«6D075100 – Информатика, есептеу техникасы және басқару»
мамандығының докторанты
Елеусінов Арман Инабатұлының
«Қазақ тілінде қолжазба мәтіндерді тану және машиналық оқыту алгоритмдері мен ақпараттық моделін құру» тақырыбындағы
диссертациялық жұмысына

АҢДАТПА

Жұмыстың өзектілігі: Қолжазба мәтінін тану (HTR - Handwritten Text Recognition) құжатты цифрландыруға байланысты көптеген қолданбалардың кілті болып табылады, соның ішінде мұрағаттық құжаттарды өңдеу, кітапхана ісі, денсаулық сақтау, банк ісі және т.б. Ең үлкен мәселе стильдердің алуан түрлілігінен туындайды. Машиналық оқыту әдістері осы мәселенің ықтимал шешімін ұсынады. Көптеген бағдарламалық қосымшалар қазірдің өзінде HTR-ді жүзеге асыруда, бірақ олар тілдің ерекшелігіне байланысты қазақ тіліне сәйкес келмейді және бұл саладағы зерттеулер әлі де өзекті.

Қазақстан Республикасы Президентінің Архиві Қазақстанның қазіргі саяси тарихы бойынша бірегей құжаттардың сақтаушысы болып табылады. 2020 жылдың 1 қаңтарындағы жағдай бойынша Мұрағатта 1918-2008 жылдар аралығындағы 830 мыңнан астам сақтау бірлігі бар. Оның 762 мың сақтау бірлігі қағаз жүзінде, оның ішінде қолжазба түрінде жазылған.

Мұрағаттық құжаттарды цифрландыру – бұл көптеген факторларға байланысты техникалық күрделі процесс: арнайы бағдарламалық-аппараттық кешеннің болуы, ұтымды жұмыс стратегиясын анықтау, сканерлеу операторларының біліктілігі және қажетті нормативтік-әдістемелік құжаттаманың болуы. Қазіргі уақытта «цифрландыру» терминінің орнына кейде ұқсас ұғым қолданылады - сканерлеу.

Цифрландыру қауіпсіз және үнемді сақтауға, қажетті ақпаратты ыңғайлы іздеуге және мұрағаттық құжаттарға жедел қол жеткізуге мүмкіндік береді. Мұрағаттық құжаттарды цифрландыру бірден бірнеше мәселені шешуге мүмкіндік береді:

- қорлар есебін ретке келтіру;
- түпнұсқаларды айналымнан кезең-кезеңімен алып тастау арқылы сақталуын қамтамасыз ету;
- мұрағаттық құжаттарға қолжетімділікті кеңейту және жеңілдету;
- электрондық қорға қолжетімділікті қамтамасыз ету;

- мемлекеттік қызмет көрсету процесін жеделдету;
- мұрағаттың ғылыми анықтамалық аппаратымен жұмыс істеу

Бұл мәселелерді шешу кешенді зерттеуді және соңғы технологиялық және зерттеу нәтижелерін енгізуді, арнайы жиналған деректердің үлкен жиынтығын қажет етеді.

Диссертациялық жұмыстың мақсаты – қазақ тіліндегі архивтік қолжазбаларды тану үшін машиналық оқыту алгоритмдері мен нейрондық желі архитектурасын құру

Зерттеудің міндеттері. Қойылған мақсатқа жету үшін келесі міндеттерді орындау қарастырылады:

1) Қазақ тіліндегі жазылған қолжаза мәтінді тануға арналған машиналық оқыту әдістерін оқытуға және тестілеуге арналған деректер қорын жинау және өңдеу.
2) Қолжазба мәтінді тануға арналған жүйелерді, әдістер мен алгоритмдерді зерттеу және талдау;
3) Архивтік құжаттарды алдын ала өңдеу алгоритмдерін және сегменттеу моделін әзірлеу;
4) Қазақ тіліне арналған қолжазба мәтіндер деректерін жасанды көбейту үшін генеративті қарсылас нейрондық желілер (GAN) негізінде машиналық оқыту моделін әзірлеу;
5) Мәтінді тану үшін трансформер негізіндегі машиналық оқыту моделін әзірлеу;
6) Құжаттардан керек мәліметті іздеу жүйесін әзірлеу

Зерттеу нысаны. Қазақ тіліндегі қолжазба мәтіндердің кескіндері.

Зерттеу пәні. Қазақ тіліндегі қолжазба мәтіндерді тануға арналған машиналық оқыту алгоритмдері, нейрондық желілер және ақпараттық модельдер

Зерттеу әдісі. Компьютерлік көру алгоритмдері, терең нейрондық желілер, конволюциялық желілер, генеративті қарсылас желі (GAN), трансформер архитектурасы

Жұмыстың ғылыми жаңалығы:

1. Қазақ тілінде қолжазба мәтінін тану үшін бейімделген CNN+Transformer гибридті архитектурасы құрылды.
2. Қолжазбаның шынайы вариацияларын тудыратын қазақ тіліндегі қолжазба мәтін деректерін толықтыруға GAN архитектурасы негізіндегі әдіс ұсынылды. Бұл әдіс деректер саны аз кезде де тану дәлдігін жоғарылатуға болатынын көрсетті.

Жұмыстың теориялық маңызы. Зерттеу машиналық оқыту әдістерін және қолжазбаны тану үшін қолданылатын нейрондық желілерді дамытуға ықпал етеді. Қазақ жазба тілінің ерекшеліктеріне архитектураның (мысалы, CNN, Transformer) дамуы мен бейімделуі мәтіндік ақпаратты интеллектуалды өңдеу саласындағы теориялық білім қорын кеңейтеді.

Кеңінен зерттелетін тілдерге қарағанда қазақ тілінің агглютинативті құрылымы, күрделі морфологиясы және әліпбилік ерекшеліктері бар. Зерттеу қазақ тіліндегі мәтінді тану мен өңдеудің жаңа әдістерін теориялық негіздеуге және ұсынуға мүмкіндік береді.

Жұмыстың практикалық маңызы. Құжаттарды интеллектуалды өңдеудің автономды жүйесін дамыту Қазақстан Республикасында ақпараттық және телекоммуникациялық технологияларды дамытудың перспективалық бағыты болып табылады. Бұл жобаның нәтижелерін жүзеге асыру бірінші кезекте қолжазба мәтіндерін цифрландыру процесінің тиімділігін арттырады. Сәйкесінше, жұмыстың нәтижелерін іске асырудың негізгі әлеуметтік-экономикалық әсері әртүрлі түрдегі және көлемдегі құжаттарды өңдеуге уақыт пен қол күшін үнемдеумен байланысты.

Қорғауға шығарылған негізгі тұжырым.

1. Қолжазба мәтінді тану үшін конволюциялық нейрондық желі мен трансформер гибридті архитектурасы wer (сөздік қате көрсеткіші) және cer (әріптік қате көрсеткіші) метрикаларын сәйкесінше 10.44% және 2.85% төмендетуге мүмкіндік береді.
2. Қолжазба мәтін деректерін жасанды көбейтудің GAN негізіндегі әдісі сөздерді тану дәлдігін 75.8%-дан 86.3%-ға арттыруға мүмкіндік береді.

Диссертация құрылымы және көлемі. Диссертациялық жұмыс кіріспеден, 3 бөлімнен, қорытынды, әдебиеттер тізімі және қосымшалардан тұрады. Диссертацияның толық көлемі: 84 бет машиналық мәтіні, оның ішіне 54 сурет, 6 кесте, 83 пайдаланылған әдебиеттер тізімі және 2 қосымша кіреді.

Кіріспеде мәселенің сипаттамасы және тақырыптық аймаққа қысқаша шолу жасалады. Диссертациялық жұмыстың өзектілігі дәлелденді, әдеби дереккөздерге шолу жасалады.

Бірінші тарауда қолжазбаны тану мәселесі және оны шешудің жалпы тәсілдері қарастырылады. Қолжазбаны тану үшін деректер жиынын жасау және өңдеу әдістері егжей-тегжейлі сипатталған. Алдын ала өңдеу әдістерін (сүзгілеу, гистограмманы теңестіру, шуды жою) қолдану арқылы кескін сапасын жақсарту және тану дәлдігін арттыру әдістері көрсетілген. Бұл кезең мәтінді тану тиімділігін арттыру үшін маңызды дайындық кезеңі болып саналады.

Екінші тарауда қолжазбаны тану үшін қолданылатын машиналық оқыту әдістері талқыланады. Біріншіден, жолдар мен сөздерді сегменттеу әдістері, оларды жүзеге асыратын нейрондық желілер және бағалау нәтижелері ұсынылады. Әрі қарай, әріптерді жіктеу және сөзді тану үшін CNN және Transformer архитектурасына негізделген моделдер сипатталады. Сонымен қатар, деректерді ұлғайту үшін GAN (Generative Adversarial Networks) әдісін қолдану және оның нәтижелері сипатталған.

Үшінші тарауда әзірленген қолжазбаны тану жүйесінің архитектурасы, құрамдас бөліктері және техникалық орындалуы сипатталған. Жүйе құрамдастары Nvidia Triton Inference Server көмегімен оңтайландырылған. Сонымен қатар, құжаттардан қажетті ақпаратты автоматты түрде іздеу және шығару жүйесі, сондай-ақ дайын жүйені біріктіру және орналастыру мүмкіндіктері қарастырылады.

Диссертациялық жұмыс нәтижелерінің апробациясы. Зерттеудің негізділігі мен сенімділігі талаптарға сәйкес келеді. Диссертацияның негізгі ережелері сынақтан өтті және 6 баспа жұмысында ұсынылған, оның ішінде:

- Scopus-те индекстелген халықаралық ғылыми журналдардағы 2 мақала (Q2, Q3):

1. **Yeleussinov, A**., Amirgaliyev, Y., Cherikbayeva, L. Improving OCR Accuracy for Kazakh Handwriting Recognition Using GAN Models. *Appl. Sci.* **2023**, *13*, 5677 (SJR 0.492, процентиль 75) https://doi.org/10.3390/app13095677

2. Amirgaliyev B., **Yeleussinov A.**, Taizo M. Kazakh handwritten recognition. Journal of Theoretical and Applied Information Technologies, 2020, 98(14), pp. 2744–2754.

- Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім беру сапасын қамтамасыз ету комитеті (КОКСНВО) ұсынған тізімнен журналда 1 мақала:

1. Амиргалиев Е. Н., Миачи Т., **Елеусинов А.И.** Методы и алгоритмы формирования базы и обработки рукописных текстов на казахском языке, Вестник КазНИТУ, 2020г., №138, сс. 653-660.

- халықаралық және республикалық конференциялар материалдарында және ғылыми семинарларда баяндамаларда 3 жарияланым:

1. **Yeleussinov A.**, Islamgozhayev T., Kozhirbayev Z. Creation and Annotation of a Handwritten Text Database for the Kazakh Language: Methodologies and Preliminary Results. In2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE) 2024 Oct 30 (pp. 1-6). IEEE. (**Scopus**)

2. **Елеусинов А.И.**, Бурибаев Ж.А. Сегментация рукописных символов с исправлением наклона текста. V-международная конференция «Информатика и прикладная математика», Алматы, ИИВТ, 29.09.2020-01.10.2020.

3. **Елеусинов А.И.,** Амиргалиев Е. Н., Мажитов Ш. С. О разработке базы изображений рукописных символов казахского языка, Конференция"IV международная научно-практическая конференция "Информатика и прикладная математика", посвященной 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию профессора Амиргалиева Е.Н., Институт информационных и вычислительных технологий МОН РК, 25.09.2019-29.09.2019.

Диссертация кафедраның ғылыми семинарларында да талқыланды.