ӘОЖ 004.934.1
ҒТАМР 20.19.00

https://doi.org/10.55452/1998-6688-2025-22-2-54-66

[1]Рахимова Д.,
PhD, ORCID ID: 0000-0003-1427-198X,
e-mail: di.diva@mail.ru
[1,2]*Жігер А.,
магистр, ORCID ID: 0000-0002-3641-8260,
*e-mail: alia_94-22@mail.ru
[3]Малых В.,
PhD, ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Карюкин В.,
PhD, ORCID ID: 0000-0002-8768-0349,
e-mail: vladislav.karyukin@gmail.com
[2]Бекарыстанқызы А.,
PhD, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com

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

АҒЫЛШЫН-ҚАЗАҚ ТІЛІ ЖҰБЫ ҮШІН НЕЙРОНДЫҚ МАШИНАЛЫҚ АУДАРМА

Аңдатпа

Қазіргі уақытта ақпараттық технологиялар қарқынды дамуда. Оның бір саласы ретінде машиналық аудармаларды айтуға болады. Әртүрлі елдегі адамдар бір-бірін түсіну үшін машиналық аудармаларды қолданып, оның қажеттілігі жыл сайын артып келеді. Қазіргі кезде жақсы аударылатын машиналық аудармалар қатарына Google және Yandex машиналық аудармаларын жатқызуға болады. Жыл сайын Yandex және Google машиналық аудармалары аударма сапасын жоғары деңгейге көтеруде. Алайда жүргізілген эксперимент нәтижесі бойынша ағылшын немесе орыс тілінен қазақ тіліне, сондай-ақ түркі тілдес тілдерге аударғанда, аударма сапасы өзге тілдермен салыстырғанда төмендеу. Оған дәлел ретінде 2024 жылдың қыркүйек айында осы екі машиналық аудармадан алынған аударма нәтижесі жұмыста көрсетілді. Өйткені нейрондық машиналық аударманы жүзеге асыратын модель тілдердің құрылымына тікелей байланысты. Осыған орай, 2000-жылдардан бастап Қазақстан ғалымдары қазақ тіліне аударылған мәтіндердің сапасын жақсарту мақсатында өзге тілдерді жақсы аударатын модельдерді салыстырып, қызу зерттеп, қазақ тіліне арнайы модельдер құрып, ғылыми жұмыстарды жариялай бастады [1–5]. Жұмыстың мақсаты – ағылшын тілінен қазақ тіліне аударылған мәтіндердің сапасын жақсарту. Ол үшін ашық нейрондық машиналық аудармада (OpenNMT) аударманы оқыту үшін қазақ және түркі тілдес тілдерге бейімделген трансформер моделі құрылды. Құрылған модель 180 000 ағылшын–қазақ параллель корпусты оқып, үйренді. Аударма нәтижесіне баға беру үшін құрылымы жағынан әртүрлі (жай, құрмалас) 20 000 сөйлемнен тұратын ағылшын тіліндегі файл қазақ тіліне аударылды. Нәтиже аударма сапасының метрикасы (BLEU) арқылы өлшеніп [6–7], жақсы деңгейді көрсетті. Жұмыста жүргізілген эксперимент алынған нәтижені одан да жоғары деңгейге көтеру үшін, модельді оқыту кезінде ағылшын-қазақ тілдер жұбынан құрылған параллель сөйлемдер санын көбейту қажет екенін көрсетті [8].

Тірек сөздер: нейромашиналық аударма, аударма көрсеткіш метрикасы, параллель корпус, ашық нейромашиналық аударма, трансформер моделі.

Кіріспе

Қазіргі уақытта әр түрлі тілде сөйлейтін адамдар өзара қарым-қатынас орнату үшін түрлі машиналық аударма жүйелерін қолданады. Қолданылып жүрген машиналық аудармалардың ең жоғары деңгейде жұмыс істейтін түрлеріне Google және Yandex аударма жүйелерін жатқызуға болады. Алайда осы машиналық аудармалардың кемшіліктеріне:

1) Кейбір күрделі құрмалас сөйлемдерді аударғанда, сөйлем мағынасын жоғалтады.
2) Тұрақты сөздерді тікелей сөзбе сөз аударады.
3) Сөйлемдерде, адам жер атауларын дұрыс аудармайды
4) Морфологиялық құрылымы қате

Нақты қателерге сипаттама берілу үшін, 2024 ж. наурыз айында гугл және яндекс машиналық аудармалар арқылы алынған аудармаларға төмендегі 1-кестеде сипаттама берілді.

Кесте 1– Машиналық аудармалар арқылы алынған аудармалар

Мәтін жанрлары	Ағылшын мәтін	Гугл машиналық аудармасында алынған	Яндекс машиналық аудармасында алынған	Аудармаға сипаттама
Көркем-әдеби стиль	Fasting is part of the practices of many religions, including Islam, Judai and Christianity.	Ораза ислам, Иудаизм және христиандық сияқты көптеген діндердің әдет-ғұрыптарының бөлігі болып табылады.	Ораза көптеген діндердің, соның ішінде ислам, иудаизм және христиандықтың тәжірибесінің бөлігі болып табылады.	Аударма қателігі: Бұл сөйлемде practice-қолданылу деп аударылу орнына, әдет-ғұрып және тәжірибе деп аударылды. Аударылған синонимдер мына сөйлемде семантикалық мағынасын жоғалтты.
Көркем-әдеби стиль	Despite all the fun I had, I also made sure to use my summer vacation to catch up on some of the work I had fallen behind on during the school year.	Қанша қызық болғанына қарамастан, мен де жазғы демалысымды оқу жылында артта қалған жұмыстарымның біразын аяқтау үшін пайдалан-дым.	Қанша қызық болғанына қарамастан, Мен де жазғы демалысты оқу жылында артта қалған жұмыстарымның біразын аяқтау үшін пайдаландым	Despite all the Fun-барлық қызықтарға қарамастан деп аударылу дұрыс сөйлем ішінде. Қазақ тілінде Қанша қызық болғанына қарамастан деген сөз тіркесі қолданылмайды.
Көркем-әдеби стиль	The 14th of April 2012 was the centenary – the 100th anniversary – of the sinking of the passenger ship Titanic in the north Atlantic.	2012 жылдың 14 сәуірінде Атлант мұхиты-ның солтүстігінде «Титаник» жолаушылар кемесі суға батқанына 100 жыл толды.	2012 жылдың 14 сәуірінде Солтүстік Атлантикадағы Титаник жолаушылар кемесінің апатқа ұшырағанына100 жыл толды.	100 жыл толды-«толды» деген сөз тарихи сәтті сипаттағанда қолданылмайды. Оның орнына 100 жыл өтті немесе 100 жыл болды деген аударма сөйлем мағынасын жоғалтпайды.

1-кестенің жалғасы

Ғылыми	To this end, we carried out a pooled analysis of current studies and evaluated the association between underlying or previous history of CVD conditions and outcomes of infection severity in COVID-19 patients	Осы мақсатта біз жүзеге асырдық ағымдағы зерттеулер-дің жиынтық талдауы және арасындағы байланысты бағалады КВД жағдайының негізгі немесе бұрынғы тарихы және COVID-19 пациенттеріндегі инфекция ауырлығы-ның нәтижелері	осы мақсатта біз ағымдағы зерттеулер-дің бірлескен талдауы және арасындағы байланысты бағалады негізгі немесе алды ңғы жүрек-қан тамырлары ауруларының тарихы және covid-19 пациенттеріндегі инфекцияның ауырлығының нәтижелері	Қате: құрмалас сөйлем семантикалық жағынан мағына жоғалтқандығында.
Ғылыми	The list of available treatment options for managing blood glucose in patients with type 2 diabetes (T2D) has grown over recent years making the task of choosing between traditional and newer glucose-lowering agents a difficult one for healthcare providers.	2 типті қант диабеті (T2D) бар науқастарда қан глюкозасын басқаруға арналған қолжетімді емдеу нұсқаларының тізімі соңғы жылдары өсті, бұл дәстүрлі және жаңа глю-козаны төмендететін агенттерді таңдау мін детін денсаулық сақтау провай-дерлері үшін қиынға соқты.	Пациенттердегі қандағы глюкозаны бақылауға арналған қол жетімді емдеу нұсқаларының тізімі соңғы жылдары 2 типті қант диабеті (T2D) кеңейіп, медицина мамандары үшін дәстүрлі және жаңа қантты төмендететін құралдарды таңдауды қиындатты.	Аударма қателігі: Гугл аудармадасында медицина құралдары орына проввайдер деп аударуы, ал яндекс машииналық аудармада пациенттердегі қандағы орнына пациенттердің қанындағы деген морфологиялық құрылым жағынан қателік кетті.

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

Жоғары кесте қорытындылай келе, ағылшын-қазақ тіл жұптары үшін алынған аударма қателерін келесідей топтастыруға болады:

- Мәтін жанрларына (ғылыми, көркем стиль) байланысты қателер;
- Құрмалас сөйлемдер;
- Тұрақсыз сөз тіркестерін сөзбе сөз аудару;
- Жалқы сөздерді дұрыс аудармау;
- Морфология құрылымы бойынша қателер;

Жоғарыда аталған қателердің пайда болу себебі – қазақ тілінің құрылымына тікелей байланысты. Қазақ тілі, басқа түркі тілдес тілдер секілді, құрылымы жағынан күрделі. Өйткені қазақ тілінде қазіргі, өткен және келер шақтар арнайы жұрнақтар арқылы беріледі. Сонымен қатар, қазақ тілінде жұрнақтар мен жалғаулардың көптеген түрлері бар. Аудармада жиі кездесетін қателердің бірі – құрмалас сөйлемдерде. Себебі қазақ тіліндегі құрмалас сөйлемдердің 10-нан астам түрі бар. Бұл түрлер мағынасы мен жасалу жолына қарай жіктеледі. Google және Yandex машиналық аудармалары құрмалас сөйлемдерді аударған кезде, оларды жай сөйлемдерге бөліп, көмекші сөздер арқылы дұрыс байланыстырмай, семантикалық тұрғыдан қателіктер жібереді.

Кесте 2 – Сөйлем құрылымына байланысты аудармада алынған қателіктер

Сөйлем құрылымы бойынша түрлері	Ағылшын тілінде	Гугл машиналық аудармада алынған аударма	Яндекс машиналық аудармада алынған аударма	Аудармаға сипаттама
Құрмалас сөйлем	When we British go on holiday in, for example, France or Spain, and we ask for a cup of tea in a hotel or cafe, the waiter brings us a cup of lukewarm water and a tea bag on the end of a piece of string.	Біз британдық-тар, мысалы, Францияға немесе Испанияға демалуға барғанда және біз қонақүйде немесе кафеде бір шыны шай сұрасақ, даяшы бізге жіптің ұшында бір кесе жылы су мен шай пакетін әкеледі.	Біз британдықтар Франция немесе Испания сияқты демалысқа шығып, қонақүйден немесе кафеден бір кесе шай сұрағанда, даяшы бізге бір кесе жылы су мен жіптің соңында бір қап шай әкеледі.	Мында қателік сөйлемнің соңғы жағында көрінеді. Негізі дұрыс аудармасы: жіптің ұшында ілінген шай пакеті мен жылы су әкелді. Бұл аудармада синтаксис құрылымы және семантика жағынан қателік бар.
Жай сөйлем	Leonardo also made rough drawings of machines that are similar to those that were invented much later, such as submarines and helicopters.	Леонардо соны-мен бірге сүңгуір қайықтар мен тікұшақтар сияқты кейінірек ойлап табылған машиналарға ұқсас машиналардың дөрекі сызбаларын жасады.	Леонардо сонымен қатар сүңгуір қайықтар мен тікұ шақтар сияқты кейінірек ойлап табылғандарға ұқсас машиналардың өрескел сызбаларын жасады.	Мында қателік: дөрекі, өрескел сызба. Сызбаның сипаттамасы күрделі сызба құрылды деп аударылғн дұрыс. Лексикалық жағынан дұрыс синоним таңдалмады.

Осы мақалада құрылымы күрделі қазақ тіліне бейімделген арнайы модельдер құрылды. Бұл модельдер ашық нейрондық машиналық аударма жүйесінде (OpenNMT) оқытылды [8–10]. Эксперимент ағылшын-қазақ тілдік жұбына жүргізілді. Модельді оқыту үшін ағылшын-қазақ тілдерінен тұратын параллель корпус «Ақорда» сайтынан жинақталды. Қазақ тіліндегі алынған аударма нәтижесі BLEU аударма метрикасы арқылы бағаланды. Мақалада жүргізілген жұмыстарды толық түсіндіріліп, ашу үшін бірнеше бөлімге бөлінді:

- Бөлім 2-де, басқа тілден қазақ тіліне аударғанда, аударма сапасын жоғары деңгейде алуға ат салысқан Қазақстан және өзге елдің ғалымдарының жұмыстарына шолу.
- Бөлім 3-те модельдерді оқыту үшін жиналған деректерге сипаттама беріледі. Трансформер модельге және аударма метрикалық өлшемге (bleu) математикалық және бағдарламалық сипаттама берілді.
- Бөлім 4-те нәтиже талданады.
- Бөлім 5-те жұмысқа жалпылама қорытынды жасалынады.
- Бөлім 6-да мақалада айтылған жұмысты қорытындыланады.

Қазақстан мемлекеті 1991 ж. тәуелсіз ел болып, қазақ тілі мемлекеттік тіл ретінде бекітілгеннен бастап, барлық ресми мәліметтер, жаңалықтар мен оқулықтарды қазақ тіліне аудару міндетті болды.

Сол себепті Қазақстан ғалымдары арасында өзге тілден қазақ тіліне аудару және алынған аударманың сапасын жақсарту тақырыбы қызу талқыланып, зерттеле бастады.

Ғалым К. Бектаев алғаш рет 1999 ж. өз еңбегінде қазақ тілінің морфологиялық құрылымын анықтайтын модель құрды. Бұл модель қазақ тіліндегі жалғаулардың 753 түрін анықтауға мүмкіндік береді. Сонымен қатар, Бектаев моделінде қазақ тілінен орыс тіліне сөздер тікелей аударылып, сөздік жасалды.

2000 жж. басында профессор У.А. Тукеев қазақ тілінің морфологиялық құрылымын зерттеп, жалғаудың 3240 түрін анықтайтын модель ұсынды.

Профессор У.А. Тукеевның шәкірттері – Д. Рахимова және Картбаев өз еңбектерінде орыс тілінен қазақ тіліне аударуда аударманың семантикалық жағына ерекше назар аударып, соған сәйкес модель құрды.

Соңғы жылдары нейрон желілері арқылы оқыту дамығандықтан, қазіргі таңда аудармаларға арнайы модельдер құрылып, олар нейрон желілері арқылы оқытылып, жақсы нәтижелерге қол жеткізуде [11–15].

Жақсы нәтижеге жеткен нейрондық машиналық аудармалар қатарына [16] жұмысын жатқызуға болады: мұнда алынған аудармалар трансформер моделі арқылы түрленіп, қытай тілінен ағылшын тіліне аударылды.

Көптеген нейрондық машиналық аудармалар трансформер модельдері негізінде жақсы көрсеткіштер көрсетті. Алайда, трансформер модельдерімен салыстырғанда қайталанатын нейрондық желілер (RNN) негізінде құрылған модельдер қытай–ағылшын тіл жұбы үшін одан да жоғары нәтиже көрсетті [17].

Трансформер және қайталанатын нейрондық желілер (RNN) әр түрлі машиналық аудармаларда оқытылып, әр түрлі нәтижелер көрсеткендіктен, мынадай қорытынды жасауға болады: аударма нәтижесі қолданылатын модельге және оны оқытатын машиналық аударма жүйесіне тікелей байланысты. Мұндай жүйелердің бірі – ашық нейрондық машиналық аударма (OpenNMT). Бұл жүйе алғаш рет [18] жұмысында зерттеле бастады. Аталған зерттеуде қайталанатын нейрондық желі (RNN) моделі құрылып, ағылшын тілінен неміс тіліне аударма жасалды. Нәтижесінде неміс тіліне жасалған аударма жоғары сапа көрсетті. Алайда, ағылшын–қазақ тілдер жұбы үшін аударма нәтижесі төмендеу болды.

Осыған байланысты [19] жұмысында ашық нейрондық машиналық аударма жүйесінде (OpenNMT) қазақ тіліне бейімделген арнайы параметрлермен трансформер моделі құрылды. Бұл модель негізінде қазақ тіліндегі аударманың семантикалық сапасы жақсарып, аударма нәтижесі жоғары көрсеткіш көрсетті.

Материалдар мен әдістер

Трансформер модельді оқыту үшін ағылшын–қазақ тіл жұптарына арналған, құрылымы жағынан әр түрлі мәтіндерден тұратын параллель корпус құрылды. Бұл корпус ғылыми стильде жазылған 109 772 параллель сөйлемді қамтиды және ол akorda.kz, mfa.gov.kz, economy.gov.kz, strategiya2050.kz ресми сайттарынан жинақталды. Сонымен қатар, көркем әдеби стильге жататын әр түрлі сөйлемдер мен сөз тіркестерінен тұратын қосымша 20 000 бірлік мәтін де енгізілді.

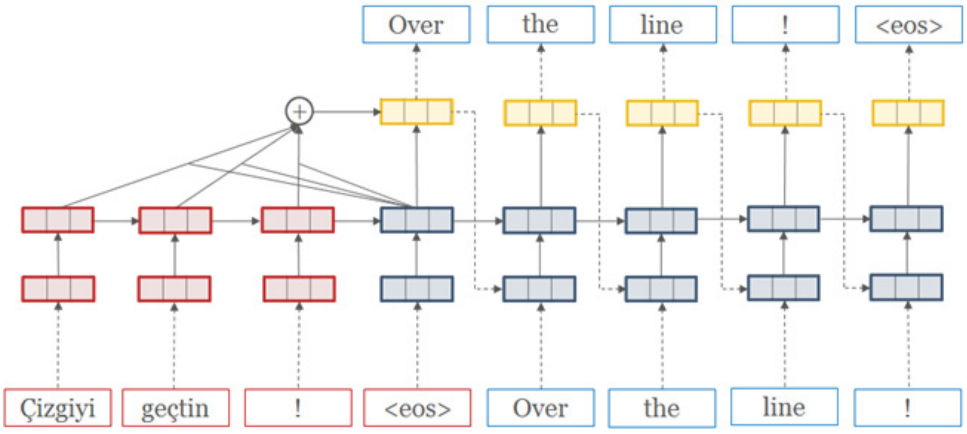
Кесте 3 – Корпустағы сөйлемдер саны

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

Жинақталған корпустағы сөйлемдер құрылымы жағынан құрмалас және жай сөйлемдер болатындай арнайы таңдалды. Сонымен қатар корпус жер, су, адам атауларын қамтиды.

Көркем әдеби стильдегі сөйлемдер арасында синтаксистік құрылымы сақталмаған диалог түрінде келген жай сөйлемдерді де кездестіруге болады.

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

1-суретте қызыл түсті тіктөртбұрышта аударылатын бастапқы сөйлемдегі сөздер енгізіледі. Сары түстегі тіктөртбұрышта жасырын қабатта сөздер аударылып, жіберіледі. Рекурентті нейрон желісі арқылы (rnn), transformer нейрон желісі негізінде сары түстен көк тіктөртбұрышқа жақын аударма сөздер жіберіледі.

Ашық нейронды машиналық аудармада (OpenNMT) қазақ тіліне аудару алгортмі:

1. Корпус дайындалды: src-train.txt атты файлға ағылшын тіліндегі сөйлемдер; tgt-train.txt атты файлға қазақ тіліндегі дұрыс ресми сайттан алынған аударма сөйлемдер жазылады;
2. Трансформер моделі құрылып, коды en-kk.yaml файл ішіне жазылды және оқытылуға арнайы парамерлер берілді:

```
#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]

# Optimization
model_dtype: "fp32"
optim: "adam"
learning_rate: 2
decay_method: "noam"
adam_beta2: 0.998
max_grad_norm: 0
label_smoothing: 0.1
param_init: 0
param_init_glorot: true
normalization: "tokens"

#Batch size
batch_size: 2048
batch_type: tokens
normalization: tokens
```

```
#Batch size
batch_size: 2048
batch_type: tokens
normalization: tokens

early_stopping: 4
early_stopping_criteria: accuracy
early_stopping_criteria: ppl

data:
    corpus_1:
        path_src: kk_en_corpora//src_train_bpe_shuffled
        path_tgt: kk_en_corpora//tgt_train_bpe_shuffled
    valid:
        path_src: kk_en_corpora//src_valid_bpe_shuffled
        path_tgt: kk_en_corpora//tgt_valid_bpe_shuffled

# Train on a single GPU
world_size: 2
gpu_ranks: [0,1]

# Where to save the checkpoints
save_model: kk_en_corpora/transformer_bpe_model/model
save_checkpoint_steps: 10000
train_steps: 100000
valid_steps: 5000
```

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.

onmt_translate -model model_step_60000.pt -src src-test.txt -output pred_1000.txt -gpu 0 -verbose. Шыққан нәтиже келесідей болды:

```
Мемлекет басшысы <unk> ААҚ басқарма төрағасы Алексей <unk> <unk>
<unk> бәсекелестікті дамыту үшін бәсекелестікті дамыту мәселелері жөніндегі Орталық Азия
мемлекеттері басшыларының бейресми саммиті болып өтті.
<unk> конституциялық заңдарға өзгертулер енгізу қаралды
<unk> соңына дейін <unk> тобына төрағалықты қабылдады
Оңтүстік Қазақстан сегіз айда 3 мың тоннаға жуық ет <unk>
Қазақстан Президенті <unk> компаниялар тобының жетекшісі <unk> <unk> кездесті
```

Аударма сапасын жақсарту үшін қазақ тілінде аударма табылмаған <unk> сөздердің жақын аудармасы табу үшін -replace_unk командасы қосып қайта оқытылды: onmt_translate -replace_unk -model model_step_60000.pt -src src-test.txt -output pred_1000.txt -gpu 0 -verbose

Бұл аударма командасынан кейін <unk> сөздердің аудармасы табылғанын және аударма көрсеткіші жақсарғанын көрдік. Олардың нәтижесі келесі бөлімде көрсетілген.

Трансформер моделінің математикалық сипаттамасы.

y вектор аударманы алу алгоритмі:

a) Аударылатын сөздер x1, x2, x3, …..xn векторлы токендермен түрленеді. Бұл токендер жасырын қабатта Wk, Wq және Wv матрицалар арқылы төмендегі (1) формуласымен ki qi және vi векторлары xi токендерінде генерацияланады.

$$k_i \leftarrow x_i W_k \in R^{1xd_k}\ q_i \leftarrow x_i W_q \in R^{1xd_k}\ v_i \leftarrow x_i W_v \in R^{1xd_k}$$

for i=1,…..,n.

б) алынған барлық нәтижелер масштабтан шықпауы үшін келесі формула бойынша түрленеді

$$\sqrt{d_k}^{-1}$$

c) softmax белсендіру функциясы есептеледі:

$$a_{i,j} \leftarrow \frac{q_i k_j^T}{\sqrt{d_k}}\ for\ j = 1, \ldots\ldots, i$$

$$a_{i,j} \leftarrow softmax\ (a_{i,j}) = \frac{\exp(a_{i,j})}{\sum_{j=1}^{i} \exp(a_{i,j})}\ for\ j = 1, \ldots\ldots, i$$

d) Шығу y векторы келесі формула бойынша алынады:

$$y_i \leftarrow \sum_{j=1}^{i} a_{i,j} v_j\ for\ j = 1, \ldots, i.$$

Нәтижесінде y_i аударма векторы алынады.

Нәтижелер мен талқылау

Құрылған модель 178000 параллель сөйлемдерден тұратын файл 6 сағат уақытта оқытылды. Бұл модельді тестілеу үшін 20000 тұратын ғылыми, көркем әдебиет стилінде жазылған, құрылымы жағынан әртүрлі сөйлемдер ағылшын тілінен қазақ тіліне аударылды. Аударма

нәтижесі BLEU (Bingual Evaluation Understudy) метрикасымен есептелінді. BLEU (Bingual Evaluation Understudy) бір тілден аударылған мәтіннің сапасын бағалауға арналған өлшеу метрикасы. Бұл көрсеткіш 0 ден 1 (0% ден 100%) аралығын қамтиды [20].

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

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

$$\text{Brevity\_Penalty} = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$

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 оқытылатын трансформер моделінде replace_unk: true командасын қосып,модель қайта оқытылды. Аударма нәтижесі 5-кесте бойынша алынды.

Кесте 5 – Replace_unk командасымен трансформер моделінде алынған нәтижелер

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

Кестеде көрсетілген нәтиже replace_unk командасымен оқытылғанда, аударма нәтижесі 0.1-ге жоғарлады. Өйткені, replace_unk командасы ашық нейронды машиналық аудармада аударылмай қалған <unk> сөздер орнына аудармасы және мағынасы жағынан жақын сөздер қойылады.

Қорытынды

Мақалада ашық нейрондық машиналық аударма жүйесінде (OpenNMT) трансформер модель негізінде қазақ тіліндегі аударма сапасын арттыратын арнайы параметрлер таңдалды. Бұл модельді оқыту аз уақыт алды. Оқыту үшін әр түрлі ресми сайттар мен көркем әдеби оқулықтардан алынған, құрылымы жағынан әртүрлі (жай және құрмалас) 180 000-ға жуық параллель сөйлемнен тұратын корпус жинақталды. Модельді тестілеу үшін 20 000 ағылшын тіліндегі сөйлем қазақ тіліне аударылды. Аударма нәтижесі BLEU метрикасы бойынша ағылшын-қазақ және қазақ-ағылшын тіл жұптары үшін 45% көрсеткіш көрсетті. Кейін аударылмаған сөздерді replace_unk командасы арқылы қайта өңдеу нәтижесінде аударма сапасы 55%-ға жетті. Аударма нәтижелері бойынша алынған көрсеткіштер Google және Yandex машиналық аударма жүйелерінен кем түспейтінін көрсетті. Жалқы есімдердің (адам, жер, су атаулары) аудармасы табылды. Сонымен қатар, аударманың морфологиялық құрылымы мен сөйлемдердің семантикалық мәні жақсы сақталғаны байқалды. Аударма сапасын одан әрі

жақсарту үшін корпус көлемін ұлғайту қажеттігі эксперимент жүзінде дәлелденді. Сонымен қатар, OpenNMT платформасында құрылған трансформер модель арнайы параметрлер арқылы қазақ тілінің аударма сапасын арттыра алды. Бұл модельді түркі тілдес басқа тілдерге де бейімдеп қолдануға болады, себебі түркі тілдер тобына ортақ морфологиялық және синтаксистік ерекшеліктер бар.

Қаржыландыру туралы ақпарат

Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігінің қолдауымен BR24993001 жобасымен қаржыландырылды.

ӘДЕБИЕТТЕР

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

2 Zhumanov Z.M., Tukeyev U.A. 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. – 2009. – Vol. 1. – pp. 356–363.

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

4 Tukeyev U., Rakhimova D. 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, 2012, pp. 1080–1085.

5 Abeustanova A., Tukeyev U. Automatic Post-editing of Kazakh Sentences Machine Translated from English // Advanced Topics in Intelligent Information and Database Systems: ACIIDS 2017. – Springer, 2017. – Vol. 710. – P. 283–295.

6 Schuster Sebastian, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval // Proceedings of the International Conference on Vision and Language (VL). – 2015/ – P. 70–80.

7 Xu Kelvin, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv, 2015, arXiv:1502.03044.

8 Shormakova A., Zhumanov Zh., Rakhimova D. Post-editing of Words in Kazakh Sentences for Information Retrieval // Journal of Theoretical and Applied Information Technology. – 2019. – Vol. 97. – No. 6. – P. 1896–1908.

9 Turganbayeva A., Tukeyev U. The Solution of the Problem of Unknown Words Under Neural Machine Translation of the Kazakh Language // Journal of Information and Telecommunication. – 2020. – P. 214–225.

10 Tukeyev U., Karibayeva A., Zhumanov Z. Morphological Segmentation Method for Turkic Language Neural Machine Translation // Cogent Engineering. – 2020. –Vol. 7. – No. 1. – P. 1–16. https://doi.org/10.1080/23311916.2020.1780271.

11 Koehn Philipp, Rebecca Knowles. Six Challenges for Neural Machine Translation // Proceedings of the First Workshop on Neural Machine Translation. – 2017. – P. 28–39.

12 Koehn Philipp. Statistical Machine Translation. Draft of Chapter 13. Neural Machine Translation. arXiv, 2017, arXiv:1709.07809v1[cs.CL], 117.

13 Papineni Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). – Philadelphia, July 2002. – P. 311–318.

14 Alvarez-Melis D., Jaakkola T.S. 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, 2017, pp. 412–421.

15 Zhou Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 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). – 2018. – Vol. 1. – P. 654–663.

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

17 Hao L., Gao W., Fang J. High-Performance English-Chinese Machine Translation Based on GPU-Enabled Deep Neural Networks with Domain Corpus // Applied Sciences. – 2021. – Vol. 11. – No. 22. – P. 10915. https://doi.org/10.3390/app112210915.

18 Quadri, Mohatesham Pasha, Pradeep Kumar. Corpus-Based Machine Translation for English to Low-Resource Language Using OpenNMT // Innovative Computing and Communications. – 2024. – P. 199–217.

19 Senellart Jean, Dakun Zhang, Bo Wang, Guillaume Klein, Jean-Pierre Ramatchandirin, Josep Crego, and Alexander Rush. 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, 2018, pp. 122–128. https://doi.org/10.18653/v1/W18-2715

20 Klein Guillaume, Francois Hernandez, Vincent Nguyen, and Jean Senellart. "The OpenNMT Neural Machine Translation Toolkit: 2020 Edition. Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, vol. 1, MT Research Track, October 6–9, 2020, pp. 102–109.

REFERENCES

1 Rakhimova D. R., and A. Zh. Zhunusova. Post-editing for the Kazakh Language Using OpenNMT. Journal of Mathematics, Mechanics and Computer Science, 113 (1), 118–122 (2022). https://doi.org/10.26577/JMMCS.2022.v113.i1.12.

2 Zhumanov Z.M., and U. A. Tukeyev. 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, 1, 356–363 (2009).

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

4 Tukeyev U., and D. Rakhimova. 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, 2012), pp. 1080–1085.

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

6 Schuster Sebastian, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. "Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. Proceedings of the International Conference on Vision and Language (VL), 70–80 (2015).

7 Xu, Kelvin, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv, 2015, arXiv:1502.03044.

8 Shormakova A., Zhumanov Zh., and D. Rakhimova. Post-editing of Words in Kazakh Sentences for Information Retrieval. Journal of Theoretical and Applied Information Technology, 97 (6), 1896–1908 (2019).

9 Turganbayeva A., and U. Tukeyev. The Solution of the Problem of Unknown Words Under Neural Machine Translation of the Kazakh Language. Journal of Information and Telecommunication, 214–225 (2020).

10 Tukeyev U., Karibayeva A., and Z. Zhumanov. Morphological Segmentation Method for Turkic Language Neural Machine Translation. Cogent Engineering, 7 (1), 1–16 (2020). https://doi.org/10.1080/23311916.2020.1780271.

11 Koehn Philipp and Rebecca Knowles. Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, 28–39 (2017).

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

13 Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, 311–318.

14 Alvarez-Melis D., and T.S. Jaakkola. 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, 2017, pp. 412–421.

15 Zhou Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 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, 2018, pp. 654–663.

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

17 Hao L., Gao W., and J. Fang. High-Performance English-Chinese Machine Translation Based on GPU-Enabled Deep Neural Networks with Domain Corpus. Applied Sciences, 11 (22), 10915 (2021). https://doi.org/10.3390/app112210915.

18 Quadri Mohatesham Pasha, and Pradeep Kumar. Corpus-Based Machine Translation for English to Low-Resource Language Using OpenNMT. Innovative Computing and Communications, 199–217 (2024).

19 Senellart Jean, Dakun Zhang, Bo Wang, Guillaume Klein, Jean-Pierre Ramatchandirin, Josep Crego, and Alexander Rush. 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, 2018, pp. 122–128. https://doi.org/10.18653/v1/W18-2715

20 Klein Guillaume, Francois Hernandez, Vincent Nguyen, and Jean Senellart. The OpenNMT Neural Machine Translation Toolkit: 2020 Edition. Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, vol. 1, MT Research Track, October 6–9, 2020, pp. 102–109.

[1]Рахимова Д.,
PhD, ORCID ID: 0000-0003-1427-198X,
e-mail: di.diva@mail.ru
[1,2]*Жігер А.,
магистр, ORCID ID: 0000-0002-3641-8260,
*e-mail: alia_94-22@mail.ru
[3]Малых В.,
PhD, ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Карюкин В.,
PhD, ORCID ID: 0000-0002-8768-0349,
e-mail: vladislav.karyukin@gmail.com
[2]Бекарыстанқызы А.,
PhD, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com

[1]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан
[2]Университет Нархоз, г. Алматы, Казахстан
[3]Санкт-Петербургский государственный университет информационных технологий, механики и оптики, г. Санкт-Петербург, Россия

НЕЙРОМАШИННЫЙ ПЕРЕВОД ДЛЯ АНГЛИЙСКО-КАЗАХСКОЙ ЯЗЫКОВОЙ ПАРЫ

Аннотация

В настоящее время в качестве одной из отраслей можно назвать машинный перевод. Люди из разных стран используют машинный перевод, чтобы понимать друг друга, поэтому потребность в нем возрастает

с каждым годом. На данный момент Гугл и Яндекс машинные переводы входят в число лучших машинных переводов. С каждым годом машинные переводы Яндекса и Гугла поднимают качество перевода на более высокий уровень. Однако по результатам проведенного эксперимента при переводе с английского или русского на казахский, а также на тюркские языки качество перевода ниже по сравнению с другими языками. В доказательство тому в сентябре 2024 года в работе был показан перевод, полученный в результате этих двух машинных переводов. Потому что модель, осуществляющая машинный перевод, напрямую связана со структурой языков. В связи с этим с 2000-х годов ученые государства Казахстан в целях совершенствования переводов на казахский язык стали сравнивать и интенсивно изучать модели, хорошо переводящиеся на другие языки, также начали создавать специальные модели для казахского языка и публиковать научные работы их результатов. Цель работы – повышение качества перевода с английского на казахский язык. С этой целью была создана модель-трансформер, подходящая для казахского и тюркского языков, для обучения переводу в Open Neural Machine Translation (OpenNMT). Созданная модель была обучена путем чтения и обучения англо-казахского параллельного корпуса объемом 180 000 предложений. Для оценки результата перевода на казахский язык был переведен файл, состоящий из 20 000 предложений различной структуры (простые, сложные). Результат был измерен по метрике индикатора перевода (Bleu) и показал хороший уровень. Проведенный в работе эксперимент показал, что для повышения полученного результата на еще более высокий уровень необходимо в ходе обучения модели увеличить количество параллельных предложений, создаваемых из пары английский-казахский языки.

Ключевые слова: нейромашинный перевод, метрика перевода, параллельный корпус, открытый нейромашинный перевод, модель-трансформер.

[1]Rakhimova D.,
PhD, ORCID ID: 0000-0003-1427-198X,
e-mail: di.diva@mail.ru
[1,2]*Zhiger A.,
master's degree, ORCID ID: 0000-0002-3641-8260,
*e-mail: alia_94-22@mail.ru
[3]Malykh V.,
PhD, ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Karyukin V.,
PhD, ORCID ID: 0000-0002-8768-0349,
e-mail: vladislav.karyukin@gmail.com
[2]Bekarystankyzy A.,
PhD, ORCID ID: 0000-0003-3984-2718,
e-mail: akbayan.b@gmail.com

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Narxoz University, Almaty, Kazakhstan
[3]Saint Petersburg State University of Information Technologies,
Mechanics and Optics, Saint Petersburg, Russia

NEURAL MACHINE TRANSLATION FOR ENGLISH-KAZAKH LANGUAGE PAIR

Abstract

Currently, information technology is rapidly developing and one of its branches can be called machine translation. The use of machine translation in the process of understanding each other by people from different countries is increasing every year. At the moment, Google and Yandex machine translations are among the best machine translations. The quality of machine translation from Yandex and Google is improving every year. However, according to the results of the experiment, when translating from English or Russian into Kazakh and Turkic languages, the quality of the translation decreases. This was shown by the translation result obtained from these two machine translations in March 2024. After all, translation has also shown that it is directly related to the

structure of language. Since 2000, scientists from the state of Kazakhstan have been actively studying translations into the Kazakh language. The goal of the work is to improve the quality of translation from English into Kazakh. For this purpose, a transforming model was created for the Kazakh and Turkic languages for learning translation in neural machine translation OpenNMT(). The created model studied and learned an English-Kazakh parallel corpus of 180,000 words. Later, the document with a structure of 20,000 different English sentences was translated into Kazakh. The result is measured using the Blue() metric. The translation result showed a high level. It is shown that in order to improve the results of the experiment carried out in the work during model training, it is necessary to increase the number of parallel corpora created from the English-Kazakh language pair.

Keywords: neural machine translation, BLEU metric, parallel corpus, OpenNMT, transformer model.

Мақаланың редакцияға түскен күні: 15.01.2025