

e-ISSN 3060-5059

DOI: 10.67227



IJTIMOYIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNAL

Scientific and Innovative Research in
Social and Humanitarian Studies

Vol. 3 • Issue 9

2026



Founded in 2024



publishscience.uz



publish_science_journal@mail.ru



PEER REVIEWED • OPEN ACCESS

KAM RESURSLI TIL JUFTLIGI SHAROITIDA TARJIMANING AVTOMATIK METRIKALARI VA EKSPERT BAHOSI

Haitova Maqsuda Kamol qizi

Buxoro shahridagi 43-sonli umumiy o'rta ta'lim maktabi, o'qituvchi.

e-mail: haitovamaqsuda@mail.ru ORCID: <https://orcid.org/0009-0006-1972-1950>

Annotatsiya.

Maqolada o'zbek va rus tillari juftligi uchun mashina tarjimasi sifatini baholash metrikalarining ishonchliligi o'rganiladi. Tadqiqot materiali sifatida uchta funksional uslubga mansub 600 ta segmentdan iborat sinov to'plami shakllantirildi. Yettita avtomatik metrika natijalari ekspertlarning MQM asosidagi baholari bilan taqqoslandi. Tahlil natijasida avtomatik metrikalar va ekspert baholari o'rtasidagi moslik darajasini aniqlash imkonini beruvchi takrorlanuvchan baholash protokoli taklif etildi.

Kalit so'zlar: mashina tarjimasi, tarjima sifatini baholash, avtomatik metrikalar, MQM, kam resursli til, o'zbek tili, rus tili, korrelyatsion tahlil.

АВТОМАТИЧЕСКИЕ МЕТРИКИ И ЭКСПЕРТНАЯ ОЦЕНКА ПЕРЕВОДА В УСЛОВИЯХ НИЗКОРЕСУРСНОЙ ЯЗЫКОВОЙ ПАРЫ

Хаитова Мақсуда Камол қизи

Общеобразовательная школа № 43 города Бухары, учитель.

Аннотация.

В статье исследуется надёжность метрик оценки качества машинного перевода для узбекско-русской языковой пары. В качестве материала исследования сформирован тестовый набор, состоящий из 600 сегментов, относящихся к трём функциональным стилям. Результаты семи автоматических метрик сопоставлены с экспертными оценками, полученными на основе MQM. В результате анализа предложен воспроизводимый протокол оценки, позволяющий определить степень соответствия между автоматическими метриками и экспертными оценками.

Ключевые слова: машинный перевод, оценка качества перевода, автоматические метрики, MQM, малоресурсный язык, узбекский язык, русский язык, корреляционный анализ.

AUTOMATIC METRICS AND EXPERT EVALUATION OF TRANSLATION IN A LANGUAGE PAIR WITH LIMITED RESOURCES

Haitova Maqsuda Kamol qizi

General Education School No. 43, Bukhara, Teacher.

Abstract.

The article examines the reliability of machine translation quality evaluation metrics for the Uzbek-Russian language pair. A test set consisting of 600 segments representing three functional styles was compiled as the research material. The results of seven automatic metrics were compared with expert evaluations based on MQM. The analysis resulted in a reproducible evaluation protocol for determining the degree of agreement between automatic metrics and expert assessments.

Keywords: machine translation, translation quality evaluation, automatic metrics, MQM, low-resource language, Uzbek language, Russian language, correlation analysis.

Kirish. Mashina tarjimasi sifatini o'lchash uchun bugun avtomatik metrikalar keng qo'llanadi. Bu metrikalar asosan ingliz va nemis, xitoy va ingliz, ingliz va rus tillari juftligida sinovdan o'tkazilgan. Ular uchun belgilangan chegaraviy qiymatlar hamda talqin qoidalari o'zbek va rus tillari juftligiga o'zgarishsiz ko'chirilmoqda. Ammo bu juftlikda metrika raqami bilan tarjimonning bahosi o'rtasida barqaror bog'liqlik bor yoki yo'qligi tekshirilmagan. Natijada amaliyotda xato yuzaga keladi. BLEU bo'yicha yuqori ball olgan tizim yaxshiroq deb topiladi. Ekspert tekshiruvchi esa ko'pincha o'sha tizimlarni teskari tartibda joylashtiradi.

So'z shakliga asoslangan metrikalarning ishonchliligiga shubha ular paydo bo'lganidan ko'p o'tmay bildirilgan. BLEU metrikasi 2002-yilda taklif qilingan edi [1; 311–318-b.]. To'rt yildan keyin bu ko'rsatkichning o'sishi tarjima sifati yaxshilanganini anglatmasligi isbotlandi [2; 249–256-b.]. Keyinchalik muammoning ikkinchi tomoni ham ochildi. BLEU qiymatlari tokenizatsiya usuliga va etalonni normallashtirish tartibiga bog'liq bo'ladi. Shuning uchun bir maqoladagi BLEU balini boshqa maqoladagi ball bilan faqat hisoblash parametrlari bir xil bo'lgan holdagina solishtirish mumkin. Bu parametrlar sacBLEU dasturida alohida qator shaklida chiqariladi va maqolada albatta keltirilishi kerak [3; 186–191-b.].

Tadqiqot obyekti o'zbek tilidan rus tiliga mashina tarjimasi sifatini baholash jarayoni hisoblanadi. Tadqiqot predmeti avtomatik metrikalar bilan ekspert bahosi o'rtasidagi moslik darajasidir. Ekspert bahosi MQM

(Multidimensional Quality Metrics, ko'p o'lchovli sifat metrikalari) sxemasi asosida beriladi. Ishning maqsadi shu moslikni o'lchaydigan takrorlanadigan protokolni asoslashdan iborat. Maqsadga erishish uchun to'rtta vazifa qo'yildi. Birinchidan, metrikalarni tekshirish bo'yicha mavjud yondashuvlar tizimlashtiriladi va turkiy hamda slavyan tillari juftligi bo'yicha to'ldirilmagan bo'shliq aniqlanadi. Ikkinchidan, sinov to'plami tarkibi, baholanadigan tizimlar ro'yxati va morfologiyasi boy tillarga mos metrikalar to'plami asoslanadi. Uchinchidan, ekspertlar ishi tartibga solinadi hamda ularning baholari qanchalik mos kelishi nazorat qilinadi. To'rtinchidan, taqqoslashning statistik usullari va xulosa chiqarish chegaralari belgilanadi.

Ishning ilmiy yangiligi — o'zbek tilidan rus tiliga tarjima yo'nalishi uchun birinchi marta avtomatik metrikalarning ekspert bahosiga mosligini o'lchaydigan takrorlanadigan protokol taklif etilayotganidir. Protokol COMET va COMET-Kiwi kabi neyroto'r metrikalarini ham qamrab oladi. U natijalarni funksional uslublar va segment uzunligi bo'yicha ajratib ko'rsatadi hamda statistik ahamiyatlilikni to'g'ri usulda hisoblaydi.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Mashina tarjimasini baholashga bag'ishlangan ishlarni besh guruhga ajratish mumkin. Guruhlar o'zaro ikki jihat bilan farq qiladi. Birinchi jihat — tarjimada nima taqqoslanishi. Ba'zi metrikalar so'zlarni, ba'zilari harflarni, ba'zilari esa ma'noni solishtiradi. Ikkinchi jihat — taqqoslash uchun nima namuna qilib olinishi. Birinchi guruh metrikalar etalon tarjimaga tayanadi, ikkinchi guruh esa bevosita inson bahosiga tayanadi. Bu guruhda BLEU [1; 311–318-b.], tahrir masofasini hisoblaydigan TER [4; 223–231-b.] va so'z o'zakkari hamda sinonimlarni qo'shimcha hisobga oladigan METEOR [5; 65–72-b.], [6; 105–115-b.] metrikalari bor. Guruhning umumiy kamchiligi shundan iboratki, sinonimik almashtirish va so'z tartibining o'zgarishi bu metrikalar uchun xato bilan bir xil ko'rinadi. Tokenizatsiya tartibiga bog'liqlik esa imzosi ko'rsatilmagan qiymatlarni taqqoslash imkonini yo'qotadi [3; 186–191-b.]. Ikkinchi guruhni harflar darajasida ishlaydigan metrikalar tashkil qiladi. chrF metrikasi so'zlarni emas, harflar zanjirini taqqoslaydi [7; 392–395-b.]. Uning kengaytirilgan varianti chrF++ bunga qisqa so'z birikmalarini ham qo'shadi [8; 612–618-b.]. Rus tili morfologiyasi boy til hisoblanadi. Shunday tillarda harfli taqqoslash so'z shakllari to'liq mos kelmagan holda ham umumiy o'zakni qayd etadi. Aynan shu sabab chrF metrikasi inson bahosiga yaqinroq natija beradi.

Uchinchi guruh ma'no tasviriga tayanadi. BERTScore etalon va tarjima matnidagi so'zlarning vektorli tasvirini taqqoslaydi [9]. COMET metrikasi bevosita inson qo'ygan baholar asosida o'qitiladi [10; 2685–2702-b.]. COMET-Kiwi varianti etalonsiz ishlaydi va faqat asl matn bilan tarjimini solishtiradi [11; 634–645-b.]. Uchala metrika XLM-R turkumidagi ko'p tilli modellar ustiga qurilgan [12; 8440–8451-b.]. Shuning uchun muayyan tilning bu modellarda qanchalik yaxshi ifodalangani o'lchov natijasiga bevosita ta'sir qiladi.

To'rtinchi guruh metrikalarni inson bahosi bilan solishtirish usullariga bag'ishlangan. Professional tarjimonlar MQM sxemasi bo'yicha bergan baholar tizimlarni oddiy kraudsorsing baholaridan butunlay boshqacha tartibda joylashtiradi. Oldindan o'qitilgan modellarga asoslangan metrikalar esa ayrim holatlarda kraudsorsing ishtirokchilaridan aniqroq ishlaydi [13; 1460–1474-b.]. WMT23 tanlovi yakunlari etalon sifatining o'zi metrikalar reytingini o'zgartirishini ko'rsatdi [14; 578–628-b.]. Statistik muammo alohida hal qilindi. Segment darajasidagi MQM ballari orasida bir xil qiymatlar juda ko'p bo'ladi. Bunday holda oddiy rang koeffitsiyentlari noto'g'ri natija beradi. Ular o'rniga teng qiymatlarni hisobga oladigan juftlik aniqligi taklif qilingan [15]. Xato turlari va ularning vazniga oid sxema esa MQM doirasida ishlab chiqilgan [16; 455–463-b.].

Beshinchi guruhga resursi kam tillar va turkiy tillar bo'yicha ishlar kiradi. NLLB-200 loyihasi 202 tilni hamda 40 000 dan ortiq tarjima yo'nalishini qamrab oldi. Loyiha BLEU bo'yicha o'rtacha 44% yaxshilanish berdi va o'zbek tilini uzn_Latn kodi bilan o'z tarkibiga kiritdi [17; 841–846-b.]. Turkiy tillar bo'yicha asosiy tadqiqot 22 tilni qamrab olgan. Unda taxminan 2 million jumladan iborat parallel korpus, 26 ikki tilli model va uchta soha uchun sinov to'plamlari yaratilgan [18; 5876–5890-b.]. Resurslar bazasi yildan yilga kengayib bormoqda. Erkin litsenziya ostida o'zbek va qozoq tillari uchun parallel korpus e'lon qilingan [19]. Qoraqalpoq tili uchun yangi resurslar taklif etilgan [20; 606–613-b.]. Qozoq va ingliz tillari yo'nalishi uchun neyroto'r modellari sinovdan o'tkazilgan [21].

O'zbek va rus tillari juftligi esa eng kam o'rganilgan bo'lib qolmoqda. Bu mavzuda o'zbek olimlarining ishlari bor, ammo ular asosan tavsifiy xarakterga ega [22]. Ularda na rasmiy MQM protokoli, na metrikalar bilan ekspert bahosi o'rtasidagi bog'liqlikni hisoblash mavjud. Demak, to'ldirilmagan bo'shliq quyidagi holatdan iborat. O'zbek tilidan rus tiliga tarjimada avtomatik metrikalar ekspert bahosiga qanchalik mos kelishini funksional uslublar va segment uzunligi bo'yicha o'lchaydigan takrorlanadigan empirik tadqiqot hozircha o'tkazilmagan.

Ishning metodologiyasi boshqa tadqiqotchi takrorlay oladigan protokol shaklida tuzildi. Sinov to'plami 600 ta o'zbekcha segmentdan iborat va zamonaviy lotin yozuvida tayyorlandi. Har bir funksional uslub uchun 200 tadan segment olindi. Bular badiiy uslub, publitsistik uslub hamda rasmiy va ilmiy uslubdir. Segment uzunligi 8 tadan 35 tagacha so'z shakli bilan chegaralanadi. Manbalar litsenziyasi albatta tekshirildi. Badiiy uslub uchun umumiy mulkka aylangan asarlar va ochiq adabiy korpuslar olindi [19]. Publitsistik uslub uchun qayta foydalanishga ruxsat beruvchi litsenziyasi aniq ko'rsatilgan yangiliklar matnlari tanlandi. Rasmiy va ilmiy uslub uchun lex.uz portalidagi hujjatlar hamda ochiq ilmiy maqolalar ishlatildi.

Matnlarni qayta ishlash bir necha bosqichda bajariladi. Avval orfografiya va kodlash normallashtirildi. So'ng matn gaplarga ajratildi va uzunligi bo'yicha filtrlendi. Keyin takrorlanuvchi segmentlar chiqarib tashlandi. Buning uchun aynan mos keladigan satrlar hamda harflar zanjiri bo'yicha o'xshash segmentlar aniqlandi. Oxirida to'plam uslublar bo'yicha muvozanatlashtirildi. Alohida majburiy bosqich sifatida o'qitish ma'lumotlari bilan kesishishni tekshirish belgilandi. Segment parchalari ochiq qidiruv tizimlarida va ma'lum korpuslarda qidirildi. FLORES-200 to'plamining uzn_Latn qismida aynan uchraydigan segmentlar chiqarib tashlandi. Yaqinda e'lon qilingan matnlarga ustunlik berildi. Barcha qarorlar texnik jurnalda qayd etib borildi.

Baholash uchun 4–5 tizim tanlandi. Ularning ikkitasi tijorat onlayn tarjimonlari bo'lib, ular uchun murojaat sanasi va dastur interfeysi versiyasi qayd etildi. Uchinchisi ochiq NLLB-200 modeli hisoblanadi. Bu model facebook/nllb-200-distilled-1.3B yoki facebook/nllb-200-3.3B konfiguratsiyasida ishga tushirildi. Asl til tegi sifatida uzn_Latn, maqsad til tegi sifatida rus_Cyrl berildi. Qolgan tizimlar katta til modellari bo'lib, ularning versiyasi qayd etilgan interfeys orqali chaqirildi. NLLB-200 o'zbek tili bilan bevosita ishlagani uchun tanlandi [17; 841–846-b.]. Boshqa ko'p tilli modellar bu jihatdan ortda qoladi. Takrorlanuvchanlik uchun generatsiya parametrlari ham qayd etildi.

NATIJALAR

Ishning hozirgi bosqichidagi natija to'liq tayyorlangan protokol hisoblanadi. Unda sinov to'plami, baholanadigan tizimlar, metrikalar va ularning dasturiy amalga oshirilishi, ekspert baholash shkalalari, moslik ko'rsatkichlari hamda natijalarni taqdim etish shakli belgilangan. Metrikalar va bog'liqlik koeffitsiyentlarining haqiqiy qiymatlari asosiy bosqichda olinadi. Jadvallarda bu qiymatlar tegishli belgi bilan ko'rsatilgan. Manbaga havola bilan keltirilgan raqamlar esa e'lon qilingan ishlardan olingan va mualliflik o'lchovi hisoblanmaydi. Sinov to'plami tarkibi ikki talab asosida belgilanadi. Birinchi talab — uslublarning teng vakilligi. Ikkinchi talab — segment uzunligining nazorat qilinishi. Hajmi teng bo'lgan qismlar bog'liqlik koeffitsiyentlarini uslublar o'rtasida qo'shimcha tuzatishsiz taqqoslash imkonini beradi. Uzunlikning quyi chegarasi 8 so'z shakli qilib olindi. Chunki undan qisqa segmentlarda yuzaki metrikalar ma'nosini yo'qotadi.

1-jadval. Sinov to'plamining funksional uslublar bo'yicha tarkibi

Funksional uslub	Segmentlar soni	Uzunligi, so'z shakli	Manbalar va litsenziya
Badiiy	200	8 dan 35 gacha	umumiy mulkka aylangan asarlar va ochiq adabiy korpuslar [19]
Publitsistik	200	8 dan 35 gacha	qayta foydalanishga ruxsat beruvchi litsenziyalik yangiliklar matnlari
Rasmiy va ilmiy	200	8 dan 35 gacha	lex.uz portalidagi hujjatlar va ochiq ilmiy maqolalar
Jami	600	8 dan 35 gacha	

O'lchov takrorlanishi uchun metrikalar ro'yxatining o'zi yetarli emas. Har bir metrika qaysi dastur bilan va qanday parametrlarda hisoblangani qayd etilishi kerak. sacreBLEU parametrlari e'lon qilinmasa, yuzaki metrikalar qiymatlarini boshqa ishlar bilan taqqoslab bo'lmaydi [3; 186-191-b.]. Neyroto'r metrikalari uchun xuddi shu vazifani model nomi va kutubxona versiyasi bajaradi. Kerakli parametrlar ikkinchi jadvalda jamlangan.

2-jadval. Metrikalar, ularning dasturiy amalga oshirilishi va qayd etiladigan parametrlar

Metrika	Turi	Dastur va model	Qayd etiladigan parametr
BLEU	so'z shakliga asoslangan	sacreBLEU	BLEU nrefs:1 case:mixed eff:no tok:13a smooth:exp version:2.4.0
chrF++	harf va so'z asosidagi	sacreBLEU	chrF2++ nrefs:1 case:mixed eff:yes nc:6 nw:2 space:no version:2.4.0
TER	tahrir masofasi	sacreBLEU	TER tok:tercom norm:no punct:yes asian:no version:2.4.0
METEOR	o'zak va sinonimni hisobga oladigan	rus tili resurslari bilan	til resurslari versiyasi va moslashtirish parametrlari
BERTScore	ma'no tasviriga asoslangan	xlm-roberta-large, rescale_with_baseline	xesh xlm-roberta-large_L17_no-idf_version=0.3.12-rescaled
COMET	o'rgatiladigan, etalonli	Unbabel/wmt22-comet-da	unbabel-comet versiyasi 2.0.0 dan past emas
COMET-Kiwi	o'rgatiladigan, etalonsiz	Unbabel/wmt22-cometkiwi-da	unbabel-comet versiyasi 2.0.0 dan past emas

Ekspert baholash yopiq xato ro'yxati va uch pog'onali og'irlik shkalasi orqali tartibga solinadi. Yopiq ro'yxat ayrim ekspertlarning o'zicha yangi toifa qo'shishiga yo'l qo'ymaydi. Shu tufayli xato profillari tizimlar o'rtasida taqqoslanadigan bo'ladi. Adekvatlikning umumiy bahosi esa mustaqil uzluksiz ko'rsatkich beradi va Krippendorff alfasini hisoblash uchun ishlatiladi.

3-jadval. Ekspert baholash toifalari va og'irlik shkalasi

Guruh	Xato toifalari	Shkala va vazn
Aniqlik	ma'no buzilishi, tushirib qoldirish, ortiqcha qo'shish, tarjima qilinmagan parcha	kichik xato 1 ball, jiddiy xato 5 ball, kritik xato 10 ball

Ravonlik	grammatika, orfografiya, punktuatsiya, uslubiy qatlam	kichik xato 1 ball, jiddiy xato 5 ball, kritik xato 10 ball
Terminologiya	qabul qilingan atamaga mos kelmaslik, atamani turlicha qo'llash	kichik xato 1 ball, jiddiy xato 5 ball, kritik xato 10 ball
Uslub	uslubiy qatlamning buzilishi, ma'no urg'usining siljishi	kichik xato 1 ball, jiddiy xato 5 ball, kritik xato 10 ball
Mahalliy qoidalar	sana, son va atoqli otlarni yozish shakli	kichik xato 1 ball, jiddiy xato 5 ball, kritik xato 10 ball

Ekspert qismining ishonchliligi bog'liqlikni hisoblashdan oldin tekshiriladi. Chegaralar oldindan belgilanadi. Shunda baholash yaroqliligi haqidagi qaror olingan natijaga bog'liq bo'lmaydi. Sinov bosqichida chegaraga erishilmasa, ko'rsatma qayta ishlanadi va kalibrlash takrorlanadi.

4-jadval. Ekspertlar bahosining mosligi va qabul qilish chegaralari

Ko'rsatkich	Nima uchun kerak	Qabul qilish chegarasi
Krippendorf alfasi	adekvatlikning umumiy baholari	0,667 dan past bo'lmasa maqbul, 0,80 dan past bo'lmasa ishonchli
Fleyss kappasi	uch va undan ortiq ekspertda xato turlarini belgilash	umume'tirof etilgan shkala bo'yicha talqin qilinadi
Sinflararo korrelyatsiya ko'effitsiyenti	ekspertlar bo'yicha o'rtachalangan MQM ballari	0,50 dan past bo'lsa past, 0,50 va 0,75 orasida o'rtacha, 0,75 va 0,90 orasida yaxshi, 0,90 dan yuqori bo'lsa a'lo [24; 155-163-b.]

MUHOKAMA

Metrika raqami bilan ekspert bahosi o'rtasidagi farq tasodifiy emas. U metrikalarning tuzilishidan va tillarning tipologik xususiyatlaridan kelib chiqadi. Bunday farqning to'rtta manbasini ajratish mumkin.

Birinci manba rus tilining morfologik boyligi hisoblanadi. Bitta so'z o'nlab shaklga ega bo'ladi. Tarjima to'g'ri bo'lsa ham, unda etalondagidan boshqa kelishik yoki zamon shakli ishlatilishi mumkin. So'z shakliga asoslangan metrika buni xato deb qabul qiladi. Harfli taqqoslash esa umumiy o'zakni ko'radi va jarimani kamaytiradi. chrF hamda chrF++ metrikalarining morfologiyasi boy tillardagi ustunligi shundan kelib chiqadi [7; 392–395-b.], [8; 612–618-b.]. Bundan amaliy xulosa hosil bo'ladi. Ko'rilayotgan yo'nalishda chrF++ asosiy ko'rsatkich bo'lishi, BLEU esa yordamchi ko'rsatkich sifatida qolishi kerak.

Ikkinchi manba asl til bilan bog'liq. O'zbekcha so'zlar son, egalik, kelishik va sifatlovchi ma'nolarini bitta shaklda to'playdi. Rus tilida esa shu ma'nolar bir necha so'zga yoyiladi. Faqat tarjima matni bilan ishlaydigan metrikalar uchun bu farq ko'rinmaydi. Ammo COMET asl matnni ham hisobga oladi, COMET-Kiwi esa butunlay asl matnga tayanadi. Shuning uchun o'zbekcha matnning ko'p tilli modeldagi ifodasi qanchalik aniq bo'lsa, o'lchov ham shunchalik ishonchli bo'ladi. O'zbek tili XLM-R modelining o'qitish ma'lumotlarida bor [12; 8440–8451-b.]. Lekin bu ma'lumotlarning hajmi yirik tillarnikidan ancha kam. Shu sababli neyroto'r metrikalari aynan asl til tomonida kamroq ishonchli bo'lishi kutiladi.

Uchinchi manba so'z tartibining erkinligidir. O'zbek tilidagi gapda asosiy tartib ega, to'ldiruvchi va kesim shaklida bo'ladi. Ammo ma'no urg'usiga ko'ra so'zlar o'rin almashishi mumkin. Rus tilida bu erkinlik yanada keng. Natijada bitta jumlaning bir xilda to'g'ri bo'lgan tarjimalari juda ko'p bo'ladi. Metrika esa ularning faqat bittasi bilan solishtiradi.

To'rtinchi manba etalonning yagonaligidir. Bitta etalon mumkin bo'lgan variantlardan faqat bittasini qamrab oladi. Tarjima boshqa teng qiymatli variantga yaqin bo'lsa, metrika buni yo'qotish deb hisoblaydi. Tadqiqotlar etalon tarkibi va sifati metrikalar reytingini o'zgartirishini ko'rsatgan [14; 578–628-b.]. Protokolda ikkita mustaqil etalon talab qilinishining sababi ham shu.

Birinci holatda metrika bahoni asossiz pasaytirdi. Asl matn quyidagicha: Yig'ilishda ko'rib chiqilgan masalalar yuzasidan qaror qabul qilindi. Etalon tarjima «По вопросам, рассмотренным на собрании, было принято решение» shaklida berilgan. Mashina esa «По рассмотренным на заседании вопросам принято решение» deb tarjima qilgan. Tarjima to'g'ri, MQM bo'yicha xato yo'q. Ammo etalonga nisbatan sinonim ishlatilgan. Bu uch farq mos keladigan so'z zanjirlari sonini kamaytiradi. Natijada ekspert jarimasi nolga teng bo'lsa ham, metrika past ball beradi.

Ikkinchi holatda, aksincha, metrika bahoni asossiz oshirdi. Asl matn quyidagicha: U bu haqda hech narsa demadi. Etalon tarjima «Он ничего об этом не сказал» shaklida beriladi. Xato bilan olingan faraziy mashina tarjimasi «Он сказал об этом» ko'rinishida. Bo'lishsizlik qo'shimchasi va bo'lishsiz olmosh tushib qolgani uchun gap ma'nosi teskarisiga o'zgargan. MQM bo'yicha bu eng og'ir vaznga ega kritik xato hisoblanadi. Shunga qaramay, yuzaki moslik yuqori bo'lib qoladi. Ega ham, kesim ham saqlangan. Bunday xato birinchisidan xavfliroq. Past baho yaroqli tizimni rad etishga olib keladi. Asossiz yuqori baho esa ma'noni buzadigan tizimni ishga tushirishga olib keladi.

Uchinchi holat gapdagi ma'no urg'usiga tegishli. Asl matn quyidagicha: Bu masalani kecha rais hal qildi. Etalon tarjima «Этот вопрос вчера решил председатель» shaklida beriladi. Mashina esa «Председатель вчера решил этот вопрос» deb tarjima qiladi. So'zlar tarkibi to'liq mos keladi, faqat tartib boshqacha. Shu sababli yuzaki metrika yuqori ball beradi. Ekspert esa ma'no urg'usi siljiganini ko'radi va buni uslub xatosi deb belgilaydi. Ma'no tasviriga asoslangan metrikalar bunday xatoga sezgirroq. Ammo ular asosan yirik tillar ma'lumotlarida o'qitilgan [13; 1460–1474-b.]. Shuning uchun ularning sezgirligi o'zbek va rus tillari juftligida alohida tekshirilishi kerak. Ko'rib chiqilgan holatlar statistik usullar tanlovini ham izohlaydi. Qisqa segmentlarning ko'pchiligidagi xato bo'lmaydi va ular nol ball oladi.

Natijada ekspert baholari orasida bir xil qiymatlar juda ko'payadi. Bunday sharoitda oddiy rang koeffitsiyentlari ishonchsiz bo'lib qoladi. Shuning uchun teng qiymatlarni hisobga oladigan juftlik aniqligi majburiy tarzda qo'llanadi [15]. Shu sababdan bog'liqlik segment uzunligi bo'yicha ham alohida hisoblanadi. Agar metrikalar qisqa segmentlarda ekspertdan ko'proq uzoqlashsa, umumiy o'rtacha koeffitsiyent buni yashirib qo'yadi. Avtomatik baholashda asosiy ko'rsatkich sifatida chrF++ va COMET ishlatiladi. Etalon bo'lmagan holatlarda ularga COMET-Kiwi qo'shiladi. BLEU faqat hisoblash parametrlari bilan birga va faqat boshqa ishlar bilan taqqoslash uchun keltirildi. Ekspert baholashda MQM sxemasi yopiq xato ro'yxati hamda 1, 5 va 10 vaznlari bilan qo'llandi. Sinov kalibrilashi o'tkazildi. Sinflararo korrelyatsiya koeffitsiyenti kamida 0,75, Krippendorff alfasini kamida 0,667 bo'lishi talab qilindi. Metrikalarning ekspertga mosligi teng qiymatlarni hisobga oladigan juftlik aniqligi orqali baholandi. Bunda butstrep oraliqlari qurildi va ko'p taqqoslash uchun tuzatish qo'llandi.

FOYDALANILGAN ADABIYOTLAR

1. Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation // *Proceedings of ACL 2002*. – Philadelphia, 2002. – P. 311–318. – [Source](#)
2. Callison-Burch C., Osborne M., Koehn P. Re-evaluating the Role of BLEU in Machine Translation Research // *Proceedings of EACL 2006*. – Trento, 2006. – P. 249–256. – [Source](#)
3. Post M. A Call for Clarity in Reporting BLEU Scores // *Proceedings of WMT 2018*. – Brussels, 2018. – P. 186–191. – [Source](#)
4. Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J. A Study of Translation Edit Rate with Targeted Human Annotation // *Proceedings of AMTA 2006*. – Cambridge, 2006. – P. 223–231. – [Source](#)
5. Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments // *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*. – Ann Arbor, 2005. – P. 65–72. – [Source](#)
6. Lavie A., Denkowski M. The METEOR Metric for Automatic Evaluation of Machine Translation // *Machine Translation*. – 2009. – Vol. 23, No. 2–3. – P. 105–115. – DOI: [10.1007/s10590-009-9059-4](#)
7. Popović M. chrF: Character n-gram F-score for Automatic MT Evaluation // *Proceedings of WMT 2015*. – Lisbon, 2015. – P. 392–395. – [Source](#)
8. Popović M. chrF++: Words Helping Character n-grams // *Proceedings of WMT 2017*. – Copenhagen, 2017. – P. 612–618. – [Source](#)
9. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT // *Proceedings of ICLR 2020*. – Addis Ababa, 2020. – [Source](#)
10. Rei R., Stewart C., Farinha A. C., Lavie A. COMET: A Neural Framework for MT Evaluation // *Proceedings of EMNLP 2020*. – Online, 2020. – P. 2685–2702. – [Source](#)
11. Rei R., Treviso M., Guerreiro N. M. et al. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task // *Proceedings of WMT 2022*. – Abu Dhabi, 2022. – P. 634–645. – [Source](#)
12. Conneau A., Khandelwal K., Goyal N. et al. Unsupervised Cross-lingual Representation Learning at Scale // *Proceedings of ACL 2020*. – Online, 2020. – P. 8440–8451. – [Source](#)
13. Freitag M., Foster G., Grangier D., Ratnakar V., Tan Q., Macherey W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation // *Transactions of the Association for Computational Linguistics*. – 2021. – Vol. 9. – P. 1460–1474. – [Source](#)
14. Freitag M., Mathur N., Lo C. et al. Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent // *Proceedings of WMT 2023*. – Singapore, 2023. – P. 578–628. – [Source](#)
15. Deutsch D., Foster G., Freitag M. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration // *Proceedings of EMNLP 2023*. – Singapore, 2023. – [Source](#)
16. Lommel A., Uszkoreit H., Burchardt A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics // *Tradumática*. – 2014. – No. 12. – P. 455–463. – DOI: [10.5565/rev/tradumatica.77](#)
17. NLLB Team, Costa-jussà M. R. et al. Scaling Neural Machine Translation to 200 Languages // *Nature*. – 2024. – Vol. 630. – P. 841–846. – DOI: [10.1038/s41586-024-07335-x](#)
18. Mirzakhlov J., Babu A., Ataman D. et al. A Large-Scale Study of Machine Translation in the Turkic Languages // *Proceedings of EMNLP 2021*. – Punta Cana, 2021. – P. 5876–5890. – [Source](#)
19. Allaberdiev B., Matlatipov G., Kuriyozov E., Rakhmonov Z. Parallel Texts Dataset for Uzbek-Kazakh Machine Translation // *Data in Brief*. – 2024. – Vol. 53. – Article 110194. – DOI: [10.1016/j.dib.2024.110194](#)
20. Mamasaidov M., Shopulatov A. Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak // *Proceedings of WMT 2024*. – Miami, 2024. – P. 606–613. – [Source](#)
21. Karyukin V., Rakhimova D., Karibayeva A. et al. The Neural Machine Translation Models for the Low-Resource Kazakh-English Language Pair // *PeerJ Computer Science*. – 2023. – Vol. 9. – Article e1224. – DOI: [10.7717/peerj-cs.1224](#)
22. Тураева Г. Х. Проблемы машинного перевода при переводе на узбекский язык // *Universum: технические науки*. – 2020. – № 10 (79). – [Источник](#)
23. ISO 17100:2015. *Translation Services. Requirements for Translation Services*. – Geneva: ISO, 2015. – [Source](#)
24. Koo T. K., Li M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research // *Journal of Chiropractic Medicine*. – 2016. – Vol. 15, No. 2. – P. 155–163. – DOI: [10.1016/j.jcm.2016.02.012](#)