O‘ZBEK TILI MATNLARIDA NOMLOVCHI BIRLIKLARNI TEGLASH MUAMMOLARI

Nilufar Abduraxmonova
O‘zbekiston Milliy universiteti professori, filologiya fanlari doktori
Kompyuter lingvistikasi va amaliy tilshunoslik kafedrasi mudiri
Kanayeva Ozoda Esirgapovna
O‘zbekiston Milliy universiteti Jurnalistika va o‘zbek filologiyasi fakulteti
Kompyuter lingvistikasi mutaxassisligi I bosqich magistranti

Annotatsiya. Ushbu maqolada o‘zbek tili matnlarida nomlovchi birliklarni (named entities) avtomatik teglash jarayonida yuzaga keladigan asosiy muammolar tahlil qilinadi. Xususan, shaxs nomlari, joy nomlari, tashkilotlar va boshqa maxsus birliklarni aniqlash va ularni to‘g‘ri klassifikatsiya qilish jarayonidagi lingvistik va texnik murakkabliklar ko‘rib chiqiladi. O‘zbek tilining agglutinativ xususiyati, erkin so‘z tartibi hamda morfologik boyligi nomlovchi birliklarni aniqlashda qo‘shimcha qiyinchiliklarni yuzaga keltiradi. Shuningdek, maqolada qoidaviy, statistik va neyron yondashuvlar asosida nomlovchi birliklarni teglash usullarining afzallik va cheklovlari muhokama qilinadi. Resurslarning cheklanganligi, belgilangan korpuslarning yetishmasligi hamda standartlashtirilgan lug‘at va ontologiyalarning kamligi ushbu sohadagi asosiy muammolar sifatida ajratib ko‘rsatiladi. Tadqiqot natijasida o‘zbek tilida nomlovchi birliklarni aniqlash va teglash samaradorligini oshirish uchun gibrid yondashuvlardan foydalanish, shuningdek, milliy lingvistik resurslarni kengaytirish zarurligi asoslab beriladi.

Kalit so‘zlar: nomlovchi birliklarni aniqlash, avtomatik teglash, o‘zbek tili, korpus lingvistikasi, neyron tarmoqlar, tabiiy tilni qayta ishlash.

Abstract. This study analyzes the current challenges of named entity recognition and automatic tagging in Uzbek language corpora. It examines the degree to which named entity recognition models are adapted to the Uzbek language and discusses the main linguistic and technical factors affecting their performance. The paper also summarizes effective approaches aimed at improving annotation accuracy and enhancing the overall quality of automatic tagging.

Keywords: named entity recognition, automatic tagging, Uzbek language, corpus linguistics, neural networks, natural language processing.

Аннотация: В данной статье рассматриваются основные проблемы автоматической разметки именованных сущностей в текстах на узбекском языке. Особое внимание уделяется трудностям идентификации и классификации таких сущностей, как имена лиц, географические названия, организации и другие типы именованных единиц. Агглютинативная природа узбекского языка, свободный порядок слов и богатая морфологическая структура создают дополнительные сложности в процессе разметки. Кроме того, анализируются правила-ориентированные, статистические и нейронные подходы к задаче распознавания именованных сущностей, а также их преимущества и ограничения. В качестве ключевых проблем выделяются ограниченность лингвистических ресурсов, недостаток размеченных корпусов и отсутствие стандартизированных словарей и онтологий. В результате исследования обосновывается необходимость применения гибридных подходов и расширения национальных лингвистических ресурсов для повышения эффективности разметки именованных сущностей в узбекском языке.

DOI: 10.5281/zenodo.19873696

Ключевые слова: распознавание именованных сущностей, автоматическая разметка, узбекский язык, корпусная лингвистика, нейронные сети, обработка естественного языка.

1. Kirish. Nomlangan birliklarni aniqlash (NER) tabiiy tilni qayta ishlash (NLP) sohasidagi muhim vazifalardan biri bo‘lib, matndan shaxs (PER), joy (LOC), tashkilot (ORG) va boshqa nomlarni aniqlash hamda tasniflashni o‘z ichiga oladi. Ushbu vazifa axborotni avtomatik tahlil qilish, bilim bazalarini yaratish, qidiruv tizimlari va ma’lumotlarni anonimlashtirish jarayonlarida keng qo‘llaniladi [1].

Yuqori resursli tillarda NER tizimlari katta hajmdagi annotatsiyalangan datasetlar asosida yuqori aniqlikka erishgan bo‘lsa-da, o‘zbek tili kabi kam resursli tillarda bu jarayon ancha murakkabdir. Bunga asosiy sabab sifatli korpuslar, belgilangan datasetlar va lingvistik vositalarning yetishmasligidir [2].

2. O‘zbek tilida NERning lingvistik muammolari

O‘zbek tili agglyutinativ til bo‘lgani sababli, so‘z shakllari juda ko‘p variantlarga ega. Bitta o‘zakka turli qo‘shimchalar qo‘shilishi natijasida yuzlab yangi shakllar hosil bo‘ladi. Bu esa NER tizimlarida token darajasida umumlashtirishni qiyinlashtiradi. Masalan: *“Toshkent”*, *“Toshkentda”, “Toshkentdan”* – bular bitta entity bo‘lsa ham model uchun turli tokenlar hisoblanadi.

Shu sababli morfologik tahlil va lemmatizatsiya NER tizimining ajralmas qismi hisoblanadi. Abduraxmonova va hammualliflarning morfologik dataset bo‘yicha ishlari o‘zbek tilida bu muammoni hal qilish uchun muhim asos yaratadi [4].

Bundan tashqari, o‘zbek tilida katta harf bilan yozish har doim ham entity belgisi bo‘la olmaydi, bu esa ingliz tilidagi kabi oddiy qoidalarga asoslangan yondashuvlarni cheklaydi.

3. Ma’lumotlar to‘plami va belgilash muammolari

Nomlovchi birliklarni aniqlash tizimining muvaffaqiyati, avvalo, unga tayanch bo‘ladigan ma’lumotlar to‘plamining sifati bilan belgilanadi. Kam resursli tillarda, jumladan o‘zbek tilida, eng katta muammo yetarli hajmdagi va ilmiy mezonlar asosida belgilangan matnlar bazasining kamligidir. Tizim qanchalik zamonaviy bo‘lmasin, agar o‘qitish uchun berilayotgan ma’lumotlar bir xil mezonda tayyorlanmagan bo‘lsa, uning natijasi barqaror bo‘lmaydi.

Ma’lumotlar to‘plamini yaratishda birinchi masala – nomlovchi birlik chegarasini to‘g‘ri belgilashdir. Ayrim hollarda bitta nom bir necha so‘zdan tashkil topadi, biroq annotatsiya jarayonida uning bir qismi tushirib qoldirilishi yoki ortiqcha so‘z ham birlik tarkibiga qo‘shib yuborilishi mumkin. Bu esa keyinchalik modelning noto‘g‘ri o‘rganishiga olib keladi. Xususan, murakkab tashkilot nomlari, davlat idoralari, oliy ta’lim muassasalari, tarixiy joy nomlari ana shunday xatolarga ko‘proq og‘uvchan bo‘ladi.

Ikkinchi muammo – ko‘p ma’nolilikdir. Bir xil so‘z turli matnlarda turlicha vazifada qo‘llanishi mumkin. Masalan, ayrim birliklar ham joy nomi, ham familiya, ham tashkilot tarkibidagi so‘z sifatida uchraydi[4]. Bunday vaziyatda faqat so‘z shakliga tayanish yetarli emas, balki uning atrofidagi kontekst, mavzu sohasi va gapdagi vazifasi ham e’tiborga olinishi kerak bo‘ladi.

Uchinchi muammo – sohaviy farqlardir. Badiiy matn, yangilik matni, rasmiy hujjat, ijtimoiy tarmoq posti yoki ilmiy maqolada nomlovchi birliklarning ifodalanish usuli bir xil emas[4]. Rasmiy hujjatlarda tashkilot va hudud nomlari nisbatan to‘liq va aniq yozilsa, ijtimoiy tarmoqlarda qisqartmalar, og‘zaki shakllar, imlo xatolari va norasmiy yozuvlar ko‘p

uchraydi. Shu sababli bir soha uchun tuzilgan model boshqa sohada sust natija berishi mumkin.

To‘rtinchi muammo – belgilovchilar o‘rtasidagi kelishuv masalasidir. Agar ikki nafar mutaxassis bir xil matndagi birliklarni har xil belgilasa, bu holat ma’lumotlar to‘plamining ichki bir xilligi buzilganini ko‘rsatadi. Shu sababli belgilashdan oldin aniq qo‘llanma ishlab chiqilishi, har bir toifaning mezonlari izchil tushuntirilishi va kamida ikki belgilovchi ishtirokida tekshiruv o‘tkazilishi zarur.

Shu nuqtai nazardan, kam sonli, lekin puxta belgilangan va tekshirilgan ma’lumotlar to‘plami katta hajmli, biroq tartibsiz bazadan ko‘ra ko‘proq ilmiy qiymatga ega bo‘ladi. Ayniqsa o‘zbek tili kabi resurslari cheklangan tillarda dastlab kichik, ammo aniqligi yuqori bo‘lgan namunaviy korpuslar yaratish eng maqbul yo‘l sanaladi.

1-jadval. O‘zbek tilida nomlovchi birliklarni belgilashdagi asosiy muammolar

Muammo turi	**Mazmuni**	**Tizimga ta’siri**	**Tavsiya etiladigan yechim**
Birlik chegarasini aniqlash	Ko‘p so‘zli nomning boshlanishi va tugashini belgilash qiyin	Noto‘g‘ri teglash, past aniqlik	Aniq annotatsiya qo‘llanmasi tuzish
Ko‘p ma’nolilik	Bir so‘z turli vazifada qo‘llanishi mumkin	Toifa adashishi ortadi	Keng kontekstni hisobga olish
Sohaviy farq	Turli janrlarda nomlar turlicha ifodalanadi	Model boshqa sohada sust ishlaydi	Turli manbalardan korpus tuzish
Imlo va yozuv farqlari	Qisqartma, norasmiy yozuv, xato shakllar uchraydi	Tanish darajasi pasayadi	Tozalash va me’yorlashtirish
Beligilovchilar kelishuvi	Turli annotatorlar turlicha belgilashi mumkin	Dataset sifati pasayadi	Ikki bosqichli tekshiruv va kelishuv ko‘rsatkichi

4. Natijalar va tahlil. O‘zbek tilida nomlovchi birliklarni avtomatik aniqlash tajribalari shuni ko‘rsatadiki, barcha toifalar bir xil darajada aniqlanmaydi. Odatda shaxs nomlari va joy nomlari nisbatan yuqoriroq aniqlik bilan topiladi. Buning asosiy sababi shundaki, ushbu birliklar ko‘pincha aniq kontekstda, izchil ko‘rinishda va boshqa toifalarga qaraganda barqarorroq shaklda qo‘llanadi. Masalan, shaxs nomlari ko‘pincha ism-familiya shaklida, joy nomlari esa ma’lum hudud, viloyat, tuman yoki davlat nomlari sifatida aniq ifodalanadi.

Tashkilot nomlari esa nisbatan murakkabroq guruh hisoblanadi. Chunki ular ko‘p hollarda uzun birikma tarzida keladi, qisqartmalar bilan ifodalanadi yoki tarkibida umumiy ma’noli so‘zlar bo‘ladi. Shu sababli model tashkilot nomining qayerdan boshlanishi va qayerda tugashini aniqlashda ko‘proq qiynaladi. Bundan tashqari, ayrim tashkilot nomlari matnda umumiy otga o‘xshab ketishi ham noto‘g‘ri tasniflashga sabab bo‘ladi.

Tahlillar shuni ko‘rsatadiki, kontekstni chuqur hisobga oluvchi modellar, ayniqsa so‘zning oldi va keyingi muhitini birgalikda ko‘ra oladigan tizimlar oddiy yondashuvlarga nisbatan yuqoriroq natija beradi. Biroq bu ustunlik asosan ma’lumotlar to‘plami sifatli tuzilgan va belgilash mezonlari izchil bo‘lgan holatlarda namoyon bo‘ladi. Aks holda, eng kuchli model ham noto‘g‘ri yoki notekis belgilangan ma’lumot ustida barqaror ishlay olmaydi.

Shuningdek, so‘zning tarkibiy qismlarini alohida tahlil qiluvchi usullar ham foydali natija beradi. O‘zbek tilida qo‘shimchalar juda faol qo‘llanilishi bois, so‘zning faqat yuzaki ko‘rinishini emas, balki uning tarkibini ham hisobga olish kerak bo‘ladi. Ayniqsa kelishik,

egalik va ko‘plik qo‘shimchalari bilan kelgan joy yoki shaxs nomlarini to‘g‘ri tanish uchun shunday bosqich zarur.

Natijalarni umumlashtirib aytganda, o‘zbek tili uchun nomlovchi birliklarni aniqlashda muvaffaqiyat ko‘p jihatdan uch omilga bog‘liq: birinchisi – sifatli va muvozanatli ma’lumotlar to‘plami, ikkinchisi – morfologik xususiyatlarni hisobga oluvchi tilga mos model, uchinchisi esa – belgilash mezonlarining izchil va aniq bo‘lishidir.

5. Xulosa. O‘zbek tilida nomlovchi birliklarni avtomatik aniqlash masalasi tilning agglyutinativ xususiyati, morfologik murakkabligi va resurslarning cheklanganligi sababli murakkab vazifa hisoblanadi. Ushbu jarayonda ma’lumotlar to‘plamining sifati, annotatsiya mezonlarining aniqligi va morfologik tahlilning mavjudligi hal qiluvchi ahamiyatga ega. Tahlillar shuni ko‘rsatadiki, shaxs va joy nomlari nisbatan yuqori aniqlikda aniqlansa, tashkilot nomlarida murakkab tuzilma va kontekstga bog‘liqlik sabab qiyinchiliklar ko‘proq uchraydi. Shu bois o‘zbek tili uchun lingvistik qoidalar, sifatli korpus va zamonaviy modellarni uyg‘unlashtirgan yondashuv eng maqbul yo‘l hisoblanadi. Kelgusida ochiq va belgilangan korpuslarni kengaytirish, annotatsiya standartlarini takomillashtirish hamda tilga mos gibrid tizimlarni yaratish ushbu yo‘nalish rivojida muhim ahamiyat kasb etadi.

ADABIYOTLAR:

1. D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd ed. Stanford University, 2026.
2. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” in Proceedings of NAACL-HLT 2016, 2016.
3. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” in Proceedings of NAACL-HLT 2018, pp. 2227–2237, 2018.
4. D. Mengliev, V. Barakhnin, N. Abdurakhmonova, and M. Eshkulov, “Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation,” Data in Brief, vol. 54, 2024.
5. E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition,” in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147, 2003.
6. L. A. Ramshaw and M. P. Marcus, “Text Chunking Using Transformation-Based Learning,” in Proceedings of the Third Workshop on Very Large Corpora, 1995.
7. N. Z. Abdurakhmonova, R. K. Shirinova, R. Sayfullayeva, D. B. Mengliev, B. B. Ibragimov, and M. Ernazarova, “An annotated morphological dataset for Uzbek word forms: Towards rule-based and machine learning approaches,” Data in Brief, vol. 61, Art. 111702, 2025.
8. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” OpenAI, 2018.
9. J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” in Proceedings of ACL 2018, pp. 328–339, 2018.