

**O‘ZBEK TILIDAGI ONOMASTIK BIRLIKLARNI BIO SXEMASI
ASOSIDA AVTOMATIK ANNOTATSIYALASH MASALASI**

Shabonova Guljamol Abdullo qizi

Buxoro viloyati Kogon tumani 11-umumta’lim maktabi

ona tili va adabiyoti fani o‘qituvchisi

shabonova95@gmail.com

Annotatsiya

Mazkur maqolada o‘zbek tilidagi onomastik birliklarni korpus asosida avtomatik annotatsiyalashning lingvistik-kompyuter modeli masalasi tadqiq etiladi. Tadqiqotning dolzarbligi o‘zbek tilini tabiiy tilni qayta ishlash (NLP) tizimlariga moslashtirish jarayonida sifatli, lingvistik jihatdan asoslangan va mashina tomonidan qayta ishlanishi mumkin bo‘lgan annotatsiyalangan ma’lumotlarga ehtiyojning ortib borayotgani bilan belgilanadi. O‘zbek tilida so‘nggi yillarda morfologik, sintaktik, so‘z turkumlari va nomlangan birliklarni belgilashga yo‘naltirilgan korpus resurslari yaratilgan bo‘lsa-da, onomastik birliklarning lingvistik tasnifi va ularni avtomatik aniqlash hamda tasniflashni yagona metodologik tizim asosida tashkil etish masalasi dolzarb bo‘lib qolmoqda. Maqolada onomastik birliklarni antroponim, toponim, gidronim, etnonim, tashkilot nomlari va boshqa guruhlariga ajratishning kompyuter lingvistikasi uchun maqbul modeli taklif etiladi. Shuningdek, ekspert annotatsiyasi, qoida asosidagi yondashuv, mashinaviy o‘qitish va transformer modellarini bosqichma-bosqich birlashtiruvchi avtomatik annotatsiya arxitekturasini ishlab chiqiladi. Tadqiqotda “sun’iy intellekt annotatsiyasi – lingvist ekspert nazorati – xatolarni tahlil qilish – modelni qayta takomillashtirish” tamoyili asosiy metodologik mexanizm sifatida qaraladi. Taklif qilinayotgan yondashuv kelgusida o‘zbek tilining ko‘p qatlamli lingvistik korpusini shakllantirish, onomastik ma’lumotlar bazasini yaratish va



o'zbek tiliga mo'ljallangan NLP modellarini takomillashtirish uchun nazariy hamda amaliy asos bo'lib xizmat qilishi mumkin.

Kalit so'zlar: *onomastika, korpus lingvistikasi, avtomatik annotatsiya, NER, BIO sxemasi, o'zbek tili.*

Raqamli texnologiyalar va sun'iy intellektning rivojlanishi tilshunoslikda matnni mashina tomonidan qayta ishlanadigan ma'lumotlar majmuasi sifatida ham o'rganishni taqozo etmoqda. Bu jarayonda atoqli otlarni avtomatik aniqlash va tasniflash (Named Entity Recognition, NER) axborot qidiruvi, savol-javob tizimlari va mashina tarjimai kabi vazifalarda muhim komponent hisoblanadi. O'zbek tilida morfologik, sintaktik va nomlangan birliklarni belgilashga yo'naltirilgan korpus resurslari yaratilayotgan bo'lsa-da, onomastik birliklarni yagona metodologik tizim asosida annotatsiyalash masalasi hamon dolzarb bo'lib qolmoqda.

O'zbek tilida bu vazifa bir qator o'ziga xos qiyinchiliklar bilan bog'liq. Birinchidan, til agglutinatив bo'lgani uchun grammatik ma'nolar affiksalar orqali ifodalanadi va bu avtomatik tahlilni murakkablashtiradi. Ikkinchidan, atoqli otlar oddiy otlardan har doim ham formal belgilar bilan farqlanavermaydi. Uchinchidan, bir xil leksik birlik kontekstga qarab turlicha onomastik maqomga ega bo'lishi mumkin. Masalan, "Navoiy" komponenti "Alisher Navoiy" birikmasida antroponim, "Navoiy viloyati" birikmasida esa toponim tarkibida keladi. Demak, annotatsiya so'zning shakliga emas, balki uning kontekstual va semantik xususiyatlariga tayanishi lozim.

Tadqiqot obyekti o'zbek tilidagi zamonaviy yozma matnlarda uchraydigan onomastik birliklar, predmeti esa ularni avtomatik aniqlash, chegaralash va semantik-onomastik guruhlariga tasniflash modellaridir. Boshlang'ich bosqichda obyektni haddan tashqari kengaytirmaslik uchun beshta kategoriya ajratiladi: PERSON (antroponimlar), LOCATION (toponimlar), ORGANIZATION (tashkilot nomlari), ETHNIC (etnonimlar) va



HYDRONYM (gidronimlar). Keyingi bosqichlarda ro'yxatni boshqa onomastik kategoriyalar bilan to'ldirish mumkin.

Annotatsiya uchun BIO tamg'alash sxemasi taklif etiladi: B – birlikning boshlanishi, I – uning davomi, O – onomastik birlikka kirmaydigan token. Masalan, “Alisher Navoiy Samarqand shahrida yashagan” gapi quyidagicha belgilanadi:

Token	Teg
Alisher	B-PERSON
Navoiy	I-PERSON
Samarqand	B-LOCATION
shahrida	O
yashagan	O

Bunday format ketma-ketlikni belgilash vazifasi uchun qulay bo'lib, modelga kontekstni hisobga olgan holda birlikning chegarasi va turini aniqlash imkonini beradi.

Maqolada avtomatik annotatsiya jarayoni to'rt tamoyilga asoslangan arxitektura sifatida ko'rib chiqiladi: qoida asosidagi lingvistik yondashuv, ekspert annotatsiyasi, mashinaviy o'qitish va transformer modellari. Metodologik mexanizm sifatida “sun'iy intellekt annotatsiyasi – lingvist ekspert nazorati – xatolarni tahlil qilish – modelni qayta takomillashtirish” tsikli qabul qilinadi. Bunda sun'iy intellekt lingvistning o'rnini bosmaydi, balki

annotatsiyani tezlashtiruvchi va katta hajmdagi ma'lumotlarni qayta ishlashga yordam beruvchi vosita sifatida qaraladi.

Xulosa qilib aytganda, taklif etilgan model o'zbek tilining ko'p qatlamli lingvistik korpusini shakllantirish, onomastik ma'lumotlar bazasini yaratish va NLP modellarini takomillashtirish uchun nazariy hamda amaliy asos bo'la oladi. Keyingi bosqichlarda onomastik qatlamni morfologik, sintaktik va semantik qatlamlar bilan integratsiyalash, birliklarni normalizatsiya qilish va bilimlar grafigini yaratish ko'zda tutiladi.

FOYDALANILGAN ADABIYOTLAR

1. Abdurahmonova N. Z. O'zbek tili elektron korpusining kompyuter modellari: filol. fanlari doktori (DSc) diss. – Toshkent, 2020.
2. Hamroyeva Sh.M. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: filol. fan. bo'yicha falsafa doktori (PhD) diss. – Qarshi, 2018. – 250 b.
3. Zaxarov V.P., Mengliyev B.R., Hamroyeva Sh.M. Korpus lingvistikasi: o'quv qo'llanma. – Toshkent: Fan, 2021. – 185 b.
4. Po'latov A.K., Muhammedova S. Kompyuter lingvistikasi: o'quv qo'llanma. – Toshkent, 2008.

