«6D070300 – Ақпараттық жүйелер» мамандығының Phd докторанты Кәрібаевва Айдана Сейілғазықызының «**Нейронды машиналық аударма үшін қазақ тіліндегі мәтіндерді морфологиялық сегментациясының модельдері мен әдістерін құру және зерттеу**» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Зерттеу тақырыбының өзектілігі. Машиналық аударма - жасанды интеллектін басты бір есебі. Нейронды машиналық аударма қазіргі таңдағы машиналық аударма түрлерінің ішіндегі танымал және кең таралған әдіснама болып табылады.

Машиналық аударманың көптеген мәселелері толық түсінілмеген және тілдің ерекшеліктеріне байланысты егжей-тегжейлі қарастыруды қажет етеді. Машиналық аударма жүйелері әрдайым мәселені дәстүрлі әдістермен шеше алмайды. Ережеге негізделген машиналық аудармада барлық ережелер ескерілмеуі мүмкін; статистикалық аудармада дұрыс аударма әрқашан контекстпен анықтала бермейді. Бүгінгі күні нейрондық желілерді қолдану көптеген пәндік аймақтарда танымал болды және машиналық аударма да кең қолданыс табуда.

Машиналық аударма мәселелерін шешудің әртүрлі тәсілдері бар, мысалы, тілдердің грамматикалық ережелеріне негізделген тәсіл; аударылған тілдердің ықтималдық фразалық кестесін табудың статистикалық тәсіліне негізделген статистикалық машиналық аударма тәсілі; аударма тілдерінің нейрондық желілерін оқытуға негізделген нейрондық машиналық аударма тәсілі. Осы тәсілдердің әрқайсысының артықшылықтары мен кемшіліктері бар. Соңғы кезде машиналық аудармада нейронды желілерге негізделген нейронды машиналық аударма ең жақсы нәтижелерді көрсетуде. Машиналық аударма мәселесі әлі жеткілікті жоғары деңгейде, кәсіби аудармаға жақын шешілмегендіктен, машиналық аударма мәселесі өте өзекті болып табылады. Машиналық аударма мәселесін шешу, табиғи тілді түсіну сияқты жасанды интеллекттің басқа да өте маңызды мәселелерін шешуге жол ашатындығын атап өткен жөн.

Нейрондық машиналық аударманың негізінде матрицалық есептеулерге негізделген қайталанатын нейрондық желілердің механизмі жатыр, бұл статистикалық машиналар аудармашыларына қарағанда едәуір күрделі ықтималдық модельдерін құруға мүмкіндік береді. Нейрондық машиналық аударма бағыты табиғи тілді өңдеудің өзекті тақырыбы болып табылады, өйткені нейронды машиналық аудармасы ережеге негізделген және статистикалық машиналық аудармадан асып түседі.

Табиғи тілді өңдеу проблемаларында бірқатар өзекті мәселелер бар. Аударма сапасын жақсарту мақсатында кең таралған және қолданылатын тәсілдің бірі: сегменттеу. Табиғи тілді сегменттеу есебі есептеуіш лингвистикадағы өзекті зерттеу тақырыбы болып табылады және бүгінгі күнге дейін ашық мәселе болып қала береді. Көптеген әдістер тілдің морфологиялық ерекшеліктерін ескермейтін жиіліктік сегменттеуді ұсынады. Бұл әдіс BPE (байт-жұптық кодтау) әдісін қамтиды. BPE негізіндегі сегментация агглютинативті тілдер үшін жақсы нәтиже бермейді. Жұмыста қазақ тілінің морфологиялық ерекшеліктеріне негізделген CSE (Complete set of endings - жалғаулардың толық жиынтығы)-модель негізінде сегменттеу әдісі ұсынылған.

Сөздіктер нейронды машиналық аудармада (NMT) маңызды рөл атқарады. Алайда үлкен сөздікке едәуір жады қажет, бұл NMT қолдану мүмкіндігін шектейді және жадыда қате тудыруы мүмкін. Бұл шектеуді әр сөзді параллель корпуста морфемаларға бөлу арқылы шешуге болады. NMT оқыту кезінде тиісті NMT сөздігінің көлемі тез артады; сондықтан компьютер жадының шамадан тыс қорын қажет етеді. Сондықтан бұл зерттеу түркі тілдері үшін жалғаулардың толық жиынтығына (CSE)

негізделген жаңа морфологиялық сегменттеу тәсілін ұсынады, бұл бастапқы корпустардың сөздігін азайтады, тиісінше, қажетті жад көлемі азаяды.

Машиналық аудармада бірнеше алдын ала өңдеу сатылары бар. Мәтінді сегментациялау - сөздік қорын азайтуға арналған машиналық аудармаға дайындық кезеңдерінің бірі. Сегменттеу мәселесі талдамалық тілдер үшін көптеген тәсілдермен зерттелген, ал агглютинативті тілдер үшін, дәлірек айтсақ түркітілдес үшін, зерттеу көлемі аз. Нейрондық желілер әдетте сөздердің көп бөлігін аударма тіліне аудару үшін үлкен сөздік құрады. Нейрондық машиналық аудармада сегменттеу мәселесі нейронды желілерін сөздік көлемін азайтуға, сонымен қатар белгісіз және сирек кездесетін сөздердің мәселелерін шешуге арналған. Берілген мәтін бойынша сегменттеу осындай шешімдердің бірі болып табылады. Сондықтан нейронды машиналық аудармада мәтінді сегменттеудің **өзектілігі** артып келеді.

Диссертациялық жұмыстың мақсаты. Лингвистикалық ерекшеліктері негізінде қазақ тілінің нейронды машиналық аудармасының сапасын жақсарту үшін модельдер, алгоритмдер мен бағдарламалық қамтамасыздандыру құру.

Зерттеудің міндеттері. Қойылған мақсатқа қол жеткізу үшін келесі міндеттерді орындау қарастырылады:

1) Жалғаулардың толық жиынтығына (CSE – Complete Set of Endings) негізделген қазақ тілінің морфологиясының тілдік моделін жетілдіру (қазақ тілінің мүмкін жалғаулар тізімін кеңейту);
2) Қазақ тілі морфологиясының CSE-моделі негізінде морфологиялық сегментация моделі мен алгоритмін құру;
3) Нейрондық машиналық аударма платформасында айқындалған тапсырмалар бойынша эксперименттер жүргізу және жасау.

Зерттеу нысаны: қазақ тілі.

Зерттеу субъектісі. Қазақ тілі үшін нейронды машиналық аударма.

Зерттеу әдістері. Зерттеу әдісі ретінде комбинаторикалық талдаудың сандық әдістері, машиналық оқыту, терең оқыту және нейрондық желілер қолданылды.

Алынған нәтижелердің ғылыми жаңалығы:

1) Тіл жалғауларының толық жиынын құрумен ерекшеленетін мүмкін болатын жалғауларды қарастырудың негізінде қазақ тілі морфологиясының жетілдірілген есептеуіш моделі әзірленді.
2) Танымал әдістер мен алгоритмдерден жалғаулардың толық жиынын шешімдер кестесін ретінде құрумен ерекшеленетін және нейрондық машиналық аударма сөздігінің көлемін қысқартуға мүмкіндік беретін қазақ тілі морфологиясының жетілдірілген есептеу моделіне негізделген морфологиялық сегменттеу әдісі мен алгоритмі әзірленді. Бұл нейронды машиналық аударманы үлкен көлемді кіріс мәліметтерде (сөздіктің үлкен көлемінде) оқытуға мүмкіндік береді.

Жұмыстың теориялық және практикалық маңызы. Бұл жұмыстың теориялық маңыздылығы қазақ тілінің тілдік ерекшеліктерін ескере отырып, морфологиялық сегменттеудің универсалды жаңа әдісін жасауда жатыр. Жасалған CSE-моделі негізіндегі морфологиялық сегменттеу әдісін басқа Түркі тілдерге де қолдануға болады.

Жұмыстың практикалық маңызында нейронды машиналық аударманы сегменттелген мәтін негізінде оқыту жады көлемін ұдайы азайтады және жадымен қателіктерді болдырмауында.

Қорғауға шығарылған негізгі тұжырымдама. Қазақ тіліндегі сөздерді морфологиялық сегменттеудің жаңа моделі мен алгоритмі, қазақ тіліндегі сөздерді сегменттеудің ұсынылған моделі мен алгоритмінің тиімділігін растайтын қазақ тілінің нейронды машиналық аудармасы бойынша тәжірибелердің нәтижелері.

Сенімділік дәрежесі мен апробациялау нәтижелері. Зерттеудің сенімділігі мен нәтижелерінің негізділігі міндеттерді қоюдың негізделген жауапкершілігімен, критерийлердің және берілген саладағы зерттеулердің жай-күйінің сарапталуымен, жүргізілген эксперименттермен, сондай-ақ олардың қазақ тілінің нейронды машиналық аудармада нәтижелерідің жақсартулар аулымен қамтамасыз етіліп дәйектеледі. Диссертация нәтижелері төмендегі жарияланымдарда жарық көрді.

Scopus базасындағы журналдық мақала:

1) Tukeyev U., Karibayeva A., Zhumanov Zh. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 2020, 1 том, номер №1. (Scopus:Q2; CiteScore-2.5; Percentile- 73%))

2) Turgangayeva A., Rakhimova D., Karyukin V., Karibayeva A, Turarbek A. Semantic Connections in the Complex Sentences for Post-Editing Machine Translation in the Kazakh Language. *Information* 2022, *13*(9), 411; https://doi.org/10.3390/info13090411(Scopus: Q2; CiteScore 4.2; Percentile-64%)

3) Rakhimova D., Karibayeva A. Aligning and extending technologies of parallel corpora for the Kazakh languge. *Eastern-European Journal of Enterprise Technologies*, 2022, 4(2-118), стр. 32–39 (Scopus: Q3; Citescore: 2.0; Percentile: 37%)

Бақылау комитеті салалық білім және ғылым саласында ұсынылған журналдарда:

1) Karibayeva A., Rakhimova D., Abduali B., Amirova. Анализ машинного перевода казахского языка. Вестник КазНИТУ №3 (127), КазНИТУ имени К. И. Сатпаева, 2018, 90 - 96 б. (CCES)

2) Рахимова Д., Тұрарбек А., Карюкин В., Карибаева А., Тұрғанбаева Ә. Қазақ тіліне арнлаған заманауи машиналық аударма технологияларына шолу. Вестник КазНИТУ, №5 (141) 2020. -стр. 103-110.

3) Абдуали Б.А., Әмірова Д.Т., Рахимова Д.Р., Кәрібаева А.С. Қазақ тіліндегі мәтінді ресурстар мен құжаттарды аналитикалық өңдеу. Вестник КазНИТУ, №2(132), 2019, стр. 356-362. (CCES)

4) Karibayeva A., Karyukin V., A. Turgynbayeva, A. Turarbek. The translation quality problems of machine translation systems for the Kazakh language. Journal of Mathematics, Mechanics and Computer Science, [S.l.], v. 111, n. 3, p. 132-140, oct. 2021. ISSN 2617-4871. (CCES)

Web Science және Scopus базасындағы конференциялар:

1) Tukeyev U., Amirova D., Karibayeva A., Sundetova A., Abduali B. Combined technology of lexical selection in rule-based machine translation. Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017, Proceedings, Part II (Lecture Notes in Computer Science) 1st ed. 2017 Edition, Springer, p. 491-500 (**Q3, SJR=0.25, CS=1.8, Percentile-50%).**

2) Tukeyev U., Karibayeva A., Abduali B. Neural machine translation system based on synthetic corpora. CMES-2018, Poland, Kazimeirz Dolny, 2018, MATEC Web of Conferences. 252. 03006. 10.1051/matecconf/201925203006 (**Web of Science**).

3) Tukeyev U., Turganbayeva A., Abduali B., Rakhimova D., Amirova D., Karibayeva A. Lexicon-free stemming for Kazakh language information retrieval. DOI:10.1109/ICAICT.2018.8747021.AICT-2018, Kazakhstan, Almaty (**Scopus)**.

4) Tukeyev U., Karibayeva A. Inferring the Complete Set of Kazakh Endings as a Language Resource. Proceedings of International Conference on Computational Collective Intelligence, 2020, p. 741-751 **(Q4, SJR=0.209, CS=0.9, Percentile – 16%)**

5) Tukeyev U., Karibayeva A., Turganbayeva A., Amirova D. Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words. Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, vol 12876. Springer,

Cham. https://doi.org/10.1007/978-3-030-88081-1_48 -p. 643–654 (**Q3, SJR=0.25, CS=1.8, Percentile-50%**).

6) Rakhimova D., Karyukin V., Karibayeva A., Turarbek A., Turganbayeva A. The Development of the Light Post-editing Module for English-Kazakh Translation. DOI: https://doi.org/10.1145/3492547.3492651. ICEMIS'21: The 7th International Conference on Engineering & MIS 2021, Almaty, Kazakhstan, October 2021 (**Scopus**)

Халықаралық конференциялар

1. Tukeyev U., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules generation, based on the complete set of Kazakh endings // Информатика және қолданбалы информатика: Халықаралық ғылыми конференция материалдары (27-30 қыркүйек 2017 ж). 2-бөлім. - Алматы, 2017, - 38 б.
2. Tukeyev U., Zhumanov Zh., Karibayeva A., Amirova D., Sundetova A., Abduali B. Формирование двуязычного словаря многозначных слов для машинного перевода казахского языка" The Vth International Conference on Computer Processing of Turkic Languages "TurkLang 2017", 18-21 October, Kazan, Tatarstan.
3. Tukeyev U., Zhumanov Zh., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules genearation, based on the complete set of Kazakh endings. The II International Conference "Computer Science and Applied Mathematics", 2017, Part II, Almaty, Kazakhstan.
4. Tukeyev U., Zhumanov Zh., Rakhimova D., Karibayvea A, Amirova D. Complex technology of machine translation resources extension for the Kazakh language. Varia Informatica 2017 №1, ISBN 978-83-936692-3-3, Lublin, 14 стр
5. Karibayeva A., Abduali B., Amirova D. Formation of the synthetic corpora for Kazakh on the base of endings complete system. Turklang-2018, Uzbekistan, Tashkent, pp. 153 – 161.
6. Кәрібаева А.С, Абдуали Б.А, Тукеев У. А. Разработка программы морфологической сегментации текста казахского языка на основе полной системы окончаний. «Фараби әдемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Казахстан, Алматы, 2020, - 53 стр.
7. Әмірова Д.Т., Кәрібаева А.С. Исследование технологии машинного перевода казахско-английской пары языков и обратно на основе трансферной модели нейронной сети. «Фараби әлемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Қазақстан, Алматы, 2020, -45 стр.
8. Рахимова Д.Р., Турарбек А., Карибаева А., Карюкин В. Технологий машинного перевода и постредактирования казахского языка. Глава в коллективной монографии «Современные методы и подходы обработки казахского языка» КГТУ, Бишкек 2021.

Зерттеушінің жеке үлесі. Ізденуші диссертациялық жұмыстың қойылған міндеттерін шешті. Қазақ тілінің нейронды машиналық аудармасында мәтінді морфологиялық сегменттеудің моделі мен әдісі әзірленді. Нейронды машиналық аударма жүйесінде оқыту мен тестілеуге арналған қазақ тіліндегі параллельді мәтіндер корпусы құрастырылды. Әзірленген модель мен әдістің тиімділігін анықтау мақсатында эксперименттер жүргізілді. Қазақ тілі үшін CSE (complete set of endings)-моделі негізінде жалғаулардың толық тізімі құрылды.

Диссертация тақырыбының ғылыми-зерттеу жұмыстарының жоспарларымен байланысы. Диссертациялық жұмыс Қазақстан Республикасы Білім және ғылым министрлігінің «Қазақ тілінің нейронды машиналық аудармасын құру және зерттеу» тақырыбы бойынша (2017-2020 жж.) гранттық зерттеу жобасы аясында жүргізілді.

Жұмыстың көлемі мен құрылымы. Диссертация кіріспеден, 4 тараудан және

қорытындыдан тұрады. Диссертацияның жалпы көлемі 172 беттен, 7 суреттен, 54 кестеден, 79 қолданылған әдебиттен тұрады.

Кіріспеде жұмыстың өзектілігі анықталып, тақырыпқа байланысты мәселелер көрсетілді. Жұмыстың идеясы, зерттеудің мақсаты мен міндеттері, зерттеудің ғылыми жаңалығы мен практикалық құндылығы, зерттеу әдістері көрсетіледі.

Бірінші тарауда нейронды машиналық аударманы жақсартудың қолданыстағы технологияларын зерттеу және талдау сипатталған. Машиналық аудармада сегменттеу және тілдің морфология моделі саласындағы зерттеулерге аналитикалық шолу жасалынады. Әдістердің негізгі артықшылықтары мен кемшіліктері анықталған.

Екінші тарауда тіл морфологиясын сипатттауда қолданылатын моделдер талданылған. Морфология модельдерін ескеріп, қазақ тілі морфологиясының жалғаулардың толық жүйесіне негізделген тілдік моделін жасау жұмыстары сипатталған.

Үшінші тарауда жалғаулардың толық жүйесінің (CSE) моделін пайдалана отырып, морфологиялық сегменттеудің моделі мен алгоритмін құру жұмыстары жүргізілді. Морфологиялық сегменттеудің қадамдық алгоритмі жасалды. Қате сегменттеуді болдырмас үшін ерекше сөздер сөздігі жасалды.

Төртінші тарауда қазақ тілінің нейронды машиналық аудармасы жүйесін оқыту үшін бағдарламалық құрал таңдалды, қазақ тіліндегі мәтіндерді сегменттеудің бағдарламасы сипатталынады. Нейронды машиналық аудар ма жүйесін оқытуда қолданылатын негізгі нейронды желі модельдері сипатталған. Оқыту процесі Tensorflow кітапханасындағы қайталанатын нейрондық желілерге негізделген seq2seq моделімен жүзеге асырылған. Құрылған әдіспен (CSE) және басқа әдіспен (BPE) тәжірибелер сипатталған. Тәжірибе нәтижелеріне талдау беріледі. Нейрондық машиналық аударманың сапасын анықтау үшін сегменттеу әдісін, атап айтқанда BPE-мен салыстыру бойынша эксперименттік жұмыс жүргізілді, BLEU метрикасында сапа нәтижелері алынды.

Қорытындыда диссертацияның негізгі нәтижелері мен қорытындылары берілген.

Алынған ғылыми нәтижелер әртүрлі оқыту конфигурациялары бар эксперименттермен расталады. Зерттеудің негізділігі мен сенімділігі әзірленген әдістің нәтижелеріне сәйкес келеді.