«6D060200 – Информатика» мамандығы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған Тойганбаева Назгуль Абеновнаның «Қазақ-орыс тілінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру» тақырыбындағы диссертациясына

АҢДАТПА

Зерттеу тақырыбының өзектілігі. Әртүрлі салалар сандық ақпарат алмасуға көшірілгеніне қарамастан, сауалнамалар, жадынамалар, емтихан сұрақтарына жауап беру және мұрағаттық деректер сияқты көптеген құжаттарда әлі де қолмен енгізуді қажет етеді. Осыған байланысты қолжазбаларды компьютер арқылы транскрипциялаудың автоматты әдісі болып табылатын қолжазба мәтінін тануды жүзеге асыру қажет. Қолмен жазылған мәтінді тану (ағылш. Handwriting recognition, HWR немесе Handwritten Text Recognition, HTR) – бұл қағаз құжаттар, суреттер және сенсорлық экрандар және басқа құралдар көздерден түсінікті қолжазба ақпаратын алу және тану қабілеті. Қолмен жазылған мәтінді оффлайн тану – кірісі екі өлшемді кескін, шығысы таңбалар тізбегі болып табылатын кескіндегі әріптерді немесе сөздерді сандық мәтінге түрлендіру әдісі. Қазіргі уақытта қол жетімді қолжазбаны тану бағдарламалық шешімдерін онлайн және оффлайн тану деп екі негізгі бөлікке бөлуге болады: Онлайн тану әртүрлі құрылғылардың экранында арнайы құрылғылармен (стилустармен) жазылған мәтінді тануды қамтиды. Оффлайн қолжазбаны тану суреттегі мәтінді компьютерлік және мәтінді өңдеу қосымшаларында қолдануға болатын әріптік кодтарға автоматты түрде түрлендіруді қамтиды. Өз кезегінде, қолжазба мәтінін оффлайн тану мәселесі әлі толық шешілген жоқ. Соңғы 10 жылда орын алған баспа және қолжазба мәтіндерін тану саласындағы ілгерілеу байқалады. Есептеулердің күрделілігіне қарамастан, кейбір өте күрделі есептеулерді машиналық оқыту және терең нейрондық желілер үшін жасалған құрылымдар арқылы шешуге болады. Кириллица негіздегі қазақ-орыс тіліндегі қолжазбаларды тану мәселесі әлі де толық зерттелмеген. Сондықтан, қазақ-орыс тілдеріндегі оффлайн қолжазба мәтіндерін тану алгоритмдерін жасау өте маңызды және осы салада зерттеулер жүргізу қажет, яғни диссертация тақырыбы өзекті болып отыр. Сондықтан оптикалық таңбаларды тану (OCR) жүйелеріне негізделген тиімді автоматты қолжазба мәтіндерін тану алгоритмдерін жасау өте маңызды және осы салада зерттеулер жүргізу қажет. Бұл, әсіресе, қазақ тіліндегі қолжазба мәтіні үшін деректер жиынтығының жоқтығын ескерсек, зерттеу жұмыстарын жүргізу қажеттігі айқын. Осыған орай, бұл тақырып бүгінгі таңда өзекті және практикалық маңызы зор. Қолжазба мәтіндерін тану мәтін жазылған қағазды сканерлеуден, алынған ақпаратты мәтіндерге және бейнелерге бөлуден, қолжазбаларды интеллектуалды тану үшін әдістерді таңдаудан және нәтижелерді өңдеуден тұратын күрделі құрылымды үдеріс. Қолжазба мәтінін тану есептерін шешу үшін нейронды желілер қолданылады. Адам миының биологиялық

құрылымына негізделген нейронды желілер өзінің жоғарғы деңгейде есептеу мүмкіндігімен басқа оқыту алгоритмдерінен бірнеше есе тиімді болып есептеледі

Соңғы жылдары тану әдістері мен жүйелерінің жетілгеніне қарамастан, кескіндердегі қолжазба мәтіндерін автоматты түрде тану әлі де күрделі мәселе болып табылады. Қолмен жазылған жан-жақты және шексіз мәліметтер жиынтығы бұрынғыдан да маңызды бола түсуде. Қолжазба мәтіндері жадынамаларда, медициналық жазбаларда, тарихи құжаттарда, стилус енгізген мәтіндерде, емтихан жауаптарында және т.б. құжаттарда болады. Сондықтан кескіндердегі қолжазба мәтінін тану мәселесін толық шешілуі қамтамасыз етілуі керек. Бірнеше тілдер үшін қолжазбаны танудың кең ауқымды жүйелерін әзірлеу саласындағы зерттеулердің қажеттілігін көрсетеді. Осылайша, қазақ және орыс тілдеріндегі қолжазба мәтіндері үшін деректер базасын құру және тиімді OCR құралдарын әзірлеу – ғылыми ізденістер тұрғысынан да, практикалық қолданыс тұрғысынан да өзекті мәселе болып табылады.

Диссертациялық жұмыстың мақсаты: тереңдетілген оқыту модельдерін қолдана отырып, қазақ-орыс тілдерінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру.

Зерттеудің міндеттері: қойылған мақсатқа қол жеткізу үшін келесі міндеттерді іске асыру қажет:

1. Мәтінді оптикалық тану есептеріне терең оқытуды қолданған шешімдерге шолу және талдау жасау.
2. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер жинағын құру.
3. Сканерленген беттердегі мәтінді жолдарға, жолдарды сөздерге бөлетін сегментация үдерісін іске асыру үшін Генетикалық алгоритмді қолдану.
4. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер қоры негізінде терең оқытудың әртүрлі әдістерімен эксперименттер жүргізу.

Зерттеу нысаны: қазақ-орыс тіліндегі қолжазба мәтіндерін тану.

Зерттеу пәні: қазақ-орыс тіліндегі қолжазба мәтіндерінің деректер жиынын құру және тану үшін терең оқыту модельдерін қолдану.

Зерттеу әдістері: зерттеу әдісі ретінде қолжазба мәтіндерін бөлу үшін генетикалық алгоритм, қолжазба мәтіндерін тану мәселесін шешуге рекурентті және конволюциялық нейрондық желілерді қолданылады.

Алынған нәтижелердің ғылыми жаңалығы:

- Нейрондық модельдерді оқыту мақсатында қолжазба мәтіндер корпусын краудсорсинг әдісі арқылы жинауға арналған автоматтандырылған технологиялық жүйе әзірленді.
- Қазақ тіліндегі қолжазба мәтіндерден сөздерді дәл және тиімді түрде сегменттеуге бейімделген генетикалық алгоритм жасалды.
- Алғаш рет қазақ тіліндегі сөз деңгейіндегі оффлайн қолжазба мәтіндерден тұратын құрылымды корпус – Kazakh Offline Handwritten Text Dataset (KOHTD) құрылды.

- Қазақ және орыс тілдеріндегі қолжазба мәтіндерді оптикалық тануға арналған Flor, Puigcerver, Bluche және Scheidl нейрондық архитектуралары қазақ тіліне бейімделіп, икемделген оқыту (fine-tuning) әдісімен үйретілді.

Жұмыстың теориялық және практикалық маңызы. Қолжазба жазбаларының цифрланған мәтіні көптеген компаниялардың бизнес-процестерін автоматтандыруға мүмкіндік береді және адамның жұмысын жеңілдетеді. Сонымен қатар, қазақ-орыс тілдеріндегі қолжазба мәтіндерін тану пошта орталықтарының жұмысын автоматтандыруға мүмкіндік береді. Қазақ тілінде алғаш құрылған қолжазба мәтіндерінің деректер жинағы KOHTD көмегімен алгоритмдердің өнімділігін бағалауға және салыстыруға болады.

Қорғауға ұсынылатын негізгі тұжырымдар:

- Қолжазба мәтіндердің деректерін жинау, алдын ала өңдеу, анотациялау және корпусқа енгізу үдерістерін автоматты түрде орындайтын толығымен автоматтандырылған жүйе жасалды және тәжірибелік негізде іске қосылды.
- Қазақ тіліндегі қолжазба мәтіндерден сөздерді дәл сегменттеуге арналған генетикалық алгоритм бейімделіп, оның эвристикалық мүмкіндіктері арқылы әріпаралық шекараларды анықтау дәлдігі арттырылды.
- Қазақ тіліндегі оффлайн қолжазба мәтіндерден тұратын Kazakh Offline Handwritten Text Dataset (KOHTD) атты ашық датасет жасалды.
- Flor, Puigcerver, Bluche және Scheidl нейрондық модельдері қазақ және орыс тілдеріндегі қолжазба мәтіндерді оптикалық тануға арналған болып, қазақ тілінің графемалық құрамының ерекшеліктеріне бейімделіп, тиісті түрде оқытылды.

Сенімділік дәрежесі және апробациялау нәтижелері: Диссертация тақырыбы бойынша зерттеу нәтижелері 5 мақала жарияланды және 1 авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік алынды. Зерттеу нәтижелері әл-Фараби атындағы ҚазҰУ компьютерлік ғылымдар кафедрасында ғылыми семинарларда, сонымен қатар, халықаралық конференцияларда баяндалды.

Scopus базасындағы журналдық мақала:

1) Toiganbayeva N., Kasem M, Abdimanap G., Bostanbekov K., Abdallah A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset. Signal Processing: Image Communication. Elsevier, 2022, vol.108, pp. 116827. doi https://doi.org/10.1016/j.image.2022.116827 (Scopus: Q2; CiteScore-2.5; Percentile- 73%)) KOHTD: Kazakh offline handwritten text dataset,

Ғылыми еңбектің негізгі нәтижелерін жариялау үшін Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім саласында сапаны қамтамасыз ету комитеті ұсынатын ғылыми басылымдар тізбесіндегі журналдардағы мақалалар:

2) Тойганбаева Н.А., Әбдіманап Ғ.С., Алимова А.Н., Жунусова Ж.Х. Қазақ тіліндегі қолжазба мәтіндерінің мәліметтер қорын құру мен тану. «Алматы энергетика және байланыс университетінің Хабаршысы», 3(62), 2023, 105-116б.

3) Тойганбаева Н.А., Алимова А.Н., Әбдіманап Ғ.С., СакыпбековаМ.Ж., Гусманова Ф.Р. Нейронды желілер негізінде қазақ-орыс тілдеріндегі қолжазба мәтіндерді тану. Абай атындағы ҚазҰПУ. Хабаршы. Физика-математика ғылымдары сериясы. Том 84 № 4 (2023). 183-191б.

4) Тойганбаева Н.А. Терең оқыту модельдері негізінде қазақ-орыс тіліндегі оффлайн қолжазба мәтіндерін тану. Қазақстан Республикасы Ұлттық инженерлік академиясының хабаршысы. № 4 (90) – 2023.

Халықаралық конференция:

5) Тойганбаева Н.А., Тойганбаев Б.А. Нейронды желілерді қолжазбаларды интеллектуалды тану есептерінде қолдану мәселелеріне шолу // САТПАЕВ ОҚУЛАРЫ – 2021, Том II, 86-89б.

Авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік:

6) Нурсеитов Д.Б, Бостанбеков К.А. Әбдіманап Ғ.С., Тойганбаева Н.А., Алимова А.Н. "Kazakh Offline Handwritten Text Dataset (KOHTD)" қазақ тіліндегі қолжазба мәтінінің деректер қорына авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік. № 34615. 2023.

Зерттеушінің жеке үлесі. Диссертациялық жұмыстың қойылған мақсаттарына сәкес міндеттер шешілді. Қолжазба мәтіндерін тану үшін алдын-ала өңдеу жұмыстары және мәтін жолдарын генетикалық алгоритм көмегімен сегментациялау, KOHTD деректор қорын құру, толтыру жұмыстары іске асты. Flor, Puigcerver, Bluche және Scheidl модельдерімен тәжірибелер жүргізіліп, CER, WER, SER стандартты тиімділік көрсеткіштері бойынша бағаланды. Барлық модельдер Python және Tensorflow, Keras, OpenMP Matplotlib кітапханаларының көмегімен жүзеге асырылды.

Жұмыстың көлемі мен құрылымы. Диссертация қазақ тілінде жазылды. Диссертация кіріспеден, 4 тараудан және қорытындыдан тұрады. Диссертацияның жалпы көлемі 97 беттен, 42 суреттен, 10 кестеден, 95 қолданылған әдебиеттен тұрады.

Кіріспеде жұмыстың өзектілігі анықталып, қолжазбаны танудың кең ауқымды жүйелерін әзірлеу саласындағы зерттеулердің қажеттілігі көрсетілді. Сонымен қатар, диссертациялық жұмыстың мақсаты, міндеттері, әдістері, алынған нәтижелердің ғылыми жаңалығы, жұмыстың теориялық және практикалық маңызы жазылып, жарияланымдар тізімі берілген.

Бірінші тарауда терең оқыту мүмкіндіктерін қолжазба мәтіндерді оптикалық тану қолданған есептердің шешімдерге шолу жасалды. Қолжазбаны тану саласындағы деректер жиынтықтарына сандық талдау жасалды. Сонымен қатар, қолжазбаны тану саласында кеңінен қолданылатын деректер жиынтықтарына сандық талдау жүргізіліп, олардың көлемдік, құрылымдық және тілдік сипаттамалары сарапталды. Бұл талдау HTR жүйелерінің тиімділігін бағалау және салыстырмалы зерттеу жүргізу үшін негіз бола алады.

Екінші тарауда алғаш құрылған қазақ тіліндегі оффлайн қолжазба мәтіндерінің деректер жинағын Kazakh Offline Handwritten Text Dataset

(KOHTD) құру, мәтінді сегменттеу, аннотация жазу қарастырылды.Осы бөлімде KOHTD деректер жиынтығына статистикалық талдау жасалды.

Үшінші тарауда қолжазбаларды тануда кеңінен қолданылатын нейрондық желілердің негізгі түрлері мен олардың жұмыс принциптері қарастырылды. Жасанды, рекуррентті (RNN), ұзақ-қысқа мерзімді жады (LSTM) және конволюциялық нейрондық желілердің (CNN) архитектуралары мен мүмкіндіктері сипатталды. Сонымен қатар нейрондық желілерді оқыту әдістері мен Connectionist Temporal Classification (CTC) алгоритмінің қолжазбаны танудағы рөлі түсіндірілді.

Төртінші тарауда Bluche ұсынған Gated-CRNN әдісі, Puigcerver ұсынған 2-LSTM әдісі, Flor ұсынған Gated-CNN әдісі, Abdallah ұсынған Attention-Gated-CNN-BGRU әдісі қарастырылды. Сонымен қатар, модельдер KOHTD деректер жиынтығында оқытылып, стандартты өнімділік көрсеткіштері CER, WER және SER бойынша бағалау жасалды.

Қорытындыда диссертацияның негізгі нәтижелері мен қорытындылары берілген.