на основе семантического анализа.

Предмет исследования

Методы семантического анализа, учитывающие лингвистические особенности и морфологическую структуру казахского языка, методы формализации законодательной терминологии и методы тонкой настройки больших языковых моделей (LLM) на основе корпуса данных в области законодательства.

Научная новизна полученных результатов:

1. Учитывая высокую сложность и терминологическую специфику юридических текстов, был разработан двухэтапный подход к семантической предварительной обработке для сопоставления вопросов на естественном языке с законодательной терминологией.

2. На основе разработанных методов была создана интеллектуальная система вопросов и ответов, объединяющая семантическую предварительную обработку, адаптивную LLM и RAG-технологии. Его эффективность была подтверждена на экспериментальных данных с использованием качественных и количественных метрик.

Положения, выносимые на защиту:

1. Был составлен специализированный набор данных, состоящий из цифрового корпуса нормативно-правовых актов Республики Казахстан объемом 5,3 ГБ и 8000 пар «вопрос-ответ».
2. В результате тонкой настройки больших языковых моделей на основе собранного набора законодательных данных, текстовая точность ответов увеличилась в 1,5-2 раза по сравнению с базовыми моделями ROUGE.
3. Предложен двухэтапный подход к семантической предварительной обработке, который сопоставляет запрос на простом языке с законодательной терминологией. Этот метод улучшил качество поиска информации по показателю Recall@5 на 11-15% и по метрике MRR на 16-18%.
4. Была разработана система, объединяющая семантическую предварительную обработку, тонкую настройку LLM и технологии RAG. Это позволило повысить семантическую точность ответов по BERTScore на 3-5% и по Expert Score на 25-30%.

Объем и структура диссертационной работы

Диссертация состоит из введения, четырех основных глав, заключения, списка использованной литературы и приложения. Общий объем работы составляет 141 страниц машинописного текста, включая 40 рисунков, 41 таблиц, 108 использованные литературные работы и 1 приложение.

В первой главе представлен аналитический обзор интеллектуальных систем вопросов и ответов в сфере законодательства: эволюция развития и современные архитектуры систем вопросов и ответов, использование больших языковых моделей (ЛЯМ) в правовых информационных системах, лингвистические и семантические особенности обработки законодательных текстов, текущее состояние развития правовых систем вопросов и ответов на казахском языке, а также систематически рассматривается сравнительный анализ отечественных и зарубежных исследований.

Во второй главе описываются методы формирования законодательной базы данных Республики Казахстан и адаптации моделей LLM: анализируются ресурсы правовой информации и цифровая инфраструктура;

предлагаются методы формирования корпуса законодательных текстов и набора данных вопросов и ответов, сбора текстов, предварительной обработки и аннотирования; рассматривается сравнительный анализ базовых моделей LLM, предметная адаптация моделей на основе методов тонкой настройки, встраивания и семантического представления; также описывается модуль поиска архитектуры RAG, модель генерации генеративных ответов и общая архитектурная схема системы.

В третьей главе описывается концептуальная и архитектурная модель предлагаемого двухэтапного метода семантической предварительной обработки — SemanticPrep, формализация вычисления вопросов и ответов, адаптация модели mT5 посредством тонкой настройки и модель преобразования вопросов на естественном языке в юридические структуры; математическая модель метода семантического расширения на основе SVD/LSA и методы оценки семантического сходства, а также общий алгоритм и вычислительная сложность предлагаемого метода.

В четвертой главе представлена экспериментальная оценка предложенного метода SemanticPrep: сравнительный анализ моделей LLM, результаты реализации системы RAG, влияние семантической предварительной обработки на качество ответа, а также результаты тестирования предложенной системы в практической среде.

Апробация работы

Результаты исследования обсуждались на научных семинарах Казахского национального университета имени аль-Фараби, докладывались на международных конференциях, а также прошли опытную апробацию в филиале № 1 ТОО «Корпорация Алмалы» и в ТОО «ART Service».

Публикации.

По теме диссертации опубликовано 9 научных работ, в том числе 4 статьи в изданиях, индексируемых в базах данных Scopus и Web of Science, 2 работы в материалах международных конференций и 3 работы в журналах, рекомендованных Министерством науки и высшего образования Республики Казахстан.

1. Рахимова Д., Карюкин В., Турарбек А., Сарсенбаева А.К, Алиев Р. (2025). Правовой ИИ в языках с ограниченными ресурсами: создание и оценка систем вопросов и ответов для казахстанского законодательства. Компьютеры (MDPI). 14(9), 354. <https://doi.org/10.3390/computers14090354> WoS (SCIE), Q1, 2025; 5-летний импакт-фактор: 2,7; SJR: 0,492; Перцентиль: 84 (Информатика, Прикладные науки)
2. Д. Рахимова, А. Сарсенбаева, Е. Адали, А. Кульжанова, Классификация вопросов в интеллектуальной системе вопросов и ответов для государственного языка, Вестник НИА РК № 4 Алматы, 2024, - С. 166-179 <https://doi.org/10.47533/2024.1606-146X.014>

3. Рахимова Д.Р., Сарсенбаева А.К., Турарбек А., Ауэзова А.С. Применение глубокого обучения для повышения точности ответов в многоязычных системах вопросов и ответов. Вестник НАН РК, физико-математическая серия, № 2, Алматы, 2025, - С. 196-211 <https://doi.org/10.32014/2025.2518-1726.353>
4. Сарсенбаева А.К., Рахимова Д.Р., Шормакова А.Н., Мансурова М. Контекстно-ориентированный подход к выявлению семантического сходства в юридических текстах агглютинативных языков: пример казахского языка. Вестник Национальной академии наук Республики Казахстан, – серия «Физика и математика», № 1, Алматы 2026 – С. 354-372 <https://doi.org/10.32014/2026.2518-1726.417>
5. Рахимова Д., Карюкин В., Карибаева А., Сарсенбаева А.К. Разработка передовых моделей ответов на вопросы для правовых казахских текстов: сравнительное исследование современных подходов. ASIIDS 2025: 17-я Азиатская конференция по интеллектуальным информационным и баз данных системам. DOI: 10.1007/978-981-96-5881-7_19 (Scopus, процентиль 39)
6. Рахимова, Д. , Карюкин, В. , Амирова, Д. , Сарсенбаева, А. Сбор и предварительная обработка данных для магистерской программы по казахскому языку в области законодательства. Communications in Computer and Information Science, 2024, 2211 CCIS, с. 129–139. Springer. DOI: 10.1007/978-3-031-72260-8_11 (Scopus, процентиль)
7. Сарсенбаева А.К., Рахимова Д.Р., Турарбек А. Многоязычные большие языковые модели в юридической сфере: анализ эффективности на казахском, турецком и английском языках. 10-я Международная конференция по информатике и инженерии (UBMK), Стамбул, Турция, 2025, стр. 410-415, doi: 10.1109/UBMK67458.2025.11207002. (Scopus, 39-й процентиль)
8. Сарсенбаева А.К., Рахимова Д.Р. Оценка качества ответов на вопросы о законодательстве на казахском, турецком и английском языках: экспериментальное исследование многоязычных моделей LLM. X Международная научно-практическая конференция «Информатика и прикладная математика», Алматы, 2025. – С. 293-302
9. Рахимова Д., Сарсенбаева А. Исследование текущей ситуации в системах вопросов и ответов, связанных с областью законодательства, информатики и прикладной математики: Мат. IX Межнаучно-практической конференции (31.10 - 01.11.2024). – Алматы, 2024, – С. 76-83.