Әйтім Әйгерім Қайратқызының 8D06101 – «Интеллектуалды жүйелер» білім беру бағдарламасы бойынша философия докторы (PhD) ғылыми дәрежесін алу үшін ұсынылған «Құрылымданбаған ақпаратты өңдеудің үлгілері мен әдістерін автоматтандыру» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Зерттеу тақырыбының өзектілігі – Қазақстан Республикасында цифрлық трансформация, ғылыми инновациялар және мәдениетті жаңғырту стратегиялық міндеттері аясында қазақ тілі үшін интеллектуалды тілдік технологияларды әзірлеу қажеттілігінің артуымен айқындалады. Жұмыс жасанды интеллект, табиғи тілді өңдеу (NLP) және ұлттық тіл саясатының түйіскен нүктесінде орналасқан және қазақ тілін заманауи цифрлық жүйелерге интеграциялау бойынша ғылыми әрі әлеуметтік-техникалық міндеттерді шешуге бағытталған.

Осы зерттеуді жүргізуге түрткі болған негізгі факторлардың бірі – 2017 жылы іске қосылған «Цифрлық Қазақстан» мемлекеттік бағдарламасы. Бағдарламаның мақсаты – цифрлық технологияларды енгізу арқылы экономиканы, мемлекеттік қызметтерді және инфрақұрылымды жаңғырту. Бағдарламаның маңызды құрамдас бөлігі – қазақ тіліндегі инклюзивті цифрлық сервистерді құру. Алайда, қазақ тіліндегі мәтіндерді автоматты өңдеуге арналған сенімді құралдардың – сөз таптарын анықтаушылардың, морфологиялық талдаушылардың және семантикалық талдау жүйелерінің болмауы бұл мақсатқа жетуде елеулі кедергі болып отыр. Осы диссертациялық жұмыс қазақ тіліндегі құрылымданбаған мәтіндерді жоғары лингвистикалық дәлдікпен өңдей алатын есептеу модельдерін әзірлеу арқылы осы олқылықтың орнын толтыруға бағытталған.

Сонымен қатар, зерттеу «Рухани жаңғыру» бағдарламасының мақсаттарына сәйкес келеді. Бұл бағдарлама қазақ тілін және мәдениетін сақтау, дамыту және цифрлық кеңістікте танымал етуге бағытталған. Зерттеу шеңберінде әзірленген табиғи тілді өңдеу құралдары – морфологиялық талдаушы, KazBERT негізіндегі таңбалау модельдері және аннотацияланған корпустар – қазақ тілін заманауи коммуникациялық технологияларда, білім беру платформаларында және мәдени мұра архивтерінде пайдалануға ықпал етеді. Ірі көлемдегі цифрлық контентті автоматты түрде тілдік талдау және түсіну мүмкіндігін қамтамасыз ете отырып, бұл зерттеу Қазақстанның тілдік егемендігі мен цифрлық болмысын нығайтуға өз үлесін қосады.

Жұмыс сондай-ақ Қазақстанның 2025 жылға дейінгі Ғылымды дамыту стратегиясында белгіленген басымдықтарға сәйкес келеді, онда қазақ тілі үшін жасанды интеллект технологияларын, үлкен деректерді өңдеу әдістерін және цифрлық тілдік ресурстарды дамыту қажеттілігі атап өтіледі.

Диссертацияда қазақ тілінің ерекшеліктеріне бейімделген тілдік ресурстар мен терең оқыту модельдерін әзірлеудің ғылыми негізделген әдістемесі ұсынылады, бұл өз кезегінде аз ресурсты тілдерді өңдеу технологиясының дамуына ықпал етеді.

Зерттеу сондай-ақ «Білімді ұлт» бастамасына сәйкес келеді, оның мақсаты – білім беруді цифрландыру және дербестендірілген оқытуға арналған инновациялық құралдарды әзірлеу. Диссертациялық жұмыста әзірленген NLP компоненттері қазақ тілін оқытуға арналған интеллектуалды жүйелерге, мобильді қосымшаларға және цифрлық оқулықтарға интеграциялануы мүмкін, бұл мемлекеттік тілде сапалы және лингвистикалық тұрғыдан негізделген білім беру контентін қамтамасыз етеді.

Осылайша, аталған ұлттық бағдарламалар қазақ тіліндегі құрылымданбаған ақпаратты автоматты түрде өңдей және түсіне алатын алдыңғы қатарлы есептеу модельдерін әзірлеуге стратегиялық сұранысты қалыптастырады. Зерттеу нәтижелері – QNLP бағдарламалық кітапханасы, лингвистикалық онтология және аннотацияланған деректер жиынтығы – қазақ тілін Қазақстанның цифрлық инфрақұрылымына интеграциялауға арналған практикалық шешімдер ұсынады. Бұл болашақтағы технологиялық жүйелердің қазақ тілін қолдап қана қоймай, оның ақпараттық дәуірдегі орнын нығайтуына мүмкіндік береді.

Зерттеу нысаны – сандық жаңалықтар, әлеуметтік желілер, көркем әдебиет, ресми құжаттар) қазақ тіліндегі құрылымданбаған мәтіндік ақпарат.

Зерттеу пәні – қазақ тіліндегі құрылымданбаған мәтіндерді автоматты өңдеу, талдау және лингвистикалық аннотациялау үшін есептеу модельдері мен әдістерін құру ерекшеліктері мен принциптері. Негізгі назар токенизация, сөз таптарын анықтау, морфологиялық талдау, синтаксистік талдау және семантикалық таңбалау сияқты табиғи тілді өңдеудің базалық міндеттеріне аударылады.

Зерттеудің мақсаты – қазақ тіліндегі құрылымданбаған мәтіндік ақпаратты автоматты және дәл өңдеуге арналған машиналық оқыту әдістері мен тілдік ережелерді біріктіретін интеграцияланған жүйе мен лингвистикалық құралдар жиынтығын әзірлеу. Ерекше назар сөйлем құрылымы әртүрлі әрі формальды үлгілерге бағынбайтын жаңалықтар мәтініндегі шынайы қазақша мәтіндерді талдауға аударылады, бұл компьютерлік лингвистикада икемді әрі бейімделетін шешімдерді қажет етеді.

Зерттеу міндеттері:

– құрылымданбаған мәтіндік деректерді автоматты өңдеу әдістері мен модельдері бойынша әдеби шолу жүргізу, агглютинативті және түркі тілдеріне ерекше назар аудару;

– қолданыстағы NLP архитектураларын қазақ тіліне қолданудағы шектеулер мен қиындықтарды анықтау;

– онлайн дереккөздерден (жаңалық порталдары, блогтар, ресми құжаттар) қазақ тіліндегі құрылымданбаған мәтіндер корпусын жинау;

– мәтіндерді тазарту, қайталанатын мәтіндерді жою, тілдік сүзгіден өткізу және алдын ала аннотациялау үшін автоматтандырылған өңдеу пайплайнын әзірлеу және іске асыру;

– ережелерге және машиналық, терең оқыту әдістеріне негізделген сөз таптарын анықтау модельдерін әзірлеу және бағалау;

– қазақ тілінің ерекшеліктерін және алдын ала оқытылған тілдік модельдерді (мысалы, KazBERT) ескере отырып, атаулы есімдіктерді тану және синтаксистік талдау модельдерін құру;

– тізбектік таңбалау тапсырмалары үшін гибридті модель архитектурасын іске асыру;

– әзірленген модельдер мен ресурстарды қамтитын модульдік бағдарламалық құрал ретінде QNLP кітапханасын жобалау және әзірлеу.

Зерттеудің ғылыми сұрақтары:

– Қазақ тілінің агглютинативтік және бай морфологиялық ерекшеліктерін ескеретін NLP модельдерін қалай құруға болады?

– Шектеулі ресурстар жағдайында аннотацияланған корпустарды құрудың қандай әдістері ең тиімді болып табылады?

– Трансферлік оқыту технологияларын және көптілді алдын ала оқытылған модельдерді (мысалы, KazBERT) қазақ тіліне қаншалықты бейімдеуге болады?

– Морфологиялық және синтаксистік міндеттерді шешу үшін ережелер мен машиналық оқыту әдістерін тиімді түрде қалай біріктіруге болады?

Зерттеу болжамы:

Қазақ тіліндегі мәтіндерді автоматты түрде өңдеудің дәлдігі мен орнықтылығын айтарлықтай арттыру үшін морфологиялық ережелерге негізделген модельдерді қазақ тіліне арналған алдын ала оқытылған трансформерлермен және терең нейрондық архитектуралармен интеграциялау қажет деп болжанады. Сондай-ақ, мұқият құрылған корпус пен лингвистикалық онтология қазақ тіліне тән ерекше құрылымдарды тиімді модельдеуге мүмкіндік береді, бұл шектеулі ресурстар жағдайында түрлі NLP міндеттеріне моделдерді жалпы қолдануға ықпал етеді.

Диссертациялық жұмыстың ғылыми жаңалығы:

Диссертациялық жұмыс барысында келесі негізгі ғылыми нәтижелер алынды:

– Қазақ тіліндегі құрылымдалмаған мәтіндердің түпнұсқа корпусы жинақталып, тазартылып, аннотацияланып, қайталанатын эксперименттер үшін дайындалды.

– Қазақ тілінің ерекшеліктеріне бейімделген морфологиялық талдау, сөз таптарын белгілеу (POS-теггинг), тәуелділік синтаксисі және атаулы

мәндерді тану (NER) үшін толық құралдар жиынтығын ұсынатын жаңа QNLP атты Python кітапханасы әзірленді.

– Қазақша аннотацияланған корпус негізінде ережелерге сүйенген және KazBERT + BiLSTM + CRF модельдерін біріктіретін алғашқы гибридтік модель жүзеге асырылды.

Қорғауға ұсынылатын қағидаттар:

– Қайталанатын эксперименттер үшін қазақ тіліндегі құрылымдалмаған мәтіндер корпус.

– Қазақ тілінің морфологиялық талдауы, сөз таптарын белгілеу, тәуелділік синтаксисі және атаулы мәндерді тану үшін Python тілінде QNLP кітапхана.

– Қазақ мәтіндерінің тізбектік белгілеу міндеттері үшін ережелерге негізделген және KazBERT + BiLSTM + CRF модельдері қолданылған гибридтік модель.

Зерттеудің практикалық құндылығы:

Зерттеу барысында қазақ тіліне арналған ашық қолжетімді QNLP кітапханасы әзірленді, ол морфологиялық талдау, сөз таптарын анықтау және атаулы есімдіктерді тану модульдерін қамтиды. Қазақ тілінің ерекшеліктеріне бейімделген лингвистикалық аннотациялау протоколдары мен аннотациялау тізбегі құрылды. Зерттеу нәтижелері қазақ тілінің цифрлануы мен сақталуына ықпал етеді және ұлттық цифрлық инфрақұрылым мен тілдік технология өнімдеріне интеграциялануы мүмкін.

Зерттеу әдістері сапалық және сандық әдістер үйлесімді қолданылды: қазақ тіліндегі ірі көлемдегі мәтіндік корпустарды жинау және алдын ала өңдеу, ережелерге негізделген морфологиялық модельдеу, сөз таптарын анықтау және тәуелділік грамматикасын құру, тізбектік таңбалау тапсырмалары үшін статистикалық және нейрондық модельдерді қолдану, модель сапасын бағалау үшін дәлдік пен F1-өлшемдерін пайдалану, сондай-ақ қолмен және автоматты аннотациялаудың гибридті әдістерін қолдану.

Жұмыстың апробациясы:

Диссертациялық жұмыстың негізгі ережелері мен ғылыми нәтижелері Халықаралық ақпараттық технологиялар университетінің «Ақпараттық жүйелер» кафедрасының семинарларында және халықаралық конференцияларда баяндалды:

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.

Жарияланымдар:

Диссертациялық жұмысты орындау барысында алынған негізгі нәтижелер он үш мақала басылымда жарияланды, оның ішінде Қазақстан Республикасы Білім және ғылым министрлігінің Білім және ғылым саласындағы сапаны қамтамасыз ету комитеті ұсынған басылымдарда 4 мақала және Scopus дерекқорында индекстелетін, cite score көрсеткіші 2.0 және перцентилі 46 болатын Eastern-European Journal of Enterprise Technologies (Q3) журналында 2 мақала жарық көрді. Сонымен қатар, халықаралық конференциялардың материалдарында 4 мақала жарияланды, олардың ішінде бір ғылыми мақала cite score көрсеткіші 4.5 және перцентилі 69 (Q2) деңгейінде бағаланған.

Диссертация тақырыбы бойынша алынған нәтижелер келесі жарияланымдарда ұсынылған:

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 resultsEastern-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.

Диссертацияның құрылымы.

Жұмыс кіріспеден, төрт бөлімнен, қорытындыдан, пайдаланылған әдебиеттер тізімінен және қосымшалардан тұрады.

Диссертацияның негізгі мазмұны. Диссертациялық жұмыс төрт негізгі тараудан тұрады.

Бірінші тарау қазақ тілін өңдеу саласындағы әдебиеттерге шолу мен концептуалды негізге арналған. Бұл бөлімде қазақ тіліне арналған табиғи тілді өңдеу (NLP) құралдарының даму тарихы мен қазіргі жағдайы талданады, агглютинативті морфологияның күрделілігі атап көрсетіледі және KazBERT, CRF және BiLSTM сияқты заманауи нейрондық архитектуралар ұсынылады. Сонымен қатар, бұрын қолданылған лингвистикалық әдістер, алғашқы тілдік өңдеу жүйелері және әмбебап модельдерді қазақ тілінің ерекшеліктеріне бейімдеудегі қиындықтар қарастырылады.

Екінші тарау интернеттен қазақ тіліндегі мәтіндерді автоматты түрде жинауға арналған арнайы әзірленген QCrawler жүйесіне арналған. Бұл бөлімде краулердің архитектурасы, алгоритмдері мен математикалық моделі сипатталады, сондай-ақ оның тиімділігі мен қамту ауқымы базалық парсинг әдістерімен салыстырылады. Сонымен қатар, жиналған корпус негізінде KazBERT+CRF моделін пайдалана отырып, атаулы есімдіктерді тану (NER) тапсырмасын бастапқы қолдану нәтижелері көрсетіледі.

Үшінші тарау QNLP платформасының лингвистикалық негізіне бағытталған және оған морфологиялық талдаушыны, қазақ тілінің тезаурусын және қазақ морфологиясының онтологиясын әзірлеу кіреді. «Түбір + жұрнақтар» моделіне негізделген сөз құрылымын өңдеу тәсілі жан-жақты сипатталады, бұл сөз формаларын жоғары дәлдікпен сегментациялауға және генерациялауға мүмкіндік береді. Бұл лексикалық және ережеге негізделген компоненттер тереңдетілген лингвистикалық талдауға және модельдердің интерпретациялануын жақсартуға ықпал етеді.

Төртінші тарау қазақ тілін өңдеуге арналған толыққанды ашық QNLP платформасын сипаттайды. Бұл бөлімде модульдік архитектура, компоненттердің өзара әрекеттесуі және ішкі математикалық формулалар баяндалады. Сонымен бірге, POS-таңбалау, атаулы есімдіктерді тану (NER) және морфология бойынша эксперимент нәтижелері, KazBERT, BiLSTM және CRF компоненттерінің әсерін зерттеу, сондай-ақ қолданыстағы құралдармен салыстыру нәтижелері ұсынылады. Тарау нәтижелерді визуализациялайтын графиктер мен дәлдік көрсеткіштері арқылы аяқталады, бұл QNLP платформасының жоғары нәтижелерін көрсетеді.