



Botir ELOV,

Toshkent davlat o'zbek tili va adabiyoti universiteti dotsenti, PhD

E-mail: elov@navoiy-uni.uz

Madina SAMATBOYEVA,

Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch doktoranti

"Alfraganus" universiteti dotsenti, PhD A.Dauletov taqrizi asosida

O'ZBEK TILI KORPUSIDA NERNI TEGTLASHNING UMUMIY QOIDALARI

Annotatsiya

Ushbu maqola o'zbek tili korpusida nomlangan obyektlarni tanib olish (NER – Named Entity Recognition) tizimlarini yaratish va ularni teglash jarayonlariga bag'ishlangan. Maqolada NER tizimlarining ahamiyati, ularning tabiiy tilni qayta ishlashdagi roli, shuningdek, o'zbek tilida ushbu tizimlarni ishlab chiqishdagi mavjud qiyinchiliklar va muammolar tahlil qilinadi. O'zbek tilining morfologik va sintaktik xususiyatlari, shuningdek, fonetik va leksik farqlari NER tizimlarining aniqligini oshirishda muhim omillar hisoblanadi. Shuningdek, o'zbek tilida NER tizimlarini yaratish va ularni samarali ishlashini ta'minlash uchun zarur bo'lgan metodologiyalar va amaliy yondoshuvlar ko'rib chiqiladi. Taqdim etilgan materiallar va tahlillar texnik xodimlar, kompyuter tilshunoslari va sohada ishlayotgan tadqiqotchilar uchun foydali manba bo'lishi mumkin.

Kalit so'zlar: NLP, NER, nomlangan obyekt, atoqli ot, korpus, annotatsiya, ML, DL, teg, model.

GENERAL RULES FOR TAGGING NER IN THE UZBEK LANGUAGE CORPUS

Annotation

This article is dedicated to the creation of Named Entity Recognition (NER) systems in the Uzbek language corpus and the process of tagging them. The article discusses the importance of NER systems, their role in natural language processing, as well as the challenges and problems encountered in developing these systems for the Uzbek language. The morphological and syntactic characteristics of the Uzbek language, as well as its phonetic and lexical differences, are key factors in improving the accuracy of NER systems. Additionally, the article covers the necessary methodologies and practical approaches to ensure the effective development and functioning of NER systems in the Uzbek language. The materials and analyses presented can be a valuable resource for technical staff, computational linguists, and researchers working in the field.

Key words: NLP, NER, named entity, proper noun, corpus, annotation, ML, DL, tag, model.

ОБЩИЕ ПРАВИЛА ТЕГИРОВАНИЯ NER В КОРПУСЕ УЗБЕКСКОГО ЯЗЫКА

Аннотация

Эта статья посвящена созданию систем распознавания именованных объектов (NER – Named Entity Recognition) в корпусе узбекского языка и процессу их тегирования. В статье рассматривается важность систем NER, их роль в обработке естественного языка, а также существующие трудности и проблемы, с которыми сталкиваются при разработке таких систем для узбекского языка. Морфологические и синтаксические особенности узбекского языка, а также фонетические и лексические различия являются важными факторами для повышения точности систем NER. Также рассматриваются необходимые методологии и практические подходы для создания эффективных систем NER на узбекском языке. Представленные материалы и анализы могут быть полезным источником для технических специалистов, компьютерных лингвистов и исследователей, работающих в этой области.

Ключевые слова: NLP, NER, именованные объекты, собственные имена, корпус, аннотация, ML, DL, тег, модель.

Kirish. Til juda murakkab ta'lim bo'lganligi tufayli kompyuter tilshunosligida avtomatik matn tahlili; avtomatik matn sintezi; avtomatik lug'atlarni yaratish va qo'llab-quvvatlash; avtomatlashtirilgan axborot-qidiruv tizimlarini yaratish; mashina tarjimasini; avtomatik til o'rganish tizimlarini yaratish; lingvistik ma'lumotlar bazalarini yaratish; nazariy va amaliy tilshunoslik masalalarini yechishning dasturiy vositalarini ishlab chiqish kabi turli yo'nalishlar rivojlanib bormoqda [1].

Atrofimizdagi dunyoda barcha mavjudotlar, narsalar, jonli va jonsiz tabiat o'zlarining farqlanish va individuallashtirish tendensiyasi natijasida nomlanadi. Bu nomlar juda ko'p bo'lib, ularni bir vaqtning o'zida tushunish qiyin. Garchi obyektlar umumiy nomlar (kategoriya) ostida birlashtirilgan bo'lsa-da, ular boshqa "nomlar" asosida ularning tuzilishi, xususiyatlari, ma'lum bir kategoriya bilan sababli qayta nomlanadi [2].

Kompyuter lingvistikasini kompyuter vositalaridan foydalanish sohasi sifatida dasturlar, ma'lumotlarni qayta ishlash texnologiyalarini tashkil etish va muayyan sharoitlarda til faoliyatini modellashtirish uchun, shuningdek, tilshunoslikda tilning kompyuter modellari va unga oid fanlar doirasida qo'llash mumkin.

O'zbek tilining xalqimiz ijtimoiy hayotida va xalqaro miqyosdagi obro'-e'tiborini tubdan oshirish, unib-o'sib

kelayotgan yoshlarimizni vatanparvarlik, milliy an'ana va qadriyatlariga sadoqat ruhida tarbiyalash, mamlakatimizda davlat tilini to'laqonli joriy etishni ta'minlash, O'zbekistondagi millat va elatlarning tillarini saqlash va rivojlantirish, davlat tili sifatida o'zbek tilini o'rganish uchun shart-sharoitlar yaratish, o'zbek tili va til siyosatini rivojlantirishning strategik maqsadlari, ustuvor yo'nalish va vazifalarini hamda istiqboldagi bosqichlarini belgilash maqsadida, shuningdek O'zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi "O'zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to'g'risida"gi PF-5850-son farmoniga muvofiq: 2020 — 2030-yillarda o'zbek tilini rivojlantirish va til siyosatini takomillashtirish konsepsiyasi tasdiqlangan va unda "davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga faol integratsiyalashuvini ta'minlash" [3] qismi alohida kiritilgan. Shuningdek, dasturiy mahsulotlarning o'zbekcha ilovalari va elektron lug'at dasturlarini yaratish, xorijliklar uchun o'zbek tilini o'rgatuvchi dasturlar yaratish ishlari ham ustuvor vazifa sifatida belgilangan.

O'tgan yarim asrlik vaqt mobaynida kompyuter lingvistikasi sohasida bir qator ilmiy va amaliy natijalarga erishilgan: tabiiy tilda avtomatik tarjima tizimi yaratildi, matndagi ma'lumotlarning avtomatik qidiruv tizimi ishlab chiqildi, og'zaki

nutqning avtomatik analiz va sintez tizimi yaratildi, bir qator lingvistik muammolarni hal etuvchi kompyuter dasturlari ishlab chiqildi, inson va mashina(kompyuter) muloqoti optimallashtirildi, tabiiy tilni qayta ishlash (Natural Language Processing) tizimi shakllantirildi. To'g'ri, bu yo'nalishlarda sezilarli yutuqlar qo'lga kiritilgan bo'lsa ham, lekin sanalgan masalalar hali to'liq o'z yechimini topgani yo'q. Chunki inson tili va tafakkuri sirli mohiyatga ega, uning biz anglamagan qirralari juda ko'p. Bu esa kompyuter lingvistikasi doirasida to'xtovsiz izlanishlar olib borishni taqozo etadi.

Mavzuga oid adabiyotlar tahlili. NER obyektlari atolqi otlarni ifodalaydi. O'zbek tili uchun atolqi otlar grammatikasiga tayanib NERlar aniqlanadi. Aniqlangan otlar nafaqat morfologik, balki semantik jihatdan ham tahlil qilinadi. Tahlil jarayonida otlarning shakliga ham e'tibor qaratiladi. Matndan ajratilgan otlar ularning "maxsus nom" xususiyatlariga ko'ra aniqlanadi. Gapdagi nomlangan otlar ajratiladi. Maxsus otlar "NER xususiyatlari"ga ko'ra tahlil qilinadi (matnda katta harf bilan yozilganlik, lug'atdan tashqari leksikalik...). Lug'atlarga asoslangan yondashuv ham NER obyektlarini aniqlashda yordam beradi. Biroq bu yondashuv ham mukammal samarali yechim bo'la olmaydi. Mukammal yechimlar barchasi ML va DL metodlari asosida amalga oshiriladi [4].

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 sohasida yanada rivojlanib, murakkab tahlil imkoniyatlarini yaratadi.

Kompyuter mutaxassislaridan biri Daniel Martin o'z tajribalari davomida bir necha bor NLPga qo'l uradi va ushbu sohada o'z fikrlarini keng ommaga e'lon qiladi. Uning fikricha, "NLP, ya'ni tabiiy tilni qayta ishlash, bu kompyuter fanlari, sun'iy intellekt va tilshunoslik chorrahasida joylashgan sohadir. U kompyuterlarga inson tilini tushunish, talqin qilish va yaratish imkonini berishga qaratilgan. NLP muhandislari qidiruv tizimlari, ovozli yordamchilar va hissiyotlarni tahlil qilish kabi ilovalarga ruxsat beruvchi katta hajmdagi matn ma'lumotlarini qayta ishlash va tahlil qilish tizimlarini ishlab chiqishda hal qiluvchi rol o'ynaydi deb ta'kidlaydi [5].

NLP tarkibida NERlar o'rganilganda, J.R. Finkel va hamkasblari o'z maqolalarida NERning tabiati va uning NLP sohasidagi ahamiyatini keng yoritadilar. Ular NERni nafaqat ma'lum obyektlarni aniqlash, balki ularning o'zaro bog'liqligini va kontekstual ma'nolarini to'g'ri tushunishni ham ta'minlaydigan tizim sifatida ko'rsatishadi. Ularning fikricha, NER texnologiyasi rivojlangan va uni mashinasozlik va lingvistik metodlarning uyg'unlashuvi orqali takomillashtirish mumkin [6].

NERni to'liq tizim sifatida ko'rib chiqqan ikki olim A. Nadeau va S. Sekine uni ikkita asosiy qismga ajratadilar: 1) Obyektni aniqlash va 2) Ma'lum tizimga ajratish. Ular shuningdek, NER tizimlarining turli turlari va yondashuvlaridan, masalan, qoidalarga asoslangan va statistik metodlardan foydalanishdan ham so'z yuritadilar. Maqolada, Nadeau va Sekine, NERni ishlab chiqishda afzallik beriladigan yondashuvlar va metodlarni ta'kidlashadi [7].

David Milne va Ian H. Witten o'z maqolalarida NERni tasniflash va aniqlashning turli metodlari bilan tanishtiradilar. Ular, ayniqsa, matnlarni annotatsiya qilishda semantik tuzilmani

tushunishning zarurligini ta'kidlaydilar. Ular NERni nafaqat to'g'ri obyektlarni aniqlash, balki bu obyektlar orasidagi o'zaro bog'lanishni ham ko'rsatish vositasi sifatida baholashadi [8].

Tadqiqot metodologiyasi. O'zbek tilidagi NER tizimlarini to'g'ri va samarali ishlab chiqish uchun zarur metodikalarini belgilash va ularni amaliyotga tatbiq etish muhimdir. Tadqiqot quyidagi metodologik yondashuvlarni o'z ichiga oladi:

Korpusni tahlil qilish va matn tozalash. Tadqiqotning birinchi bosqichida o'zbek tili korpusini shakllantirish va uni grammatik tahlil qilish kerak. O'zbek tilidagi matnlarni korpus yordamida yig'ish uchun mavjud bo'lgan elektron matnlar, gazeta maqolalari, adabiy asarlar va boshqa manbalar tahlil qilinadi. Korpusda mavjud bo'lgan turli xildagi nomlangan obyektlarni aniqlash uchun grammatik va leksik analizlardan foydalaniladi. Korpus tarkibiga kiritiluvchi matnlar ham "matn tozalash" jarayonidan o'tkaziladi. Bunda matn va gaplar keraksiz unsurlar (Removing extra spaces and empty lines – ortiqcha bo'shliqlar va bo'sh qatorlar, Cleaning special characters (excluding certain necessary ones) – maxsus belgilar; Removing URLs, emails, and HTML tags – internet manzillari va teglari) dan tozalanadi.

NER tizimlarining teglash qoidalarini ishlab chiqish. NER tizimlarining to'g'ri ishlashini ta'minlash uchun, o'zbek tiliga xos bo'lgan teglash qoidalarini ishlab chiqiladi. Bu qoidalar orasida so'zlarning morfologik tahlili, ularning sintaktik vazifalari, semantik ma'nolari va kontekstual xususiyatlari hisobga olinadi. Teglash jarayonida nomlar, joylar, tashkilotlar va boshqa obyektlar aniqlanadi. Shuningdek, ko'p ma'noli so'zlar va sinonimlar uchun qo'shimcha teglar kiritiladi.

Aniqlangan NERlarga toifa berish. NER tizimlarining samarali ishlashini ta'minlash uchun matndagi NERlar tasnifi amalga oshiriladi. Har bir NER turini (shaxs nomi, joy nomi, tashkilot nomi va boshqa) alohida teglash ko'rinishlari bilan belgilash kerak. Shuningdek, o'zbek tilining morfologik xususiyatlarini hisobga olib, so'z birikmalari bilan bog'liq alohida qoidalar ishlab chiqiladi.

Teglash usullarini tanlash. Ushbu metod orqali NER tizimlarining teglash usullari aniq belgilanadi. Bir qancha usullar orasida qoidali yondashuv, mashinali o'qitish (ML - Machine Learning) va chuqur o'rganish (DL - Deep Learning) metodlari sinovdan o'tkaziladi. MLni o'rgatishda eng ko'p ishlatiladigan algoritmlar, masalan, Naive Bayes, Support Vector Machines (SVM), va Conditional Random Fields (CRF) kabi usullarni sinovdan o'tkazish rejalashtirilgan. Shu bilan birga, DL asosida ishlovchi modellar, xususan, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) va Transformer metodlar samaradorligi ham tahlil qilinadi.

Tahlil va natijalar. NERlarni aniqlashda, avvalo, NERlar ro'yxati (NER objects list) shakllantiriladi. NER deb qabul qilingan obyektlar tavsifiga (ya'ni nimani ifodalab kelishiga) ko'ra ularga maxsus teglar belgilanadi. Jahon NLP sohasida NER ro'yxati bir qancha turlarni o'z ichiga oladi. Ular tarkibida hamma obyektlar ham o'zbek tili uchun "NERlar ro'yxatini" shakllantirilmaydi. Bunga sabab, turli til oilalari, turli grammatik qurilishlar, turli til tarixi va so'zlarning paydo bo'lishidagi farqlardir (Qarang:1-rasm). Tadqiqot obyektimiz sifatida, o'zbek tili grammatik qoidalaridan kelib chiqib ular orasidan "O'zbek tili uchun NERlar" sifatida bir qancha obyektlarni ajratib oldik. Ushbu obyektlar asosiy katta turlarni hosil qilib, ularga maxsus teglar modellar asosida taqdim etiladi.

PERSON	People including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements such as weight or distance
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type

1-rasm. Ingliz til asosida yaratilgan DeepPavlov NER-modelida qo'llanilgan NER obyektlari [9].

O'zbek tili uchun mos bo'lgan NER obyektlari ajratib olindi. Ular asosiy katta tur NERlar bo'lib, inson ismlari, joy nomlari, tashkilot nomlari, qisqartma shaklidagi atoqli otlar qabul qilindi. NER yordamida aniqlangan obyektlar, odatda, joylar yoki tashkilotlarni anglatadi. Biroq, ular maxsus narsalarni ham anglatishi mumkin. Bir obyekt bir so'z yoki doimo bir xil narsani

anglatadigan so'zlar ketma-ketligi bo'lishi mumkin. NERni amalga oshirishda maxsus obyektlarni ajratib olish va ularga qaysi so'zlar birikishi to'g'risidagi qoidalarni ishlab chiqish muhimdir [10]. Ularga teglar tanlashda dunyo standartlariga mos holatda taqdim etildi (Qarang: 1-jadval).

№	Inglizcha	O'zbekcha	Misol	Teglar
1	Person	Inson ismlari	Lola, Diyor	PER
2	Location	Joy nomlari (manzillar)	Toshkent, Zarafshon	LOC
3	Organization (with abbreviation)	Tashkilot nomlari	"Asaka" bank, BMT, SamSI	ORG
4	Facility	Qulaylik	Toshkent xalqaro aeroporti	FAC
5	Work of art	San'at asari (adabiyot, san'at...)	"O'tkan kunlar" asari, "Oqqushlar" kompozitsiyasi	ART

1-jadval. O'zbek tilidagi NERlar uchun taqdim etilgan teglar.

Xulosa va takliflar. 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.

O'zbek tili korpusi uchun NER tizimlarini samarali yaratish va ularning ishlashini optimallashtirishda muhim ahamiyatga ega. Belgilangan metodlar, usullar va tahlillar natijasida olingan bilimlar nafaqat ilmiy, balki amaliy jihatdan ham muhimdir.

ADABIYOTLAR

1. Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари. Филол. фан. доктори (DSc) диссертацияси. – Тошкент, – В.4
2. Elov, Botir B., and Madina T. Samatboyeva. "NER: Methods FOR Automatic Identification of Anthroponyms in Uzbek Texts." 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). IEEE, 2024. <https://doi.org/10.1109/PIERE62470.2024.10805025>
3. <https://lex.uz/uz/docs/-5058351>
4. Elov B., M.Samatboyeva. Identifying NER (Named Entity Recognition) objects in uzbek language texts. science and innovation international scientific journal volume 2 issue. Toshkent., 44-57. UIF-2022: 8.2 | ISSN: 2181-3337 | Scientists.Uz.
5. Martin D. Tabiiy tilni qayta ishlash (NLP) muhandisi: mas'uliyat va yo'l xaritasi. 2023. 13-b.
6. <https://www.unite.ai/uz/tabiiy-tilni-qayta-ishlash-nlp-muhandisi-mas%27uliyati-yo%27l-xaritasi/>
7. J.R. Finkel, T. Grenager, C. Manning. Understanding Named Entity Recognition. 2005 – p 5.
8. A. Nadeau, S. Sekine. A Survey of Named Entity Recognition and Classification. 2007 – p 3.
9. David Milne, Ian H. Witten. Named Entity Recognition and Classification. 2008 – p 8.
10. <https://docs.deeppavlov.ai/en/master/features/models/NER.html>
11. Maria L. Coppola. Named entity recognition tagging: How to tag data & recognise entities [+ Tools]. Datavid published. 2023 – p 55.