O‘ZBEK TILI KORPUSIDA NER INSTRUMENTARIYSI VA ALGORITMI

Nazokat O‘rinboyeva

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

Kalit so‘zlar.
Korpus, token, nomlovchi birlik, qoidaga asoslangan yondashuv, NER algoritmi, API, BRAT Rapid Annotation Tool.

Annotatsiya. Mazkur 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. Eksperimental tahlillar 26 000 dan ortiq token asosida amalga oshirilib, algoritmning aniqlik va to‘liqlik ko‘rsatkichlari baholangan.

Kirish. Instrumentariy loyihasi — bu NER tizimini ishlab chiqish uchun talab qilinadigan dasturiy vositalar, modullar va texnologiyalar majmuasidir. Ushbu loyiha NER tizimini ishlab chiqish, sinovdan o‘tkazish va amaliyotga joriy qilish uchun zarur bo‘lgan barcha texnik jihatlarni qamrab oladi.

Instrumentariy algoritmi[1] NER tizimining ishlash jarayonini tartibga soluvchi qoidalar va bosqichlar majmuasi bo‘lib, u quyidagi asosiy bosqichlarni o‘z ichiga oladi:

1. Dastlab matn ma’lumotlari to‘planadi va oldindan ishlov berilgach, ma’lumotlardagi keraksiz belgilar olib tashlanadi, so‘zlar standart shaklga keltiriladi (lemmanizatsiya yoki stemming qo‘llaniladi). Annotatsiyalangan (belgilangan) korpus yaratiladi.

2. So‘zlarning kontekstual ma’nolari aniqlangach, lingvistik va statistik xususiyatlar (masalan, qaysi harf bilan boshlanishi, so‘z uzunligi, qaysi qismlarga bo‘linishi) aniqlanadi hamda kontekstual so‘z vektorlari (BERT, Word2Vec, FastText) kabi yondashuvlar qo‘llaniladi.

3. Qoidaga asoslangan yondashuv asosida oldindan belgilangan qoidalar asosida entitilarni aniqlash, mashina ta’limining CRF, SVM, HMM kabi usullar

[1] https://store-restack.vercel.app/p/entity-recognition-answer-types-of-ner-algorithms-cat-ai

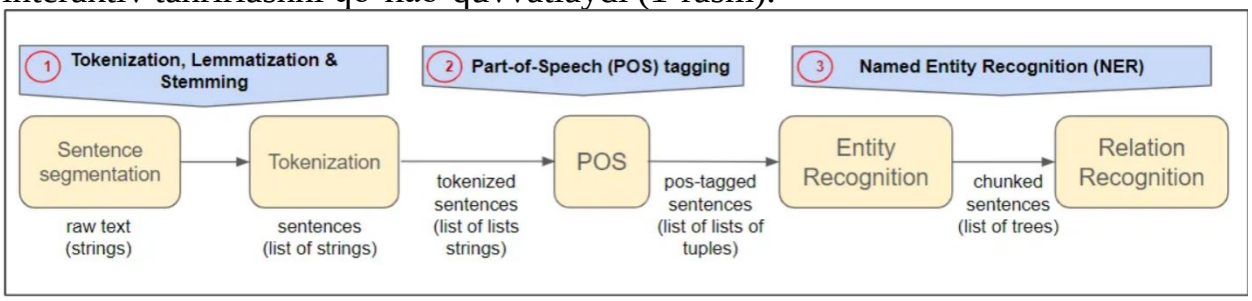
yordamida modellarni o‘qitish, Chuqur o‘rganish (Deep Learning) yondashuvi usulida RNN, LSTM, Transformer, BERT kabi modellaridan foydalaniladi.

4. Sinov to‘plami yordamida NER tizimi aniqligi, takrorlanuvchanligi va samaradorligi baholanib, F1 mezon, aniqlik (Precision) va qamrov (Recall) ko‘rsatkichlari aniqlanadi hamda noto‘g‘ri aniqlangan entitilar tahlil qilinib, model takomillashtiriladi.

5. Tayyor bo‘lgan NER tizimi real ma’lumotlarga qo‘llanilib, API yoki veb-xizmat sifatida integratsiya qilinadi. Ma’lumotlar bazasi yoki hujjatlar tizimida foydalanish uchun moslashtiriladi.

Shunday qilib, instrumentariy loyihasi va algoritmi NER tizimini ishlab chiqish va amaliyotga tatbiq qilishning asosiy tamoyillarini o‘z ichiga oladi. BRAT Rapid Annotation Tool[2] matnlarni belgilash va entitilarni aniqlash uchun qulay interfeysga ega bo‘lib, ko‘plab tillarni qo‘llab-quvvatlaydi.

INCEpTION platformasi[3] NER uchun ilg‘or annotatsiya va mashina ta’limi imkoniyatlarini taqdim etadi, shuningdek, foydalanuvchi interfeys orqali interaktiv tahrirlashni qo‘llab-quvvatlaydi (1-rasm).



1-rasm. NER tizimining asosiy bosqichlari.

Rasmda[4] oddiy axborot tizimining asosiy arxitekturasi ko‘rsatilgan. Bu matnni qayta ishlashdan boshlanadi, dastlab jumlalar, so‘ngra har bir jumla tokenizer yordamida so‘zlarga bo‘linishi jarayonning birinchi bosqichi bo‘ladi. Ikkinchi bosqich nutqning bir qismini belgilashni amalga oshirishi sabab nomlangan birlikni tanib oladi.

Natijalar. O‘zbek tilida NER tizimini yaratish uchun katta hajmdagi belgilangan (annotatsiyalangan) matnlar to‘planadi. Ushbu ish doirasida 2000 ta jumla va 25 865 so‘zdan iborat ma’lumotlar to‘plami yaratildi. Manbalar BIOES sxemasidan foydalangan holda huquqiy hujjatlar va qo‘lda tayyorlangan jumlalardan tashkil topgan. Bu esa modelning yuqori aniqlikka erishishiga xizmat qildi.

NER modeli test majmuasida yuqori natijalarga erishdi [5] (2-rasm):

[2] https://brat.nlplab.org/

[3] https://github.com/inception-project/inception-external-recommender?ysclid=m83btlkp7842896134

[4] Gutierrez F. Information extraction algorithms using Named Entity Recognition Felix Gutierrez // Level Up Coding. 2023.

[5] Mengliev D., Abdurakhmonova N., Barakhnin V., Eshkulov M.. Developing named entity recognition algorithms for Uzbek: Dataset Insights and Implementation // Data in Brief.. 2024.–№ 54. – B. 4.

E	Quantity of samples which were proceed	F1-Score	Precision	Recall
0	0	89.76	86.73	93.02
1	200	91.95	88.85	95.28
2	400	92.75	89.62	96.11
4	600	92.70	89.57	96.06
....	...	..	...	...
122	3400	92.68	89.55	96.03
135	3600	92.75	89.62	96.11

2-rasm. NER modeli natijasi.

Modelning samaradorligini aniqlash uchun Precision, Recall va F1-score ko‘rsatkichlari hisoblanadi. Buning uchun maxsus sinov korpuslari yaratiladi va model natijalari baholanadi (3-rasm):

No.	Quantity of correctly recognized entities / Quantity of recognized entities	Precision	Quantity of recognized entities / Quantity of all entities in the document	Recall
Dictionary-based algorithm results				
1	182 / 182	100%	182 / 201	91%
2	74 / 104	71%	104 / 199	52%
SpaCy model's (neural network) results				
1	172 / 193	89.5%	193 / 201	96%
2	162 / 176	87%	186 / 199	93%

3-rasm. NER modeli natijasi.

Modelni takomillashtirish uchun transfer learning, data augmentation kabi texnikalar qo‘llaniladi.

O‘zbek tiliga moslashtirilgan Nomlovchi birliklarni aniqlash (NER) instrumentariysi tabiiy tilni qayta ishlash (NLP) sohasida turli ilovalar uchun muhim ahamiyat kasb etadi. Uni samarali va amaliy jihatdan foydali qilish uchun quyidagi funksional imkoniyatlar mavjud bo‘lishi lozim[6]:

1. Matnni qayta ishlash funksiyalari:

-matnni so‘zlar va iboralarga ajratish;

-so‘zlarning lug‘aviy shaklini, so‘z turkumlarini va grammatik xususiyatlarini aniqlash;

-nomlovchi birliklarni shaxslar (PER), joy nomlari (LOC), tashkilotlar (ORG), sanalar (DATE), raqamli birliklar (NUM) kabi kategoriyalarga ajratish kabi vazifalarni bajarishi lozim.

2. Lingvistik va semantik aniqlash imkoniyatlari:

-ko‘p qismli nomlarni ajratishda uch yoki undan ortiq so‘zdan iborat nomlarni to‘g‘ri belgilash (Muhammad Ali, O‘zbekiston Respublikasi Markaziy banki);

-sinfiy farqlashda bir xil so‘zlarning kontekst bo‘yicha turli ma’nolarini ajratish (Toshkent – shahar yoki viloyat nomi sifatida);

-sinonim va qisqartmalarga moslashish (O‘zbekiston Respublikasi = O‘zbekiston yoki AKT vazirligi = Axborot texnologiyalari va kommunikatsiyalar vazirligi);

-bir xil so‘z turli kontekstlarda turli nomlovchi birliklarga (Asadov – familiya yoki kompaniya nomi) tegishli bo‘lishi mumkin.

[6] Nadeau D., Sekine S. A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes.2007.–№30(1). – B. 2.

3. Mashina ta'limi (ML) va moslashuv imkoniyatlari[7]:

-lug'at va qoidalar asosida tayyor o'zbek tilidagi bazalar yordamida aniqlash;

-Mashina ta'limi modeli bilan ishlashda Transformerlar (BERT, XLM-R, mBERT) yoki CRF (Conditional Random Fields) kabi ilg'or modellar yordamida tahlil qilish;

-annotatsiyalangan ma'lumotlar bilan qayta o'qitishda NER modelini yangi ma'lumotlar asosida yaxshilash imkoniyati;

-o'zbek tilining morfologik xususiyatlariga moslashish;

4. Integratsiya va qo'llash imkoniyatlari[8]:

- API interfeys orqali JSON, XML yoki boshqa formatda ma'lumot olish va jo'natish;

- chatbotlar, avtomatlashtirilgan xizmatlar va qidiruv tizimlariga osongina moslashish;

-veb va mobil ilovalarga Django, Flask yoki boshqa server orqali moslashib ishlash;

- PDF, DOCX yoki TXT formatidagi hujjatlarni tahlil qilish.

5. Ilovalar va qo'llash sohalari:

-qidiruv tizimlarini yaxshilash matn bo'yicha aniq natijalar berishadi;

-chatbot va mijozlarni qo'llab-quvvatlash tizimlari savollarga mos javob topish imkonini beradi;

- axborot xavfsizligi va monitoring qilishda muayyan shaxslar yoki tashkilotlarga oid matnlarni tahlil qilishi mumkin;

-hujjatlarni avtomatik teglash jarayonida katta hajmdagi ma'lumotlarni tartibga solish va sinflarga ajratish kabi imkoniyatlarga ega.

Xulosa. O'zbek tili uchun ishlab chiqilayotgan NER instrumentariysi lingvistik nozikliklarga moslashgan, real vaqtda ishlay oladigan, ko'p sohaga integratsiyalash mumkin bo'lgan va sun'iy intellekt yutuqlaridan foydalangan holda ishlovchi bo'lishi kerak. Ushbu imkoniyatlar NER tizimining samaradorligini oshirishga va uni ilmiy, texnologik hamda amaliy jihatdan yetuk tizimga aylantirishga xizmat qiladi.

ADABIYOTLAR RO'YXATI:

1. 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.

[7] Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL. 2019. –№ 1. – B. 4180.

[8] Lample G., Ballesteros M., Subramanian S. Neural Architectures for Named Entity Recognition.Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. –№ 1. – B. 263.

2. Mengliev D., Barakhnin V., Abdurakhmonova N., Eshkulov M. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation // Data in Brief. 2024. – № 54.

3. 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).

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.

6. Gutierrez F. Information extraction algorithms using Named Entity Recognition Felix Gutierrez // Level Up Coding. 2023.

7. Nadeau D., Sekine S. A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes.2007.–№30(1). – B. 2.