ӘОЖ 004.934.1
ҒТАХР 20.19.00

https://doi.org/10.55452/1998-6688-2026-23-2-205-217

[1]Рахимова Д.,
доцент, ORCID ID: 0000-0003-1427-198X,
e-mail: drakhimova060@gmail.com
[1,2,3]*Жігер А.,
магистр, ORCID ID: 0000-0002-3641-8260,
*e-mail: aliya.zhunussova.zh@narxoz.kz
[3,4]Малых В.,
т.ғ.к., ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Мансурова М.,
физ.-мат.ғ.к., профессор,
ORCID ID:0000-0002-9680-2758,
e-mail: mansurova.madina@gmail.com
[1]Карюкин В.И.,
PhD, ORCID ID: 0000-0002-8768-0349,
email: vladislav.karyukin@gmail.com

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

ҚАЗАҚ ТІЛІНДЕГІ МЕДИЦИНАЛЫҚ МӘТІНДЕРДІ ПОСТРЕДАКТОРЛЕУДІҢ ӘДІСТЕРІ МЕН САПАЛЫҚ ТАЛДАУЫ

Аңдатпа

Қазіргі таңда медицина саласында ағылшын тілінен қазақ тіліне сапалы және дәл аудару – медициналық ақпараттың қолжетімділігі мен қауіпсіздігін қамтамасыз етуде маңызды мәселелердің бірі. Бұл ғылыми жұмыста кеңінен қолданылатын Google Translate және Yandex Translate секілді машиналық аударма жүйелерінің медициналық мәтіндерге қатысты дәлдігі мен тиімділігі зерттелді. Зерттеудің негізгі мақсаты – ағылшын тіліндегі медициналық терминдер мен күрделі сөйлемдерді қазақ тіліне мағыналық және стилистикалық тұрғыдан дұрыс аудару жолдарын қарастыру. Осы мақсатта халықаралық медициналық мақалалар, клиникалық зерттеулер және дәрі-дәрмек сипаттамаларынан тұратын 102 374 сөйлем жинақталып, арнайы корпус жасалды. Корпус MarianNMT нейронды машиналық аудармада оқытылып, қазақ тіліне аударылды. Аударма нәтижелерін жеңіл постредакторлеу үшін Kaz-RoBERTa трансформерлік моделі және толық постредакторлеу үшін Үлкен Тілдік Модельдердің (LLM) бірі GPT-4.1 моделі қолданылып, оның медициналық мәтіндерге бейімделу мүмкіндіктері қарастырылды. Аударма сапасы BLEU, TER және METEOR метрикалық бағалау көрсеткіштері арқылы өлшенді. Бастапқы MarianNMT машиналық аудармадан кейін алынған аудармалар мен Kaz-RoBERTa мен GPT-4.1 модельдерінен кейінгі нәтижелер салыстырмалы түрде ұсынылды. Талдау нәтижелері көрсеткендей Kaz-RoBERTa моделінен кейін алынған аудармалар бастапқы MarianNMT машиналық аудармадан кейінгі алынған аудармадан BLEU метрика бойынша 9%-ға жоғары нәтиже, ал GPT-4.1 моделінен кейін 23%-ға жоғары нәтиже көрсетті. Қазақ тілі лингвист мамандар баға беруі бойынша, Kaz-RoBERTa моделі қолданылғаннан кейінгі алынған аудармаларда термин, лексика бойынша қателер саны азайғаны байқалды. Ал GPT-4.1 моделінен кейін аударма стильдік және семантикалық жағынан жақсарғанын көруге болады. Мақалада қарастырылған модельдер қазақ тіліне ұқсас түркі тілдеріндегі медицина саласындағы аудармаларды постредакторлеу саласында қолдануға болады.

Түйін сөздер: нейронды машиналық аударма, MarianNMT, Kaz-RoBERTa, трансформер моделі, медицина мәтіндері, BLEU, TER, METEOR.

Кіріспе

Қазақстан Республикасында медицина саласындағы сөйлемдер мен терминдерді ағылшын тілінен қазақ тіліне қолжетімді әрі сапалы аударылуы – қоғамдық денсаулық сақтау, ғылыми-зерттеу, клиникалық тәжірибе және фармацевтикалық қызметтің тиімділігін арттыруда шешуші рөл атқарады.

Халықаралық емханалық нұсқаулықтар, ғылыми мақалалар және дәрілік заттар жөніндегі ресми материалдарды қазақ тіліне дәл жеткізу медициналық мамандардың өзекті ақпаратқа қолжетімділігін қамтамасыз етіп, медициналық шешім қабылдау сапасын жақсартады және ұлттық тілдегі кәсіби медициналық контенттің кеңеюіне ықпал етеді [1].

Жетілдірілген машиналық аударма жүйелерінің (Google Translate, Yandex Translate және т.б.) және жасанды интеллект технологияларының дамуына қарамастан, медициналық мәтіндерді ағылшын тілінен қазақ тіліне немесе керісінше аудару кезінде елеулі қателіктер жиі кездеседі. Мұндай қателіктердің басты себептерінің бірі – қазақ тілінің морфологиялық, синтаксистік және семантикалық тіл құрылымдарының күрделілігі, оның өзге тілдерден айқын ерекшеленуі болып табылады.

Мысалы, қазақ тіліндегі күрделі морфология жүйесі, сөздердің жуан-жіңішке үндестігі, сөйлемдердің жиі құрмалас құрылымы, сондай-ақ терминдердің контекстке тәуелді көпмағыналылығы – машиналық аударма жүйелері үшін күрделі кедергілер тудырады. Әсіресе медицина саласындағы арнайы терминдер, қысқартулар, латынша сөздер дәлдігіне айтарлықтай әсер етеді [2].

Әр түрлі корпустар үшін, сонымен қатар медициналық корпустар үшін ағылшын-қазақ және орыс-қазақ тілдеріндегі машиналық аудармасы (МТ) саласында бірнеше әдістемелік тәсілдер қолданылған. Атап айтқанда, Apertium жүйесінде құрылымдық ауысым ережелерін құру тәжірибесі жүргізілген [3], сондай-ақ қазақ тілінің бай морфологиясын ескеру мақсатында морфологиялық сегментация әдістері қолданылған [4]. Сонымен қатар, орыс тіліндегі леммatизация және екі тілдік сөздіктер арқылы сөйлемдерді туралау әдістері зерттелген [5]. Бұл тәсілдер қазақ тілінің синтаксистік және морфологиялық ерекшеліктерін есепке ала отырып, алғашқы машиналық аудармасы жүйелерінің дамуына негіз қалады.

Осы қиындықтарды ескере отырып, қазақ тілінде постредакторлеу жүйесін дамыту қажеттілігі туындайды.

Постредакторлеу дегеніміз – машиналық аудармадан кейін мәтінді адам арқылы өңдеу процесі [6]. Қазіргі таңда көптеген лингвистикалық қызмет көрсетушілер редакторларды оқыту және постредакторлеу әдістерін дамыту жолдарын іздестіріп жатыр. Постредакторлеу және оның қолданылуы бойынша алғашқы зерттеулер 1980-жылдары пайда болды. 1999 ж. Американдық машиналық аударма ассоциациясының және Еуропалық машиналық аударма ассоциациясының мүшелері постредакторлеуге арналған арнайы қызығушылық тобы (Special Interest Group on Post-Editing) құрды [7].

Машиналық аударманы постредакторлеу (PEMT) – машиналық аударма жүйелерінің жіберген қателіктерін түзету арқылы соңғы аударманың сапасын арттыру процесі. Бұл әдіс көбінесе ағылшын тілінен басқа тілдерге аударманы жақсарту үшін қолданылады. Алайда, машиналық аударманы постредакторлеуде (PEMT) әдістерінің аз ресурсты тілдерге, мысалы қазақ тіліне тиімділігі мәселесі әлі де толық зерттелмеген. Қазақ тілі агглютинативті тіл болғандықтан, машиналық аудармада өзіндік қиындықтар туындайды. Қазақ тіліндегі машиналық аударманы постредакторлеу (PEMT) морфология, синтаксис және лексикаға байланысты бірнеше ерекше қиындықтарға тап болады.

Машиналық аударманы постредакторлеуді зерттеу барысында үлкен көлемдегі параллель корпустарды құру жүзеге асырылды [8], олар қоғамдық қолжетімді болып, BLEU көрсеткішін 0,49 дейін жеткізе алады [9]. Соның ішінде осы деректерде оқытылған нейронды машиналық аударма (NMT) үшін постредакторлеу модельдері құрылды [10]. Мәтіндік ресурстардың басым бөлігі мемлекеттік көздерден алынды. Алайда бұл корпуста медицина саласы қарастырылмады.

Бұл мақалада біз постредакторлеу әдістерін арнайы жинақталған медицина саласындағы корпусқа қолдануды қарастырамыз. Зерттеудің мақсаты – аударманы постредакторлеу (PEMT) әдістері медицина саласында қазақ тіліндегі машиналық аударманың сапасын жақсарта ала ма, сонымен қатар аз ресурсты тілдермен жұмыс істегенде зерттеушілердің кездесетін негізгі мәселелері мен шектеулерін анықтау болып табылады.

1.1. Машиналық аудармадағы постредакторлеу түрлеріне шолу

Автоматты машиналық аударманы постредакторлеу-мақсатты тілде еркін сөйлейтін маман арқылы жүзеге асырылады. Машиналық аударманы постредакторлеу (PEMT) – машиналық аударма нәтижелерін түзету және жетілдіру арқылы олардың сапасын арттыру және табиғи тіл стандартына жақындату процесі [11].

Екі негізгі постредакторлеу түрі қолданылады:

1) Жеңіл постредакторлеу мәтіндегі медициналық терминдерді, сөздердегі морфология, грамматика, лексика тіл бөліктері бойынша қателіктерді азайтады [12]; Жеңіл постредакторлеу деңгейі аз ресурсты тілдер үшін машиналық аудармадан шыққан бастапқы нәтижені алып, мәтінді түсінікті, мазмұндық дәл және грамматикалық дұрыс ету үшін минималды түзетулер енгізуді қамтиды [13;14]. Медицина саласындағы корпустарға жеңіл постредакторлеу (лексика, грамматика) ағылшын-француз тілдеріне тіл мамандары көмегімен бастапқы DeepL жүйесінен алынған сөйлемдер мен сөз тіркестеріне жүргізілген [15]. Бұл жұмыста постредакторлеу қажет ететін қателіктерге лексикалық қателіктер анықталды: терминдердің дұрыс таңдалмауы, әсіресе медициналық терминдердің қате қолданылуы; морфологиялық қателіктер: тілдің ерекшеліктеріне сәйкес келмейтін сәйкестендіру: септік жалғаулар дұрыс жалғанбауы;

2) Толық постредакторлеу медицина саласындағы синтаксис, семантикалық және стильдік қателерді жоюға бағытталған. Нәтижесінде мәтін түсінікті және толық қабылданатын болады. Тілдік сападан гөрі түсініктілікке мән береді және тілдік әрі стильдік мәселелерге рұқсат береді .

Қолданыстағы зерттеулерді талдау негізінде келесі зерттеу тапсырмалары анықталды:

- Медицин саласында корпус жинау , өңдеу және құрылымдау: Бұл үшін ағылшын – қазақ тілдер жұбынан тұратын медицина саласындағы сөйлемдер мен сөз тіркестері жинақталды;
- Ағылшын – қазақ тілдер жұбынан тұратын медицина саласындағы корпус үшін постредакторлеу әдісін дамыту: Жеңіл постредакторлеу үшін Kaz-RoBERTa трансформерлік моделі және толық постредакторлеу үшін Үлкен Тілдік Модельдер (LLM) бірі GPT-4.1 модельдер негізінде жүзеге асырылады;
- Сапаны бағалау: BLEU, TER, METEOR аударма сапалар көрсеткіші бойынша бағалау жүргізілді.

Осы тапсырмаларды жүзеге асыру үшін қазақ тіліндегі медицина саласында сөйлемдерге машиналық аударманы постредакторлеу жүйесін дамыту әдістемесі мен кезеңдері төменде ұсынылады.

Материалдар мен әдістер

Медицина саласындағы сөз тіркесстері мен сөйлемдерді ағылшын тілінен қазақ тіліне аудару сапасын арттыру үшін келесі архитектура кезеңдері қарастырылды:

1. Бірінші кезең: Машиналық аударма процесі (ағылшын тілінен қазақ тіліне аудару). Бұл процесс токенизацияны қамтиды, яғни ағылшын тіліндегі сөйлемдерді жеке сөздерге немесе токендерге бөлу; морфологиялық талдау жүргізу; тілдің морфологиялық ерекшеліктерін анықтау; бастапқы аударма үшін оқытылған MarianNMT моделін (трансформер архитектурасына негізделген) пайдалану; және детокенизация, яғни токенделген мәтінді қайта оқылатын қазақ тіліндегі форматқа келтіру [16;17].
2. Постредакторлеу кезеңі: Алғашқы тексеру барысында қазақ тіліндегі машиналық аударма нәтижелерінде айқын қателер бар-жоғы анықталады. Қазақ тіліндегі машиналық аудар-

ма жеңіл постредакторлеуден өтеді, яғни тек айқын қателер түзетіледі (медициналық термин, лексика, морфология). Кейін жоғары сапалы аударма алу үшін, толық постредакторлеуден өтеді және барлық қателер: стильдік, семантикалық, синтаксис жағынан түзетіледі.

3. Адаптивті кері цикл: Постредакторлеу кезінде қолданылған түзетулер болашақ аудармаларды жақсарту үшін машиналық аударма жүйесіне қайтарылады.

4. Шығыс деректер және тексеру: Сапаны бақылау жүргізіледі, атап айтқанда аударылмаған сегменттердің, сәйкессіздіктердің немесе форматқа қатысты мәселелердің бар-жоғы автоматты түрде тексеріледі. Соңғы тексеру кезеңі аударманың талап етілген сапа стандарттарына сәйкес келетінін қамтамасыз ету үшін жүргізіледі. Ақырында нәтиже дайын аударма түрінде ұсынылып, қажетті форматта немесе ортада шығарылады. Сурет 1-де медицина саласында аз ресурсты тілдер үшін постредакторлеудің жалпы архитектурасы көрсетілген

Сурет 1 – Медицина саласында аз ресурсты тілдер үшін машиналық аударма постредакторлеу жүйесінің архитектурасы

Сурет 1-де көрсетілген архитектурада орындалу процесі келесі қадамдардан тұрады:

1) медицина саласындағы мәтіндерді ағылшын тілінен қазақ тіліне аудару процесі MarianNMT нейронды машиналық аудармада орындалды.

2) Алынған қазақ тіліндегі модельді жеңіл постредакторлеу үшін Kaz-RoBERTa моделі қолданылды.

3) Жеңіл постредакторлеуден кейін алынған аударманы толық постредакторлеу процесі GPT-4.1 моделі арқылы жүзеге асырылды. Төменде әрбір қадам орындалу түсініктемесі беріледі.

MarianNMT. MarianNMT жүйесі негізінен PyTorch және Hugging Face Transformers фреймворктерін пайдаланып жүзеге асырылған. MarianNMT – ашық дереккөзді нейроаударма жүйесі, оның дайын модельдері Hugging Face репозиторийінде қолжетімді (https://huggingface.co/models, қол жеткізілген мерзім: 10 желтоқсан 2025). Ағылшын тілінен қазақ тіліне аудару барысында MarianNMT моделін оқыту үшін ағылшын-қазақ тілдер жұбынан тұратын деректер жиынтығы оқу, тексеру және тестілеу үшін 80%/10%/10% пропорциясына бөлінді. Корпус токенизациядан кейін мәтіндік файлдарда (txt) сақталды. Модельдің энкодер-декодер трансформерлік архитектурасы әр токеннің контекстін self-attention арқылы талдайды [18], ал декодер cross-attention арқылы энкодерден шыққан контекстті пайдаланып аударма токендерін жасайды. Бұл тәсіл медициналық мәтіндердің терминологиясы мен грамматикалық құрылымын салыстырмалы түрде дәл аударуға мүмкіндік береді.

Kaz-RoBERTa моделі. Kaz-RoBERTa – қазақ тіліне арнайы үйретілген, RoBERTa негізіндегі трансформерлік модель. Оның дайын модельдері Hugging Face репозиторийінде қолжетімді (https://huggingface.co/kz-transformers/kaz-roberta-conversational).

Kaz-RoBERTa моделі жеңіл автоматты постредакторлеу кезінде жоғары көрсеткіш көрсетті [19]. Жұмыста қарастырылған медициналық корпустар үшін толық постредакторлеуде Kaz-RoBERTa моделі қолданылды.

Kaz-RoBERTa модель архитектурасы. Модельге кіріс ретінде қазақ тіліндегі MarianNMT-ден алынған қазақ тіліндегі бастапқы мәтін беріледі. Мәтін субсөздік токенизация арқылы,

SentencePiece әдісімен бөлінеді, және әр токенге оның мәнін және сөйлемдегі орнын көрсететін векторлық ұсыну беріледі. Содан кейін токеннің векторы трансформер қабаттарынан өтеді, мұнда әр токен self-attention механизмі арқылы барлық қалған токендерді «қарап», контексті түсінеді. Көпқабатты назар (multi-head attention) модельге бір уақытта контексттің әртүрлі аспектілерін, соның ішінде көрші сөздерді ескеруге мүмкіндік береді. Назардан кейін әр токен екі қабатты нейрондық желі Feed Forward Network арқылы өтеді, бұл оның ерекшеліктерін күшейтеді және түрлендіреді. Қалдық байланыс (residual connection) және қабаттық нормализация (LayerNorm) қолданылады, бұл ақпаратты сақтауға және оқытуды тұрақтандыруға көмектеседі. Оқыту кезінде бастапқы мәтіннен кездейсоқ токендер маскаланған болып табылады, және модель қалған сөздердің контекстін пайдаланып осы маскаланған токендерді болжауға тырысады, бұл оған мәтіндегі қателерді түзетуге мүмкіндік береді. Модель өзінің болжамдарын эталондық пост-редакторленген мәтінмен салыстырады, осылайша грамматика, лексика және морфология тіл бөлшектерін түзетуді үйренеді. Сурет 2-де Kaz-RoBERTa модель архитектурасы мысалмен көрсетілген.

Сурет 2 – Kaz-RoBERTa моделінің
архитекурасын мысалмен көрсету

Kaz-RoBERTa моделінің архитектурасын математикалық моделі бойынша түсіндіретін болсақ:

1. Енгізу қабаты. Kaz-RoBERTa алдымен мәтінді токендерге бөледі және әр токенді вектор түріне түрлендіреді:

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

мұнда: ω_i – i – ші токен, $TokenEmb(\omega_i)$ – токен эмбеддингі, $PosEmb(i)$ – позициялық эмбеддинг (сөздердің ретін ескеру үшін)

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

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

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

$$Q = XW^Q, K = XW^K, V = XW^V$$

Назар функциясы:

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

b) Көпқабатты назар функциясы:

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

Мұндағы n назар қабаттар саны.

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

Әр токен үшін екі қабатты нейрондық желі қолданылады:

$$FFN(x) = (0, xW_1 + b_1)\, W_2 + b_2$$

d) Қалдық байланыс + Нормализация (Residual + LayerNorm)

Әр бөлікке LayerNorm қолданылады:

$$x' = LayerNorm(x + MultiHead(x))$$

$$x'' = LayerNorm(x' + FFN(x'))$$

3. Маска қойылған тілдік модельдеу (Masked Language Modeling)

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

$$\sigma_{MLM} = -\sum_{i \in \mu}^{N} log\ log\ P(\omega_i | x_i)$$

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

GPT-4.1 моделі. Толық постредакторлеу үшін аз ресурсты тілдерге арналған LLM модельдерінің бірі (GPT-4.1) моделі қолданылды. Толық постредакторлеуге GPT-4.1 моделі қолданылу себебі төменде зерттеу жұмыстарына байланысты таңдалды.

Xu және т.б. ғалымдар жұмысында үлкен тілдік модельдерді (LLM) машиналық аударма үшін қайта оқытуда маңызды жаңалықтарды енгізді, параметрлік тиімді әдістер аз есептеу ресурстарын пайдалана отырып айтарлықтай өнімділікті арттыра алатынын көрсетті [20]. Бұл әдістер әсіресе төмен ресурсты тіл жұптары үшін тиімді.

Одан кейін бірнеше зерттеулер LLM-дердің аударма қабілеттерін жүйелі түрде бағалады. Hendy және т.б. ғалымдар жұмысында GPT модельдерін жан-жақты бағалап, бұл модельдер нөлдік үлгідегі (zero-shot) көрсеткіштерде жақсы нәтижелер көрсетсе де, әдетте жоғары ресурсты тілдер үшін мамандандырылған аударма жүйелерінен артта қалатынын анықтады [21]. Алайда төмен ресурсты тілдерде өнімділік айырмашылығы әдетте аз болады, бұл LLM-дердің мұндай жағдайларда қолдануға жарамды балама бола алатынын көрсетеді.

Jiao жұмысында ChatGPT-нің аударма өнімділігін нақты талдауға назар аударды және GPT-4 бұрынғы нұсқаларға қарағанда айтарлықтай жақсы аударма сапасына қол жеткізетінін, әсіресе төмен ресурсты тілдер үшін екенін хабарлады [22]. Олардың нәтижелері аударманы оңтайландыру үшін нұсқаулыққа бейімделген модельдердің және мұқият жасалған prompting стратегияларының маңыздылығын көрсетеді.

Модельге қол жеткізу және баптау OpenAI fine-tuning API арқылы жүзеге асырылды, ол бақыланатын бейімдеуге ыңғайлы интерфейс ұсынады. Бастапқы кіріс мәтін ретінде

жеңіл постредакторлеуден Kaz-RoBERTa моделі арқылы алынған аударма қолданылады. Ал шығыс мәтін ретінде аудармашылар көмегімен дұрыс аударылған мәтін беріледі. Бұл тілдік жұп арқылы бақыланатын қайта оқытылған модель (fine-tuning SFT) процедурасы модельді қателері бар қазақ тіліндегі аудармаларды түзетілген нұсқаларға айналдыруға үйреттіледі.

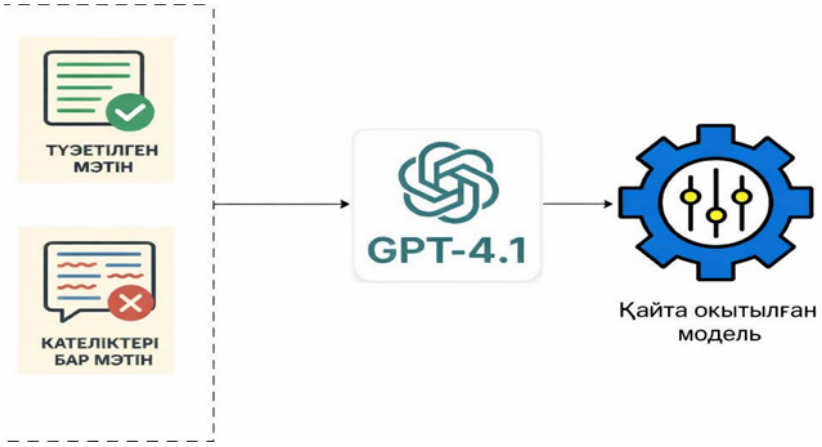
Оқу деректері сұхбат форматы (prompt-response) түрінде құрылды, бұл модельдің өзара әрекет стиліне сәйкес келеді:

User: [ҚАТЕЛІКТЕРІ БАР ҚАЗАҚ СӨЙЛЕМІ]

Assistant: [ТҮЗЕТІЛГЕН ҚАЗАҚ СӨЙЛЕМІ]

GPT-4.1 моделі орындалу алгоритмі Сурет 3-те көрсетілген.

Бұл формат модельге мақсатты түзетулерді орындауға және бастапқы мәтіннің мағынасын сақтауға нақты нұсқаулық береді.

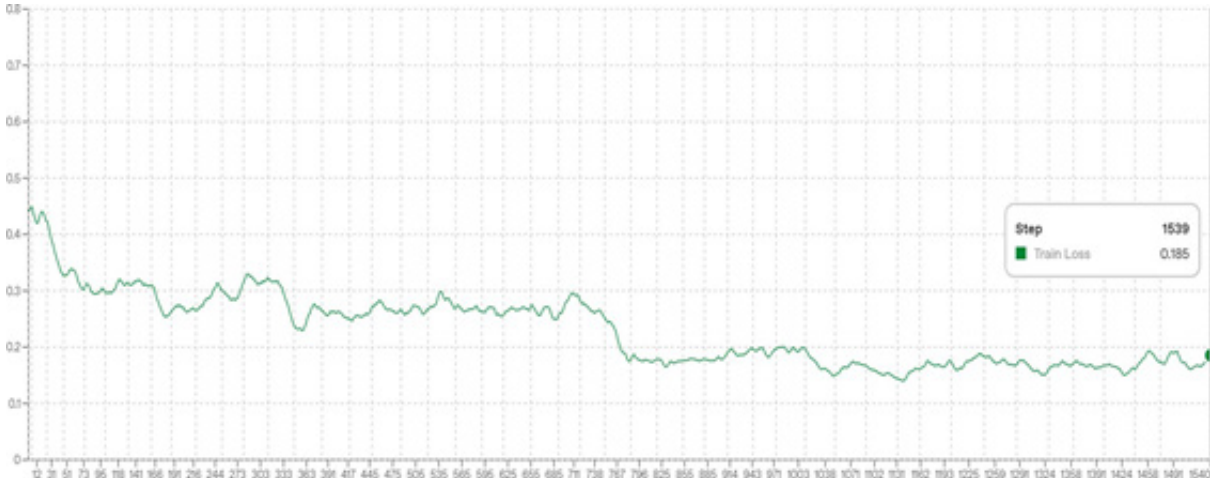


Сурет 3 – GPT-4.1 моделінің орындалу алгоритмі

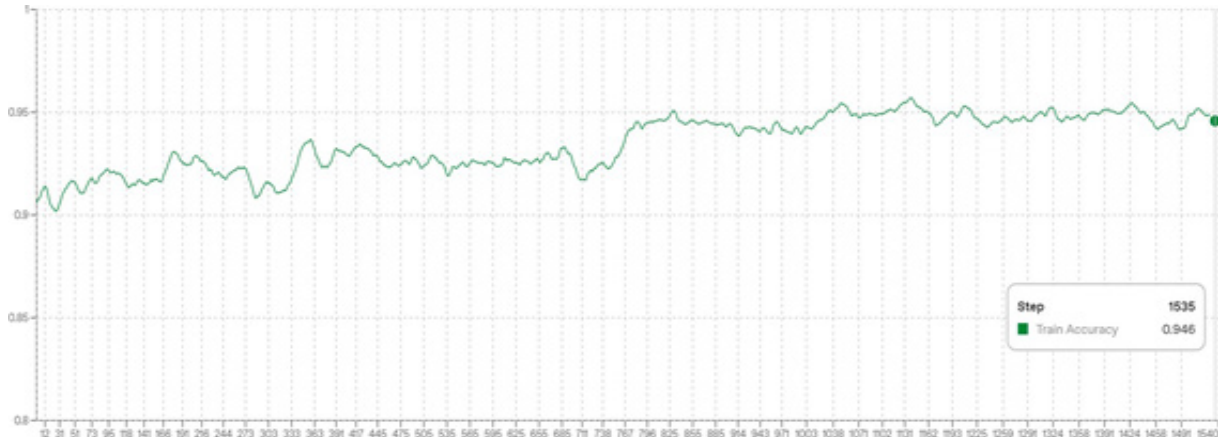
Fine-tuning келесі гиперпараметрлермен жүргізілді:

- Оқыту жылдамдығы мультипликаторы: 2
- Batch өлшемі: 13 мысал
- Оқу кезеңдері (epochs): 2 (валидация нәтижесіне сүйене отырып анықталды)

Біз оқыту барысын валидация шығыны арқылы бақыладық және overfitting қаупін азайту үшін ерте тоқтатуды (early stopping) қолдандық. Сурет 4 және Сурет 5-те оқыту шығыны мен дәлдіктің эволюциясы көрсетілген.



Сурет 4 – Итерациялар бойынша оқыту шығыны



Сурет 5 – Итерациялар бойынша оқыту дәлдігі

Деректер жиынтығы. Медицина саласынан 102374 ағылшын-қазақ тілдер жұбынан тұратын корпус әзірленді. Корпус сөйлемдер мен сөз тіркестерін дереккөздерден ағылшын тілінде жинап, аудармашылардың көмегімен қазақ тіліне аударды. Корпусқа алынған дереккөздер мыналар: Халықаралық клиникалық хаттамалар; Ғылыми медициналық мақалалар; Фармацевтикалық нұсқаулықтар мен дәрілік аннотациялар; Ресми медициналық веб-сайттар. Кесте 1-де әрбір дереккөздің сілтемесі мен одан алынған сөйлемдер саны көрсетілген.

Кесте 1– Медициналық корпустағы сөйлемдер саны

Алынған дереккөздері	Деректер саны
Халықаралық клиникалық хаттамалар (WHO, NICE, CDC) (https://www.who.int/publications/who-guidelines?utm_source=chatgpt.com)	28,435
Ғылыми медициналық мақалалар (PubMed, ResearchGate, т.б.) (https://pmc.ncbi.nlm.nih.gov/search/?term=%223%20Biotech%22%5Bjournal%5D)	25,210
Фармацевтикалық нұсқаулықтар мен дәрілік аннотациялар (https://open.fda.gov/apis/drug/label/download/?utm_source=chatgpt.com)	18,950
Ресми медициналық веб-сайттар (денсаулық сақтау министрлігі, NIH, FDA) (https://www.cdc.gov/respiratory-viruses/about/index.html)	29,779
Барлығы	102,374

Корпустағы сөйлемдер мен сөз тіркестерді жинау үшін арнайы бағдарлама құрылды. Бұл бағдарлама PDF форматындағы құжаттардан және медицина саласындағы ресми сайттардан терминдер, сөз тіркестері мен сөйлемдерді автоматты түрде таңдап, корпус құрамын толықтырады.

Нәтижелер мен талқылау

Эксперимент келесі сипаттамалары бар компьютерде жүргізілді: CPU – Core i7 4790K, 32 GB RAM, 1 TB SSD, GPU: RTX 2070 Super, GTX 1080.

Ағылшын–қазақ тіліндегі аударма сапасын бағалау үшін келесі метрикалар қолданылды: BLEU [23], METEOR [24] және TER [25].

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

$$BLEU = BP * EXP\left(\sum_{1}^{N} \omega_n \, log \, log \, p_n\right)$$

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.

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

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

$Penalty = 0.5 \cdot \left(\frac{chunks}{matches}\right)^3$ – 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.

$$TER = \frac{\text{Редакция қадамдарының саны}}{\text{эталондағы сөздер саны}}$$

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

Модельдерді тестілеу үшін медицина саласынан 10000 сөйлем мен сөз тіркестерінен тұратын корпус құрылады.

Осы корпус MarianNMT нейронды машиналық аудармада ағылшын тілінен қазақ тіліне және салыстыру мақсатында қазақ тілінен ағылшын тіліне аударылады. Кесте 2-де аудармалардың нәтижелері BLEU, TER, METEOR метрикалық көрсеткіштерімен көрсетіледі.

Кесте 2 – MarianNMT нейронды машиналық аудармадан алынған аударма нәтижесі

Тіл жұптары	BLEU	TER	METEOR
Ағылшын-қазақ	0.594	0.294	0.752
Қазақ-ағылшын	0.585	0.251	0.746

Нәтижелерге сүйенсек, модельдің ағылшын-қазақ бағытындағы BLEU көрсеткіші 0.594, ал қазақ-ағылшын бағытында 0.585. Бұл мәндер аударма сапасының екі бағытта да салыстырмалы түрде жоғары екенін көрсетеді. TER шамасы сәйкесінше 0.294 және 0.251 құрап, аудармалардағы қателік деңгейінің төмен екендігін білдіреді. Ал METEOR метрикасы 0.752 және 0.746 мәтіннің мазмұндық және лексикалық дәлдігінің жоғары екенін айғақтайды.

Жалпы алғанда, MarianNMT моделі медициналық саласындағы сөйлемдерді аударуда ағылшын және қазақ тілдері арасындағы машиналық аудармада қанағаттанарлық нәтиже көрсетіп, екі бағытта да аударманың сапасын сақтай отырып жұмыс істейтінін байқауға болады.

Кесте 3-те MarianNMT нейронды машиналық аудармада алынған қазақ тіліндегі аудармаларын сапасын жақсарту үшін жеңіл постредакторлеу (Kaz-RoBERTa моделі) және толық постредакторлеу (GPT-4.1-mini моделі) нәтижелері алынды.

MarianNMT нейрондық машиналық аудармаларын Kaz-RoBERTa моделі арқылы постредакторлеу аударманың сапасын айтарлықтай жақсартады. BLEU көрсеткіші 9%-ға, METEOR 5%-ға өсті, мамандардың бағалауынша лексикалық сәйкестік деңгейі де артты. TER метрикасы бойынша көрсеткіш 0,294-тен 0,145-ке дейін төмендеді, бұл қателер санын екі есеге

жуық қысқарғанын көрсетеді. Жеңіл постредакторлеуден кейін аудармалар эталондық мәтінге әлдеқайда жақындайды.

Кесте 3 – Қазақ тілінде алынған аударманы жеңіл және толық постредакторлеуден кейінгі нәтижелер

Модельдер	BLEU	TER	METEOR
Kaz-RoBERTa	0.682	0.145	0.805
GPT-4.1-mini	0.820	0.078	0.885

Толық постредакторлеу нәтижесінде BLEU көрсеткіші жеңіл постредакторлеуге қарағанда 14%-ға жоғары, TER 7%-ға төмен, METEOR 8%-ға өскен. Бұл көрсеткіштер GPT-4.1-mini моделі арқылы алынған аударманың Kaz-RoBERTa моделінен алынған аудармадан сапасы жоғары екенін дәлелдейді. Тіл мамандарының пікірінше, толық постредакторлеуден кейінгі аударма жеңіл постредакторлеуден кейінгі аудармамен салыстырғанда семантикасы мен стилі бойынша айтарлықтай жақсарған.

Қорытынды

Жүргізілген зерттеу нәтижелері ағылшын тіліндегі медициналық мәтіндерді қазақ тіліне аударуда машиналық аударма мен постредакторлеу модельдерін кешенді қолданудың тиімді екенін көрсетті. MarianNMT негізінде алынған бастапқы аудармалар жалпы мағынаны жеткізгенімен, терминологиялық, лексикалық және стилистикалық қателердің кездесетіні анықталды. Бұл қателерді азайтуда постредакторлеу кезеңінің маңызы зор екені дәлелденді.

Kaz-RoBERTa трансформерлік моделін қолдану нәтижесінде медициналық терминдердің дәлдігі артып, лексикалық қателер саны едәуір қысқарды. Ал GPT-4.1 үлкен тілдік моделін пайдалану аударманың семантикалық тұтастығы мен стильдік сапасын айтарлықтай жақсартып, мәтінді кәсіби медициналық ғылыми тіл талаптарына жақындатты. Метрикалық бағалау нәтижелері де бұл тұжырымды растап, GPT-4.1 моделінің аударма сапасын бастапқы MarianNMT нәтижелерімен салыстырғанда BLEU метрикасы бойынша 23%-ға дейін арттырғанын көрсетті.

Зерттеу қорытындылары медициналық мәтіндерді қазақ тіліне аударуда толық немесе жартылай автоматтандырылған постредакторлеу жүйелерін енгізудің өзектілігін айқындайды. Ұсынылған тәсілдер қазақ тілімен типологиялық тұрғыдан жақын түркі тілдері үшін де қолданбалы маңызға ие болып, медицина саласындағы көптілді ақпарат алмасудың сапасын арттыруға үлес қоса алады. Болашақта доменге бейімделген тілдік модельдерді жетілдіру және медициналық корпус көлемін ұлғайту аударма сапасын одан әрі жақсартуға мүмкіндік береді.

Қаржыландыру туралы ақпарат

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

ӘДЕБИЕТТЕР

1 Issakova, S., Khassangaliyeva, B., Issakova, A., Taganova, A., Toxanbayeva, T., Kamarova, N., and Kuzdybaeva, A. The frame representation of medical terminological system in Kazakh and English. Forum for Linguistic Studies, 7 (5), 739–747 (2025). https://doi.org/10.30564/fls.v7i5.9233

2 Kucherenko, O.F., Kuanysheva, A.B., and Keller Deditskaya, E.R. The corpus of borrowings and their functioning in the medical terminology of Kazakhstan (on the material of professional periodicals). Bulletin of the Karaganda University. Philology Series, 105 (1), 54–61 (2022). https://doi.org/10.31489/2022Ph1/54-61

3 Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F.M. Apertium: A free/opensource platform for rule-based machine translation. Machine Translation, 25, 127–144 (2011). https://doi.org/10.1007/s10590-011-9090-0

4 Assylbekov, Z., and Nurkas, A. Initial explorations in Kazakh to English statistical machine translation. In: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014, pp. 12–16 (2014). Available at: https://clic2014.fileli.unipi.it/proceedings/vol1/CLICIT201413.pdf

5 Assylbekov, Z., Myrzakhmetov, B., and Makazhanov, A. Experiments with Russian to Kazakh sentence alignment. In: Proceedings of the 4th International Conference on Computer Processing of Turkic Languages (2016). Available at: https://nur.nu.edu.kz/handle/123456789/1694

6 Vieira, L.N., Alonso, E., and Bywood, L. Introduction: Post-editing in practice—Process, product and networks. Journal of Specialized Translation, 31, 2–13 (2019). https://doi.org/10.26034/cm.jostrans.2019.173

7 Shterionov, D., do Carmo, F., Moorkens, J., Hossari, M., Wagner, J., Paquin, E., Schmidtke, D., Groves, D., and Way, A. A roadmap to neural automatic post-editing: An empirical approach. Machine Translation, 34, 67–96 (2020). https://doi.org/10.1007/s10590-020-09249-7

8 Rakhimova, D., and Karibayeva, A. Aligning and extending technologies of parallel corpora for the Kazakh language. Eastern European Journal of Enterprise Technologies, 4 (2(118)), 32–39 (2022). https://doi.org/10.15587/1729-4061.2022.259452

9 Zhumanov, Z., and Tukeyev, U. Integrated technology for creating quality parallel corpora. In: Advances in Computational Collective Intelligence (Springer, Cham, 2021), pp. 511–524. https://doi.org/10.1007/978-3-030-88113-9_41

10 Karyukin, V., Rakhimova, D., Karibayeva, A., Turganbayeva, A., and Turarbek, A. The neural machine translation models for the low-resource Kazakh–English language pair. PeerJ Computer Science, 9, e1224 (2023). https://doi.org/10.7717/peerj-cs.1224

11 Rakhimova, D., Sagat, K., Zhakypbaeva, K., and Zhunussova, A. Development and study of a post-editing model for Russian-Kazakh and English-Kazakh translation based on machine learning. In: Advances in Computational Collective Intelligence (Springer, Cham, 2021), pp. 525–534. https://doi.org/10.1007/978-3-030-88113-9_42

12 Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y. Pointing the unknown words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 140–149 (2016). https://doi.org/10.18653/v1/P16-1014

13 Li, X., Zhang, J., and Zong, C. Towards zero unknown word in neural machine translation. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2852–2858 (New York, NY, USA, 2016). Available at: https://www.ijcai.org/proceedings/2016

14 Rakhimova, D., Turarbek, A., Karyukin, V., Karibayeva, A., and Turganbayeva, A. The development of the light post-editing module for English-Kazakh translation. In: Proceedings of the 7th International Conference on Engineering & MIS (Almaty, Kazakhstan, 2021). https://doi.org/10.1145/3492547.3492651

15 Martikainen, H. Post editing neural MT in medical LSP: Lexico grammatical patterns and distortion in the communication of specialized knowledge. Informatics, 6 (3), 26 (2019). https://doi.org/10.3390/informatics6030026

16 Lee, W., Park, J., Go, B.-H., and Lee, J.-H. Transformer-based automatic post-editing with a context-aware encoding approach for multi-source inputs. arXiv preprint arXiv:1908.05679 (2019).

17 Rubino, R., Huet, S., Lefèvre, F., and Linarès, G. Statistical post editing of machine translation for domain adaptation. In: Cettolo, M., Federico, M., Specia, L., and Way, A. (Eds.), Proceedings of the 16th Annual Conference of the European Association for Machine Translation, pp. 221–228 (2012). Available at: https://aclanthology.org/2012.eamt-1.55/

18 Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., and Birch, A. Marian: Fast neural machine translation in C++. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 116–121 (2018). https://doi.org/10.18653/v1/P18-4020

19 Baitenova, L., Tussupova, S., Mambetov, S., Munaitbas, G., and Mukhamejanova, G. Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language. Frontiers in Artificial Intelligence, 8, Article 1708566 (2025). https://doi.org/10.3389/frai.2025.1708566

20 Xu, H., Kim, Y.J., Sharaf, A., and Awadalla, H.H. A paradigm shift in machine translation: Boosting translation performance of large language models. arXiv preprint arXiv:2309.11674 (2023).

21 Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., and Awadalla, H.H. How good are GPT models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023).

22 Jiao, W., Wang, W., Huang, J.-T., Wang, X., and Tu, Z. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745 (2023).

23 Reiter, E. A structured review of the validity of BLEU. Computational Linguistics, 44 (3), 393–401 (2018). https://doi.org/10.1162/COLI_a_00322

24 Lavie, A., and Agarwal, A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231 (2007). URL: https://aclanthology.org/W07-0734

25 Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231 (2006). URL: https://aclanthology.org/2006.amta-papers.25/

[1]Рахимова Д.,
доцент, ORCID ID: 0000-0003-1427-198X,
e-mail: drakhimova060@gmail.com
[1,2,3]*Жігер А.,
магистр, ORCID ID: 0000-0002-3641-8260,
*e-mail: aliya.zhunussova.zh@narxoz.kz
[3,4]Малых В.,
к.т.н., ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Мансурова М.,
к. ф.-м. н., профессор, ORCID ID:0000-0002-9680-2758,
e-mail: mansurova.madina@gmail.com
[1]Карюкин В.И.,
PhD, ORCID ID: 0000-0002-8768-0349,
e-mail: vladislav.karyukin@gmail.com

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

МЕТОДЫ ПОСТРЕДАКТИРОВАНИЯ МЕДИЦИНСКИХ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ И КАЧЕСТВЕННЫЙ АНАЛИЗ

Аннотация

В настоящее время в сфере медицины качественный и точный перевод с английского языка на казахский является одной из важных задач для обеспечения доступности и безопасности медицинской информации. В данной научной работе была исследована точность и эффективность систем машинного перевода, широко используемых на практике, таких как Google Translate и Yandex Translate, применительно к медицинским текстам. Основной целью исследования является рассмотрение способов семантически и стилистически корректного перевода медицинских терминов и сложных предложений с английского языка на казахский. В рамках исследования был сформирован специальный корпус, включающий 102,374 предложения, собранные из международных медицинских статей, клинических исследований и описаний лекарственных препаратов. Корпус был обучен на нейронной системе машинного перевода MarianNMT и переведен на казахский язык. Для легкого постредактирования результатов перевода использовалась трансформерная модель Kaz-RoBERTa, а для полного постредактирования – одна из больших языковых моделей (LLM), а именно GPT-4.1, при этом были рассмотрены ее возможности адаптации к медицинским текстам. Качество пере-

вода оценивалось с использованием метрик BLEU, TER и METEOR. Результаты переводов, полученные после первоначального машинного перевода MarianNMT, были сравнены с результатами после обработки моделями Kaz-RoBERTa и GPT-4.1. Анализ показал, что переводы после применения модели Kaz-RoBERTa продемонстрировали улучшение на 9% по сравнению с исходным переводом MarianNMT, а после использования модели GPT-4.1 – на 23%.

Ключевые слова: нейронный машинный перевод, MarianNMT, Kaz-RoBERTa, трансформерная модель, медицинские тексты, BLEU, TER, METEOR.

[1]Rakhimova D.,
Associate Professor, ORCID ID: 0000-0003-1427-198X,
e-mail: drakhimova060@gmail.com
[1,2,3]*Zhiger A.,
MSc, ORCID ID: 0000-0002-3641-8260,
*e-mail: aliya.zhunussova.zh@narxoz.kz
[3,4]Malykh V.,
Cand. Tech. Sc., ORCID ID: 0009-0008-5632-6188,
e-mail: valentin.malykh@phystech.edu
[1]Mansurova M.,
Cand. Phys.-Math. Sci., Professor, ORCID ID: 0000-0002-9680-2758,
e-mail: mansurova.madina@gmail.com
[1]Karyukin V.,
PhD, ORCID ID: 0000-0002-8768-0349,
email: vladislav.karyukin@gmail.com

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

METHODS OF POST-EDITING MEDICAL TEXTS IN THE KAZAKH LANGUAGE AND QUALITATIVE ANALYSIS

Abstract

At present, in the field of medicine, high-quality and accurate translation from English into Kazakh is one of the key challenges in ensuring the accessibility and safety of medical information. This scientific study investigates the accuracy and effectiveness of machine translation systems widely used in practice, such as Google Translate and Yandex Translate, when applied to medical texts. The main objective of the study is to explore methods for achieving semantically and stylistically correct translation of medical terminology and complex sentences from English into Kazakh. For this purpose, a specialized corpus consisting of 102,374 sentences was compiled from international medical articles, clinical studies, and drug descriptions. The corpus was processed using the MarianNMT neural machine translation system and translated into Kazakh. For light post-editing of the translation results, the transformer-based Kaz-RoBERTa model was employed, while full post-editing was carried out using one of the large language models (LLMs), namely GPT-4.1, whose adaptability to medical texts was also examined. Translation quality was evaluated using the BLEU, TER, and METEOR metrics. The translations obtained after the initial MarianNMT machine translation were compared with the results after post-editing using the Kaz-RoBERTa and GPT-4.1 models. The analysis showed that translations processed with the Kaz-RoBERTa model achieved an 9% improvement over the baseline MarianNMT translations, while the use of the GPT-4.1 model resulted in a 23% improvement.

Keywords: neural machine translation, MarianNMT, Kaz-RoBERTa, transformer model, medical texts, BLEU, TER, METEOR.

Received August 10, 2025; revised January 23, March 2, 17, 2026; accepted March 27, 2026.