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

ӘОЖ 004.934.1
ҒТАХА 20.19.00
https://doi.org/10.53364/24138614_2025_39_4_14

Д. Рахимова[1], **А. Ж. Жігер**[1,2*],**В. Малых**[3]
[1]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
[2]Нархоз университеті, Алматы, Қазақстан
[3]Санкт-Петербург мемлекеттік ақпараттық технологиялар, механика және оптика университеті, Санкт-Петербург, Ресей

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

ҚҰҚЫҚТЫҚ МӘТІНДЕРДІ ҚАЗАҚ, ОРЫС ТІЛДЕРІНЕ НЕЙРОНДЫ МАШИНАЛЫҚ АУДАРУДЫҢ ӘДІСТЕРІ МЕН САПАЛЫҚ ТАЛДАУЫ

Аңдатпа. *Қазіргі таңда Қазақстан Республикасында құқық саласындағы мәтіндерді қазақ тілінен орыс және ағылшын тілдеріне, сондай-ақ осы тілдерден қазақ тіліне сапалы аудару – өзекті мәселелердің бірі болып отыр. Бұл ғылыми жұмыста кеңінен қолданылатын Яндекс пен Гугл секілді машиналық аударма жүйелері арқылы арнайы құқықтық дереккөздерден алынған мәтіндердің қазақ-орыс тіл жұбы негізінде аударылып, аударма сапасындағы қателіктерге талдау жүргізілді.*

Зерттеудің негізгі мақсаты – құқық саласына тән сөйлемдер мен терминдерді дәл әрі мағыналық тұрғыдан дұрыс аудару жолдарын қарастыру. Осы мақсатта құқықтық құжаттар, сот шешімдері мен ресми сайттардан арнайы бағдарлама көмегімен 96 555 сөйлем мен сөз тіркестерінен тұратын корпус жинақталды.

Аталған корпус MarianMT нейронды машиналық аударма жүйесінде оқытылып, қазақ-орыс тіл жұбында аударма сапасы тәжірибе арқылы тексерілді. MarianMT моделінің нәтижелерін жақсарту үшін қосымша KazRobert трансформерлік моделі қолданылды. Жұмыста KazRobert моделінің архитектурасы мен оның математикалық негізі жан-жақты сипатталады.

Аударма сапасы BLEU, TER және METEOR секілді халықаралық деңгейде мойындалған өлшемдер арқылы бағаланды. Жұмыста екі түрлі нәтиже салыстырмалы түрде көрсетілді: тек MarianMT жүйесінде алынған нәтиже және KazRobert моделінде оқытылған MarianMT жүйесінің нәтижесі. Талдау қорытындысы бойынша, ұсынылған әдіс OpenNMT негізіндегі бұрынғы аударма моделіне қарағанда сапалы нәтижелер көрсетті.

Жүргізілген тәжірибелер корпус көлемі мен терминдердің саны артқан сайын аударма сапасының да жақсара түсетінін көрсетті. Сонымен қатар, зерттеу нәтижелері бұл әдісті құрылымдық жағынан қазақ тіліне жақын түркі тілдеріне де тиімді түрде бейімдеуге болатынын дәлелдеді.

Түйін сөздер: *нейронды машиналық аударма, MarianMT машиналық аударма, KazRobert моделі, трансформер моделі, құқық саласындағы корпус, BLEU аударма көрсеткіші, TER аударма көрсеткіші, METEOR аударма көрсеткіші.*

Кіріспе.

Қазақстан Республикасында әртүрлі ұлт өкілдерінің өмір сүруіне байланысты заң аясындағы іс-қағаздар мемлекеттік тіл – қазақ тілімен қатар, орыс және ағылшын тілдерінде де жүргізіледі. Қазіргі таңда, дамыған машиналық аударма жүйелері, мысалы, Yandex және Google, сондай-ақ жасанды интеллект технологияларына қарамастан, құқық саласындағы мәтіндер орыс және ағылшын тілдерінен қазақ тіліне және керісінше қазақ тілінен орыс және ағылшын тілдеріне аударылғанда қателіктердің бар екендігі анықталды. Бұл қателіктердің негізгі себептері қазақ тілінің құрылымдық ерекшеліктерімен тікелей байланысты.

Атап айтқанда, қазақ тілінің морфологиялық құрылымы (жалғаулар мен жұрнақтар түрлерінің көптігі және түбірге жалғану ережелерінің түрлілігіне байланысты), мысалы, егер түбір қатаң дыбысқа аяқталса, жалғанатын қосымшаның да қатаң дыбыстан басталуы міндетті. Сонымен қатар, түбірдің жуан не жіңішке болуына байланысты оның жалғаулары да сәйкес дыбыстармен қолданылуы қажет. Синтаксистік құрылым тұрғысынан алғанда, қазақ тілінде көп сөйлемдер күрделі және құрмалас түрінде болады, ал бұл орыс және ағылшын тілдерінде кеңінен қолданылатын қарапайым құрылымдарға қарағанда өзгеше. Құрмалас сөйлемдерде жай сөйлемдердің арасындағы байланыстың түрлі ережелері бар, сондықтан олардың дұрыс аударылуы қиындық туғызады. Семантикалық аспект бойынша да кейбір қарапайым сөйлемдер орыс немесе ағылшын тілінен қазақ тіліне аударылғанда, мағыналық тұрғыдан толық үйлеспейтін жағдайлар орын алады. Бұл әсіресе сөйлемдерде қысқартулар, тұрақты тіркестер, болымсыз сөйлемдер мен құқықтық публицистикалық стильдегі мәтіндерде жиі кездеседі.

Қазақ тілінің құрылымдық ерекшеліктері мен осы ерекшеліктерге байланысты туындайтын аударма қателіктері жоғарыда көрсетілген. Сонымен қатар, машиналық аударма модельдері қазақ тілінің ерекшеліктеріне бейімделмеген жағдайда, аударма сапасының төмендігі байқалады. Нейронды машиналық аударманың осы мәселелерді шешудегі тиімділігіне қатысты, көп корпус жинақтап, қазақ-орыс немесе қазақ-ағылшын тілдеріндегі мәтіндер мен сөз тіркестерінің оқылмауы процесі де қателіктерге себеп болады.

Төменде құқық саласындағы мәтіндердің Яндекс және Google машиналық аударма жүйелері арқылы аударылған нәтижелері мен олардың дұрыс аудармалары салыстырылып, аударма қателіктері сипатталады. Бұл жұмыста, құқық саласына арналған ағылшын-қазақ және орыс-қазақ тілдеріндегі аудармаларға негізделген деректі сөз тіркестері мен сөйлемдер арнайы құқықтық ақпарат көздерінен алынып, құқықтық терминдерді анықтап, оларды бөлектеп шығаратын бағдарлама құрылды. Осы бағдарлама арқылы 90 мыңнан аса корпус жинақталып, MarianMT ашық нейронды кодымен арнайы қазақ тіліне бейімделген трансформер моделі арқылы аударылды. Аудармадан алынған қателіктер KazRobert моделінің көмегімен постредакторленіп, аударма сапасы жақсартылады.

Кесте 1 – Яндекс және Google машиналық аударма жүйелерінен алынған құқық саласындағы мәтіндердің аударма сапасын көрсету

Негізгі мәтін	Яндекс машиналық аударма	Гугл машиналық аударма	Негізгі аударма	Аударма қателігі
Стороны заключили договор аренды, регулируемый положениями	Тараптар Ресей Федерациясының Азаматтық кодексінің ережелерімен реттелетін	Тараптар Ресей Федерациясының Азаматтық кодексінің ережелерімен реттелетін	Тараптар Ресей Федерациясының Азаматтық кодексінің	Аударма жағынан қателік жоқ. Публицисти-калық стиль жағынан ғана

Гражданского кодекса Российской Федерации.	жалдау шартын жасады.	жалдау шартын жасады.	нормаларына сәйкес реттелетін жалдау шартын жасасты.	жақсарды.
В случае нарушения условий настоящего договора одной из сторон, другая сторона вправе в одностороннем порядке отказаться от исполнения обязательств, предварительно уведомив об этом контрагента не менее чем за 10 (десять) календарных дней до предполагаемой даты расторжения договора.	Тараптардың бірі осы Шарттың талаптарын бұзған жағдайда, екінші Тарап бұл туралы контрагентті Шартты бұзудың болжамды күніне дейін кемінде күнтізбелік 10 (он) күн бұрын алдын ала хабардар ете отырып, міндеттемелерді орындаудан біржақты тәртіппен бас тартуға құқылы.	Тараптардың бірі осы шарттың талаптарын бұзған жағдайда, екінші тарап бұл туралы контрагентті шартты бұзудың болжамды мерзіміне дейін кемінде 10 (он) күнтізбелік күн бұрын хабардар ете отырып, міндеттемелерді орындаудан біржақты тәртіппен бас тартуға құқылы.	күніне дейін кемінде 10 (он) күнтізбелік күн бұрын ескерте	«Болжамды күніне дейін» → «Күтілетін күніне дейін» "Болжамды" сөзі жалпы мағынада қолданылады және кейде нақты емес мағына береді. Ал "күтілетін" сөзі заң тілінде жиі қолданылады және неғұрлым нақты, ресми стильге сәйкес келеді. екеуі де рұқсат етіледі.
В случае нарушения одной из сторон существенных условий настоящего договора, повлёкшего невозможность дальнейшего исполнения обязательств другой стороной, последняя вправе в	Тараптардың бірі екінші Тараптың міндеттемелерді одан әрі орындай алмауына әкеп соққан осы шарттың елеулі талаптарын бұзған жағдайда, соңғысы жазбаша хабарлама жібере	Тараптардың бірі осы шарттың маңызды талаптарын екінші тараптың міндеттемелерді одан әрі орындау мүмкін еместігіне әкеп соққан бұзған жағдайда, соңғысы жазбаша хабарлама жіберу арқылы біржақты	Егер Тараптардың бірі осы Шарттың елеулі талаптарын бұзып, нәтижесінде екінші Тараптың шарттық міндеттемелерін әрі қарай орындауы мүмкін болмаса, екінші Тарап	Соңғысы" — дұрыс қолданылған, бірақ күрделі заң мәтіндерінде түсініксіздік болмас үшін "екінші Тарап" деп нақтылап жазған дұрыс. "Бастамашылық жасауға құқылы" — мағынасы дұрыс, бірақ сәл артық айтылған. Жай ғана

одностороннем внесудебном порядке инициировать расторжение договора с направлением письменного уведомления, при этом нарушившая сторона обязуется компенсировать все документально подтверждённые убытки, включая прямой ущерб, упущенную выгоду, а также иные расходы, понесённые в связи с невыполнением или ненадлежащим выполнением договорных обязательств, в соответствии с положениями действующего гражданского законодательства.	отырып, шартты біржақты соттан тыс тәртіппен бұзуға бастамашылық жасауға құқылы, бұл ретте бұзған Тарап тікелей залалды, жоғалған пайданы, сондай-ақ басқа да шығыстарды қоса алғанда, құжатталған барлық залалдарды өтеуге міндеттенеді. қолданыстағы азаматтық заңнаманың ережелеріне сәйкес шарттық міндеттемелерді орындамау немесе тиісінше орындамау.	тәртіпте шартты соттан тыс бұзуға бастамашылық жасауға құқылы, бұл ретте бұзушы тарап барлық құжатталған залалдарды, оның ішінде тікелей байланысты жоғалтқан пайданы, сондай-ақ жоғалған залалды өтеуге міндеттенеді. қолданыстағы азаматтық заңнаманың ережелеріне сәйкес шарттық міндеттемелерді орындау немесе тиісінше орындамау.	жазбаша түрде хабарлай отырып, шартты біржақты соттан тыс тәртіппен бұзуға құқылы. Бұл ретте, шартты бұзған Тарап қолданыстағы азаматтық заңнамада көзделген тәртіпке сәйкес, тікелей залалды, жоғалған пайданы, сондай-ақ өз міндеттемелерін орындамауына немесе тиісінше орындамауына байланысты туындаған өзге де құжатпен расталған шығыстарды өтеуге міндетті.	"шартты бұзуға құқылы" деп жазған жеткілікті, себебі "бастамашылық" деген сөз бұл жерде артық. Соңғы сөйлем "қолданыстағы азаматтық заңнаманың ережелеріне сәйкес..." деп басталады, бірақ негізгі сөйлеммен грамматикалық түрде дұрыс байланыспаған. Оны толық сөйлем ретінде біріктірген жөн.

Бұл қателіктерді салыстыра отырып, келесі қорытынды жасауға болады: Яндекс және Google машиналық аударма жүйелерінде орын алған аударма қателіктерінің арасында публицистикалық стильдің дұрыс қолданылмауы байқалады. Аударма мәтіндерінде жиі кездесетін мәселелердің бірі — мәтіндер мен сөйлемдердің ауызекі сөйлеу стилінде берілуі, бұл өз кезегінде ғылыми немесе құқықтық мәтіндерге тән ресми стильге сәйкес келмейді. Сонымен қатар, эксперимент барысында морфологиялық құрылымы жағынан сөздер мен сөйлемдер дұрыс байланыспаған жағдайлар да анықталды.

Бұл қателіктерді азайтып, аударма сапасын жақсарту үшін аударма мәтінін постредакторлеу қажет екендігі анықталды. Яғни, корпустағы сөз тіркестері мен сөйлемдердің сапасын арттыру, құқық саласындағы терминдердің санын көбейту арқылы аударма нәтижелерін жетілдіруге болады. Бұл мақалада келесі тараулар қарастырылады:

2-тарау: Материалдар мен зерттеу әдістері.
3-тарау: Нәтижелер және оларды талқылау.
4-тарау: Қорытынды.

Материалдар мен зерттеу әдістері.

Қазіргі таңда Қазақстан мемлекетінде құқық саласындағы орыс, ағылшын тілдерінен қазақ тіліне және қазақ тілінен орыс пен ағылшын тілдеріне дұрыс аударма жасау өзекті мәселелердің бірі болып табылады. Бұл мәселе, әсіресе, көптілді қоғамда құқықтық мәтіндердің түсініктілігі мен дәлдігін қамтамасыз ету үшін маңызды. Соңғы 5 жыл ішінде көптеген ғалымдар құқық саласындағы мәтіндерді аударуда нейронды машиналық аударма технологияларын пайдаланған [1-4]. Мұндай аударма тәсілдері, әсіресе, тілдер арасындағы морфологиялық ерекшеліктер мен контекстік факторларды ескеру арқылы жоғары сапалы аудармалар алуға мүмкіндік береді.

Қазақстандық зерттеушілер ағылшын-қазақ және қазақ-ағылшын тіл жұптарында аударма сапасын жақсарту үшін тілдің морфологиялық құрылымын зерттеп, нейронды машиналық аудармаларға негізделген әдістерді қолдануда [5-7]. Бұл бағытта OpenNMT жүйесі арқылы тіл жұптары үшін жақсы нәтижелер алынғанын атап өтуге болады [8-10].

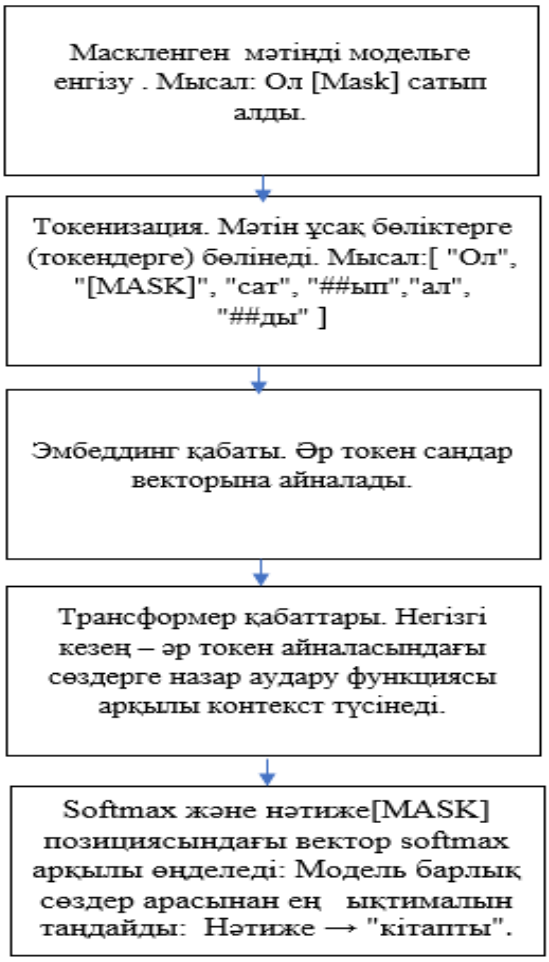
Аталған жұмыс аясында құқық саласындағы мәтіндерді аудару үшін арнайы корпус жасалды. Орыс-қазақ тіл жұптарынан құралған, 96 555 сөйлем мен сөз тіркестерінен тұратын корпус келесі Кесте 2 -де көрсетілген дереккөздерден алынды.

Кесте 2- Корпустағы сөйлемдер саны

Алынған дерек көздер	Сөйлемдер саны
Заң актілері	30561
Сот шешімдері	10005
Ресми веб-сайттар (үкіметтік, соттық, министрлік сайттары)	55989

Сонымен қатар корпустарды жинау үшін арнайы бағдарлама құрылды. Бұл бағдарлама pdf құжаттардан, құқық саласында сайттардан осы саладағы терминдер мен сөзтіркестері мен сөйлемдер алыну арқылы корпус жинақталды.

Осы жиналған корпустар MarianMT нейрондық машиналық аударма моделі арқылы оқытылды [11-13]. Аудармадан кейін қазақ тіліндегі аудармада қателер болды. Бұл аударма сапасын жақсартып, қателерден арылу үшін KazRobert моделінде оқытылды. KazRobert моделінің орындалу алгоритмі келесі қадамдардан тұрады:



Сурет 1 – KazRobert моделінің архитекурасын мысалмен көрсету

Жоғарыда көрсетілген KazRobert моделінің архитекурасын математикалық моделі бойынша түсіндіретін болсақ:

1. Енгізу қабаты. KazRoBERTa алдымен мәтінді токендерге бөледі және әр токенді вектор түріне түрлендіреді:

$$x_i = TokenEmb(\omega_i) + PosEmb(i) \quad (1)$$

мұнда:

$\omega_i -$ i – ші токен

$\mathrm{TokenEmb}(\omega_i) -$ токен эмбеддингі

$\mathrm{PosEmb(i)} -$ позициялық эмбеддинг (сөздердің ретін ескеру үшін)

2. Трансформер қабаттары (энкодер). Модель L қабаттан тұрады. Әр қабатта n-төрт негізгі бөлік бар [14-16]:

a) Көпқабат өзіндік назар (Multi-Head Self-Attention)

Әрбір қабатта сұраныс (Q), кілт (K) және мән (V) векторлары есептеледі:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \quad (2)$$

Назар функциясы:

$$\mathrm{Attention(Q,K,V)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V} \quad (3)$$

b) Көпқабатты назар функциясы:

$\mathrm{MultiHead(Q,K,V)} = \mathrm{Concat}(head_1, \ldots . head_n)W^Q$

Мұндағы n — назар қабаттар саны.

c) Толық байланысқан тор (Feed Forward Network, FFN)

Әр токен үшін екі қабатты нейрондық желі қолданылады:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \quad (4)$$

d) Қалдық байланыс + Нормализация (Residual + LayerNorm)

Әр бөлікке LayerNorm қолданылады:

$$x' = \mathrm{LayerNorm}(x + \mathrm{MultiHead}(x)) \quad (5)$$

$$x'' = \mathrm{LayerNorm}(x' + \mathrm{FFN}(x')) \quad (6)$$

3. Маска қойылған тілдік модельдеу (Masked Language Modeling)

Модель кездейсоқ таңдалған токендерді "маскалап", соларды болжауға үйренеді:

$$\sigma_{\mathrm{MLM}} = -\sum_{i\in\mu} \log P(\omega_i|x_i) \quad (7)$$

Мұндағы, μ- маскаланған токендердің орындары, ω_i- шынайы токен, x_i-маскаланған токендерді вектор түріне айналдыру.

Нәтижелер және оларды талқылау.

Алынған нәтижеге баға беру үшін, эксперимент ретінде құқық саласынан 1000 сөйлемнен тұратын корпус құрылған бағдарламада орыс-қазақ және қазақ-орыс тілдер жұптары үшін аудармалар алынды. Аударма көрсеткіші BLEU, METEOR, TER көрсеткіштерімен есептелді [17-19]. Көрсеткіш метрика түрлеріне жеке тоқталсақ:

BLEU (Bilingual Evaluation Understudy). Машиналық аударманың сапасын n-граммдар бойынша эталонмен салыстыра отырып бағалайды.

$$BLEU = BP * EXP(\sum_1^N \omega_n \log p_n) \quad (8)$$

Мұндағы, p_n-n-грамм дәлдігі (мысалы, unigram, bigram және т.б.); ω_n- n-граммдарға берілетін салмақ ($\omega_n = \frac{1}{N}$); BP — Brevity Penalty, қысқалық жазасы. BLEU жоғары болған сайын (максимум = 1.0), аударма сапасы жақсы деген мағынаны білдіреді.

METEOR (Metric for Evaluation of Translation with Explicit ORdering). Сөздердің дәлдігімен қоса, синонимдер, түбірлес сөздер (стемминг) және сөздердің реті ескереді.

$$\mathrm{METEOR} = (1 - \mathrm{Penalty}) \cdot F_{mean} \quad (9)$$

Мұндағы: $F_{mean} = \frac{10 \cdot P \cdot R}{9R+P}$ — Precision мен Recall-дың үйлесімді орташа мәні;

$\mathrm{Penalty} = 0.5 \cdot \left(\frac{\mathrm{chunks}}{matches}\right)^3$ — сөздердің ретінің бұзылғанына берілетін айып

P — Precision (гипотеза ішіндегі дұрыс сөздер үлесі)

R — Recall (эталондағы сөздермен салыстырғанда)

METEOR жоғары болған сайын (максимум = 1.0), аударма сапасы да жақсы болады.

TER (Translation Edit Rate). Гипотезаны эталонға айналдыру үшін қажет редакциялау қадамдарының санын өлшейді.

$$TER = \frac{\text{Редакция қадамдарының саны}}{\text{эталондағы сөздер саны}} \quad (10)$$

TER неғұрлым төмен болса, соғұрлым жақсы (минимум = 0). Бұл аудармаға аз түзету қажет дегенді білдіреді.

Осы жұмыста зерттелген құқық саласындағы мәтіндер орыс тілінен қазақ тіліне және қазақ тілінен орыс тілдерінде алынған аудармалардың нәтижесі төмендегі кестелерде көрсетілген. Кесте 3-те MarianMT нейронды машиналық аудармада алынған аударма нәтижесі көрсетілсе, Кесте 4- KazRobert моделінде MarianMT нейронды машиналық аудармада алынған аударманы оқытып, аударса сапасын жақсарғанын көруге болады.

Кесте 3- MarianMT нейронды машиналық аудармада алынған аударма нәтижесі

Тіл жұптары	Bleu	Ter	METEOR
Орыс-қазақ	0,59	0.3	0.65
Қазық-орыс	0,59	0.3	0.65

Кесте 4- Алынған аударманы KazRobert моделінде пост-редакторлеу кейінгі нәтиже

Тіл жұптары	Bleu	Ter	METEOR
Орыс-қазақ	0,7	0.15	0.8
Қазық-орыс	0,7	0.15	0.8

Бұл нәтижелер Opennmt ашық нейронды машиналық аудармадан алынған нәтижеден жоғары көрсеткіш көрсетті [20].

Қорытынды.

Бұл ғылыми жұмыста құқық саласына қатысты мәтіндерді қазақ және орыс тілдері арасында сапалы аудару мәселесі қарастырылып, машиналық аударма жүйелерін жетілдіру жолдары тәжірибе жүзінде сараланды. Қазіргі таңда құқықтық терминология мен сөйлем құрылымдарын дұрыс және дәл аудару — аударма сапасына қойылатын негізгі талаптардың бірі. Осы тұрғыдан алғанда, дәстүрлі машиналық аударма құралдарының (мысалы, Гугл, Яндекс) құқықтық мәтіндерді аудару кезінде жиі қателіктерге жол беретіні айқындалды.

Зерттеу барысында арнайы құқық саласына бағытталған 96 555 сөйлем мен сөз тіркестерінен тұратын корпус жинақталып, осы деректер MarianMT нейронды машиналық аударма жүйесінде оқытылды. Бұл модель орыс-қазақ тіл жұбына бейімделіп, әрі қарай аударма сапасын жақсарту мақсатында KazRobert трансформерлік моделімен біріктірілді. KazRobert моделінің архитектурасы мен математикалық негіздемесі жұмыста жан-жақты сипатталып, модельдің құқықтық мәтіндер контексінде тиімді жұмыс істейтіні дәлелденді.

Аударма сапасын бағалау BLEU, TER және METEOR секілді кеңінен танылған метрикалар арқылы жүргізілді. Жүргізілген эксперименттер корпус құрамындағы сөйлемдер мен терминдердің көлемі неғұрлым үлкен болған сайын, модельдің дәлдігі мен аударма сапасы да соғұрлым жоғарылайтынын көрсетті. Сонымен қатар, бұл модельдің жалпы архитектурасы мен әдіснамасы тілдік құрылымы ұқсас түркі тілдес тілдерге бейімдеуге мүмкіндік беретіндігі анықталды. Мұндай тәсіл болашақта түркі халықтарының құқық саласына арналған мультитілді аударма жүйелерін жасауға негіз бола алады.

Осылайша, зерттеу жұмысы құқықтық мәтіндерді аударуда MarianMT нейронды машиналық аударма мен Kazrobert моделінде интеграциялау арқылы сапаны арттырудың ғылыми тұрғыдан тиімді жолдарын ұсынды. Бұл бағытта жүргізілген зерттеу келешекте салалық бағыттағы аудармаларды автоматтандыруда, сондай-ақ мемлекеттік тілдің кәсіби қолданысын кеңейтуде маңызды үлес қоса алады.

Алғыс.

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

Әдебиеттер

1. Bajčić, M., & Golenko, D. (2023). *Large Language Models in Legal Translation: An Interdisciplinary Research Agenda.* Journal of Language and Law, 10(1). Retrieved from https://www.languageandlaw.eu/jll/article/view/172

2. Ding, L. (2024). *An Empirical Study on Legal Text Translation from the Perspective of Translation Quality Assessment: A Comparison between ChatGPT and Neural Machine Translation.* Theory and Practice in Language Studies, 14(2), 297–308. https://doi.org/10.17507/tpls.1402.14

3. Elnaggar, A., Schafer, U., & Gurevych, I. (2018). *Multi-task Deep Learning for Legal Document Translation, Summarization and Multi-label Classification.* arXiv, arXiv:1810.07513 [cs.CL]. Retrieved from https://arxiv.org/abs/1810.07513

4. Zhu, J. (2024). *Application of Machine Translation Technology in Legal Interpretation.* American Journal of Education and Information Technology, 8(1), 18–22. https://doi.org/10.11648/j.ajeit.20240801.13

5. 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.

6. 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.

7. 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.

8. Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation." *arXiv preprint arXiv:1701.02810.* Available at: https://arxiv.org/abs/1701.02810

9. Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2018). "OpenNMT: Neural Machine Translation Toolkit (2018 Edition)." *Proceedings of the 2018 Conference on Neural MachineTranslation.* https://doi.org/10.48550/arXiv.1805.11462

10. Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., Pedrycz, W. (2023). "A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks." *arXiv preprint arXiv:2306.07303*. https://doi.org/10.48550/arXiv.2306.07303

11. Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., & Aue, A. (2018). *Marian: Cost-effective High-Quality Neural Machine Translation in C++*. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (pp. 129–135). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2716

12. Kim, Y. J., Junczys-Dowmunt, M., Hassan, H., Aji, A. F., Heafield, K., Grundkiewicz, R., & Bogoychev, N. (2019). *From Research to Production and Back: Ludicrously Fast Neural Machine Translation.* In Proceedings of the Third Workshop on Neural Generation and Translation. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5632

13. Gowda, T., Grundkiewicz, R., Rippeth, E., Post, M., & Junczys-Dowmunt, M. (2024). *PyMarian: Fast Neural Machine Translation and Evaluation in Python.* arXiv. https://doi.org/10.48550/arxiv.2408.11853

14. (2024). "Survey of Transformers and Towards Ensemble Learning Using Transformers for Natural Language Processing." *Journal of Big Data*, vol. 11, no. 3, pp. 123–137. https://doi.org/10.1186/s40537-023-00842-0

15. Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., Shen, C. (2023). "A Survey on Efficient Training of Transformers." *arXiv preprint arXiv:2302.01107*. https://doi.org/10.48550/arXiv.2302.01107

16. Dong, Z., Tang, T., Li, L., Zhao, W. X. (2023). "A Survey on Long Text Modeling with Transformers." *arXiv preprint arXiv:2302.14502*. https://doi.org/10.48550/arXiv.2302.14502

17. Glushkova, T., C. Zerva, and A. F. T. Martins. (2023). "BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation." *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 47–56. https://aclanthology.org/2023.eamt-1.6/

18. ElNokrashy, M., and T. Kocmi. (2023). "eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings." *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)*, pp. 756–765.

19. Saadany, H., and C. Orasan. (2021). "BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text." *Proceedings of the Translation and Interpreting Technology Online Conference (TRITON 2021)*, pp. 50–59. https://aclanthology.org/2021.triton-1.6/

20. 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.

References

1. Bajčić, M., & Golenko, D. (2023). *Large Language Models in Legal Translation: An Interdisciplinary Research Agenda.* Journal of Language and Law, 10(1). Retrieved from https://www.languageandlaw.eu/jll/article/view/172

2. Ding, L. (2024). *An Empirical Study on Legal Text Translation from the Perspective of Translation Quality Assessment: A Comparison between ChatGPT and Neural Machine Translation.* Theory and Practice in Language Studies, 14(2), 297–308. https://doi.org/10.17507/tpls.1402.14

3. lnaggar, A., Schafer, U., & Gurevych, I. (2018). *Multi-task Deep Learning for Legal Document Translation, Summarization and Multi-label Classification.* arXiv, arXiv:1810.07513 [cs.CL]. Retrieved from https://arxiv.org/abs/1810.07513

4. Zhu, J. (2024). *Application of Machine Translation Technology in Legal Interpretation.* American Journal of Education and Information Technology, 8(1), 18–22. https://doi.org/10.11648/j.ajeit.20240801.13

5. 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.

6. 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.

7. 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.

8. Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation." *arXiv preprint arXiv:1701.02810.* Available at: https://arxiv.org/abs/1701.02810

9. Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2018). "OpenNMT: Neural Machine Translation Toolkit (2018 Edition)." *Proceedings of the 2018 Conference on Neural MachineTranslation*. https://doi.org/10.48550/arXiv.1805.11462

10. Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., Pedrycz, W. (2023). "A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks." *arXiv preprint arXiv:2306.07303*. https://doi.org/10.48550/arXiv.2306.07303

11. Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., & Aue, A. (2018). *Marian: Cost-effective High-Quality Neural Machine Translation in C++*. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (pp. 129–135). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2716

12. Kim, Y. J., Junczys-Dowmunt, M., Hassan, H., Aji, A. F., Heafield, K., Grundkiewicz, R., & Bogoychev, N. (2019). *From Research to Production and Back: Ludicrously Fast Neural Machine Translation.* In Proceedings of the Third Workshop on Neural Generation and Translation. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5632

13. Gowda, T., Grundkiewicz, R., Rippeth, E., Post, M., & Junczys-Dowmunt, M. (2024). *PyMarian: Fast Neural Machine Translation and Evaluation in Python.* arXiv. https://doi.org/10.48550/arxiv.2408.11853

14. (2024). "Survey of Transformers and Towards Ensemble Learning Using Transformers for Natural Language Processing." *Journal of Big Data*, vol. 11, no. 3, pp. 123–137. https://doi.org/10.1186/s40537-023-00842-0

15. Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., Shen, C. (2023). "A Survey on Efficient Training of Transformers." *arXiv preprint arXiv:2302.01107*. https://doi.org/10.48550/arXiv.2302.01107

16. Dong, Z., Tang, T., Li, L., Zhao, W. X. (2023). "A Survey on Long Text Modeling with Transformers." *arXiv preprint arXiv:2302.14502*. https://doi.org/10.48550/arXiv.2302.14502

17. Glushkova, T., C. Zerva, and A. F. T. Martins. (2023). "BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation." *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 47–56. https://aclanthology.org/2023.eamt-1.6/

18. ElNokrashy, M., and T. Kocmi. (2023). "eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings." *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)*, pp. 756–765.

19. Saadany, H., and C. Orasan. (2021). "BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text." *Proceedings of the Translation and Interpreting Technology Online Conference (TRITON 2021)*, pp. 50–59. https://aclanthology.org/2021.triton-1.6/

20. 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.

МЕТОДЫ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА ЮРИДИЧЕСКИХ ТЕКСТОВ НА КАЗАХСКИЙ, РУССКИЙ ЯЗЫКИ И ИХ КАЧЕСТВЕННЫЙ АНАЛИЗ

Аннотация. *В настоящее время в Республике Казахстан одной из актуальных задач является качественный перевод текстов в области права с казахского языка на русский и английский языки, а также с этих языков на казахский. В данной научной работе проведён анализ ошибок при переводе текстов, полученных из специализированных правовых источников, с использованием широко распространённых систем машинного перевода, таких как Яндекс и Гугл, на языковой паре казахский–русский.*

Основной целью исследования является точный и смысловой корректный перевод юридических предложений и терминов. Для этого был сформирован корпус из 96 555 предложений и словосочетаний, собранных с помощью специальной программы из правовых документов, судебных решений и официальных сайтов.

Сформированный корпус был использован для обучения нейронной системы машинного перевода MarianMT, применённой на языковой паре казахский–русский. Для улучшения качества перевода дополнительно использовалась трансформерная модель

KazRobert, архитектура и математические основы которой подробно рассмотрены в работе.

Качество перевода оценивалось с использованием международно признанных метрик BLEU, TER и METEOR. В исследовании сравнительно проанализированы два результата: переводы, полученные только с помощью MarianMT, и переводы, дообученные с применением модели KazRobert. В результате установлено, что предложенный метод обеспечивает более высокое качество перевода по сравнению с моделью на основе OpenNMT.

Проведённые эксперименты показали, что увеличение объёма корпуса и количества терминов напрямую влияет на улучшение качества перевода. Кроме того, результаты исследования подтвердили возможность адаптации данного подхода к тюркским языкам, близким к казахскому по структуре.

Ключевые слова: *нейронный машинный перевод, машинный перевод MarianMT, модель KazRobert, трансформер-модель, корпус в области права, показатель перевода BLEU, показатель перевода TER, показатель перевода METEOR.*

NEURAL MACHINE TRANSLATION METHODS FOR LEGAL TEXTS INTO KAZAKH AND RUSSIAN LANGUAGES AND THEIR QUALITY ANALYSIS

Abstract*.Currently, one of the pressing issues in the Republic of Kazakhstan is the accurate translation of legal texts from Kazakh into Russian and English, as well as from these languages into Kazakh. This scientific work analyzes translation errors using widely known machine translation systems such as Yandex and Google, based on legal texts sourced from specialized legal databases in the Kazakh–Russian language pair.*

The main goal of the study is to achieve precise and semantically accurate translation of sentences and terminology specific to the legal field. To this end, a corpus of 96,555 sentences and phrases was compiled using a specialized program, collecting data from legal documents, court decisions, and official websites.

This corpus was used to train the MarianMT neural machine translation system within the Kazakh–Russian language pair. To further improve translation quality, the KazRobert transformer model was applied. The study provides a comprehensive explanation of the KazRobert model's architecture and its mathematical foundations.

Translation quality was evaluated using internationally recognized metrics such as BLEU, TER, and METEOR. The study presents a comparative analysis of two outcomes: the initial results from the MarianMT model alone, and the improved results from the same model fine-tuned with KazRobert. The findings indicate that the proposed approach outperforms previous models, including the OpenNMT-based system developed by the same authors.

The experiments demonstrated that increasing the corpus size and the number of legal terms positively impacts translation quality. Furthermore, the research suggests that this method can be effectively adapted for other Turkic languages that share structural similarities with Kazakh.

Keywords: *neural machine translation, MarianMT machine translation, KazRobert model, transformer model, legal domain corpus, BLEU translation metric, TER translation metric, METEOR translation metric.*

Авторлар туралы мәлімет

Рахимова Диана Рамазановна	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, Санкт-Петербург мемлекеттік ақпараттық технологиялар, механика және оптика университеті, Санкт-Петербург, Ресей, E-mail: valentin.malykh@phystech.edu

Сведение об авторах

Рахимова Диана Рамазановна	PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан E-mail: di.diva@mail.ru
Жігер Алия Жігерқызы	Магистр, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан и Университет Нархоз, Алматы, Казахстан и E-mail: alia_94-22@mail.ru
Валентин Малых	PhD Санкт-Петербургский государственный университет информационных технологий, механики и оптики (Санкт-Петербург, Россия). E-mail: valentin.malykh@phystech.edu

Information about the authors

Rakhimova Diana Ramazanovna	PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan E-mail: di.diva@mail.ru
Zhiger Aliya Zhigerkyzy	Master, Al-Farabi Kazakh National University, Almaty, Kazakhstan and Narxoz University, Almaty, Kazakhstan and E-mail: alia_94-22@mail.ru
Valentin Malykh	PhD, St. Petersburg State University of Information Technologies, Mechanics and Optics (St. Petersburg, Russia). E-mail: valentin.malykh@phystech.edu