КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

ӘОЖ 004.93.1
ҒТАХА 20.19.00
https://doi.org/10.53364/24138614_2025_37_2_12

Д. Рахимова[1] , А.Ж. Жігер [1, 2*], В. Малых[3,4],
В. Карюкин[1], А. Бекарыстанқызы[2]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
[2] Нархоз университеті, Алматы, Қазақстан
[3]Халықаралық ақпараттық технологиялар университет, Алматы, Қазақстан
[4]Санкт-Петербург мемлекеттік ақпараттық технологиялар, механика және оптика университеті, Санкт-Петербург, Ресей

[2]*E-mail: alia_94-22@mail.ru**

ТІЛДЕР АРАСЫНДАҒЫ АУДАРМА САПАСЫН АРТТЫРУ: АҒЫЛШЫН-ҚАЗАҚ ТІЛДЕРІНДЕГІ ЖЕТІСТІКТЕР МЕН МҮМКІНДІКТЕР

Аңдатпа. *Машиналық аударма – қазіргі таңда қарқынды дамып келе жатқан және кеңінен қолданылатын заманауи технологиялық салалардың бірі. Әлемдік жаһандану үдерісі мен көптілді коммуникация қажеттілігі бұл саланың маңыздылығын айтарлықтай арттыра түсті. Түрлі мемлекеттер мен мәдениеттер арасындағы ақпарат алмасу мен өзара түсіністікті жеңілдету мақсатында машиналық аударма құралдары кең көлемде қолданылуда. Атап айтқанда, Google Translate және Яндекс Аудармашы сияқты жүйелер халықаралық деңгейде ең танымал әрі тиімді платформалар қатарына жатады. Бұл жүйелер жыл сайын жаңа алгоритмдер мен тілдік модельдерді енгізу арқылы өз аударма сапасын жетілдіруде. Алайда соңғы зерттеулер нәтижесі бойынша, бұл платформаларда ағылшын тілінен қазақ тіліне немесе басқа түркі тілдеріне жасалатын аудармалардың сапасы әлі де төмен деңгейде қалып отыр. Мұндай нәтиже, ең алдымен, қазақ тілінің күрделі морфологиялық және синтаксистік құрылымымен, сондай-ақ сөз тәртібі мен контекстуалдық мағынаның ерекшеліктерімен байланысты.*

Зерттеудің мақсаты – ағылшын тілінен қазақ тіліне бағытталған нейромашиналық аударманың сапасын арттыру үшін трансформер модельдерін бейімдеу және пост-редакторлеу әдістерін қолдана отырып тиімді тәсілдер ұсыну.

Осы мақсатта OpenNMT платформасында қазақ және басқа түркі тілдеріне бейімделген трансформер үлгісі әзірленіп, 180 000 сөйлемнен тұратын параллель корпус негізінде оқытылды. Алынған аударма нәтижелерін бағалау BLEU метрикасы арқылы жүзеге асырылды. Сонымен қатар, аударма сапасын арттыру үшін пост-редакторлеу кезеңінде Kaz-RoBERTa моделі қолданылды. Зерттеу қорытындылары көрсеткендей, параллель мәліметтердің сапасы мен көлемін ұлғайту, сондай-ақ трансформер моделін нақты тілдік ерекшеліктерге бейімдеу аударма нәтижелерінің дәлдігі мен түсініктілігін едәуір жақсартады.

Түйін сөздер: *нейромашиналық аударма, BLEU аударма метрикасы, параллельді корпус, ашық нейромашиналық аударма, трансформер моделі, пост-редакторлеу, BLEU метрикасы, Kaz-RoBERTa моделі.*

Кіріспе

Қазіргі таңда әртүрлі тілде сөйлейтін адамдар өзара байланыс орнату үшін түрлі машиналық аударма жүйелерін пайдаланады.

Қазақ тіліндегі машиналық аударма жүйелерінің қазіргі кездегі даму деңгейі, жалпы түркі тілдері арасындағы ерекшеліктерді ескере отырып, көптеген қиындықтар мен мүмкіндіктерді қарастырады.

Машиналық аударма жүйелерінің тиімділігі мен дамуы Google және Yandex сияқты әлемге әйгілі аударма жүйелері қазақ тілінде жоғары нәтижелер көрсеткенімен, олар әлі де белгілі бір қателіктерге тап болуда, әсіресе құрмалас сөйлемдер мен сөз тіркестерін аударғанда. Сондықтан арнайы қазақ тілі үшін әзірленген машиналық аударма модельдері қажеттілігін көрсетті. Мұндай модельдер, әсіресе, тілдің грамматикалық және лексикалық ерекшеліктерін ескере отырып, аударманы тиімдірек және дәл етіп жасайды. Қазіргі уақытта нейрондық желілер мен трансформерлер секілді жаңа әдістер қолданылып, бұл мәселелерді шешуге бағытталған модельдер дамытуда үлкен жетістіктерге қол жеткізілген.

Қазақ тілінің құрылымына негізделген нейрондық модельдер Қазақ тілінің грамматикалық құрылымы мен сөздердің мағынасының өзгеру ерекшеліктерін ескере отырып, нейрондық машиналық аударма жүйелерін жетілдіру бағытында маңызды зерттеулер жүргізілуде. Ағылшын-қазақ тілдеріндегі машиналық аударма жүйелерінің дамуы қазақ тілінің күрделі морфологиялық құрылымдарын түсіну үшін тереңірек модельдер мен алгоритмдерді қолдануды талап етеді. Атап айтқанда, қазақ тіліндегі жалғаулар мен жұрнақтар сөйлемнің мағынасына терең әсер етеді, және бұл факторлар жүйенің дәлдігін арттыру үшін маңызды болып табылады.

Мұндай аудармалардың ең танымал және кең қолданылатын түрлеріне Google және Yandex машиналық аударма жүйелері жатады. Алайда, бұл аударма жүйелерінің кейбір кемшіліктері де бар. Олардың негізгі мәселелері мыналар болып табылады:

a) Кейбір күрделі құрмалас сөйлемдерді аударғанда мағынаның толық жоғалуы.
b) Тұрақты тіркестерді сөзбе-сөз аудару.Жергілікті атаулардың дұрыс аударылмауы.
c) Морфологиялық құрылымда қателіктер болуы.

Бұл қателіктерге нақты сипаттама беру мақсатында, 2024 жылдың наурыз айында Гугл және Яндекс аударма жүйелері арқылы алынған аудармалар төмендегі кесте 1-де көрсетілген.

Кесте 1- Машиналық аудармалар арқылы алынған аудармалар

Мәтін жанрлары	Ағылшын мәтін	Гугл машиналық аудармасында алынған	Яндекс машиналық аудармасында алынған	Дұрыс Аударма	Аударма қателіктеріне сипаттама
Көркем-әдеби стиль	The moon hung low in the sky, casting a soft, ethereal glow over the silent forest. The air was thick with the scent of pine and damp earth, and the stillness was broken only by the occasional	Ай аспанда төмен салбырап, үнсіз орманның үстіне жұмсақ, эфирлік нұрын шашып тұрды. Ауаны қарағай мен дымқыл жердің хош иісі аңқып тұрды, ал тыныштықты түнгі самал қозғаған	Ай аспанда төмен ілініп, үнсіз орманның үстіне жұмсақ, эфирлік сәуле шашып тұрды. Ауа қарағай мен ылғалды жердің хош иісіне толы болды, ал тыныштық түнгі	Ай аспанда төменде ілініп, орманға жұмсақ, елес сияқты жарық шашып тұр. Ауа бұтақтардың иісінен және дымқыл жердің иісінен қанық, тыныштықты	Ай аспанда төмен салбырап" — бірінші мәтінде "салбырап" сөзі қолданылған, бірақ ол бұл контексте толық дұрыс емес.

	rustle of leaves stirred by the night breeze. It was a time when the world seemed to hold its breath, as though the very land was waiting for something—though what, no one could say.	жапырақтардың анда-санда сыбдыры ғана бұзды. Бұл жердің өзі бірдеңені күтіп тұрғандай, дүние демін басып тұрғандай болды, бірақ нені ешкім айта алмады.	самалмен араласқан жапырақтардың анда-санда сыбдырымен ғана бұзылды. Бұл әлем тынысын ұстап тұрғандай көрінетін уақыт болды, жердің өзі бір нәрсені күтіп тұрғандай болды—бірақ не болса да, ешкім айта алмады.	тек кейде ағаштардың жапырақтары желмен сырылдап қана бұзады. Бұл әлемнің тыныштықты ұстанып, өз тынысын ұстап тұрғандай кез. Барлығы бір нәрсе үшін күтіп тұрғандай, бірақ ол не екенін ешкім айта алмайды.	Екінші мәтінде "ілініп" сөзі қолданылып, ол айдың аспандағы қозғалысын немесе жай-күйін дұрыс сипаттайды. Негізгі қателіктер — сөздер мен тіркестердің тым ауыр әрі айқын емес болуы.
Көркем-әдеби стиль	In the distance, the faint sound of a creek gurgled, its waters rushing over smooth stones, a timeless melody that spoke of nature’s quiet endurance. Each tree, each blade of grass seemed to whisper its secrets, ancient and unspoken, as if they were old friends sharing stories beneath the cloak of night. The stars above twinkled like distant fireflies, casting their light down upon the earth,	Алыстан сылдырлаған бұлақтың әлсіз дыбысы, оның суы тегіс тастардың үстінен ағып жатыр, табиғаттың тыныш шыдамдылығын білдіретін мәңгілік әуен. Әр ағаш, әр тал шөп өзінің көне және айтылмаған сырларын сыбырлап, түн жамылғысының астында сыр бөлісетін ескі достар сияқты. Жоғарыдағы жұлдыздар алыстағы ұшқыштар сияқты жымыңдап, жер	Алыстан бұлақтың сылдырлаған әлсіз дыбысы, оның суы тегіс тастардың үстінен ағып жатқан табиғаттың сабырлы шыдамдылығын білдіретін мәңгілік әуен. Әрбір ағаш, Әрбір Тал шөбі өзінің ежелгі және айтылмаған құпияларын сыбырлайды, мысалы, түн жамылғысының астында құпияларды бөлісетін ескі достар. Жоғарыдағы жұлдыздар алыстағы	Алыста бір бұлақтың сылдырлаған дыбысы естіледі, оның суы тегіс тастардың үстімен ағып, уақытсыз әуенін шығарады, ол табиғаттың тыныш шыдамдылығын айтады. Әрбір ағаш, әрбір шөп өзінің құпияларын сыбырлап, айтар сөздерін түннің көлеңкесінде ескі достар сияқты бөлісетіндей. Жоғарыдағы жұлдыздар алыстан	Ал "үстінен" деген сөз біршама артық болып, стилистикалық тұрғыдан аздап үйлеспейді. Гугл және яндекс машиналық вудармаларда мәтін поэтикалық тұрғыда терең әрі абстрактылы сипаттамаларға ие, Ал негізгі дұрыс мәтінде нақты әрі түсінікті.

	illuminating the path that lay before the wanderer.	бетіне нұрын шашып, кезбенің алдында жатқан жолды нұрландырды.	ұшқыштар сияқты жыпылықтап, жерге нұрын шашып, Саяхатшының алдында жатқан жолды нұрландырды.	жарқырап, жерге жарық түсіреді, жолды ашып көрсетіп, саяхатшыға бағыт береді.	
Көркем-әдеби стиль	As the night deepened, a hush fell over the land, and for a brief moment, the world stood still—suspended in time, caught between the past and the future, between what was and what could be.	Түн тереңдей бергенде, жер бетінде тыныштық орнады және аз уақытқа әлем тоқтап қалды - уақыт ішінде тоқтап, өткен мен болашақтың, бар мен болатынның арасында қалды.	Түн тереңдей түскенде, жер бетінде тыныштық орнап, бір сәтке әлем бір орында тұрды—уақыт өте келе тоқтап, өткен мен болашақтың, болған мен болуы мүмкін нәрсенің арасында қалып қойды.	Түн тереңдегенде, жер үстінде тыныштық орнады, ал бір сәтте әлем тұрып қалды — уақытқа шектелген, өткен мен болашақтың арасында, болған мен болар нәрселердің арасында.	Гугл және яндекс машиналық аудармаларда мәтін аудармасы күрделендіріп аударылған. Негізгі аудармасы қарапайым әрі түсінікті аударылған.
Ғылыми	The rapid advancements in artificial intelligence (AI) and machine learning (ML) have dramatically transformed a wide array of fields, from natural language processing (NLP) to healthcare and autonomous systems. At the core of these innovations lies the development of sophisticated	Жасанды интеллект (AI) және машиналық оқытудағы (ML) жылдам жетістіктер табиғи тілді өңдеуден (NLP) денсаулық сақтау және автономды жүйелерге дейін кең ауқымды күрт өзгертті. Бұл инновацияларды ың негізінде деректердің үлкен көлемін бұрын-соңды болмаған дәлдікпен талдауға	Жасанды интеллект (AI) және машиналық оқыту (ML) саласындағы қарқынды жетістіктер табиғи тілдерді өңдеуден (NLP) денсаулық сақтау мен автономды жүйелерге дейінгі көптеген салаларды түбегейлі өзгертті. Бұл инновациялардың негізінде бұрын-соңды	Жасанды интеллект (AI) және машиналық оқыту (ML) саласындағы жылдам жетістіктер табиғат тілін өңдеуден (NLP) денсаулық сақтау мен автономды жүйелерге дейінгі көптеген салаларды түбегейлі өзгерткен. Бұл жаңалықтардың негізінде деректердің үлкен көлемін	кең ауқымды күрт өзгертті" деген тіркес қолданылған, бұл сөздер мен контекст жалпы әсердің күшті екенін көрсетеді, бірақ "күрт" сөзі аздап айқын емес, көп нәрсені қамтитын сияқты көрінуі

	algorithms capable of analyzing vast amounts of data with an unprecedented level of precision. These algorithms, often underpinned by neural networks, have significantly enhanced the efficiency of complex systems, allowing for real-time decision-making in environments that were once deemed too unpredictable or opaque for such automation.	қабілетті күрделі алгоритмдерді әзірлеу жатыр. Көбінесе нейрондық желілермен негізделген бұл алгоритмдер күрделі жүйелердің тиімділігін айтарлықтай арттырып, бір кездері мұндай автоматтандыру үшін тым болжау мүмкін емес немесе мөлдір емес деп есептелген орталарда нақты уақытта шешім қабылдауға мүмкіндік берді.	болмаған дәлдік деңгейінде деректердің үлкен көлемін талдауға қабілетті күрделі алгоритмдерді әзірлеу жатыр. Көбінесе нейрондық желілермен қамтамасыз етілетін бұл алгоритмдер күрделі жүйелердің тиімділігін едәуір арттырды, бұл бір кездері мұндай автоматтанды ру үшін тым болжау мүмкін емес немесе мөлдір емес деп саналған жағдайларда нақты уақыт режимінде шешім қабылдауға мүмкіндік берді.	бұрын-соңды болмаған дәлдікпен талдай алатын күрделі алгоритмдерді ң дамуы жатыр. Бұл алгоритмдер, көбінесе нейрондық желілермен негізделген, күрделі жүйелердің тиімділігін айтарлықтай арттырып, бұрын қолмен шешуге болатын немесе болжауға қиын болып табылған ортада нақты уақыт режимінде шешім қабылдауға мүмкіндік берді.	мүмкін. Мәтіндерді ң мазмұны бірдей болғаныме н,дұрыс аударманы ң кейбір сөздері мен құрылымд ары табиғи әрі түсінікті.

Жүргізілген эксперимент бойынша, машиналық аудармалардың қателігінің типі сөйлемдер құрылымына байланысты екенін келесі кестеден көруге болады.

Кесте 2 – Аудармадағы құрмалас сөйлемдерден алынған қателіктер

Сөйлем құрылымы бойынша түрлері	Ағылшын тілінде	Гугл машиналық аудармада алынған аударма	Яндекс машиналық аудармада алынған аударма	Дереккөзден алынған дұрыс аударма	Аудармаға сипаттама
Құрмалас сөйлем	Although the initial stages of the project were fraught	Жобаның бастапқы кезеңдерінде күтпеген	Жобаның бастапқы кезеңдері күтпеген	Жобаның бастапқы кезеңдері күтпеген	Гугл және машиналық

	with unforeseen challenges and delays, the team managed to overcome these obstacles by implementing innovative solutions, which not only streamlined the workflow but also enhanced the overall efficiency, thereby ensuring the timely completion of the final deliverables.	қиындықтар мен кідірістерге толы болғанымен, команда инновациялық шешімдерді енгізу арқылы бұл кедергілерді еңсере алды, бұл жұмыс процесін оңтайландырып қана қоймай, жалпы тиімділікті арттырды, осылайша соңғы нәтижелердің уақтылы аяқталуын қамтамасыз етті.	қиындықтар мен кідірістерге толы болғанымен, команда бұл кедергілерді инновациялық шешімдерді енгізу арқылы жеңе алды, бұл жұмыс процесін оңтайландырып қана қоймай, сонымен бірге жалпы тиімділікті арттырды, осылайша түпкілікті нәтижелердің уақтылы аяқталуын қамтамасыз етті. жеткізілімдер	қиындықтар мен кідірістерге толы болғанымен, команда осы кедергілерді жаңашыл шешімдер енгізу арқылы жеңіп шықты, олар тек жұмыс процесін оңтайландырып қана қоймай, жалпы тиімділікті де арттырып, осылайша соңғы нәтижелердің уақытында аяқталуын қамтамасыз етті.	аудармалардан алынған мәтінде “олар” деген сөздер орнына “бұлар” сөздері қолданылып, сөлем мағынасы құрылымын өзгертті.

Жоғарыда көрсетілген кесте негізінде, ағылшын-қазақ тіл жұптары бойынша алынған аударма қателіктерін келесі топтарға бөлуге болады:

Мәтін жанрларына (ғылыми, көркем стиль) байланысты қателіктер;

Құрмалас сөйлемдердегі қателіктер;

Тұрақты сөз тіркестерін сөзбе-сөз аудару;

Жалқы есімдерді дұрыс аудармау;

Морфологиялық құрылым бойынша қателіктер.

Бұл қателіктердің пайда болуы қазақ тілінің құрылымдық ерекшеліктеріне байланысты. Қазақ тілі, басқа түркі тілдері сияқты, күрделі құрылымға ие. Мысалы, қазақ тілінде қазіргі, өткен және болашақ шақтар жұрнақтар арқылы көрсетіледі. Қазақ тілінде көптеген жұрнақтар мен жалғаулар бар, бұл аударма процесінде қиындық туғызуы мүмкін. Аудармадағы негізгі қателіктер құрмалас сөйлемдерде кездеседі, себебі қазақ тіліндегі құрмалас сөйлемдер әртүрлі түрлерге бөлінеді, және олардың құрылымы мен мағынасы күрделірек. Google және Yandex машиналық аударма жүйелері құрмалас сөйлемдерді дұрыс аудармайды, көбінесе олар жай сөйлемдерді тікелей аударып, көмекші сөздермен байланыстырмай, мағынасына сай келмейтін қателіктер жібереді.

Мақалада қазақ тіліне тән құрылымдық ерекшеліктерді ескеріп, арнайы модельдер жасалды. Бұл модельдер ашық нейрондық машиналық аударма жүйесі (OpenNMT) негізінде оқытылды [1-10]. Эксперимент ағылшын-қазақ тілдері жұптары бойынша жүргізілді, және модельдерді оқыту үшін параллельді корпус Ақ Орда сайтынан жиналды. Қазақ тілінде алынған аударма нәтижелері аударма метрикалық өлшемі (BLEU) арқылы бағаланды.

Материалдар мен зерттеу әдістері.

Қазіргі уақытта нейрондық желілердің дамуына байланысты аударма жүйелерін жетілдіру үшін арнайы модельдер құрастырылып, нейрондық желілер арқылы оқыту нәтижесінде жақсы көрсеткіштерге қол жеткізілді [11-15]. Нейрондық машиналық аудармалар арасында жақсы нәтижелерге жеткен жұмыстардың бірі [16] болып табылады, онда машиналық аударма трансформер моделін қолдану арқылы қытай тілінен ағылшын тіліне аударма жасалды. Көптеген нейрондық машиналық аударма жүйелері трансформер моделінің негізінде жоғары нәтижелер көрсетті. Сонымен қатар, қайталанатын нейрондық желілер (RNN) негізіндегі модель де жетілдіріліп, қытай-ағылшын тіл жұптарында жоғары нәтижелерге қол жеткізді [17].

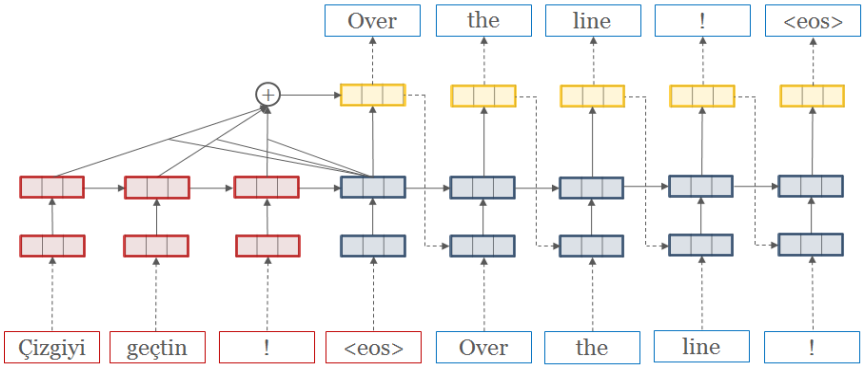
Трансформер мен RNN модельдері әртүрлі машиналық аударма жүйелерінде оқытылып, әртүрлі нәтижелер көрсеткендіктен, олардың аударма нәтижелері оқыту әдісіне байланысты екенін айтуға болады. Осы контексте ашық нейрондық машиналық аударма (OpenNMT) [18] зерттеліп, алғаш рет қайталанатын нейрондық желілер (RNN) моделі қолданысқа енгізілді. Бұл модель арқылы ағылшын тілінен неміс тіліне аударма жасалып, нәтиже жоғары болды. Алайда, ағылшын-қазақ тілдері жұбында аударма нәтижесі төменірек болды.

Осы жұмыстың [19] негізінде OpenNMT жүйесінде қазақ тіліне арнайы параметрлермен трансформер моделі жасалды. Бұл модель қазақ тілінің семантикалық құрылымын ескере отырып, аударма нәтижелерін жақсартты. Трансформер моделін оқыту үшін ағылшын-қазақ тіл жұбында құрылымы әртүрлі ресми сайттардан (ақорда, mfa.gov.kz, ekonomi.gov.kz, strategiya2050.kz) алынған 109 772 параллель сөйлемдер мен 20 000 көркем әдеби стильдегі сөйлемдер мен сөз тіркестерінен тұратын корпус пайдаланылды. Әр сайттан алынған нақты сөйлемдер санын келесі кестеден көруге болады.

Кесте 3 – Корпустағы сөйлемдер саны

Корпус аты	Сөйлемдер саны
Ақорда	40661
Премьер-министр	6680
mfa.gov	9895
Economy.gov	6550
Стратегия 2050	45986
Әдеби оқулықтардан алынған	20000

Алдымен OpenNMT-ашық нейронды машиналық аудармаға тоқталып өтсек. Оның негізгі қасиеттеріне: 1) үлкен көлемді корпустарды тез оқиды; 2) оқыту нәтижесінде үлкен көлемді мәліметтерді тез әрі сапалы аударады.



Сурет 1 – Ашық нейронды машиналық аударма (OpenNMT) үлгісі

Сурет 1-де қызыл түсті тіктөртбұрышта бастапқы аударылатын сөйлемнің сөздері енгізіледі. Сары түстегі тіктөртбұрышта жасырын қабатта сөздер аударылып, жіберіледі. Рекуррентті нейрон желісі (RNN) мен трансформер нейрон желісі негізінде сары түстен көк тіктөртбұрышқа аударылған сөздер жеткізіледі.

Ашық нейронды машиналық аударма жүйесінде (OpenNMT) қазақ тіліне аудару алгоритмі келесі қадамдардан тұрады:

1. Корпус дайындау: Ағылшын тіліндегі сөйлемдер src-train.txt файлында, ал қазақ тіліндегі сәйкес аударма сөйлемдер tgt-train.txt файлында сақталады. Қазақ тіліндегі аударма мәтіндер ресми сайттардан алынған.

2. Трансформер моделін құру: Модель құрып, оның коды en-kk.yaml файлына жазылып, оқыту үшін арнайы параметрлерден құрылды.

```
#Transformer model
encoder_type: transformer
decoder_type: transformer
position_encoding: true
enc_layers: 6
dec_layers: 6
heads: 8
rnn_size: 512
word_vec_size: 512
transformer_ff: 2048
dropout_steps: [0]
dropout: [0.1]
attention_dropout: [0.1]
```

Сурет 2 – Трансформер моделінде параметрлер сипаттамасы

Осы модель ашық нейрон машиналық аударманың арнайы командасы onmt_train -config en-kk.yaml арқылы оқытылды. Корпустағы 180000-нан тұратын сөйлемдерді аудару үшін оқытуға 6 сағаттай уақыт жұмсалды.

3) Осы модельде ағылшын тілінен қазақ тіліне аударма орындау үшін алдын ала дайындалған 20000 -нан тұратын ағылшын тіліндегі src-test.txt файлда сақталған сөйлемдерді модельдің 60000 қадамында оқытылған нұсқасында нәтижесі pred_1000.txt файлға қазақ тіліне аударма жазылды. Аударма ашық нейрон машиналық аударманың арнайы командасы арқылы жүргізілді.

Аударма сапасын жақсарту үшін қазақ тілінде аударма табылмаған <unk> сөздердің жақын аудармасы табу үшін Kaz-RoBERTa моделінде қайта оқытылып аударма сапасы жақсартылды. Kaz-RoBERTa — қазақ тілінде табиғи тіл өңдеу (NLP) тапсырмалары үшін арнайы жасалған тілдік модель. Ол RoBERTa (A Robustly Optimized BERT Pretraining Approach) архитектурасына негізделіп [20], қазақ тіліндегі мәтіндермен тиімді жұмыс істеу үшін бейімделген.

Kaz-RoBERTa қазақ тілінің ерекшеліктерін ескеріп, сол тілдегі үлкен көлемдегі деректермен оқытылған. Бұл модель мәтіннің контекстін түсіну, сұрақ-жауап жүйелері, мәтіндерді классификациялау және талдау сияқты тапсырмаларды орындауда жоғары нәтижелер көрсетеді.

Kaz-RoBERTa-ның айрықша қасиеттері:

1. Қазақ тілінің ерекшеліктеріне бейімделу: Модель қазақ тілінің грамматикалық құрылымы мен сөз құрылымын терең түсіну үшін арнайы дайындалған.

2. Үлкен мәтіндермен оқыту: Модель қазақ тіліндегі ауқымды деректерді пайдаланып оқытылған.

3. Жоғары дәлдік: Kaz-RoBERTa түрлі NLP тапсырмаларында жоғары нәтиже көрсетеді.

Kaz-RoBERTa моделін Opennmt-ден аударылмай қалған текғсттер аудару үшін процесс келесі қадамдардан тұрады:

a) Модель мен токенизаторды жүктеу:
b) Аудармасы белгісіз сөздердің индексін табу:
c) Аудармасы белгісіз сөздердің ықтималдықтар алу:
Ең ықтимал токендерді алу:
d) Токендерді сөздерге айналдыру:

OpenNMT ашық нейрон машиналық аудармада транформер моделі арқылы алынған текст Kaz-RoBERTa моделінде қайта оқытылып, қазақ тілі пост-редакторленіп, аударма сапасының жақсарғанын көруге болады.

Нәтижелер және оларды талқылау.

Құрылған модель 178 000 параллель сөйлемдерден тұратын деректер файлының негізінде 6 сағат ішінде оқытылды. Модельді тексеру мақсатында 20 000 сөйлемнен тұратын ғылыми және көркем әдебиет стиліндегі, құрылымы жағынан әртүрлі ағылшын тіліндегі мәтіндер қазақ тіліне аударылды. Аударма нәтижелері BLEU (Bilingual Evaluation Understudy) метрикасы арқылы бағаланды. BLEU метрикасы аударылған мәтіннің сапасын бағалауға арналған құрал болып табылады, оның мәні 0 мен 1 (яғни 0% пен 100%) аралығында болады.

$$BLEU = Brevity\_Penalty\ X\ \prod_{n=1}^{N} p_n^{\omega_n} \quad (1)$$

мұнда,ы, p_n –N-gram дәлдік мәндері болып табылады; N-қабаттардың максималды деңгейі ; w_n N -gram салмақтары;

$$Brevity\_Penalty = \{1, c > r \;\; e^{(1-r/c)}, c \leq r \quad (2)$$

мұндағы c –модель арқылы оқытылып, аударылған аударма ұзындығы және r – дұрыс нақты аударма ұзындығы. BLEU ұпайының мәні 1-ге (100%) жақын болған сайын, аударма сапасы соғұрлым жоғары болады.

Кесте 4- Трансформер моделінде алынған нәтижелер

Тіл жұптары	Жылдамдық ток/сек	BLEU
Ағылшын-қазақ	4300	0,45
Қазақ-ағылшын	4300	0,45

Анықталмай қалған <unk> сөздердің аудармасын анықтау үшін Kaz-RoBERTa моделі қолданылды. Аударма нәтижесі Кесте 5 бойынша алынды.

Кесте 5 - Kaz-RoBERTa моделінде алынған нәтижелер

Тіл жұптары	Жылдамдық ток/сек	BLEU
Ағылшын-қазақ	4300	0,55
Қазақ-ағылшын	4300	0,55

Нейронды машиналық аудармада аударылмай қалған <unk> сөздер орнына аудармасы және мағынасы жағынан жақын сөздер қойылады.

Қорытынды.

Мақалада ашық нейронды машиналық аударма жүйесінде (OpenNMT) трансформер негізіндегі модельдің көмегімен қазақ тіліндегі аударма сапасын жақсартатын арнайы параметрлер таңдалды. Бұл модельді оқыту үшін аз уақыт қажет болды. Оқыту процесі кезінде ресми сайттардан, сондай-ақ көркем әдеби шығармалардан алынған параллельді сөйлемдерден тұратын 180 000-ға жуық әртүрлі құрылымдағы (жай және құрмалас) сөйлемдерден құралған корпус қолданылды. Модельді тексеру үшін 20 000-нан астам ағылшын тіліндегі сөйлемдер қазақ тіліне аударылды. Нәтижесінде аударма көрсеткіші ағылшын-қазақ және қазақ-ағылшын тіл жұптары үшін BLEU метрикасы бойынша 45% құрады. Кейін аударылмаған сөздер Kaz-RoBERTa моделінде қайта өңделіп, аударылғанда нәтижесі 55%-ға жетті. Бұл нәтижелер бойынша, аударма сапасы Google және Yandex машиналық аударма жүйелерінен кем түспейді. Жалқы зат есімдер (адам, жер, су) атауларының аудармасы да дұрыс орындалды. Аударма барысында сөйлемдердің морфологиялық құрылымы мен семантикасы жақсы сақталды. Аударма сапасын одан әрі жақсарту үшін корпусты толықтыру қажеттігі эксперимент жүзінде дәлелденді. Сонымен қатар, ашық нейронды машиналық аударма жүйесінде құрылған трансформер негізіндегі модель арнайы параметрлермен қазақ тіліндегі аударма сапасын арттырды. Бұл модельді басқа түркі тілдерінде де қолдануға болады, себебі түркі тілдеріндегі тілдік құрылымдар (морфология, синтаксис) бір-біріне ұқсас келеді.

Алғыс: *Бұл зерттеу Қазақстан Республикасының жоғары білімі және Ғылым министрлігінің қолдауымен BR24993001 жобасымен қаржыландырылды.*

Әдебиеттер

1. Rakhimova, D. R., and A. Zh. Zhunusova. (2022). "Post-editing for the Kazakh Language Using OpenNMT." *Journal of Mathematics, Mechanics and Computer Science*, vol. 113, no. 1, pp. 118–122. https://doi.org/10.26577/JMMCS.2022.v113.i1.12.

2. Zhumanov, Z. M., and U. A. Tukeyev. (2009). "Development of Machine Translation Software Logical Model (Translation from Kazakh into English Language)." *Reports of the Third Congress of the World Mathematical Society of Turkic Countries*, edited by Bakhytzhan T. Zhumagulov, vol. 1, pp. 356–363.

3. Tukeyev, U., Zh. Zhumanov, and D. Rakhimova. (2010). "Features of Development for Natural Language Processing." *ICT - From Theory to Practice*, edited by M. Milosz, Polish Information Processing Society, pp. 149–174.

4. Tukeyev, U., and D. Rakhimova. (2012). "Augmented Attribute Grammar in Meaning of Natural Languages Sentences." *Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2012*, Kobe, Japan, pp. 1080–1085.

5. Abeustanova, A., and U. Tukeyev. (2017). "Automatic Post-editing of Kazakh Sentences Machine Translated from English." *Advanced Topics in Intelligent Information and Database Systems: ACIIDS 2017*, vol. 710, Springer, pp. 283–295.

6. Schuster, S., R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. (2015). "Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval." *Proceedings of the International Conference on Vision and Language (VL)*, pp. 70–80.

7. Xu, K., et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *arXiv*, arXiv:1502.03044.

8. Shormakova, A., Zh. Zhumanov, and D. Rakhimova. (2019). "Post-editing of Words in Kazakh Sentences for Information Retrieval." *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 6, pp. 1896–1908.

9. Turganbayeva, A., and U. Tukeyev. (2020). "The Solution of the Problem of Unknown Words Under Neural Machine Translation of the Kazakh Language." *Journal of Information and*

Telecommunication, pp. 214–225.

10. Tukeyev, U., A. Karibayeva, and Z. Zhumanov. (2020). "Morphological Segmentation Method for Turkic Language Neural Machine Translation." *Cogent Engineering*, vol. 7, no. 1, pp. 1–16. https://doi.org/10.1080/23311916.2020.1780271.

11. Koehn, P., and R. Knowles. (2017). "Six Challenges for Neural Machine Translation." *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39.

12. Koehn, P. (2017). "Statistical Machine Translation. Draft of Chapter 13. Neural Machine Translation." *arXiv*, arXiv:1709.07809v1[cs.CL], 117.

13. Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July, pp. 311–318.

14. Alvarez-Melis, D., and T. S. Jaakkola. (2017). "A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, Sept.9-11,pp.412-421.

15. Zhou, Q., N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao. (2018). "Neural Document Summarization by Jointly Learning to Score and Select Sentences." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 654–663.

16. Shah, R., M. K. Gupta, and A. Kumar. (2022). "Ancient Sanskrit Line-Level OCR Using OpenNMT Architecture." *Proceedings of the 2021 Sixth International Conference on Image Information Processing (ICIIP)*, pp. 347–352. https://doi.org/10.1109/ICIIP53038.2021.9702666.

17. Hao, L., W. Gao, and J. Fang. (2021). "High-Performance English-Chinese Machine Translation Based on GPU-Enabled Deep Neural Networks with Domain Corpus." *Applied Sciences*, vol. 11, no. 22, p. 10915. https://doi.org/10.3390/app112210915.

18. Quadri, M. P., and P. Kumar. (2024). "Corpus-Based Machine Translation for English to Low-Resource Language Using OpenNMT." *Innovative Computing and Communications*, pp. 199–217.

19. Senellart, J., D. Zhang, B. Wang, G. Klein, J.-P. Ramatchandirin, J. Crego, and A. Rush. (2018). "OpenNMT System Description for WNMT 2018: 800 Words/Sec on a Single-Core CPU." *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Association for Computational Linguistics, pp. 122–128. https://doi.org/10.18653/v1/W18-2715.

20. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

УЛУЧШЕНИЕ КАЧЕСТВА ПЕРЕВОДА МЕЖДУ ЯЗЫКАМИ: ДОСТИЖЕНИЯ И ВОЗМОЖНОСТИ В АНГЛИЙСКО-КАЗАХСКОМ ПЕРЕВОДЕ

Аннотация. *Машинный перевод – это одна из быстро развивающихся и широко применяемых современных технологий. Процесс глобализации и необходимость многоязычной коммуникации значительно повышают важность этой области. Для облегчения обмена информацией и взаимопонимания между различными странами и культурами активно используются инструменты машинного перевода. В частности, такие системы, как Google Translate и Яндекс Переводчик, являются наиболее популярными и эффективными платформами на международном уровне. Эти системы ежегодно внедряют новые алгоритмы и языковые модели, улучшая качество перевода. Однако, согласно последним исследованиям, качество перевода с английского языка на казахский и другие тюркские языки по-прежнему остается на низком уровне. Этот*

результат в первую очередь связан с особенностями морфологии и синтаксиса казахского языка, а также с порядком слов и контекстуальными значениями. Цель исследования – предложить эффективные методы улучшения качества нейромашинного перевода с английского на казахский язык с использованием адаптации трансформерных моделей и методов постредактирования.

С этой целью на платформе OpenNMT был разработан трансформер, адаптированный для казахского и других тюркских языков, который обучался на параллельном корпусе из 180 000 предложений. Оценка полученных результатов перевода была проведена с использованием метрики BLEU. Также для улучшения качества перевода был использован этап постредактирования с применением модели Kaz-RoBERTa. Результаты исследования показали, что увеличение качества и объема параллельных данных, а также адаптация трансформерной модели к особенностям конкретного языка значительно улучшает точность и понятность перевода.

Ключевые слова: *нейромашинный перевод, метрика BLEU, параллельный корпус, открытый нейромашинный перевод, трансформерная модель, постредактирование, Kaz-RoBERTa модель.*

IMPROVING TRANSLATION QUALITY BETWEEN LANGUAGES: ACHIEVEMENTS AND OPPORTUNITIES IN ENGLISH-KAZAKH TRANSLATION

Abstract. *Machine translation is one of the rapidly developing and widely used modern technological fields. The process of globalization and the need for multilingual communication have significantly increased the importance of this area. To facilitate information exchange and mutual understanding between different countries and cultures, machine translation tools are being widely used. Specifically, systems such as Google Translate and Yandex Translator are among the most popular and effective platforms on an international level. These systems annually introduce new algorithms and language models to improve translation quality. However, recent research has shown that translations from English to Kazakh and other Turkic languages still remain at a low level. This result is primarily related to the complex morphological and syntactic structure of the Kazakh language, as well as word order and contextual meaning.*

The aim of this research is to propose effective methods for improving the quality of neural machine translation from English to Kazakh through the adaptation of transformer models and post-editing techniques.

For this purpose, a transformer model adapted for Kazakh and other Turkic languages was developed on the OpenNMT platform and trained on a parallel corpus of 180,000 sentences. The evaluation of the translation results was carried out using the BLEU metric. Additionally, the post-editing phase was implemented with the Kaz-RoBERTa model to improve translation quality. The results of the study demonstrated that increasing the quality and volume of parallel data, as well as adapting the transformer model to the linguistic characteristics of a specific language, significantly enhances the accuracy and clarity of the translation.

Keywords: *neural machine translation, BLEU translation metric, parallel corpus, open neural machine translation, transformer model, post-editing, Kaz-RoBERTa model.*

Авторлар туралы мәлімет

Рахимова Диана Рамазанқызы	PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: di.diva@mail.ru
Жігер Алия Жігерқызы	Магстр, Әл-Фараби атындағы Қазақ ұлттық университеті, Нархоз университеті, Алматы, Қазақстан, E-mail: alia_94-22@mail.ru
Валентин Малых	PhD, Халықаралық ақпараттық технологиялар университетінің (Алматы, Қазақстан) және Санкт-Петербург мемлекеттік ақпараттық

	технологиялар, механика және оптика университетінің ғылыми қызметкері, Санкт-Петербург, Ресей, E-mail: valentin.malykh@phystech.edu
Карюкин Владислав Игоревич	PhD, Әл-Фараби атындағы Қазақ ұлттық университеті,Алматы, Қазақстан, E-mail: vladislav.karyukin@gmail.com
Бекарыстанқызы Ақбаян	PhD, Нархоз университеті, Алматы, Қазақстан E-mail:akbayan.b@gmail.com

Сведение об авторах

Рахимова Диана Рамазанқызы	PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: di.diva@mail.ru
Жігер Алия Жігерқызы	Магистр, преподаватель Университета Нархоз, Алматы, Казахстан и Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: alia_94-22@mail.ru
Валентин Малых	PhD, научный сотрудник Международного университета информационных технологий (Алматы, Казахстан) и Санкт-Петербургского государственного университета информационных технологий, механики и оптики (Санкт-Петербург, Россия). E-mail: valentin.malykh@phystech.edu
Карюкин Владислав Игоревич	PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: vladislav.karyukin@gmail.com
Бекарыстанқызы Ақбаян	PhD, Университет Нархоз, Алматы, Казахстан E-mail: akbayan.b@gmail.com

Information about the authors

Rakhimova Diana Ramazankyzy	PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan E-mail: di.diva@mail.ru
Zhiger Aliya Zhigerkyzy	Master, teacher at Narxoz University, Almaty, Kazakhstan and Al-Farabi Kazakh National University, Almaty, Kazakhstan E-mail: alia_94-22@mail.ru
Valentin Malykh	PhD, Researcher at the International University of Information Technologies (Almaty, Kazakhstan) and St. Petersburg State University of Information Technologies, Mechanics and Optics (St. Petersburg, Russia). E-mail: valentin.malykh@phystech.edu
Karyukin Vladislav Igorevich	PhD,Al-Farabi Kazakh National University, Almaty, Kazakhstan E-mail: vladislav.karyukin@gmail.com
Akbayan Bekarystankyzy	PhD, Narxoz University, Almaty, Kazakhstan E-mail: akbayan.b@gmail.com