

Сарсенбаева Асияның 8D06101 – Ақпараттық жүйелер білім беру бағдарламасы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған «Қазақстан Республикасының заңнамасы саласындағы мемлекеттік тілге арналған семантикалық тәсілдер негізінде модельдерді зерттеу және интеллектуалды сұрақ-жауап жүйесін әзірлеу» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Зерттеу тақырыбының өзектілігі.

Қазақстан Республикасының заңнамалық ақпараттарына мемлекеттік тілде қолжетімділікті қамтамасыз ету қажеттілігімен айқындалады. Қазіргі таңда азаматтардың заңнамалық ақпаратты алу процесі күрделі, көп уақытты талап етеді және көбінесе құқықтық мәтіндердің терминологиялық күрделілігіне байланысты түсінуді қиындатады. Осыған байланысты мағынасы бойынша талдап, нақты және дәл құқықтық жауаптарды автоматты түрде ұсынатын интеллектуалды сұрақ-жауап жүйелерін әзірлеу өзекті ғылыми мәселе болып табылады.

Жасанды интеллект технологияларының, әсіресе генеративті тілдік модельдердің (LLM) қарқынды дамуы заңнамалық мәтіндерді өңдеу мен талдаудың жаңа мүмкіндіктерін ашуда. GPT, LLaMA, Gemini сияқты модельдер мәтіндерді түсіну және генерациялау міндеттерінде жоғары нәтижелер көрсеткенімен, оларды қазақ тілі сияқты агглютинативті және ресурсы шектеулі тілдерге бейімдеу жеткілікті деңгейде зерттелмеген. Бұл өз кезегінде морфологиялық күрделілік, сөз түрленуінің көптүрлілігі және заңнамалық терминологияның ерекшеліктерін ескере отырып, арнайы әдістерді әзірлеуді талап етеді.

Қазіргі қолданыстағы жүйелер (Adilet, eGov, LawVision) негізінен кілт сөздерге негізделген іздеу немесе ережелік тәсілдерді қолданады және пайдаланушы сұрағын семантикалық тұрғыдан толық түсінбейді. Сонымен қатар, заманауи LLM модельдерінде нақты заңнамалық дереккөздермен байланыс болмауына байланысты hallucination мәселесі орын алады, бұл құқықтық жауаптардың дәлдігіне кері әсер етеді.

Осыған байланысты, қазақ тілінің морфологиялық ерекшеліктерін ескеретін, семантикалық талдау әдістерін, Retrieval-Augmented Generation (RAG) тәсілін және ірі тілдік модельдерді біріктіретін интеллектуалды сұрақ-жауап жүйесін әзірлеу қажеттілігі туындайды. Әсіресе табиғи тілдегі сұрақтарды заңнамалық терминологияға сәйкестендіретін семантикалық препроцессинг әдістерін енгізу ғылыми тұрғыдан өзекті болып табылады.

Осы зерттеу жұмысы қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің дәлдігін арттыруға, құқықтық ақпараттың қолжетімділігін

кеңейтуге және ұлттық деңгейдегі интеллектуалды жүйелерді дамытуға бағытталған.

Диссертациялық жұмыстың мақсаты

Қазақ тілінің морфологиялық ерекшеліктерін ескеретін, семантикалық талдау мен генеративті модельді біріктіретін заңнамалық сұрақ-жауап жүйесін зерттеу және әзірлеу.

Зерттеу міндеттері:

- Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің қазіргі жағдайын талдап, ғылыми зерттеу мәселелерін анықтау.
- Ресми порталдардан (adilet.kz, zan.gov.kz) құжаттарды жинап, мамандандырылған заңнамалық деректер қорын құрастыру.
- Сұрақтың мағыналық байланысын анықтап, оны заңнамалық формаға келтіретін екі кезеңді семантикалық препроцессинг тәсілін ұсыну.
- Ірі тілдік модельдерді (LLM) заңнамалық корпуста fine-tuning арқылы олардың қазақ тіліндегі құқықтық терминологияны түсіну деңгейін арттыру
- Интеллектуалды заңнамалық сұрақ-жауап жүйесінің пайплайнын әзірлеу
- Ұсынылған жүйені baseline модельдермен салыстырып, эксперименттік түрде тиімділігін дәлелдеу

Зерттеу әдістері

Табиғи тілдерді өңдеу (NLP), Transformer архитектурасына негізделген ірі тілдік модельдер (LLM), Retrieval-Augmented Generation (RAG) технологиясы, SVD/LSA семантикалық талдауы, mT5 трансформері, бағалау метрикалары және эксперттік бағалау әдістері.

Зерттеу нысаны

Қазақстан Республикасының заңнамалық базасына мемлекеттік тілде қолжетімділікті қамтамасыз ету және құқықтық ақпаратты семантикалық талдау негізінде іздеу мен өңдеуге арналған интеллектуалды сұрақ-жауап жүйелері.

Зерттеу пәні

Қазақ тілінің лингвистикалық ерекшеліктері мен морфологиялық құрылымын ескеретін семантикалық талдау әдістері, заңнамалық терминологияны формализациялау тәсілдері және ірі тілдік модельдерді (LLM) заңнама саласындағы деректер корпусына fine-tuning әдістемелері.

Алынған нәтижелердің ғылыми жаңалығы:

1. Құқықтық мәтіндердің жоғары күрделілігі мен терминологиялық нақтылығын ескере отырып, табиғи тілдегі сұрақтарды заңнамалық терминологиямен сәйкестендіруге арналған екі кезеңді семантикалық препроцессинг тәсілі әзірленді.

2. Әзірленген тәсілдер негізінде семантикалық препроцессинг, бейімделген LLM және RAG технологияларын біріктіретін интеллектуалды сұрақ-жауап жүйесінің пайплайны құрылды. Оның тиімділігі эксперименттік деректерде сапалық және сандық метрикалар арқылы расталды.

Қорғауға шығарылатын тұжырымдар:

1. ҚР нормативтік-құқықтық актілерінің 5.3 Гб көлеміндегі цифрлық корпусы мен 8 000 сұрақ-жауап жұбынан тұратын мамандандырылған датасет жинақталды.
2. Ірі тілдік модельдерді жинақталған заңнамалық датасетте fine-tuning жасау нәтижесінде жауаптардың мәтіндік сәйкестігі ROUGE базалық модельдермен салыстырғанда 1.5-2 есеге артты.
3. Қарапайым тілдегі сұрақты заңнамалық терминологияға сәйкестендіретін екі кезеңді семантикалық препроцессинг тәсілі ұсынылды. Бұл әдіс ақпаратты іздеу сапасын Recall@5 11-15%-ға, ал MRR метрикасын 16-18%-ға жоғарылатты.
4. Семантикалық препроцессинг, fine-tuning LLM және RAG технологияларын біріктіретін пайплайн әзірленді. Бұл арқылы жауаптардың мағыналық дәлдігі BERTScore 3-5%-ға және Expert Score 25-30% дейін жақсарды.

Диссертациялық жұмыстың көлемі мен құрылымы

Диссертация кіріспеден, төрт негізгі тараудан, қорытындыдан, пайдаланылған әдебиеттер тізімінен және қосымшалардан тұрады. Жұмыстың жалпы көлемі — 141 бет машиналық мәтін, оның ішінде 40 сурет, 41 кесте, 108 пайдаланылған әдебиет және 1 қосымша.

Бірінші тарауда заңнама саласындағы интеллектуалды сұрақ-жауап жүйелеріне аналитикалық шолу жасалады: сұрақ-жауап жүйелерінің даму эволюциясы және заманауи архитектуралары, үлкен тілдік модельдердің (LLM) құқықтық ақпараттық жүйелердегі қолданылуы, заңнамалық мәтіндерді өңдеудің лингвистикалық және семантикалық ерекшеліктері, қазақ тіліндегі құқықтық сұрақ-жауап жүйелерінің қазіргі даму жағдайы, сондай-ақ отандық және шетелдік зерттеулерге салыстырмалы талдау жүйелі түрде қарастырылады.

Екінші тарауда Қазақстан Республикасының заңнамалық деректер қорын қалыптастыру және LLM модельдерін бейімдеу әдістері баяндалады: құқықтық ақпараттық ресурстар мен цифрлық инфрақұрылым талданады; заңнамалық мәтіндер корпусын және сұрақ-жауап деректер жиынтығын қалыптастыру, мәтіндерді жинау, алдын ала өңдеу және аннотациялау әдістері ұсынылады; базалық LLM модельдерінің салыстырмалы талдауы, fine-tuning негізінде модельдерді домендік бейімдеу, embedding және семантикалық ұсыну тәсілдері қарастырылады; сондай-ақ RAG архитектурасының retrieval модулі, генеративті жауап қалыптастыру моделі және жүйенің жалпы архитектуралық схемасы сипатталады.

Үшінші тарауда ұсынылған екі кезеңді семантикалық препроцессинг — SemanticPrep әдісінің концептуалдық және архитектуралық моделі, сұрақ-жауап есебінің формализациясы, mT5 моделін fine-tuning арқылы бейімдеу және табиғи тілдегі сұрақтарды құқықтық құрылымға түрлендіру моделі баяндалады; SVD/LSA негізіндегі семантикалық кеңейту әдісінің математикалық моделі және семантикалық ұқсастықты бағалау тәсілдері, сондай-ақ ұсынылған әдістің жалпы алгоритмі мен есептеу күрделілігі ұсынылады.

Төртінші тарауда ұсынылған SemanticPrep әдісінің эксперименттік бағалауы жүргізіледі: LLM модельдерінің салыстырмалы талдауы, RAG жүйесінің іске асырылу нәтижелері, семантикалық препроцессингтің жауап сапасына әсері, сондай-ақ ұсынылған жүйенің практикалық ортада апробациясының нәтижелері ұсынылады.

Жұмыстың апробациясы

Зерттеу нәтижелері әл-Фараби атындағы Қазақ ұлттық университетінің ғылыми семинарларында талқыланып, халықаралық конференцияларда баяндалды, сондай-ақ ЖШС «Корпорация Алмалы» №1 филиалында және ЖШС «ART Service» тәжірибелік апробациядан өтті.

Жарияланымдар.

Диссертация тақырыбы бойынша 9 ғылыми жұмыс жарияланды, оның ішінде Scopus және Web of Science базаларына индексделетін басылымдарда 4 мақала, 2 жұмыс халықаралық конференция жинағында, 3 жұмыс ҚР Ғылым және жоғары білім министрлігі ұсынатын журналдарда жарияланды.

1. Rakhimova D., Karyukin V., Turarbek A., Sarsenbayeva A.K., Alieyev R. (2025). Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation. Computers (MDPI). 14(9), 354. <https://doi.org/10.3390/computers14090354> WoS (SCIE), Q1, 2025; 5-Year IF: 2.7; SJR: 0.492; Percentile: 84 (Computer Science, Applications)
2. D.Rakhimova, A.Sarsenbayeva, E. Adali, A. Kulzhanova, Classification of questions in the intellectual question-answer system for the state language, Вестник НИА РК №4 Алматы, 2024, – С. 166-179 (КОКСНВО МНВО РК) <https://doi.org/10.47533/2024.1606-146X.014>
3. Рахимова Д.Р., Сарсенбаева А.К., Турарбек А., Ауезова А.С. Применение глубокого обучения для повышения точности ответов в мультязычных вопросно-ответных системах. Вестник НАН РК, физико-математическая серия, № 2, Алматы, 2025, – С. 196-211 (КОКСНВО МНВО РК) <https://doi.org/10.32014/2025.2518-1726.353>
4. Сарсенбаева А.К., Рахимова Д.Р., Шормакова А.Н., Мансурова М. Агглютинативті тілдердің заңды мәтіндеріндегі мағыналық ұқсастықты анықтаудың контекстке негізделген тәсілі: қазақ тіліндегі жағдайды зерттеу. Қазақстан Республикасы Ұлттық ғылым академиясының

Хабаршысы, - физика математикалық сериясы, № 1, Алматы 2026 – С. 354-372 <https://doi.org/10.32014/2026.2518-1726.417>

5. Rakhimova D., Karyukin V., Karibayeva A., Sarsenbayeva A.K. Developing advanced question-answering models for Legal Kazakh texts: A comparative study of modern approaches. ACIIDS 2025: 17th Asian Conference on Intelligent Information and Database Systems. DOI: 10.1007/978-981-96-5881-7_19 (Scopus, проценти́ль 39)
6. Rakhimova, D., Karyukin, V., Amirova, D., Sarsenbayeva, A. Collection and Preprocessing of Data for LLM in the Kazakh Language in the Field of Legislation. Communications in Computer and Information Science, 2024, 2211 CCIS, P.129–139. Springer. DOI:10.1007/978-3-031-72260-8_11 (Scopus, проценти́ль 39)
7. Sarsenbayeva A.K., Rakhimova D.R., Turarbek A. Multilingual Large Language Models in the Legal Domain: An Effectiveness Analysis in Kazakh, Turkish, and English. 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025, pp. 410-415, doi: 10.1109/UBMK67458.2025.11207002. (Scopus, проценти́ль 39)
8. Сарсенбаева А.К., Рахимова Д.Р. Қазақ, түрік және ағылшын тілдеріндегі заңнама сұрақтарына жауап беру сапасын бағалау: көптілді LLM үлгілерінің тәжірибелік зерттеуі. X Международная научно-практическая конференция «Информатика и прикладная математика», Алматы, 2025. – С. 293-302
9. Рахимова Д., Сарсенбаева А. Заңнама саласына байланысты сұрақ-жауап жүйелеріндегі заманауи жағдайды зерттеу Информатика и прикладная математика: Мат. IX Межд. науч.-практ. конф. (31.10 - 01.11.2024 г.). – Алматы, 2024, – С. 76-83.