UDK: 81'33

NER: TEGLASH MUOMMOLARIGA YECHIMLAR

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent,
elov@navoiy-uni.uz
ToshDO‘TAU

Madina Samatboyeva To‘lqinjon qizi,
tayanch doktorant
samatboyevamadina@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. O‘zbek tili korpusida NER obyektlarini teglash jarayonidagi asosiy muammolar va ularning mumkin bo‘lgan yechimlari tahlil qilish muhimdir. Maqolada, o‘zbek tilidagi NER tizimini rivojlantirishda duch kelinadigan asosiy muammolar – tegishli korpus yetishmovchiligi, so‘z shakllarining ko‘p variantliligi, ommaviy ochiq NER ma’lumotlar bazasining kamligi va semantik noaniqlik kabi jihatlar yoritiladi.

Shuningdek, maqolada ushbu muammolarni hal qilish uchun zamonaviy texnologiyalar, xususan, mashinani o‘rganish (Machine Learning), chuqur o‘rganish (Deep Learning) va tabiiy tilni qayta ishlashga oid ilg‘or metodlar tahlil qilinadi. Bundan tashqari, o‘zbek tiliga moslashtirilgan, maxsus annotatsiyalangan katta hajmdagi korpus va uni xalqaro standartlarga moslashtirish muhim ekani qayd etiladi.

Abstract. It is important to analyze the main challenges in the process of tagging NER entities in the Uzbek language corpus and their possible solutions. The article highlights the main challenges in developing an NER system for the Uzbek language, including the lack of an appropriate corpus, the high variability of word forms, the scarcity of publicly available NER databases, and semantic ambiguity.

Additionally, the article analyzes modern technologies used to address these challenges, particularly machine learning, deep learning, and advanced natural language processing (NLP) methods. Furthermore, it emphasizes the importance of creating a large, specifically annotated corpus adapted for the Uzbek language and aligning it with international standards.

Аннотация: Важно проанализировать основные проблемы процесса разметки объектов NER в корпусе узбекского языка и возможные пути их решения. В статье рассматриваются основные проблемы, возникающие при развитии системы NER для узбекского языка, такие как нехватка соответствующего корпуса, наличие множества вариантов словоформ,

ограниченное количество общедоступных баз данных NER и семантическая неопределенность.

Кроме того, в статье анализируются современные технологии, используемые для решения этих проблем, в частности, методы машинного обучения (Machine Learning), глубокого обучения (Deep Learning) и обработки естественного языка (NLP). Также подчеркивается важность создания большого аннотированного корпуса, адаптированного для узбекского языка, и его соответствия международным стандартам.

Kalit so‘zlar: *NER, nomlangan obyekt, ot, atoqli ot, NLP, teg, annotatsiya.*

KIRISH

Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida o‘zbek tili korpusida nomli obyektlarni aniqlash (Named Entity Recognition, NER) dolzarb muammolardan biri hisoblanadi. O‘zbek tili morfologik jihatdan murakkab va boy til bo‘lgani uchun NER obyektlarini aniqlash va to‘g‘ri teglash jarayoni qiyinchilik tug‘diradi. Shuningdek, xalqaro tajribalarni o‘rganish va o‘zbek tiliga moslashtirilgan, aniq va mustahkam ishlaydigan NER tizimini yaratish, annotatsiya jarayonida yaratilgan ma’lumotlar to‘plamining sifatini oshirish, tokenizatsiya va morfologik analiz orqali ma’lumotlarni yaxshiroq qayta ishlash kabi metodlar muhim hisoblanadi. Shu bilan birga, o‘zbek tili uchun maxsus tayyorlangan lug‘atlar va semantik tahlil vositalaridan foydalanish NER tizimining aniq va samarali ishlashini ta’minlashi mumkin.

Named Entity Recognition (NER) – bu tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining bir bo‘limi bo‘lib, matndagi aniq va muhim ma’lumotlarni aniqlash va ularni tasniflashga qaratilgan. NER matndagi **nomlangan obyektlar** (ya’ni shaxs nomlari, joy nomlari, tashkilotlar va h.k) **so‘z, ibora yoki so‘zlar ketma-ketligi** shaklida uchraydigan elementlarni topadi va ularni oldindan belgilangan toifalarga ajratadi [1].

NER texnologiyasi zamonaviy matn tahlili tizimlarining ajralmas qismi bo‘lib, jahon miqyosida chuqur o‘rganilgan. O‘zbek tilida bu boradagi izlanishlar endigina jadallashmoqda va global tajribalarga asoslanib, mahalliy xususiyatlarga moslashtirilgan modellar ishlab chiqilmoqda. Bu esa o‘zbek tilidagi sun’iy intellekt tizimlarini rivojlantirishda muhim qadam hisoblanadi.

#Misol:

“Xalq deputatlari Toshkent viloyati Bekobod tumani Kengashining navbatdan tashqari sessiyasida Zaynilobiddin Shahobiddinovich Nizomiddinov tuman hokimi lavozimiga tayinlandi.” [2]

Bu jumladagi NER orqali quyidagilar aniqlanadi:

- **[Toshkent] (viloyati)** — *Joy nomi* (Location)
- **[Bekobod] (tumani)** — *Joy nomi* (Location)
- **[Zaynilobiddin] [Shahobiddinovich] [Nizomiddinov]** – *Shaxs* (Person)

ADABIYOTLAR SHARHI

NER texnologiyalari jahon ilmiy adabiyotida ko‘plab olimlar tomonidan keng o‘rganilgan va muhokama qilingan. Olimalarning fikriga ko‘ra, NER tizimlarini muvaffaqiyatli amalga oshirish uchun morfologik, sintaktik, semantik va kontekstual tahlilning hammasi o‘rganilishi kerak. Har bir olim NERni turli nuqtai nazardan ko‘rib chiqadi, ammo umumiy fikr shuki, bu tizimlar kelajakda tabiiy tilni qayta ishlash (NLP) sohasida yanada rivojlanib, murakkab tahlil imkoniyatlarini yaratadi.

Jehangir, S. Radhakrishnan, R. Agarwal kabi hind olimlari o‘zlarining 2023-yilda nashr etilgan “**A Survey on Named Entity Recognition — Datasets, Tools, and Challenges”** nomli maqolasida [3] qoidalarga asoslangan yondashuvlar to‘g‘risida kengroq tahlil jarayonlarini o‘tkazadilar va amqolada o‘z fikrlarini berib o‘tadilar. Shuninghdek, ushbu maqolada NER sohasidagi turli metodologiyalar, jumladan, qoidaga asoslangan yondashuvlar, ML (Machine Learning) va DL (Deep Learning) usullari tahlil qilinadi. Mualliflar NERning rivojlanishi uchun mavjud bo‘lgan ma’lumotlar to‘plamlari va vositalarni ko‘rib chiqib, sohadagi asosiy muammolarni muhokama qiladilar.

Xitoylik olimlarning “Learn and Review: Enhancing Continual Named Entity Recognition” (U Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, Dai Dai - 2022) nomli maqolasida NER tizimlarining uzluksiz o‘rganish qobiliyatini oshirish uchun “Learn-and-Review” (L&R) deb nomlangan ikki bosqichli yangi modelni taklif etadilar. Ushbu yondashuv avvalgi qoidalarni yangi ma’lumotlar bilan birlashtirish orqali NER tizimlarining samaradorligini oshirishga qaratilgan [4].

Annotatsiyalangan ma’lumotlar, albatta, juda muhim baza hisoblanadi. Annotatsiya jarayonidan o‘tgan matngini mahsus modellar asosida teglanadi. Dastur ishlash jarayoni mukammal bo‘lishi uchun annotatsiya qilingan matnlar muhimdir. Imed Keraghel, Stanislas Morbieu, Mohamed Nadifning **“A Survey on Recent Advances in Named Entity Recognition”** (2023) nomli tadqiqot ishida NER sohasidagi so‘nggi yutuqlar, xususan, transformer va til modellari (LLM) asosidagi yondashuvlar tahlil qilinadi [5]. Mualliflar, shuningdek, annotatsiyalangan ma’lumotlar yetishmovchiligi sharoitida NER tizimlarini rivojlantirish usullarini ham ko‘rib chiqadilar.

Sakher Khalil Alqaaidi, Elika Bozorgi, Afsaneh Shams, Krzysztof Kochut kabi olimlarning **“A Few-Shot Learning Focused Survey on Recent Named Entity**

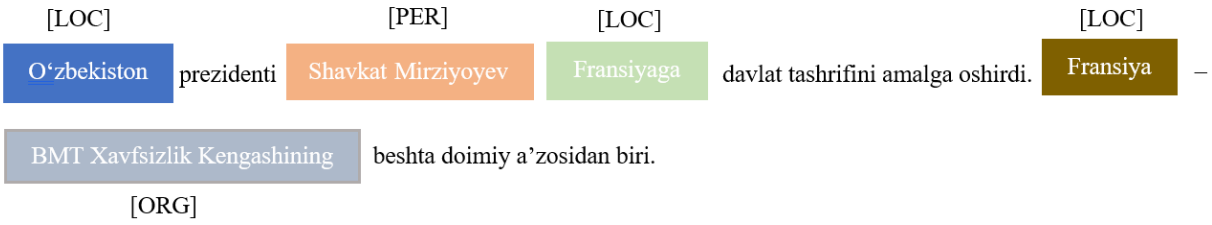
Recognition and Relation Classification Methods” (2023) maqolasida NER va munosabatlarni tasniflash sohasidagi so‘nggi chuqur o‘rganish modellari, ayniqsa, “few-shot learning” yondashuvlariga e'tibor qaratiladi. Tadqiqotchilar kam miqdordagi ma’lumotlar bilan samarali NER tizimlarini yaratish usullarini tahlil qiladilar [6].

ASOSIY QISM

“Teg” (ingl. tags) − matndagi ishchi izoh, u matn haqidagi ma’lumotni qamrab oladi. Korpus yordamida statistik hisob jarayonida tilimizda mavjud so‘zlarning faqat chastotasini aniqlash emas, balki yana bir qancha ma’lumotlarni olishimiz mumkin. Masalan, har bir so‘z bilan bilan yonma-yon uning turkumi belgilangan bo‘lsa, tilda turli nutqiy vaziyatda so‘z turkumlarining qo‘llanilish darajasini aniqlash ham mumkin. Lingvistik teglash har bir so‘zning ma’lum kodga ega bo‘lishi bilan xarakterlanadi. Ushbu kod teg, so‘zni kodlash esa tegging (ingl.tagging) [7] deyiladi.

NER tizimlari tomonidan ishlatiladigan teglar (labels) quyidagi asosiy turlarga bo‘linadi (Qarang: 1-rasm):

- **PER** (Person) – Shaxs nomlari uchun
- **LOC** (Location) – Geografik joylar uchun
- **ORG** (Organization) – Tashkilotlar nomlari uchun
- **MISC** – Boshqa nomlangan obyektlar uchun, masalan, mahsulot nomlari yoki boshqa aniq obyektlar.



1-rasm. NERlarni matn tarkibida teglar bilan belgilash jarayoni [8]

Teglash atamasi bir vaqtning o‘zida *annotatsiya* atamasi bilan yonma-yon qo‘llaniladi.

Annotatsiya – bu ma’lumotlarni qayta ishlash uchun kerakli teglashlarni amalga oshirish jarayonidir. O‘zbek tilida annotatsiya qilishda quyidagi amallar bajariladi:

- Matndagi nomlangan obyektlarni izlash.
- Ularni teglar bilan belgilash.

• Teglarni yuqori aniqlik bilan joylashtirish uchun ma’lum metodologiyalarni qo‘llashdir.

Ba’zi adabiyotlarda annotatsiya termini razmetkalash leksikasi bilan yonma-yon ham ishlatiladi. Bir vaqtning o‘zida annotatsiyalash bu – teglash jarayoni deb ham tushuniladi.

Korpusni teglash (annotatsiyalash) dasturlashtirilgan yo‘llar bilan amalga oshiriladi. Bunda, avvalo, vaqtni tejash, mehnatni kamaytirish nazarda tutilsa, ikkinchidin, matnga avtomatik ishlov berish muammosiga yechim topiladi [9].

O‘zbek tilida NER tizimlarini samarali ishlatish uchun, asosan, mashinali o‘qitish (ML) va chuqur o‘rganish (DL) metodlarini qo‘llash, maxsus korpuslar yaratish va annotatsiya qilish jarayonida yetarli darajada sifatli va keng qamrovli ma’lumotlar to‘plash metodologiyalarni qo‘llash tavsiya etiladi.

Korpusda NERlarni teglashda quyidagi jihatlarga alohida e’tibor qaratish lozim:

1. Korpusda tozalangan, turli uslub matnlari (baza) mavjudligiga;
2. Korpus matnlaridan erkin foydalanishga;
3. Korpusdagi matnlarda ishtirok etgan boshqa birliklarning teglariga e’tibor qaratish;
4. Boshqa birliklarga biriktirilgan maxsus teglarning qaytarilmasligi (bir xil bo‘lmasligi);
5. NERni teglashda ularni, avvalo, matndan to‘g‘ri aniqlash;
6. Aniqlangan NERni to‘g‘ri klassifikatsiya qilish;
7. NERlarning omonimiya holatiga e’tibor qaratish lozim.

Shuningdek, NERlarni teglash muammolarini hal qilish uchun bir nechta asosiy yo‘nalishlar taklif etiladi:

1. **Annotatsiyalangan katta hajmdagi korpus yaratish** – O‘zbek tiliga xos NER tizimini rivojlantirish uchun katta hajmdagi, aniq va tuzilmali ma’lumotlar to‘plamini yaratish zarur.
2. **Transformer modellaridan foydalanish** – BERT, CNN, LSTM kabi ilg‘or chuqur o‘rganish (DL) metodlarini o‘zbek tiliga moslashtirish va ushbu modellar yordamida NER obyektlarini aniq ajratish texnikalarini rivojlantirish.
3. **Morfologik va sintaktik tahlilni kuchaytirish** – So‘zlarning turli shakllarda qo‘llanilishi sababli tokenizatsiya va lemmatizatsiya kabi metodlardan foydalanish NER tizimining samaradorligini oshirishga yordam beradi.

4. **Ochiq ma’lumotlar bazalarini kengaytirish** – O‘zbek tilidagi ochiq kodli NER korpuslarini yaratish va ulardan ilmiy hamda amaliy loyihalarda foydalanish orqali tahlillarni takomillashtirish.

5. **O‘zbek tiliga xos maxsus lug‘atlar bazasini ishlab chiqish** – NER tizimining aniq va samarali ishlashi uchun maxsus lug‘at va ma’lumotlar to‘plamlarini ishlab chiqish zarur.

XULOSA

O‘zbek tilida NER teglash tizimlarini ishlab chiqish va rivojlantirish katta imkoniyatlarga ega, lekin bir qancha muammolarni ham o‘z ichiga oladi. Morfologik va sintaktik xususiyatlar, kontekstni to‘g‘ri tahlil qilish, resurslarning yetishmasligi kabi qiyinchiliklarga qaramay, ilg‘or texnologiyalar va innovatsion yondashuvlar yordamida bu masalalarni hal qilish mumkin. NER tizimlarini takomillashtirish va yangi resurslar yaratish o‘zbek tilini kompyuter yordamida qayta ishlash sohasida katta yutuqlarga erishishga imkon beradi.

Maqolada O‘zbek tili korpusida NER obyektlarini teglash jarayonida duch kelinadigan asosiy muammolar va ularni hal etish yo‘llari atroflicha yoritildi. O‘zbek tilining morfologik va sintaktik murakkabligi, tegishli annotatsiyalangan ma’lumotlar bazasining yetishmovchiligi va ochiq manbalardan foydalanish imkoniyatlarining cheklanganligi NER tizimlarini yaratish jarayonini qiyinlashtiruvchi omillar sifatida e’tirof etilgan.

Maqolada ilgari surilgan yechimlar O‘zbek tilida NER tizimlarini rivojlantirish va tabiiy tilni qayta ishlash bo‘yicha global standartlarga mos keluvchi model yaratish imkonini beradi. Ushbu yo‘nalishdagi tadqiqotlar va amaliy tajribalar davom ettirilsa, kelajakda o‘zbek tili uchun yuqori samarali NER tizimi ishlab chiqish imkoniyati yuzaga keladi.

Foydalanilgan adabiyotlar ro‘yhati:

1. https://www.altexsoft.com/blog/named-entity-recognition/
2. https://qalampir.uz/news/prezident-administratsiyasining-sobik-ra%D2%B3bari-tuman-%D2%B3okimi-buldi-117080
3. Jehangir, S. Radhakrishnan, R. Agarwal. **A Survey on Named Entity Recognition – Datasets, Tools, and Challenges. 2023. P-3**
4. U Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, Dai Dai. **Learn and Review: Enhancing Continual Named Entity Recognition.** 2022. P-5
5. Keraghel, Stanislas Morbieu, Mohamed Nadifning. A Survey on Recent Advances in Named Entity Recognition. 2023. P-12

6. Sakher Khalil Alqaaidi, Elika Bozorgi, Afsaneh Shams, Krzysztof Kochut kabi olimlarning. A Few-Shot Learning Focused Survey on Recent Named Entity Recognition and Relation Classification Methods. 2023. P-10

7. https://rykov-cl.narod.ru/c.html

8. https://kun.uz/news/2025/03/13/ozbekiston-va-fransiya-strategik-sherik-boldi-maqsadlar-va-manfaatlar-nima

9. V. Zaxarov, B. Mengliyev, Sh.Hamroyeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish, o‘quv qo‘llanma. Toshkent. 2021. 139