O‘ZBEK TILI KORPUSIDA NER INSTRUMENTARIYSI ALGORITMINING FUNKSIONAL IMKONIYATLARI

Nazokat O‘rinboyeva
E-mail: nnazokat71@gmail.com

Qabul qilindi: 02.11.2025
Ko‘rib chiqildi: 03.11.2025
Tasdiqlandi: 04.11.2025

Kalit so‘zlar:

O‘zbek tili korpusi, nomlovchi birliklarni aniqlash, NER algoritmi, tabiiy tilni qayta ishlash, BIOES formati, morfologik tahlil, lingvistik resurslar, sun’iy intellekt.

.

Annotatsiya: Ushbu maqolada o‘zbek tili korpusida nomlovchi birliklarni aniqlash (Named Entity Recognition — NER) jarayonini avtomatlashtirishga qaratilgan algoritmning nazariy asoslari va amaliy qo‘llanilishi yoritilgan. Tadqiqotda ishlab chiqilgan NER instrumentariysi o‘zbek tilining morfologik, sintaktik va semantik tuzilishiga moslashtirilgan bo‘lib, u korpus lingvistikasi hamda tabiiy tilni qayta ishlash (NLP) yo‘nalishlari kesishmasida yaratilgan innovatsion yechim sifatida baholanadi. Mazkur instrumentariy avtomatik teglash, so‘z turkumini aniqlash, matnni BIOES formatida belgilash, foydalanuvchi interfeysi orqali natijalarni tahrirlash kabi funksional imkoniyatlarga ega. Eksperimental tahlillar 2000 dan ortiq annotatsiyalangan matn (26 000 dan ortiq token) asosida amalga oshirilib, algoritmning aniqlik va to‘liqlik ko‘rsatkichlari baholangan. Tadqiqot natijalari o‘zbek tili korpusini boyitish, milliy til uchun zamonaviy NLP vositalarini yaratish hamda sun’iy intellekt tizimlarida o‘zbek tilini qo‘llash imkoniyatlarini kengaytirishga xizmat qiladi.

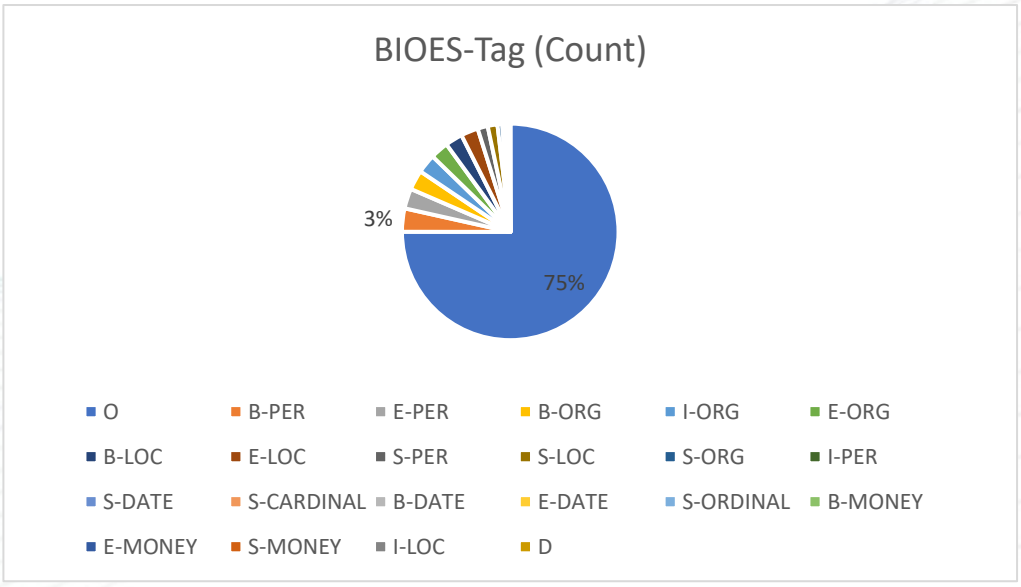
KIRISH. NLPning ajralmas yo‘nalishi hisoblangan nomlovchi birliklarni aniqlash (NER) texnologiyasi tabiiy tildagi muhim elementlarni avtomatik tarzda ajratib olish va qayta ishlash imkonini beradi.

Nomlovchi birliklarni aniqlash (NER – Named Entity Recognition) matnda uchraydigan ma’lum birliklarni aniqlash va ularni oldindan belgilangan sinflarga ajratish jarayoni bo‘lib, turli sohalarda, jumladan, avtomatik tarjima, axborot izlash tizimlari va hujjatlarni tahlil qilishda keng qo‘llaniladi. NERda **nomlovchi birliklar** odatda maxsus ma’lumotlarga ega, o‘zgarishsiz bo‘ladigan birliklardir.

O‘zbek tilidagi ma’lumotlar to‘plami NLP (tabiiy tilni qayta ishlash) sohasida ilg‘or natijalarga erishish va resurslari cheklangan lingvistik muhitlarda ilmiy tadqiqotlar o‘tkazish uchun foydalidir. To‘plam hisoblash tili modellarini yaratish jarayonida muhim va aniq ma’lumotlar taqdim etadi. Ushbu ma’lumotlar to‘plamidan NER (nomlangan birliklarni aniqlash) algoritmlarini ishlab chiqishda va baholashda, shuningdek, o‘zbek tilidagi matnlarning morfologik va sintaktik tahlilini o‘rganishda foydalanishlari mumkin. Agar lug‘atni kengaytirilsa, o‘zbek tilidagi matnlarni chuqurroq o‘rganish uchun mashina o‘rganish texnologiyalaridan ham foydalanish imkoniyati mavjud bo‘ladi. O‘zbek tilidagi ma’lumotlar to‘plamining BIOES -Tag bo‘yicha statistikasi quyidagicha (1-jadval, 2-rasm):

BIOES-Tag	BIOES-Tag (Count)
O	19404
B-PER	900
E-PER	774
B-ORG	762
I-ORG	735
E-ORG	699
B-LOC	669
E-LOC	666
S-PER	392
S-LOC	376
S-ORG	183
I-PER	81
S-DATE	80
S-CARDINAL	43
B-DATE	32
E-DATE	31
S-ORDINAL	12
B-MONEY	8
E-MONEY	8
S-MONEY	5
I-LOC	5
D	1

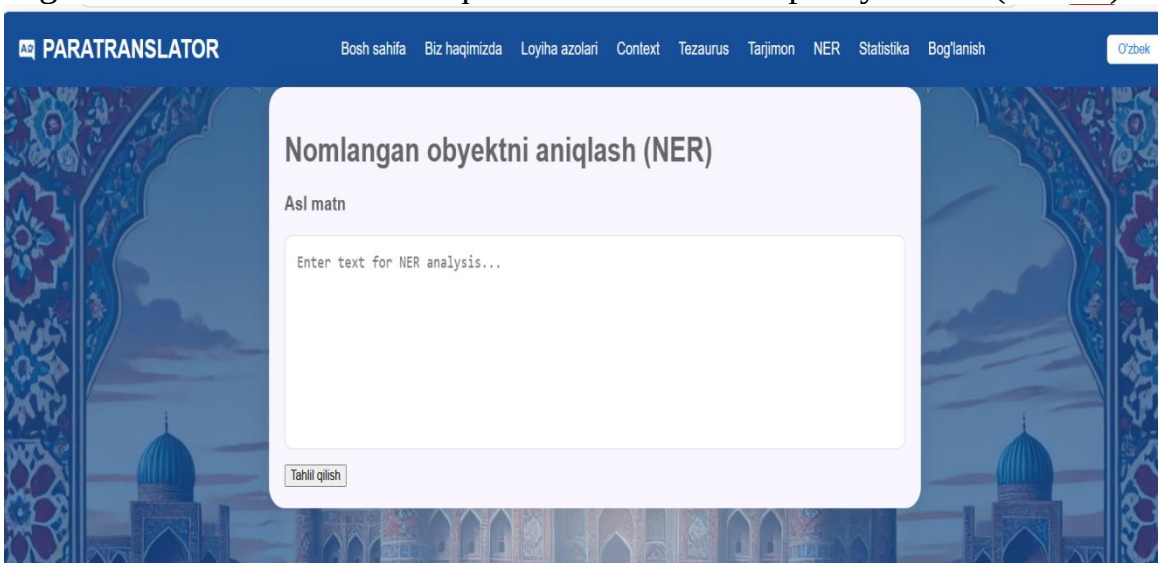
1-jadval. BIOES teglash statistikasi.



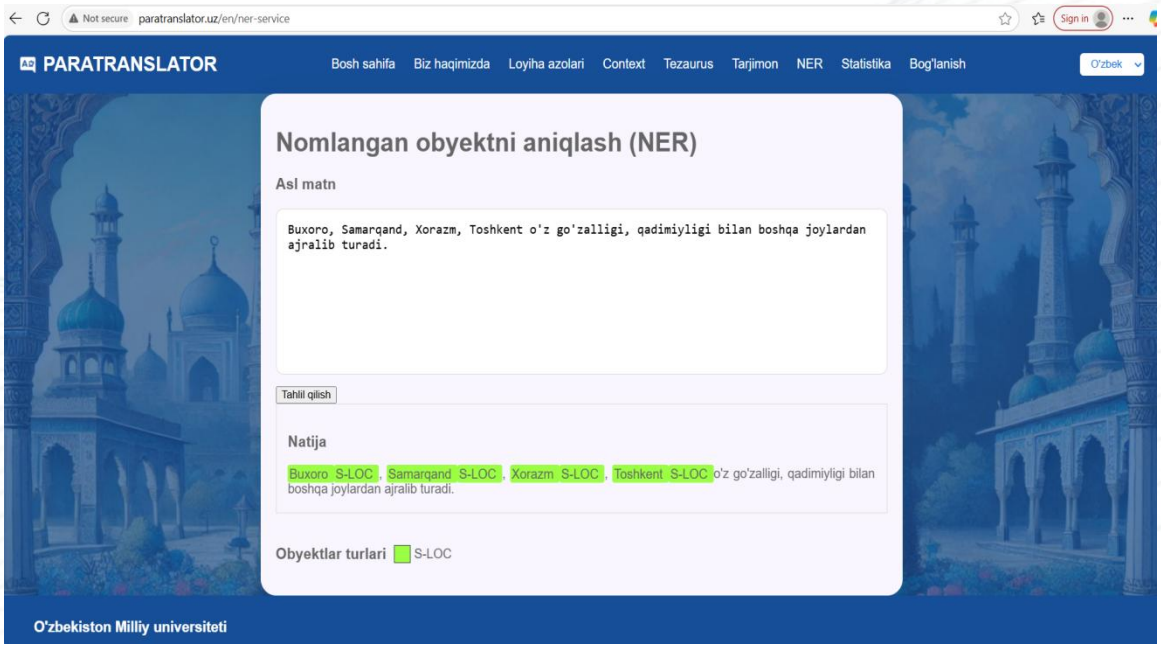
2-rasm. BIOES teglash diagrammasi.

To‘plamda yuridik hujjatlardan olingan ma’lumotlar mavjud bo‘lib, bu jumlalarining grammatik jihatdan to‘g‘ri ekanligini ta’minlaydi. Taklif etilgan ma’lumotlar to‘plami tilshunoslik tadqiqotlari, tabiiy tilni qayta ishlash tizimlarini yaratish va avtomatik tarjima modellarini rivojlantirishda ham muhim manba bo‘lib xizmat qiladi.

Shuningdek, ilk bor https://paratranslator.uz/uz/ner-recognition saytida matndagi nomlovchi birliklarni aniqab beruvchi NER korpusi yaratildi (3-rasm):

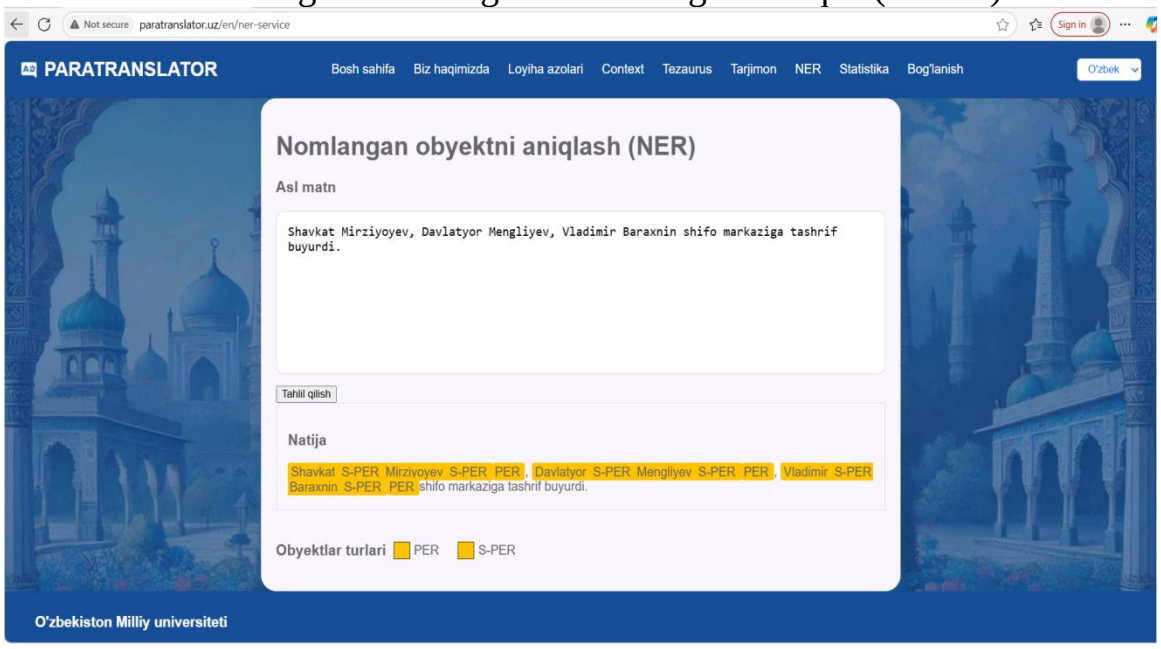


3.-rasm. Paratranslator.uz saytida NER interfeysi.



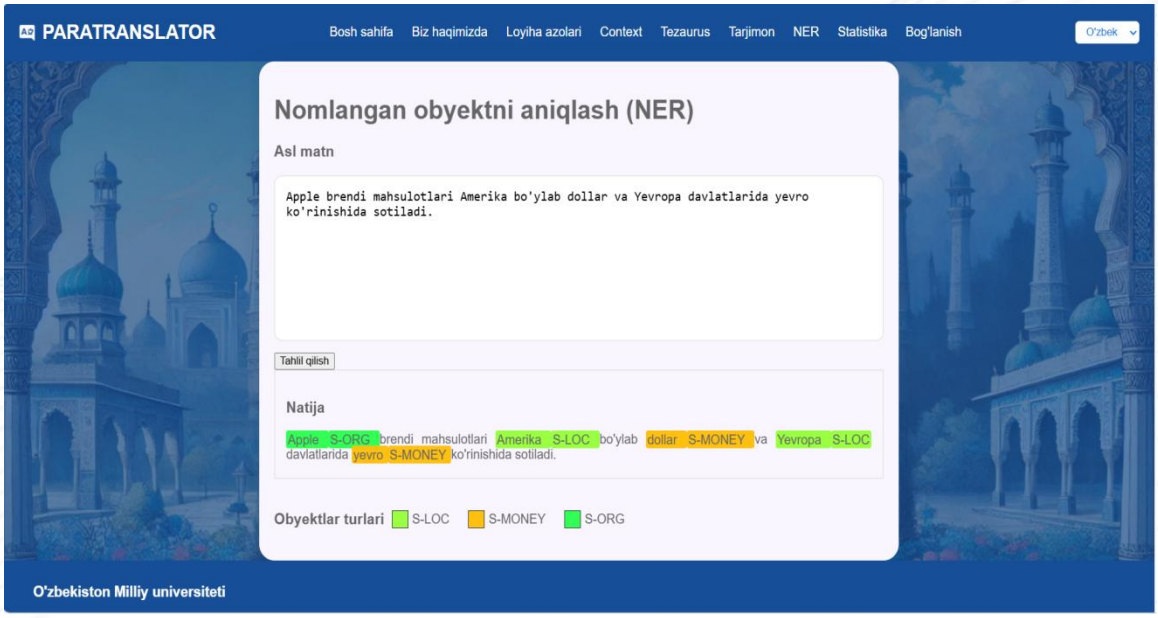
4-rasm. Paratranslator.uz saytida joy nomlarini BIOES formatida aiqlash.

Paratranslator.uz platformasida o‘zbekcha matndan nomlovchi birliklarni avtomatik aniqlash jarayonida Buxoro, Samarqand, Xorazm va Toshkent kabi joy nomlari BIOES formatidagi S-LOC teglari bilan belgilanmoqda (5-rasm).



5-rasm. Paratranslator.uz saytida shaxs nomlarini BIOES formatida aniqlash.

Paratranslator.uz saytida Shavkat Mirziyoyev, Davlatyor Mengliyev va Vladimir Baraxin kabi shaxs nomlarini avtomatik aniqlash jarayonida har bir komponent S-PER tegi bilan BIOES formatida belgilanmoqda (6-rasm).



6-rasm. Paratranslator.uz saytida NERni aniqlash.

Paratranslator.uz platformasida "Apple", "Amerika", "Yevropa", "dollar" va "yevro" kabi tashkilot, joy nomlari va valyuta birliklarini aniqlashda har bir birlik BIOES formatida teglar (S-ORG, S-LOC, S-MONEY) bilan belgilangan.

Tajriba tahlilidan ko'rinadiki, yaratilgan NER dasturiy ta'minoti orqali o'zbek tilidagi matnlarni avtomatik ravishda tahlil qilish natijalari ko'rib chiqildi. Ushbu tahlil jarayonida o'zbek tilida nomlangan birliklarning aniqlanishi va ular tasniflanishi haqida aniq ko'rsatkichlar taqdim etildi.

Tajriba natijalari:

1.NER algoritmining o'zbek tilidagi matnlar bilan ishlashdagi aniqligi va to'g'riligi baholanadi.

2.Dasturiy ta'minot foydalanuvchilar uchun qulay va samarali ekanligi sinovdan o'tkaziladi.

3.O'zbek tilidagi morfologik va sintaktik xususiyatlar, masalan, shevalar, fe'llarning turlicha ishlatilishi va boshqa tilga xos xususiyatlar, NER algoritmining samaradorligiga qanday ta'sir qilishi tahlil qilinadi.

Xulosa. Tajriba natijalari shuni ko'rsatadiki, o'zbek tili korpusida yaratilgan NER instrumentariysi va algoritmi muvaffaqiyatli ishlamoqda va o'zbek tilidagi matnlarni to'g'ri tahlil qilish imkoniyatini yaratadi. Shuningdek, bu algoritmlar kelgusida o'zbek tilining turli shevalari va sintaktik strukturalariga moslashtirilgan holda yanada takomillashtirilishi mumkin.

Ushbu bo'limda o'zbek tili korpusida nomlangan birliklarni aniqlash (NER) algoritmining loyihasi, so'rov yaratish jarayoni va tajriba tahlili natijalari ko'rib chiqildi. NER algoritmlari va dasturiy ta'minotlari, o'zbek tilidagi matnlarni tahlil qilishda samarali natijalar berdi va bu sohada yanada chuqurroq tadqiqotlar va rivojlanishlar uchun imkoniyatlar yaratdi.

ADABIYOTLAR RO'YXATI:

1. Mengliev D., Urinboeva N., Sharipov S., Polatova S., Atakhanov M., Khamraeva S., Boltayev N. Assessing the poetry of a text and its emotional content using a hybrid approach // AIP Conference Proceedings. 2024. – № 3244(1).
2. Abdurakhmonova N., Ismailov A., Mengliev D. Developing NLP Tool for Linguistic Analysis of Turkic Languages // IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). – Yekaterinburg, 2022. – B.1790-1793.
3. Mengliev D., Barakhnin V., Abdurakhmonova N., Eshkulov M. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation // Data in Brief. 2024. – № 54.
4. O'rinboyeva N. Turkiy tillar doirasida nomlovchi birliklarni aniqlashga oid tadqiqotlar tahlili // Academic research in edicational science. 2025. – № 6(1).
5. Abduraxmonova N., O'rinboyeva N. Nomlovchi Birliklarni Aniqlovchi (NER) tizimlar tavsifi // So'z san'ati xalqaro jurnali. 2023. – № 6 (6). – B. 27-32.