PARALLEL KORPUSLARNI TARJIMA XOTIRASIGA INTEGRATSIYA QILISH TAMOYILLARI

Shamsiyeva Gulshoda Asliddin qizi,
O‘zMU tadqiqotchisi
shamsiyeva_g@nuu.uz
O‘zMU

Annotatsiya. Mashina tarjimasi sohasida parallel korpuslardan foydalanish an’anaviy tarjima jarayonini sezilarli darajada osonlashtiradi. Tarjima xotirasi (Translation Memory, TM) tizimlari tarjima jarayonini avtomatlashtirish va samaradorligini oshirish uchun muhim vosita hisoblanadi. Ushbu maqolada parallel korpuslarni tarjima xotirasiga integratsiya qilish tamoyillari, metodologiyasi va samaradorligi tahlil qilinadi.

Abstract. The utilization of parallel corpora in the field of machine translation significantly streamlines the traditional translation process. Translation Memory (TM) systems are considered crucial tools for automating and enhancing the efficiency of the translation process. This article analyzes the principles, methodology, and effectiveness of integrating parallel corpora into translation memory.

Аннотация. Использование параллельных корпусов в области машинного перевода значительно облегчает традиционный процесс перевода. Системы памяти переводов (Translation Memory, TM) являются важным инструментом для автоматизации и повышения эффективности процесса перевода. В данной статье анализируются принципы, методология и эффективность интеграции параллельных корпусов в память переводов.

Kalit so‘zlar: *tarjima xotirasi, parallel korpus, filtrlash, moslashtirish, indeksatsiya qilish.*

Tarjima xotirasi – bu tarjima jarayonida avvalgi tarjimalarni saqlab qoladigan va takroriy matnlarni avtomatik tarjima qilish imkonini beradigan dasturiy ta’minotdir [1]. Parallel korpus esa turli tillardagi tarjima qilingan matnlarning muvofiqlashtirilgan bazasi bo‘lib, mashina tarjimasi uchun asosiy resurslardan biridir [2]. Zamonaviy tarjima jarayonida ushbu ikki element birgalikda ishlatilishi natijasida tarjima sifati va tezligi sezilarli darajada oshadi.

Bir qancha olimlar, xususan, H.L.Somers [3], A.Way [4], P.Koehn [5], B.Maegaard [6], D.Marcu, W.Wong [7], M.Federico [8], A.Toral [9] v.b. avtomatik tarzda parallel korpuslardan tarjima xotirasini yaratish haqida izlanishlar olib borgan

bo‘lib, ularning asarlarida tarjima xotiralari qanday ishlashi va parallel korpuslardan qanday foydalanilish haqida chuqur tahlillar berilgan, shuningdek, ularda parallel korpus va tarjima xotirasining moslashuvi, parallel korpuslardan foydalanib tarjima modellari yaratish hamda tarjima xotirasini statistika asosidagi mashina tarjimasi bilan integratsiya qilish usullari haqida so‘z boradi.

Tarjima xotiralari zamonaviy tarjimonlar va tarjima agentliklari uchun katta ahamiyatga ega bo‘lib, ular takroriy iboralarni aniqlash, aniq va izchil tarjimalarni ta’minlash uchun foydalaniladi. Shu bilan birga, parallel korpuslar sun’iy intellektga asoslangan tarjima tizimlari uchun asosiy ma’lumot bazasi sifatida xizmat qiladi.

Parallel korpuslarni tarjima xotirasiga integratsiya qilish quyidagi bosqichlarda amalga oshirildi:

1. Korpus yig‘ish va tozalash (filtrlash) – turli manbalardan tarjima matnlarini yig‘ish va strukturaga keltirish (maslan, KazParC). Korpus sifati tarjima jarayonining asosiy elementlaridan biri bo‘lib, ularning sifatini nazorat qilish, jumladan, noaniqliklarni kamaytirish, noto‘g‘ri moslangan segmentlarni chiqarib tashlash va umumiy sifatsiz ma’lumotlarni filtrlash uchun maxsus tozalash algoritmlaridan foydalaniladi.

1.1. Lingvistik tozalash (filtrlash) (Linguistic Filtering). Bunday tozalash usulida “Google Compact Language Detector (CLD3)” yoki “langid.py” kabi vositalardan foydalanilib, ularning vazifasi parallel korpus ichida noto‘g‘ri tillarda kiritilgan matnlarni chiqarib tashlash (tilni aniqlash – Language Identification), belgilar ketma-ketligi, masalan, @@@@@ , XXXXX yoki noto‘g‘ri formatlangan matnlarni olib tashlash (shovqinli ma’lumotlarni filtrlash – Noise Removal) hamda bitta til segmentining tarjima qilingan varianti bilan sintaktik va semantik bog‘lanmaganligi aniqlansa, ushbu segmentni chiqarib tashlash (mantiqiy bog‘lanmagan tarjimalarni aniqlash – Mismatched Translations) dan iborat [10].

1.2. Masofaga asoslangan tozalash (Distance-Based Filtering). Ushbu metod yordamida parallel segmentlarning o‘lchami va muvofiqligi tekshiriladi. Bunda asl va tarjima qilingan jumla orasidagi o‘zgarishlar sonini hisoblab, juda katta farqlarni filtrlash uchun “Levenshtein masofasi (Levenshtein Distance)”, tarjima segmentlarining vektorli ifodalariga asoslangan o‘zaro o‘xshashlikni tekshirish uchun “Kosinus o‘xshashligi (Cosine Similarity)” algoritmlaridan foydalaniladi [5,11].

1.3. Statistika asosida tozalash (Statistical Filtering) usulida tarjima sifati statistik usullar orqali baholanadi. Masalan,

• BLEU Score Filtering – parallel segmentlarning BLEU skori hisoblanadi, agar u juda past bo‘lsa (masalan, 0.1 dan past bo‘lsa), segment olib tashlanadi.

• Language Model Perplexity (Til modelining chalkashligi) – agar tarjima korpusining perplexity ko‘rsatkichi juda baland bo‘lsa, ya’ni, tarjima natijalarining kutilmagan va noto‘g‘ri bo‘lish ehtimoli katta bo‘lsa, segment chiqarib tashlanadi.

• TF-IDF Filtering usuli kam ishlatiladigan yoki juda tez-tez uchraydigan so‘zlar bo‘yicha segmentlarni filtrlash uchun ishlatiladi [12].

1.4. Neyron tarmoqlar asosida tozalash (Neural Filtering) da chuqur o‘rganish (deep learning) asosida noto‘g‘ri parallel segmentlarni avtomatik filtrlash amalga oshiriladi [13]. Parallel korpus sifati tarjima jarayonida hal qiluvchi ahamiyatga ega. Korpus sifatini nazorat qilish uchun tilshunoslik, statistika va neyron tarmoqlarga asoslangan tozalash algoritmlari qo‘llaniladi. Bu algoritmlar orqali noto‘g‘ri tarjimalar, shovqinli ma’lumotlar va semantik jihatdan mos kelmaydigan segmentlar filtrlab tashlanadi.

2. Parallel korpuslarni tarjima xotirasiga integratsiya qilishning keyingi bosqichi moslashtirish va indekslashtirishdir. Parallel korpuslarda tarjima sifatini oshirish va tarjima xotiralarining samaradorligini ta’minlash uchun bu jarayonlar muhim ahamiyat kasb etadi. Parallel matn segmentlarini bog‘lash va ularni tezkor qidirish imkoniyatini ta’minlash orqali tarjima tizimlarining natijalari sezilarli darajada yaxshilanadi. Moslashtirishning asosiy maqsadi har bir tarjima segmentini uning asl nusxasi bilan sintaktik, semantik va pragmatik jihatdan bog‘lashdan iborat. Buning uchun quyidagi metodlardan foydalaniladi:

2.1. To‘g‘ridan-to‘g‘ri moslashtirish (Alignment). Moslashtirishning bu turida manba tilidagi jumlalar va ularning tarjimalari bir-biriga mos kelishi lozim. Bunda quyidagi modellardan foydalaniladi:

• Qoidalarga asoslangan yondashuv (Rule-Based) – jumlalar uzunligi, tinish belgilari va so‘zlarning tarqalishiga asoslanadi [14].
• Statistik yondashuv – parallel korpusdagi so‘z va iboralar chastotasiga asoslanib segmentlarni bog‘laydi [15].
• Mashina o‘rganish metodlari – neyron tarmoqlar va transformer modellar yordamida yanada aniqroq moslashtirishni amalga oshiradi [16].

2.2. So‘z va iboralar darajasida moslashtirish. Ba’zan jumlalar to‘g‘ridan-to‘g‘ri mos kelmaydi, chunki tarjimada iboralar tuzilishi o‘zgarishi mumkin. Bu holatda dinamik dasturlash algoritmlari, xususan, Levenshtein masofasi va DICE koeffitsienti asosida so‘z moslashuvi tekshiriladi [17].

2.3. Indeksatsiya qilish. Parallel matnlarni samarali qidirish va tarjima xotirasidan tezkor foydalanish uchun indeksatsiya qilish (indexing) zarur. Indeksatsiya natijasida katta hajmdagi matnlar tez va samarali qidiriladi, bu esa tarjima tizimining tezligini oshiradi. Indeksatsiyaning eng samarali usullaridan biri – *teskari indeks* (Inverted Index) texnikasidir. Bu usulda har bir so‘zning qaysi hujjatlarda va qaysi segmentlarda uchrashi to‘g‘risidagi *qidiruv jadvali* yaratiladi [18]. Ushbu metod Google va boshqa qidiruv tizimlarida ham keng qo‘llaniladi.

2.4. TF-IDF va vektorli modellar. Indeksatsiyani yaxshilash uchun *TF-IDF (Term Frequency - Inverse Document Frequency)* va *vektorli ifodalash* metodlaridan foydalaniladi. TF-IDF – muhim terminlarni ajratish va parallel matndagi asosiy tarjima birliklarini belgilashga yordam beradi [19]. *Word2Vec*, *FastText* va *BERT* kabi *neyron modellar* qidiruv natijalarini yanada moslashuvchan qilishga yordam beradi [20].

3. Parallel korpuslarni tarjima xotirasiga integratsiya qilishning keyingi bosqichida tarjima jarayonida tarjima xotirasidan foydalaniladi. Tarjima jarayoni murakkab lingvistik va texnologik bosqichlarni o‘z ichiga oladi. Tarjima xotirasi va avtomatik tarjima tizimlari tarjima jarayonini avtomatlashtirish va samaradorligini oshirishga xizmat qiladi. Ushbu texnologiyalar tarjimonlarga vaqtni tejash, natijalarning barqarorligini oshirish va inson xatolarini kamaytirish imkonini beradi [21].

Tarjima xotirasining ishlash tamoyili quyidagi bosqichlar ketma-ketligida amalga oshiriladi:

⇒ Segmentatsiya – matn tarjima birliklariga bo‘linadi (masalan, jumlalar yoki iboralar).

⇒ Mos kelish tekshiruvi – yangi segmentlar oldin tarjima qilingan ma’lumotlar bilan taqqoslanadi.

⇒ Qayta ishlatish va takroriy tarjima – tizim mavjud tarjimalarni taklif qiladi.

⇒ Qo‘lda tahrirlash – tarjimon TM tomonidan taqdim etilgan natijani tekshiradi va tahrirlaydi.

L. Bowkerning ta’kidlashicha [1], TM tizimlari ayniqsa bir xil formatdagi va takroriy jumlalardan iborat hujjatlarda yuqori samaradorlikni ko‘rsatadi.
Zamonaviy avtomatik tarjima tizimlari TM bilan integratsiyalashgan holda ishlaydi. Ushbu tizimlarning evolyutsiyasi uch asosiy modelni o‘z ichiga oladi:
• Qoidalarga asoslangan tarjima (RBMT – Rule-Based Machine Translation) – lingvistik qoidalar va lug‘atlarga asoslangan tarjima texnologiyasi [22].
• Statistik tarjima (SMT – Statistical Machine Translation) – katta hajmdagi parallel korpuslardan foydalangan holda probabilistik usullar yordamida tarjima qiladi [23].
• Neyron tarjima tizimlari (NMT – Neural Machine Translation) – sun’iy neyron tarmoqlari yordamida matnning kontekstini chuqur o‘rganadi va yanada tabiiy tarjima hosil qiladi [24].

Neyron tarjima tizimlari (NMT) transformer modellarini qo‘llagan holda so‘zlar orasidagi bog‘liqlikni tushunadi va kontekstni hisobga olgan holda tarjima qiladi [16].

Tarjima tizimlarining samaradorligini oshirish uchun TM va MT integratsiyasi amalga oshirilmoqda. Bunda:

⇒ Tarjima xotirasidan mavjud tarjimalar ajratib olinadi.

⇒ NMT faqat yangi segmentlarni tarjima qiladi.

⇒ Natijalar TM bazasiga qayta yuklanadi va kelgusida foydalanish uchun saqlanadi [25].

Tadqiqotchilarning ta’kidlashicha, bu yondashuv tarjima jarayonining barqarorligini oshirish va inson aralashuvini minimallashtirish imkonini beradi [8]. Tarjima xotirasi tizimlari hozirda turli sohalarda, jumladan, davlat idoralari, xalqaro tashkilotlar va IT kompaniyalarda keng qo‘llanilmoqda [26].

Kelajakda sun’iy intellekt va chuqur o‘rganish texnologiyalarining rivojlanishi tarjima xotiralarining yanada takomillashishiga olib keladi [21].

Kelajak istiqbollari quyidagi yo‘nalishlarda rivojlanishi kutilmoqda:

⇒ Yuqori sifatli avtomatik tarjima tizimlari – chuqur o‘rganish va neyron tarmoqlari asosida ishlaydigan ilg‘or tizimlar tarjimalarni yanada tabiiy qilishga yordam beradi.

⇒ Ko‘p tilli tarjima xotiralari – mavjud tizimlar faqat ikki tilli parallel korpuslarga asoslangan bo‘lsa, kelajakda ko‘p tilli parallel korpuslar va tarjima xotiralari ommalashishi mumkin.

⇒ Interaktiv va adaptiv tarjima tizimlari – tarjimonlarning ishlash uslubiga moslashadigan va ularning tarjima uslubini o‘rganadigan tizimlar rivojlanishi kutilmoqda.

Tadqiqotlar shuni ko‘rsatadiki, tarjima xotirasi tizimlaridan foydalanish tarjimonlarning mahsuldorligini sezilarli darajada oshiradi. Misol uchun, 371 902 ta parallel jumlalarni o‘z ichiga oluvchi KazParC: Qozoq parallel korpusidan foydalangan holda davlat idoralari uchun maxsus ishlab chiqilgan tarjima tizimlarida tarjima jarayoni 30-40% tezlashgani qayd etilgan. Shuningdek, KazParC loyihasi doirasida Tilmash nomli neyron mashinaviy tarjima modeli ham ishlab chiqilgan bo‘lib, ushbu model BLEU va chrF kabi standart baholash metrikalariga ko‘ra, Google Translate va Yandex Translate kabi yirik kompaniyalar bilan tenglashadi yoki ba’zi hollarda ularni ortda qoldiradi. [https://github.com/IS2AI/KazParC].

Parallel korpusga asoslangan tarjima xotirasi tizimlari zamonaviy tarjima jarayonining muhim elementlaridan biri hisoblanadi. Ular tarjima jarayonini tezlashtirish, sifatini oshirish va iqtisodiy samaradorligini ta’minlash imkonini beradi. Ushbu maqolada ko‘rib chiqilgan metodologiya va yondashuvlar ushbu sohada tadqiqotlarni davom ettirish uchun muhim asos bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Bowker L. Computer-Aided Translation Technology: A Practical Introduction. – Ottawa: University of Ottawa Press, 2002. – 243 p.

2. Dagan I., Church K., Gale W. Robust Bilingual Word Alignment for Machine Aided Translation // ACL Anthology. – 1993.

3. Сомерс Х.Л. Translation Memory Systems // Hutchins J. (ed.) Computers and Translation: A Translator’s Guide. – Amsterdam: Benjamins, 2003. – С. 31–48.
4. Way A. Traditional and Emerging Use-Cases for Machine Translation // Machine Translation. – 2014. – Vol. 28, No. 3-4. – P. 243–254.
5. Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation // Proceedings of the MT Summit X. – Phuket, Thailand, 2005. – P. 79–86.
6. Maegaard B. Translation Memory: An Efficient Way to Reuse Translations // Proceedings of MT Summit VII. – Singapore, 1999. – P. 331–334.
7. Marcu D., Wong W. A Phrase-Based, Joint Probability Model for Statistical Machine Translation // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2002. – P. 133–139.
8. Federico M., Bertoldi N., Barbaiani M. Integrating Translation Memories into Phrase-Based Statistical Machine Translation // Proc. of ACL Conf. – 2014.
9. Toral A., Muñoz R. A Proposal to Automatically Build and Use Parallel Corpora within the OpenTrad Project // Proceedings of the Workshop on Building and Using Parallel Texts. – New York, USA, 2006. – P. 149–152.
10. Baldwin T., Lui M. Language Identification: The First Step Towards NLP // Computational Linguistics. – 2010.
11. Kocmi T., Bojar O. An Exploration of Sentence Embeddings for Neural Machine Translation. – 2017.
12. Papineni K., Roukos S., Ward T., Zhu W. BLEU: A Method for Automatic Evaluation of Machine Translation // ACL. – 2002.
13. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL. – 2018.
14. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. – 1993.
15. Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S. The Mathematics of Statistical Machine Translation: Parameter Estimation // Computational Linguistics. – 1993. – Vol. 19, No. 2. – P. 263–311.
16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is All You Need // Proc. of NeurIPS 2017. – Vol. 30. – P. 5998–6008.
17. Melamed I. D. Models of Translational Equivalence among Words // Computational Linguistics. – 2000.
18. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008.
19. Sparck Jones K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation. – 1972.
20. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // ICLR. – 2013.

21. Somers H. Translation Memory Systems: Enlightenment or Feudalism? // Machine Translation Journal. – 2003.
22. Hutchins W. J., Somers H. L. An Introduction to Machine Translation. – London: Academic Press, 1992. – 362 p.
23. Koehn P. Statistical Machine Translation. – Cambridge: Cambridge University Press, 2010. – 487 p.
24. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // ICLR. – 2015.
25. Koehn P., Knowles R. Six Challenges for Neural Machine Translation // Proc. of the 1st Conf. on Machine Translation (WMT 2017). – P. 28–39.
26. Gough N., Way A. Example-Based Machine Translation Using a Memory-Based Learning Approach // Computational Linguistics. – 2004.
27. Abduraxmonova N. Kompyuter lingvistikasi: darslik. – Toshkent: Nodirabegim, 2021. – 312 b.
28. Abduraxmonova N. Mashina tarjimasining lingvistik ta’minoti: monografiya. – Toshkent, 2018. – 185 b.
29. Abdurakhmonova N., Hayrullayeva G., Urdishev K. Using a Corpus for Teaching Synonymy in Uzbek // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы: материалы II Междунар. науч. конф. – Уфа, 11–12 декабря 2019 г. – Уфа: Изд-во БашГУ, 2019.
30. Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., Mamasaidov M. UZWORDNET: A Lexical-Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference. – University of South Africa (UNISA), 2021. – P. 8–19.
31. Gatiatullin A., Suleymanov D., Prokopyev N., Abdurakhmonova N. Turkic Morpheme: From the Portal to the Linguistic Platform // Aliev R.A. et al. (Eds.) Lecture Notes in Networks and Systems. – Vol. 912. – Springer, Cham, 2024. – https://doi.org/10.1007/978-3-031-53488-1_22
32. KazParC. Kazakh Parallel Corpus for Machine Translation // arXiv preprint. – 2024.
33. Mengliev D. B., Abdurakhmonova N., Hayitbayeva D., Barakhnin V. B. Automating the Transition from Dialectal to Literary Forms in Uzbek Language Texts: An Algorithmic Perspective // 2023 IEEE XVI Int. Sci. and Tech. Conf. APEIE. – Novosibirsk, 2023. – P. 1440–1443. DOI: 10.1109/APEIE59731.2023.10347617.
34. Mengliev D., Barakhnin V., Abdurakhmonova N. Development of Intellectual Web System for Morph Analyzing of Uzbek Words // Applied Sciences. – 2021. – Vol. 11, No. 19. – Art. no. 9117. https://doi.org/10.3390/app11199117