QAYTA SOZLANGAN TRANSFORMER MODELLAR YORDAMIDA INGLIZCHA-O‘ZBEKCHA MASHINA TARJIMASINI TAKOMILLASHTIRISH

IMPROVING ENGLISH-UZBEK MACHINE TRANSLATION WITH FINE-TUNED TRANSFORMER MODELS

УЛУЧШЕНИЕ АНГЛО-УЗБЕКСКОГО МАШИННОГО ПЕРЕВОДА С ПОМОЩЬЮ ТОНКО НАСТРОЕННЫХ МОДЕЛЕЙ ТРАНСФОРМЕРОВ

Ph.D U.Y.Tuliyev[1], A. M. Murodov [2],
O’zbekiston Milliy universiteti,
Toshkent, O’zbekiston
E-mail: [1]u.tuliyev@gmail.com, [2]murodov_a@nuu.uz

Annotatsiya. Tabiiy tilga tegishli matnlarni aniq va sifatli tarjima qilish vazifasi tabiiy tilga ishlov berishning (NLP) o‘ta muhim vazifalaridan biri hisoblanadi. Mashinaviy o‘rganish yordamida aniq va tez amalga oshiriladigan avtomatlashtirilgan tarjima ko‘pincha mashinaviy o‘rganish va sun’iy intellekt fanlari hamjamiyatlarida katta qiziqish uyg‘otadi. Ushbu tadqiqot doirasida mashina tarjimasini amalga oshirish uchun o’zbek tilining mahalliy internet manbaalaridan, kitob, darslik va ilmiy ishlardan yig’ilgan matnlarning parallel korpusi yordamida Generative Pretrained Transformer (GPT) modellaridan foydalanishni ko‘rib chiqamiz. Biz Hugging Face katta til modellari (LLM) ro’yxatidagi o’zbek tili uchun o’qitilgan 2 xil modelni qayta sozlash orqali mashina tarjimasini amalga oshirish va ularning turli xil metrikalar bo’yicha tarjima sifatining baholarini qiyosiy tahlil qilamiz. Tanlangan modellarni sozlash Google colab muhitida A100 Grafikani qayta ishlash bloki (GPU) yordamida amalga oshirilgan.

Kalit so’zlar: Katta til modellari, Hugging Face, Mashina tarjimasi, Generative Pretrained Transformers, Tabiiy tilni qayta ishlash

Abstract. The task of accurate and high-quality translation of texts related to natural language is one of the most important tasks of natural language processing (NLP). Automated translation, performed accurately and quickly with the help of machine learning, is often of great interest in the scientific communities of machine learning and artificial intelligence. Within the framework of this study, we will consider the use of Generative Pretrained Transformer (GPT) models for machine translation using a parallel corpus of texts collected from local Internet sources of the Uzbek language, books, textbooks, and scientific works. We carry out machine translation by reconfiguring 2 different models trained for the Uzbek language from the list of Hugging Face Large Language Models (LLM) and conduct a comparative analysis of their translation quality assessments according to various metrics. The configuration of the selected models was carried out using the A100 Graphics Processing Unit (GPU) in the Google colab environment.

Keywords: Large Language Models, Hugging Face, Machine Translation, Generative Pretrained Transformers, Natural Language Processing

Аннотация. Задача точного и качественного перевода текстов, связанных с естественным языком, является одной из важнейших задач обработки естественного языка (ОЕЯ). Автоматизированный перевод, выполняемый точно и быстро с помощью машинного обучения, часто представляет большой интерес в научных сообществах машинного обучения и искусственного интеллекта. В рамках данного исследования мы рассмотрим использование моделей Generative Pretrained Transformer (GPT) для машинного перевода с использованием параллельного корпуса текстов, собранных из местных интернет-источников узбекского языка, книг, учебников и научных трудов. Мы осуществляем машинный перевод путем перенастройки 2 различных моделей, обученных узбекскому языку, из списка Hugging Face Large Language Models (LLM) и проводим сравнительный анализ их оценок качества перевода по различным показателям. Конфигурация выбранных моделей осуществлялась с помощью A100 Graphics Processing Unit (GPU) в среде Google colab.

Ключевые слова: Большие языковые модели (LLM), Hugging Face, Машинный перевод, Generative Pretrained Transformers, Обработка естественного языка

1. Kirish

Mashina tarjimasi (MT) chuqur o‘rganish, xususan, transformer modellarga [3,6] asoslangan arxitekturalardagi yutuqlar tufayli sezilarli yutuqlarga erishdi. Biroq, o‘zbek tili kabi kam resursli tillar tarjimasi ko‘pincha katta va sifatli parallel korpuslarning yo‘qligi sababli ortda qoladi. Ushbu tadqiqot tarjima samaradorligini sezilarli oshirish uchun ingliz-o‘zbek korpusini to‘plash va ushbu ma’lumotlar asosida oldindan o‘qitilgan modellarni qayta sozlash orqali ushbu bo‘shliqni to‘ldirishga qaratilgan. 35 milliondan ortiq aholi so‘zlashadigan turkiy til bo‘lgan o‘zbek tili yuqori sifatli neyron mashina tarjimasi (NMT) uchun yetarli resurslarga ega emas. Helsinki-NLP [4,7] va M2M100 [2] kabi transformer modellar ko‘p tilli tarjimani qo‘llab-quvvatlasa-da, ularning ingliz-o‘zbek tilidagi samaradorligi hali ham kam o‘rganilgan. Ushbu maqolada yuqorida ta’kidlangan bo‘shliq quyidagicha to‘ldiriladi:

1. Mahalliy saytlar, kitoblar va ilmiy maqolalardan 50K inglizcha-o‘zbekcha parallel korpusni tahrirlash.
2. Ushbu ma’lumotlar to‘plamida Helsinki-NLP [4,7] va M2M100 [2] modellarini qayta sozlash.
3. BLEU [5] va METEOR [8] ko‘rsatkichlari yordamida tarjima sifatini baholash.
4. Natijalarni Google Tarjima [1] bilan taqqoslash.

Bizning natijalarimiz kam resursli tillar uchun NMT tadqiqotlariga hissa qo‘shadi va ingliz-o‘zbek tarjima tizimlarini joriy etish uchun amaliy tushunchalar beradi.

2. Muammoning o’rganilganligi

Kam resursli tillar uchun MT bo‘yicha oldingi tadqiqotlar tarjima aniqligini oshirish uchun transformer modellar va ko‘p tilli tarjimonlarni oldindan o‘qitishdan foydalangan. Helsinki-NLPning MarianMT [4] va Facebookning M2M100 [2] kabi loyihalari bir nechta tillarga tarjima qilish uchun asos yaratadi, ammo keng ko‘lamli

baholashlarda o‘zbek tili hali ham quyi o‘rin egallaydi. Ushbu tadqiqot yo‘naltirilgan ma’lumotlar to‘plami bilan maqsadli qayta sozlashni amalga oshirish orqali ushbu sa’y-harakatlarga asoslanadi.

3. Berilganlar bazasi (Dataset)

Tadqiqot doirasida quyidagi manbaalardan 50 000 juft gapdan iborat parallel korpus tuzildi:

• O‘zbekiston mahalliy veb-saytlari

• Kitoblar (adabiy-ma’rifiy)

• Ikki tilli formatdagi ilmiy maqolalar

Korpus filolog mutaxassislar tomonidan qo‘lda tozalandi va tahrirlandi, bu esa lingvistik izchillik va muvozanatli domen aralashmasini ta’minladi. Quyida berilganlar bazasining batafsil statistik tahlili, uning lingvistik xususiyatlari va umumiy tuzilishi haqida ma’lumotlar berilgan.

Ingliz tilidagi gaplar uchun har bir gapdagi so‘zlarning o‘rtacha soni **19,7** ta bo‘lib, eng qisqa gapda **1 ta so‘z**, eng uzun gapda esa **282** tagacha so‘z mavjud.

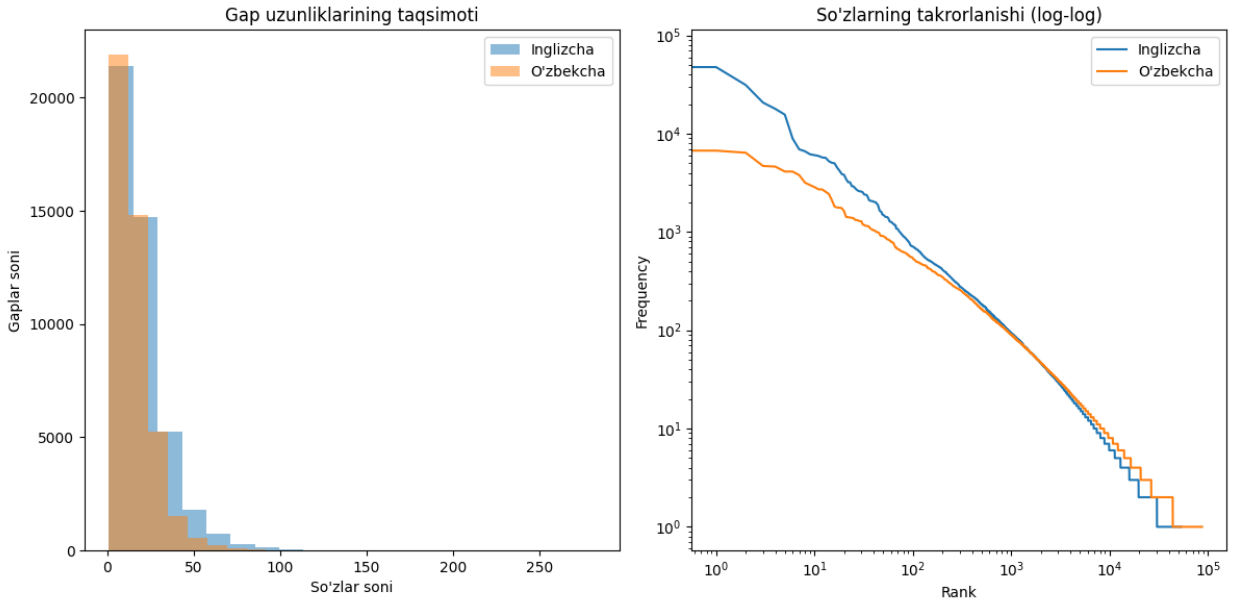
Taqqoslaganda, o‘zbek tilidagi gaplar biroz ixchamroq bo‘lib, gapning o‘rtacha uzunligi **15,4 so‘zni** tashkil etadi, eng qisqa gapda **1 ta so‘zdan** eng uzun gapda **228** tagacha so‘z bor.

Leksik statistika bo’yicha quyidagicha natijalarni ta’kidlash mumkin. Ingliz tilidagi so‘zlarning umumiy soni 876875 ta bo‘lib, ular 53601 ta o‘ziga xos so‘z turlari bo‘yicha taqsimlangan. Butun berilganlar bazasida har bir so‘z o‘rtacha 16,4 marta uchraydi. Eng ko‘p uchraydigan inglizcha so‘zlarning beshtaligi quyidagilar:

’the’ (65 345), ’of’ (47 590), ’and’ (31 273), ’to‘ (20 722), ’in’ (17 874).

O‘zbekcha qismida jami 684063 ta so‘z mavjud bo‘lib, bu so‘z boyligi o’zbek tilida boyroq ekanligini ko’rsatadi. Lekin bu o’zbek tilidagi so’zlarga qo’shimchalar qo’shilish xususiyati bilan bog’liq. Ya’ni o’zbek tilida bitta so’zga bir nechta qo’shimchalar qo’shilishi orqali yangi tokenlar hosil bo’lishiga olib keladi. Bazadagi o’zbek tili gaplarida jami 87472 ta o‘ziga xos so‘zlar mavjud. So‘zlarning o‘rtacha chastotasi 7,8 martani tashkil etadi. O‘zbek tilidagi eng keng tarqalgan beshta so‘z:

va (17489), bilan (6720), - (6403), bu (4694), ham (4626).



1-rasm. Parallel korpus statistikasi bo'yicha tahlil natijalari.

4. Masalani yechish metodologiyasi

4.1 Transformer modellar

• Helsinki-NLP/opus-mt-en-uz: Ko'p tilli ma'lumotlar asosida oldindan o'qitilgan MarianMT modeli.

• Facebook/m2m100_418M: Ko'pdan-ko'p tarjima vazifalariga o'rgatilgan ko'p tilli transformator modeli.

4.2 Modellarni qayta sozlash (Fine-Tuning)

• Platforma: A100 GPU bilan Google Colab Pro+

• Kutubxona: Hugging Face transformatorlari

• Mashg'ulot bosqichlari: 25 000

• Baholash chastotasi: har 5000 qadamda

• Tokenizatsiya: SentencePiece (ikkala model uchun)

• Optimizator: Adam isinish jadvali bilan

5. Tarjima sifatini baholash

5.1 Baholash ko'rsatkichlari

• BLEU [5]: n-gramm ustma-ust tushish o‘lchovlari

• METEOR [8]: Sinonimlar mosligi va kelib chiqishi bilan semantik o‘xshashlikni o‘lchaydi

5.2 Qiyosiy tahlil

• Asosiy tizim sifatida Google Tarjima tarjimalaridan foydalanildi.

5.3 Natijalar

Model	BLEU	METEOR
Helsinki (Fine-Tuned)	0.14	45.4
M2M100 (Fine-Tuned)	0.18	52.6
Google Translate	0.08	17.3

6. Tadqiqot natijalarining tahlili

Ikkala modelning qayta sozlanishi Google Tarjimaga nisbatan sezilarli samaradorlikni oshirdi, ayniqsa domenga xos va morfologik jihatdan murakkab jumlalarda. MarianMT modeli tezroq konvergensiyani ko‘rsatdi, M2M100 esa ko‘p tilli o‘qitish asosi tufayli yaxshiroq umumlashtirishni taklif qildi.

Ba’zi muammolar oldindan o‘rgatilgan ma’lumotlar va o‘zbek tiliga xos sintaksis o‘rtasidagi domen nomuvofiqligini o‘z ichiga oldi. Shunga qaramay, korpusning sifati va xilma-xilligi turli janrlarda kuchli modellik ijrosiga hissa qo‘shdi.

7. Xulosa

Ushbu tadqiqot maxsus parallel korpuslardan foydalangan holda kam resursli tarjima vazifalari uchun transformator modellarini nozik sozlashning samaradorligini ko‘rsatadi. MarianMT va M2M100 modellari ham tijorat asoslaridan ustunlik qiladi, bu esa domenga xos ma’lumotlar va sinchkovlik bilan baholash o‘zbek tili kabi kam ifodalangan tillar uchun tarjima sifatini sezilarli darajada oshirishi mumkinligini ko‘rsatadi.

• Korpus hajmini 500k+ gacha kengaytirish

- Teskari tarjima va sintetik ma’lumotlarni kiritish
- Qardosh turkiy tillar bilan ko‘p tilli o‘qitish strategiyalarini o‘rganish
- Ravonlik va yetarlilik uchun insonni baholash

Foydalanilgan adabiyotlar

1. Amilia, Ika & Yuwono, Darmawan. (2020). A study of the translation of google translate. Lingua : jurnal ilmiah. 16. 1-21. 10.35962/lingua.v16i2.50.
2. Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. J. Mach. Learn. Res. 22, 1, Article 107 (January 2021), 48 pages.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
4. Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.
5. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
6. Liu et al., "Multilingual Denoising Pre-training for Neural Machine Translation", 2020.
7. Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki Neural Machine Translation System. In Proceedings of

the Second Conference on Machine Translation, pages 338–347, Copenhagen, Denmark. Association for Computational Linguistics.

8. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.