UDC 004.93

ATOQLI OTLARNI ANIQLASHNING ANNOTATSIYA QOIDALARI VA MATEMATIK MODELLARI

Allaberdiyev Bobur Baxtiyorovich
Mirzo Ulug'bek nomidagi O'zbekiston milliy universiteti, Toshkent, O'zbekiston
O'zbekiston xalqaro islomshunoslik akademiyasi
allaberdiyev_91@mail.ru
Matlatipov San'atbek G'ayratovich
Mirzo Ulug'bek nomidagi O'zbekiston milliy universiteti, Toshkent, O'zbekiston
s.matlatipov@nuu.uz
Mavlonova Mujgonabonu Mirkomilovna
Mirzo Ulug'bek nomidagi O'zbekiston milliy universiteti, Toshkent, O'zbekiston
mujgonmavloova@gmail.com

ANNOTATSIYA:

Ushbu maqolada matnlardagi atoqli otlarni (Named Entity) aniqlash uchun annotatsiya qoidalari, BIO markalash tizimi, matematik modellar (CRF, BiLSTM-CRF, Transformer), agglutinativ tillarga xos xususiyatlar, hamda real O'zbek matnlari misolida amaliy misollar bayon qilinadi. Model qurilishining formal ifodasi, ehtimollik asosidagi yondashuv, annotatorlar o'rtasidagi kelishuv (Cohen's Kappa) va annotatsiya sifatini oshirish bo'yicha usullar ham yoritiladi. Maqola natural tilni qayta ishlash (NLP) yo'nalishida NER tizimi yaratish istagidagi tadqiqotchilar uchun metodik qo'llanma sifatida xizmat qiladi.

Kalit so'zlar: token, indekslash, agglutinativ, annotatsiya, obyekt.

Kirish

Atoqli otlarni aniqlash (Named Entity Recognition, NER) zamonaviy NLP tizimlarining asosiy komponentlaridan biridir. U matndan shaxslar (PER), joylar (LOC), tashkilotlar (ORG), mahsulotlar yoki maxsus obyektlar (MISC) kabi birliklarni aniqlashni maqsad qiladi [1].

NER tizimlari quyidagi sohalarda keng qo'llanadi:

- qidiruv tizimlarida kontekstga mos natija chiqarish,
- chat-botlarda mazmunni tushunish,
- avtomatik tarjima tizimlarida atoqli otlarni to'g'ri tarjimasiz qoldirish,
- hujjatlarni indekslash,
- axborotlarni avtomatik tahlil qilish.

O'zbek tilida NER bo'yicha tadqiqotlar nisbatan kam, tilning agglutinativ tuzilishi va ko'p shaklli qo'shimchalari modelni murakkablashtiradi. Shu sababli, aniq annotatsiya qoidalari va matematik modellardan foydalanish talab etiladi [3].

BIO markalash tizimi – bu NER (Named Entity Recognition) jarayonida matndagi har bir tokenning (so'zning) qaysi qismga tegishli ekanini belgilash uchun ishlatiladigan eng mashhur va standart belgilash formati.

Annotatsiya qoidalari va BIO markalash tizimi

BIO markalashning formal ta'rifi

Label	Mazmuni	Misol
PER	Shaxs ismlari	Islom Karimov, Alisher Navoiy
LOC	Joy nomlari	Toshkent, Yevropa, Sirdaryo
ORG	Tashkilot, muassasa	Oliy Majlis, Samsung, To‘palang HES AJ
MISC	Mahsulot, asar, tadbir	iPhone 15, Qur’oni Karim, Expo-2025

Jadval 3: Entity toifalari

Matn tokenlar to‘plami sifatida belgilansin:

$$T = (t_1, t_2, \ldots, t_3).$$

Har bir token uchun annotatsiya funksiyasi:

$$L(t_i) \in \{B\text{-}X,\ I\text{-}X,\ O\},$$

Bu yerda: X - PER, LOC, ORG yoki MISC .

Misol 1. Oddiy jumla annotatsiyasi

“Alisher Navoiy Samarqandga safar qildi.”

Token	Label
Alisher	B-PER
Navoiy	I-PER
Samarqandga	B-LOC
safar	O
qildi	O

Jadval 4: Entity toifalari

Agglutinativ tillarda qo‘shimchali shakllarni belgilash

O‘zbek tilida atoqli otlarga qo‘shimchalar qo‘shilishi odatiy hol:

“Toshkentdan”, “Turkiyaning”, “Germaniyaga”, “Navoiyda”.

Annotatsiya qoidasiga ko‘ra, qo‘shimchalar mavjud bo‘lsa ham, entitining ildiz qismi label belgilanadi.

Formal ta’rif:

Agar token ildiz va qo‘shimchadan tashkil topgan bo‘lsa:

$$t_i = r + s,$$

bu yerda r – ildiz (entity), s – qo‘shimcha.

U holda:

$$L(t_i) = L(r).$$

Misol 2. Qo‘shimchali entitilar

“O‘zbekistonning iqtisodiyoti barqaror o‘smoqda.”

NERning matematik modellari

Token	Label
O‘zbekistonning	B-LOC
iqtisodiyoti	O
barqaror	O
o‘smoqda	O

Jadval 5: Entity toifalari

NER – ketma-ketlik belgilash masalasi bo‘lib, modelning vazifasi:

$$\hat{Y} = \arg\max_{Y} P(Y \mid X)$$

Bu yerda:

- $X = (x_1, \ldots, x_n)$ – tokenlar ketma-ketligi,
- $Y = (y_1, \ldots, y_n)$ – BIO label ketma-ketligi.

CRF (Conditional Random Fields)

NER uchun keng qo‘llaniladigan ehtimollik modeli.

CRF (Conditional Random Fields) – bu ketma-ketlikdagi ma’lumotlarga (masalan, matn so‘zlariga) teglar berish uchun ishlatiladigan model bo‘lib, u har bir so‘zni alohida emas, balki butun jumla kontekstida baholaydi. Ya’ni CRF nafaqat so‘zning o‘ziga, balki undan oldin va keyin kelgan so‘zlarga, ularning teglariga va ular orasidagi mantiqiy bog‘lanishga qarab qaror chiqaradi. Shu sababli u “Alisher Navoiy” deganda ikkala so‘zni birga PERSON deb belgilash yoki “Apple kompaniyasi” da Apple ORGANIZATION bo‘lishi kerakligini kontekst orqali aniqlay oladi. Oddiy tasniflovchilardan farqli ravishda, CRF eng mos, uyg‘un va mantiqiy teglar ketma-ketligini tanlaydi, shuning uchun NER, POS tagging, chunking kabi vazifalarda juda aniq ishlaydi.

CRFning taqsimot funksiyasi odatda quyidagicha ifodalanadi:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\Big(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, X, t)\Big),$$

bu yerda:

- f_k – xususiyat funksiyalari (so‘z shakli, suffiks, POS tag va h.k.),
- λ_k – ularning vaznlari,
- $Z(X)$ – normallashuvchi koeffitsient.

CRF ketma-ket tokenlarga to‘g‘ri label berishda juda samarali.

Misol 3. CRF uchun xususiyatlar

Token	Feature
Toshkentga	suffix=ga
O‘zbekistonning	suffix=ning, capital=True
Navoiy	capital=True

Jadval 6: Entity toifalari

BiLSTM + CRF modeli

BiLSTM+CRF modeli – bu ketma-ketlikni belgilash vazifalarida, ayniqsa NER kabi kontekstga juda bog‘liq masalalarda eng samarali ishlaydigan chuqur o‘rganish arxitekturasi bo‘lib, unda BiLSTM qatlamlar matnning har ikki yo‘nalishdan – chapdan o‘ngga va o‘ngdan chapga qarab o‘qilgan kontekstini chuqur tahlil qiladi va har bir so‘z uchun boy semantik ma’lumotga ega yashirin vektor yaratadi, CRF esa shu vektorlardan chiqqan ehtimolliklar asosida faqat har bir so‘zning individual belgisinigina tanlab qo‘ymay, balki butun jumla bo‘yicha eng mantiqiy, ketma-ketlik jihatidan to‘g‘ri keladigan label zanjirini tanlaydi. Bunda BiLSTM matndagi uzun masofali bog‘lanishlarni – masalan, shaxsning familiyasi oldin kelishi yoki tashkilot nomi bir nechta so‘zdan tarkib topishi – kabi murakkab til hodisalarini o‘rganadi, CRF esa label-lar orasidagi mantiqiy cheklovlarni – masalan, B-PER ortidan I-PER kelishi yoki LOCATION ketidan ORGANIZATION kelmasligi kabi qoidalarga rioya qilgan holda yakuniy qarorni optimallashtiradi. Shu ikki mexanizm birgalikda ishlagani sababli model matnning semantikasini chuqur tushunadi, lekin natijada belgilangan teglar ketma-ketligi mantiqan to‘g‘ri, silliq va izchil bo‘ladi; shuning uchun BiLSTM+CRF arxitekturasi ko‘plab tillar uchun an’anaviy NER tizimlarida eng barqaror va yuqori aniqlikka ega yondashuvlardan biri hisoblanadi.

LSTM chiqishi:

$$h_t = [\overrightarrow{h_t}\ ;\ \overleftarrow{h_t}],$$

emissiya ballari:

$$s_t = W h_t + b,$$

CRF layer esa bu ballar asosida yakuniy label ketma-ketligini tanlaydi:

$$\hat{Y} = \arg\max_{Y}\ \mathrm{CRF}(s_1, \ldots, s_n).$$

Modelning ustunligi:

- uzoq kontekstni o‘rganadi,
- agglutinativ tillarda yuqori natija beradi.

Transformer modellari (BERT, XLM-R)

Transformer arxitekturasiga asoslangan modellar – BERT, XLM-R kabi tizimlar – matnni chuqur va ikki yo‘nalishda (bidirectional) kontekst asosida tushunish qobiliyati bilan NLPda inqilob qilgan modellar hisoblanadi; ular matnni ketma-ketlik sifatida emas, balki to‘liq kontekstli blok sifatida qayta ishlaydi va self-attention mexanizmi yordamida har bir so‘zning mazmunini jumladagi barcha boshqa so‘zlar bilan o‘zaro bog‘lab baholaydi, bu esa murakkab semantik va sintaktik munosabatlarni juda aniq o‘rganishga imkon beradi. BERT ingliz tili uchun juda katta hajmdagi matnlarda oldindan o‘qitilgan bo‘lib, Masked Language Modeling orqali so‘zlarning yashirilgan qismlarini topishni o‘rganadi, natijada model kontekstni chuqur anglay oladi; XLM-R esa ko‘p tilli (cross-lingual) yondashuvga ega bo‘lib, yuzlab tillardan iborat ulkan korpuslarda o‘qitilgan va turli tillar o‘rtasida bilimlarni uzatish qobiliyati bilan mashhur – ya’ni bitta tilda o‘rgatilgan vazifa boshqasida ham samarali ishlashi mumkin. Transformer modellaridan foydalanishning eng katta afzalligi – ular uzoq masofali bog‘lanishlarni BiLSTMga qaraganda ancha kuchli ushlab turadi, murakkab kontekstlarni aniq ajratadi va oldindan o‘qitilgan yirik til modellari sifatida turli NLP vazifalarida, jumladan NER, sentiment tahlili, mashina tarjimasi, matn tasnifi va savol-javob tizimlarida minimal moslashtirish bilan juda yuqori natijalar beradi.

Self-Attention mexanizmi asosida ishlaydi.

Attention formulasi:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^t}{\sqrt{d_k}}\Big)V.$$

Misol 4. BERT yordamida NER

“Bobur Baxtiyorovich Toshkent shahrida nutq so‘zladi.”

Natija:

Bobur Baxtiyorovich → PER

Toshkent → LOC

Annotatsiya jarayoni: formal yondashuv

Annotatsiya jarayoni formal yondashuvda matn korpusiga izohlar (teglar)ni oldindan belgilangan qoidalar asosida qo‘llashni anglatadi va u qat’iy struktura, aniq yo‘riqnomalar hamda annotatorlar o‘rtasidagi izchillikni ta’minlashga qaratilgan metodik bosqichlardan iborat bo‘ladi. Ushbu jarayon avvalo annotatsiya sxemasini ishlab chiqishdan boshlanadi, bunda obyektlar toifalari (masalan, PERSON, LOCATION, ORGANIZATION), ularning chegaralari va ularni belgilash qoidalari normativ ravishda tavsiflanadi; sxema barcha holatlar uchun birxillashtirilgan misollar va kontrmisollar bilan boyitiladi. Keyingi bosqichda annotatsiya instruktsiyasi tayyorlanadi – bu annotatorlar amal qilishi kerak bo‘lgan formal hujjat bo‘lib, unda teglar hierarxiyasi, cheklovlar, ko‘p ma’noli holatlarni hal qilish qoidalari, nominativ birikmalarni belgilash tartibi, kontekstga bog‘liq istisnolar va murakkab lingvistik vaziyatlar bo‘yicha izohlar qat’iy qayd etiladi. Annotatorlar ish jarayoniga kirishdan oldin mazkur yo‘riqnomaga asoslangan trening va kalibrlash bosqichidan o‘tadi; bunda ularning teg qo‘llashdagi izchilligi tekshiriladi va inter-annotator agreement (IAA) ko‘rsatkichlari – masalan, Cohen’s Kappa yoki F1 bo‘yicha mosligi baholanadi. Rasmiy annotatsiya bosqichida annotatorlar matnga qat’iy ko‘rsatmalar asosida teglar qo‘llaydi, jarayon davomida esa barcha izohlar qayta tekshiriladi va muvofiqlashtiriladi [5]. Yakunda verifikatsiya va validatsiya bosqichlari amalga oshirilib, annotatsiyalarning formal talablar bilan mosligi baholanadi, aniqlangan nomuvofiqliklar tahrir qilinadi va korpusning yakuniy versiyasi sifat jihatidan tasdiqlanadi. Shu tarzda formal yondashuv annotatsiya jarayonini me’yorlashtirilgan, nazorat qilinadigan va takror ishlab chiqarilishi mumkin bo‘lgan holga keltiradi.

Annotatsiya – matnni to‘g‘ri belgilangan label ketma-ketligiga xaritalash:

$$A(X) = Y.$$

Bir nechta annotator bo‘lsa:

$$A_i(X) = Y_i.$$

Annotatorlar o‘rtasidagi moslik:

Cohen’s Kappa statistic ko‘rsatkichidan foydalanamiz.

Cohen’s Kappa – bu ikki annotator (yoki ikki mutaxassis) bir xil ma’lumotni qanday darajada bir xil belgilaganini baholaydigan statistik ko‘rsatkich [2].

U ayniqsa NER, sentiment analysis, tasniflash, tibbiy diagnostika kabi sohalarda annotatorlar orasidagi kelishuv darajasini (inter-annotator agreement) o‘lchash uchun ishlatiladi.

Oddiy aniqlikdan farqli ravishda, Cohen’s Kappa tasodifiy mos kelishni ham hisobga oladi, ya’ni annotatorlar tasodifiy ravishda bir xil belgi qo‘ygan bo‘lishi mumkinligini tuzatadi.

Kappa quyidagicha aniqlanadi:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

bu yerda:

- P_o – annotatorlar o‘rtasidagi real kelishuv (observed agreement),
- P_e – tasodifiy kelishuv ehtimoli (expected agreement).

Bu qiymat 0.8 dan yuqori bo‘lsa – annotatsiya yuqori sifatga ega [4].

Misol:

Quyida haqiqiy o‘zbek matnidan olingan namunalar annotatsiya bilan beriladi.

Misol 5. Rasmiy matn

“Oliy Majlis bugun Toshkent shahrida yangi qonunni muhokama qildi.”

Token	Label
Oliy	B-ORG
Majlis	I-ORG
bugun	O
Toshkent	B-LOC
shahrida	O
yangi	O
qonunni	O
muhokama	O
qildi	O

Jadval 7: Entity toifalari

Misol 6. Tarixiy matn

“Amir Temur 1395-yilda Moskva tomon yurish boshlagan.”

Token	Label
Amir	B-PER
Temur	I-PER
1395-yilda	O
Moskva	B-LOC
tomon	O
yurish	O
boshlagan	O

Jadval 8: Entity toifalari

Misol 7. Texnologik yangilik

“Samsung kompaniyasi Galaxy S25 Ultra modelini taqdim etdi.”

Token	Label
Samsung	B-ORG
kompaniyasi	O
Galaxy	B-MISC
S25	I-MISC
Ultra	I-MISC
modelini	O
taqdim	O
etdi	O

Jadval 9: Entity toifalari

Xatoliklar tahlili

NER tizimlarida keng uchraydigan xatoliklar mavjud bo‘lib, ular quyidagilar:

Bunday xatoliklarni bartaraf etish uchun xatoliklarni statistik jihatdan baholash (F1-score tahlili) amalga oshiriladi.

Xato turi	Misol	Sabab
Boundary error	“Oliy Majlis binosi”da “Oliy Majlis binosi”ni ORG deb olish	Entity chegarasi noto‘g‘ri
Wrong label	“Yevropa Ittifoqi” → MISC deb belgilash	Murakkab tashkilot nomi
Missing entity	“Samarqandlik” tokenida rootni ajratmaslik	Agglutinativ tuzilma

Jadval 10: Entity toifalari

Har bir toifa bo‘yicha Precision, Recall va F1-score quyidagicha hisoblanadi:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2PR}{P + R}.$$

Agar:

- Precision past bo‘lsa → model juda ko‘p FP qiladi,
- Recall past bo‘lsa → model juda ko‘p FN qiladi.

Bu tahlil modelni qaysi yo‘nalishda yaxshilash kerakligini ko‘rsatadi.

XULOSA

Atoqli otlarni aniqlash (Named Entity Recognition – NER) tabiiy tilni qayta ishlash (NLP) sohasining eng muhim komponentlaridan biri bo‘lib, u matnlardan informatsion obyektlarni – shaxs, joy, tashkilot, sana, mahsulot, ilmiy atama va boshqa maxsus kategoriyalarni avtomatik ajratishga xizmat qiladi. Ushbu maqolada NER tizimi uchun zarur bo‘lgan barcha asosiy jarayonlar – annotatsiya qoidalarini shakllantirishdan tortib, matematik modellarni qo‘llash va xatoliklarni tahlil qilishgacha bo‘lgan bosqichlar batafsil yoritildi.

Maqolada NER modellari uchun qo‘llaniladigan zamonaviy yondashuvlar keng ko‘lamda tahlil qilindi. CRF modelining ketma-ketlikdagi bog‘liqlikni hisobga olishi, BiLSTM-CRF arxitekturasining uzun kontekstni chuqur anglash imkoniyati, Transformer asosidagi modellar – BERT va XLM-Rning ikki yo‘nalishli kontekstni oqilona qayta ishlashi kabi jihatlar keng yoritildi. Ayniqsa, self-attention mexanizmi orqali matnning barcha tokenlari o‘zaro bog‘lanishi va semantik ma’noni chuqurroq aks ettirishi NERda yuksak natijalar berishi ilmiy jihatdan asoslab berildi.

Annotatsiya jarayonining formal yondashuv asosida tashkil qilinishi NER tizimining muvaffaqiyatini keskin oshiradi. Annotatorlar uchun ishlab chiqilgan qat’iy qoida, yo‘riqnomalar, murakkab holatlar bo‘yicha talqinlar, inter-annotator agreement (Cohen’s Kappa) orqali moslikni baholash, annotatsiyaning ishonchliligi va qayta tiklanish imkoniyatini ta’minlaydi. Aynan shu jihatlar yuqori sifatli korpuslar yaratishda muhim o‘rin tutadi.

Shuningdek, maqolada model ishlash jarayonida uchraydigan xatoliklar – boundary error, misclassification, missing entity va spurious entity kabi holatlar tizimli ravishda tasniflandi. NER tizimini takomillashtirish uchun xatoliklarni statistik tahlil qilish, ya’ni Precision, Recall, F1-score ko‘rsatkichlarini hisoblash zarurligi qayd etildi. Bu ko‘rsatkichlar modelning qaysi yo‘nalishda kuchsiz ishlashini aniqlashga yordam beradi va yanada puxta NLP tizimi yaratish uchun asos bo‘ladi.

Umuman olganda, ushbu tadqiqot o‘zbek tilida NER tizimini yaratish uchun ilmiy asoslarni, amaliy metodlarni va texnik yondashuvlarni bir butun shaklda bayon qiluvchi kompleks ish bo‘lib, til texnologiyalarini rivojlantirish yo‘lida muhim qadamlardan biridir. Kelajakda ushbu yondashuvlar asosida kengroq korpuslar yaratish, modelni ko‘p tillilik (multilingual) imkoniyatlari bilan kengaytirish, shuningdek, o‘zbek tiliga xos lingvistik xususiyatlarni chuqurroq modellarga integratsiya qilish NER tizimlarining yanada yuqori aniqlikka ega bo‘lishiga xizmat qiladi.

ADABIYOTLAR RO‘YXATI:

1. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. NAACL-HLT. (BiLSTM-CRF asosidagi mashhur NER modeli)

2. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991. (NER va boshqa tegishli masalalar uchun LSTM-CRFning klassik varianti)

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. (Transformer asosida kontekstli model – zamonaviy NERning asosi)

4. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. Prentice Hall. (NLP bo‘yicha eng mashhur darslik, NER bo‘limi mavjud)

5. Rajabov J.Sh., Formalizing the Uzbek Language: A Comprehensive Exploration Using Backus-Naur Forms, Acta NUUz, vol. 1(1), 2023 (01.00.00. №- 8).

АННОТАЦИЯ

В статье описываются правила аннотирования для выделения именованных сущностей в текстах, система разметки BIO, математические модели (CRF, BiLSTM-CRF, Transformer), признаки, специфичные для агглютинативных языков, и практические примеры на основе реальных узбекских текстов. Также рассматриваются формальное выражение построения модели, вероятностный подход, соглашение между аннотаторами (каппа Коэна) и методы повышения качества аннотирования. Статья служит методическим руководством для исследователей, стремящихся создать систему NER в области обработки естественного языка (NLP).

Ключевые слова: токен, индексация, агглютинативный, аннотация, объект.

ABSTRACT

This article describes annotation rules for identifying Named Entities in texts, the BIO tagging system, mathematical models (CRF, BiLSTM-CRF, Transformer), features specific to agglutinative languages, and practical examples using real Uzbek texts. The formal expression of model construction, a probability-based approach, agreement between annotators (Cohen's Kappa), and methods for improving annotation quality are also covered. The article serves as a methodological guide for researchers who want to create a NER system in the field of natural language processing (NLP).

Keywords: token, indexing, agglutinative, annotation, object.