QO‘QON DAVLAT PEDAGOGIKA INSTITUTI ILMIY XABARLARI
(2025-yil 2-son)

FILOLOGIYA

PHILOLOGY

PARALLEL KORPUSGA MA’LUMOT YIG‘ISHNING AVTOMATLASHTIRILGAN USULLARI

Xolmonova Iqbola Alisher qizi
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi tayanch doktoranti
E-mail: iqbolabintualisher@gmail.com
Orcid: 0009-0006-8844-6569

Annotatsiya: Ushbu maqolada parallel matnlarni yig‘ish jarayonida sun’iy intellekt (AI) va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanish ko‘rib chiqiladi. Veb-skraping, mashinali o‘rganish va metama’lumotlar bo‘yicha filtrlash usullari orqali parallel korpus yaratish va tarjima sifatini avtomatik aniqlash imkoniyatlari tahlil qilinadi. Ushbu yondashuvlar parallel korpus yaratishni avtomatlashtirish va tarjima tizimlarining sifatini yaxshilashga xizmat qiladi.

Tayanch so‘z va iboralar: parallel korpus, veb-skraping, mashinali o‘rganish, tabiiy tilni qayta ishlash, metama’lumotlar, tarjima sifati, sun’iy intellect, nlp texnologiyalari, avtomatlashtirilgan matn yig‘ish.

Annotation: This research discusses the use of artificial intelligence (AI) and natural language processing (NLP) technologies in the process of parallel text collection. Methods such as web scraping, machine learning, and metadata-based filtering are explored to automate the creation of parallel corpora and automatically assess translation quality. These approaches aim to automate parallel corpus creation and improve translation system quality.

Keywords and phrases: parallel corpus, web scraping, machine learning, natural language processing, metadata, translation quality.

Аннотация: В данном исследовании рассматривается использование искусственного интеллекта (ИИ) и обработки естественного языка (ОЯЯ) в процессе сбора параллельных текстов. Исследуются методы такие как веб-скрапинг, машинное обучение и фильтрация на основе метаданных, которые позволяют автоматизировать создание параллельных корпусов и автоматически оценивать качество перевода. Эти подходы направлены на автоматизацию создания параллельных корпусов и улучшение качества переводческих систем.

Ключевые слова и фразы: параллельный корпус**,** веб-скрапинг**,** машинное обучение**,** обработка естественного языка**,** метаданные**,** качество перевода.

So‘nggi yillarda tabiiy tilni qayta ishlash (NLP) va mashinada tarjima (MT) sohalarida parallel korpuslar muhim manba sifatida e’tirof etilmoqda. Parallel korpus — bu ikki yoki undan ortiq tillarda bir-biriga mos keluvchi matnlarning to‘plami bo‘lib, neyron mashina tarjimasi (NMT), statistika asosidagi tarjima (SMT) va boshqa tarjima tizimlarini yaratishda muhim rol o‘ynaydi. Ushbu korpuslarning sifatli va keng ko‘lamli bo‘lishi tarjima tizimlarining samaradorligini oshirishda muhim omillardan biridir. Shu sababli, parallel korpus tuzish jarayonida avtomatlashtirilgan usullardan foydalanish tadqiqotchilar e’tiborini tortmoqda.

An’anaviy yondashuvlarda parallel matnlarni yig‘ish va moslashtirish jarayonlari ko‘p vaqt va resurs talab qilgan bo‘lsa, hozirgi zamonaviy texnologiyalar yordamida ushbu jarayonni samarali tarzda avtomatlashtirish mumkin. Internet resurslaridan parallel matnlarni avtomatik yig‘ish, ularni moslashtirish va filtrlash uchun turli vositalar, shu jumladan veb-skraping, avtomatik moslashtirish algoritmlari, NLP asosidagi filtrlash metodlari hamda sun’iy intellekt modellari qo‘llanilmoqda. Ushbu texnologiyalar parallel korpus yaratish jarayonini tezlashtirib, sifatli va yuqori aniqlikdagi ma’lumotlar bazasini shakllantirishga xizmat qiladi.

Parallel matnlarni yig‘ish jarayonida sun’iy intellekt va NLP texnologiyalaridan foydalanish mumkin. Bu jarayonni samarali tashkil etish uchun quyidagi usullar qo‘llanadi:

Veb-skraping (Web Scraping) – internet saytlardan avtomatik ravishda ma’lumotlarni to‘plash.

Mashinali o‘rganish (Machine Learning) asosida mos matnlarni tanlash – neyron tarmoqlar yordamida parallel bo‘lgan matnlarni avtomatik aniqlash.

Metama’lumotlar bo‘yicha filtrlash – avtomatik ravishda til va tarjima sifatini aniqlash uchun algoritmlar yaratish.

Veb-skraping – bu internet saytlardan avtomatik ravishda ma’lumotlarni to‘plash jarayoni.[1] Ushbu jarayon maxsus dasturlar yoki kodlar yordamida amalga oshiriladi va veb-sahifalardagi kerakli matn, rasm, havola kabi ma’lumotlarni ajratib olish imkonini beradi. Axborot texnologiyalarining jadal rivojlanishi natijasida internet ulkan ma’lumot manbayi sifatida shakllandi. Har kuni millionlab veb-sahifalar yangilanib, turli sohalar bo‘yicha qimmatli ma’lumotlar e’lon qilinadi. Ushbu ma’lumotlarni qo‘lda yig‘ish katta vaqt va resurs talab qilishi sababli, veb-skraping (Web Scraping) — internet saytlardan avtomatik ravishda ma’lumotlarni yig‘ish texnologiyasi rivojlanib bormoqda.

Veb-skraping quyidagi bosqichlarda amalga oshiriladi:

1. Manba saytni aniqlash – qaysi saytlardan ma’lumot olinishi belgilanadi.
2. HTML va DOM tahlili: Ko‘pgina veb-saytlar ma’lumotlarni HyperText Markup Language (HTML) formatida saqlaydi. Veb-skraping jarayonida Document Object Model (DOM) orqali HTML elementlarini tahlil qilish asosiy bosqich hisoblanadi.
3. Dastur yordamida ma’lumotlarni yuklab olish – maxsus skript yoki bot sahifani skanerlab, kerakli qismlarni ajratadi. Veb-skraping jarayonida GET va POST kabi HTTP so‘rovlar orqali saytlardan ma’lumot olinadi.

4. Ma’lumotlarni ajratish va tozalash: Saytlardan olingan ma’lumotlar ko‘pincha keraksiz kodlar yoki formatlash elementlari bilan bo‘ladi. Ushbu ma’lumotlarni tozalash uchun Regular Expressions (Regex) va XPath kabi texnologiyalar ishlatiladi.
5. Ma’lumotlarni saqlash: Yig‘ilgan ma’lumotlar odatda CSV, JSON, XML yoki SQL kabi formatlarda saqlanadi va tahlil qilinadi.

Veb-skrapingni amalga oshirish uchun bir necha texnik yondashuvlar mavjud. Quyida eng ommaviy va samarali usullar keltirilgan:

Statik veb-skraping. Bu usulda oddiy HTML kod tahlil qilinadi va kerakli ma’lumotlar ajratib olinadi. Bu usul BeautifulSoup (Python) kabi kutubxonalar yordamida amalga oshiriladi.

Dinamik veb-skraping. Ba’zi veb-saytlar ma’lumotlarni JavaScript orqali yuklaydi. Oddiy HTML analiz bu holatda yetarli bo‘lmagani sababli, Selenium, Playwright, yoki Puppeteer kabi vositalardan foydalaniladi.

API orqali ma’lumot yig‘ish. Ko‘plab zamonaviy saytlar RESTful API orqali ma’lumot almashish imkoniyatini taqdim etadi. Veb-skraping jarayonida, agar sayt API taqdim etsa, bevosita API so‘rovlar orqali ma’lumot olish tavsiya etiladi.

XPath va CSS selektorlar orqali skraping. Bu usul Lxml yoki Scrapy kabi kutubxonalar yordamida amalga oshiriladi. XPath yordamida ma’lumotlarni aniq tanlab olish mumkin.

Veb-skraping internet resurslaridan avtomatik ravishda ma’lumot yig‘ish texnologiyasi bo‘lib, parallel korpuslar yaratish va neyron mashina tarjimasi (NMT) tizimlari uchun asosiy manba hisoblanadi. Ushbu texnologiya o‘zbek-turk parallel korpusini tuzish, tarjima sifatini yaxshilash hamda lingvostatistik va lingvomadaniy tadqiqotlar uchun katta ahamiyatga ega. Veb-skraping mashina tarjimasi va lingvistik tadqiqotlarda o‘zbek va turk tillarida parallel matnlarni yig‘ish, lug‘atlar va terminologik bazalar yaratish, neyron tarjima modellarini o‘qitish uchun katta hajmdagi ma’lumotlarni yig‘ish uchun muhim ahamiyat kasb etadi. Sun’iy intellekt va tabiiy tilni qayta ishlash sohalarida NLP modellarini rivojlantirish uchun korpus yaratish, til o‘zgarishlarini statistik jihatdan tahlil qilish, o‘zbek-turk tarjima tizimlarini optimallashtirish uchun qulay vosita hisoblanadi. Akademik va ilmiy tadqiqotlar uchun Google Scholar, ilmiy maqola bazalaridan tarjima tadqiqotlari uchun ma’lumot yig‘ish, mashina tarjimasi uchun real matnlar to‘plash va tahlil qilish uchun xizmat qiladi.

Ko‘plab saytlar veb-skrapingdan himoyalanish uchun robots.txt faylini ishlatadi yoki maxsus xavfsizlik choralarini ko‘radi. Shu sababli, ma’lumot yig‘ishda sayt qoidalariga rioya qilish muhimdir. Veb-skraping jarayoni texnik va huquqiy jihatdan to‘g‘ri tashkil etilganda, katta hajmdagi ma’lumotlarni samarali to‘plash va qayta ishlash imkonini beradi.

Veb-skraping – zamonaviy axborot texnologiyalarida katta ahamiyatga ega bo‘lgan samarali ma’lumot yig‘ish vositasi hisoblanadi. U mashinada tarjima, iqtisodiy tahlil, elektron tijorat, ilmiy tadqiqotlar va ijtimoiy tarmoqlarni kuzatish kabi ko‘plab sohalarda qo‘llaniladi. Kelajakda veb-skrapingning sun’iy intellekt va mashinada o‘rganish bilan integratsiyalashuvi yanada rivojlanib, ma’lumotni avtomatik yig‘ish va tahlil qilish tizimlarining samaradorligini

oshiradi. Shu bilan birga, axloqiy va huquqiy me’yorlarga rioya qilish bu texnologiyaning barqaror rivojlanishini ta’minlaydi.

Tabiiy tilni qayta ishlash (Natural Language Processing – NLP) va neyron mashina tarjimasi (Neural Machine Translation – NMT) sohalarida sifatli parallel korpus yaratish muhim bosqich hisoblanadi. O‘zbek-turk parallel korpusini tuzish jarayonida mos matnlarni tanlash va ularning lingvistik mosligini baholash asosiy muammolardan biridir. An’anaviy usullar qo‘lda saralash va lingvistik qoidalar asosida tahlil qilishga asoslangan bo‘lsa, mashinali o‘rganish (Machine Learning – ML) yondashuvlari ushbu jarayonni avtomatlashtirish va optimallashtirishga yordam beradi. Mashina o‘rganish (Machine Learning, ML) – bu kompyuter tizimlariga tajriba asosida o‘z-o‘zidan yaxshilanish imkonini beradigan sun’iy intellekt sohasidir. O‘zbek-turk parallel korpusini yaratishda ML algoritmlari parallel matnlarni avtomatik ravishda aniqlash, saralash va moslashtirish uchun qo‘llaniladi. Bunda neyron tarmoqlar, statistik modellar va maxsus algoritmlar ishlatiladi. Parallel matnlarni tanlash jarayonida TF-IDF va Kosinus o‘xshashligi kabi statistik usullar dastlabki filtrlashni amalga oshiradi. Keyinchalik, BERT, XLM-R va S-BERT kabi transformer modellar matnlarning semantik mosligini chuqur tahlil qilib, noto‘g‘ri juftlarni chiqarib tashlash va sifatli parallel korpus shakllantirishga yordam beradi.[2] Ushbu yondashuv neyron mashina tarjimasi (NMT) modellarining sifatini oshirish, tarjima natijalarining aniq va tabiiy bo‘lishini ta’minlashga xizmat qiladi. Shu bilan birga, lingvostatistik va lingvomadaniy tadqiqotlar uchun aniq va to‘liq ma’lumot bazasi yaratish imkonini beradi. Natijada, avtomatlashtirilgan parallel korpus yaratish orqali tarjima tizimlari va tabiiy tilni qayta ishlash (NLP) sohasidagi tadqiqotlar yanada rivojlanadi.[3]

Mashina o‘rganish asosida parallel matnlarni tanlash – o‘zbek-turk parallel korpusini yaratishda muhim texnologik yondashuvlardan biridir. Neyron tarmoqlar, statistik modellar va so‘nggi NLP texnologiyalari yordamida matnlarni aniqlash va moslashtirish ancha samarali kechadi. Bu esa sifatli parallel korpus yaratish va keyingi ilmiy-tadqiqot ishlari uchun mustahkam asos bo‘lishiga yordam beradi.

Metama’lumotlar bo‘yicha filtrlash. Metama’lumotlar – bu matn yoki hujjat haqida qo‘shimcha ma’lumot beruvchi ma’lumotlar bo‘lib, ularning tuzilishi, mazmuni va sifati haqida tahlil yuritish imkonini beradi.[4] O‘zbek-turk parallel korpusini yaratishda metama’lumotlar matnlarning tili, tarjima sifati, uslubi, manbasi, uzunligi, yozuv normasi kabi jihatlarini avtomatik tahlil qilishga yordam beradi. Metama’lumotlar bo‘yicha filtrlash – bu parallel korpus yaratish jarayonida til va tarjima sifatini avtomatik ravishda aniqlash va filtrlash uchun qo‘llaniladigan algoritmlar to‘plamidir. Metama’lumotlar – matnlarning tili, uzunligi, formatlari, tarjima sifatini baholash va boshqa tegishli atributlarini ifodalaydi. Ular yordamida matnlar orasidagi moslik darajasi aniqlanadi va noto‘g‘ri yoki sifatsiz matnlar avtomatik tarzda ajratib olinadi.

Metama’lumotlar bo‘yicha filtrlash quyidagi usullarni o‘z ichiga oladi:

Tilni aniqlash algoritmlari. NLP texnikalari yordamida matnning tilini avtomatik aniqlash. Bu usul, ayniqsa, parallel korpusda bir nechta tillar mavjud bo‘lganda, noto‘g‘ri tarjimalarni ajratib olishda samarali usul hisoblanadi. Statistik modellar yordamida matnlarning tiliga mos keluvchi tarjimalarni toppish uchun xizmat qiladi.

Tarjima sifatini baholash. BLEU, TER, METEOR kabi metrikalar yordamida tarjima sifatini avtomatik baholash. Bu ko‘rsatkichlar matnning to‘g‘riligini, o‘zaro mosligini va qismlarining mosligini aniqlashda ishlatiladi. Neyron tarmoqlar (CNN, RNN) orqali tarjima sifatini aniqlash va xatoliklarni filtrlash. Ular matnlarni kontekstual ravishda tahlil qilib, ularning semantik to‘g‘riligini tekshiradi.

Matn uzunligi va strukturaviy tekshirish. Matnning uzunligi va hajmi (jumla uzunligi, so‘zlar soni, bo‘lishli va bo‘linmaydigan iboralar) metama’lumotlariga asoslanib, noto‘g‘ri formatdagi matnlarni ajratish lozim. Sintaktik va semantik tuzilmani tahlil qilish orqali matnning to‘g‘ri tarjima qilinganligini tekshirish kerak bo‘ladi.

Mahalliy va global kontekstni tahlil qilish. Transformer modellar (BERT, GPT, XLM-R) yordamida matnlarni kontekstual tahlil qilish, ular orasidagi semantik moslikni aniqlash va noto‘g‘ri tarjimalarni avtomatik ravishda chiqarib tashlash hisoblanadi.[5]

Ushbu algoritmlar orqali avtomatik filtrlash tizimi yaratiladi, bu esa parallel korpusni yaratishda sifatni yaxshilash va noaniq yoki noto‘g‘ri tarjimalardan saqlanish imkonini beradi. Natijada, tarjima tizimlarining sifatini oshirish va lingvistik tadqiqotlarni yanada samarali o‘tkazish imkoniyatlari yaratiladi.

Metama’lumotlarga asoslangan filtrlash bir qancha sohalarda muvaffaqiyatli qo‘llaniladi. Masalan, mashina tarjimasi tizimlarini o‘qitish – sifatli parallel matnlarni tanlash orqali neyron tarjima modellarining aniqligini oshirish; Lingvistik tadqiqotlar – til va tarjima sifatini chuqur o‘rganish va statistik tahlil qilish; Ma’lumotlar qazib olish – katta hajmdagi matnlar ichidan sifatli parallel ma’lumotlarni ajratib olish kabi.

Metama’lumotlar asosida filtrlash algoritmlarini yaratish parallel korpusni to‘g‘ri shakllantirishda muhim ahamiyat kasb etadi. Tilni avtomatik aniqlash, tarjima sifatini baholash, matn uzunligini tekshirish va terminologik moslikni ta’minlash kabi yondashuvlar yordamida sifatli parallel korpus yaratish imkoniyati oshadi. Bu esa o‘z navbatida neyron mashina tarjimasi, sun’iy intellekt va tilshunoslik sohalaridagi tadqiqotlar uchun mustahkam asos bo‘lib xizmat qiladi.

Parallel matnlarni yig‘ish jarayonida sun’iy intellekt (AI) va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanish jarayonni yanada samarali qilishga yordam beradi.[6] Ushbu texnologiyalar yordamida internetdan ma’lumot yig‘ish, parallel matnlarni aniqlash va filtrlash jarayonlarini avtomatlashtirish mumkin. Veb-skraping texnologiyasi yordamida internetdagi mavjud resurslardan matnlarni tez va samarali yig‘ish, mashina o‘rganish (ML) asosida mos matnlarni tanlash orqali parallel matnlarning sifatini avtomatik baholash va metama’lumotlar bo‘yicha filtrlash orqali tarjima sifatini aniqlash va noto‘g‘ri matnlarni ajratish mumkin. Bu yondashuvlar parallel korpus yaratishda tarjima sifatini yaxshilash va

ma’lumotlarni samarali yig‘ish imkonini beradi. Natijada, AI va NLP texnologiyalarini birlashtirish orqali parallel korpus yaratish jarayonini avtomatlashtirish, ayniqsa, katta hajmdagi ma’lumotlar bilan ishlashda samaradorlikni oshirish mumkin.

FOYDALANILGAN ADABIYOTLAR:

1. Bergsma, S., & Dredze, M. (2009). No Mention of Parallel Corpora: Web-Based Corpus Construction for Multilingual Language Processing. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

2. Ranzato, M., Boureau, Y. L., & LeCun, Y. (2007). Learning Labeled and Unlabeled Structures for Text Classification. In Proceedings of the 24th International Conference on Machine Learning.

3. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.

4. Koehn, P., & Knight, K. (2003). Empirical Methods for Scalable Parallel Text Processing. In Proceedings of the HLT-NAACL 2003 Workshop.

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems.

6. Ruder, S. (2017). Neural Transfer Learning for Natural Language Processing. In Proceedings of the 8th International Conference on Natural Language Processing.