JAHON TAJRIBALARIDA NERNI ANIQLASH METODLARI

Samatboyeva Madina To‘lqinjon qizi,
Tayanch doktorant (PhD)
msamatboyeva@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasi, uning nazariy asoslari hamda amaliy yo‘nalishlari keng yoritiladi. NLP informatika va sun’iy intellektning muhim bo‘limi sifatida inson tilini kompyuterlar tomonidan tushunish va qayta ishlash imkonini yaratadi. Tadqiqotda NLP ning asosiy vazifalari sifatida axborotni indekslash, axborotni qidirish, matnlarni tasniflash, avtomatik tarjima, savol-javob tizimlari va matn generatsiyasi kabi yo‘nalishlar tahlil qilinadi. Ayniqsa, axborot ekstraksiyasi (Information Extraction) va Named Entity Recognition (NER) texnologiyasining ahamiyati alohida ko‘rib chiqiladi. NER tizimlarining rivojlanish bosqichlari, qoidaga asoslangan va leksik yondashuvlari, shuningdek, CoNLL-2003 va WNUT-2017 kabi keng qo‘llaniladigan datasetlar tavsiflanadi. Tadqiqotda NER ning tuzilmagan matnlardan strukturaviy axborotni ajratib olishdagi roli ilmiy manbalar asosida tahlil qilinadi. Shuningdek, zamonaviy NLP tizimlarining transformatsion arxitekturasi va ularning bosqichma-bosqich ishlash prinsiplari ham yoritilgan. Maqola natijalari NLP va NER texnologiyalarining zamonaviy axborot tizimlaridagi muhim o‘rnini ko‘rsatadi.

Kalit so‘zlar: *NLP, NER, axborot, metodlar, Information Retrieval, data-ma’lumot, matn tahlili.*

Abstract. This article presents a comprehensive overview of Natural Language Processing (NLP), its theoretical foundations, and its practical applications. NLP is an important subfield of computer science and artificial intelligence that enables machines to understand, interpret, and process human language. The study highlights key tasks of NLP, including information indexing,

information retrieval, text classification, automatic translation, question-answering systems, and text generation. Special attention is given to Information Extraction and Named Entity Recognition (NER) as core components of modern NLP systems. The development stages of NER are analyzed, including rule-based approaches and lexical matching techniques. Furthermore, widely used datasets such as CoNLL-2003 and WNUT-2017 are described in terms of their structure and significance for training NER models. The paper also discusses the transformational architecture of information extraction systems, where each processing stage adds structure to unstructured text. The findings emphasize that NER plays a crucial role in extracting structured information from unstructured data and serves as an essential component of modern NLP pipelines. Overall, the study demonstrates the importance of NLP and NER technologies in advancing intelligent information systems and improving automated text analysis in various domains.

Keywords: *NLP, NER, information, methods, Information Retrieval, data, text analysis.*

NLP – (Natural Language Processing – Tabiiy tilni qayta ishlash) informatika sohasining bir bo‘lagi bo‘lib, u sun’iy intellekt va mashinaviy o‘rganish usullaridan foydalanib, kompyuterlarga inson tilini tushunish va undan muloqot qilish uchun foydalanish imkonini beradi[3:56].

Chowdhary (2020) esa tabiiy tilni mashinaviy qayta ishlashni bir nechta yo‘nalishlarga ajratadi. Shuningdek, U axborot degan so‘z doirasida, bir necha yonma-yon terminlarni keltiradi[1:602-603]. Bu bilan NLP tarkibini ham ko‘rsatadi:

1. Katta hajmdagi axborotlarni indekslash – *Information Indexing;*
2. Axborotni qayta topish – *Information Retrieval;*
3. Axborotni turli kategoriyalarga tasniflash, axborot ajratib olish – *Information Extraction;*
4. Avtomatik tarjima – *Automatic Translation;*

5. Qisqa savol-javob tizimlari ishlab chiqish – *Question-answering System Development;*

6. Bilim bazasini yaratish hamda matn yoki dialoglarni generatsiya qilish – *Text Generation Systems.*

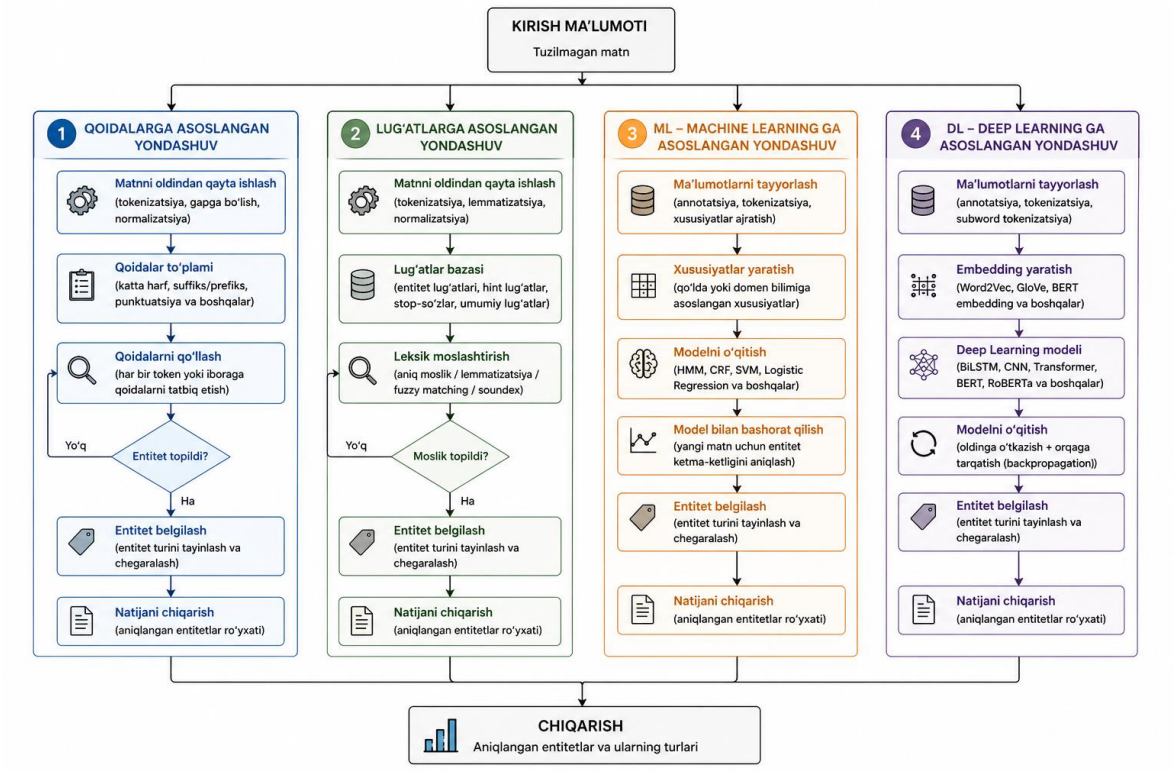
Axborot ekstraksiyasi, ushbu ish doirasida asosiy o‘rinni egallaydi va u tizimning tuzilmagan matndan strukturaviy ma’lumotlarni ajratib olish qobiliyati sifatida tushuniladi [5:94]. Shuningdek, axborotni tasniflash qayta ishlangan data-ma’lumotni kategoriyalarga ajratish ishini amalga oshiradi. Bu jarayon bir nechta transformatsion tizimlardan o‘tishi kerak. Bunda har bir bosqich ma’lumotlar ko‘lamini kengaytirib boradi va har bir bosqichda yangi obyektlarni ajratib boradi. Ma’lumotlarni ajratish jarayoning muhim bosqichlaridan biri bu – NERdir.

So‘nggi yillarda tuzilmagan matnli ma’lumotlardan strukturaviy axborotni avtomatik tarzda ajratib olishga qaratilgan yondashuvlar sezilarli darajada rivojlandi[7:7]. Mazkur yo‘nalishda muhim metodlardan biri sifatida nomlangan obyektlarni aniqlash (Named Entity Recognition – NER) texnologiyasi alohida o‘rin egallaydi. Ushbu tadqiqot doirasida ham NER yondashuvi matnlardan texnologik terminlarni aniqlash va ajratib olish maqsadida qo‘llaniladi. Ko‘pgina olimlar fikricha mazkur standart ma’lumotlar o‘z domen xususiyatiga ega bo‘ladi va bu jihat bilan bir-biridan keskin farq qiladi. Eng keng qo‘llaniladigan datasetlar qatoriga CoNLL-2003, WNUT-2017 hamda OntoNotes kiradi. Ushbu to‘plamlar mazmun jihatidan turlicha bo‘lib, natijada nomlangan obyektlarni aniqlash (NER) modellarini turli sharoit va kontekstlarda o‘qitish imkonini beradi[4:2-3].

CoNLL-2003 – yangilik matnlari to‘plami (ingliz va nemis tillari uchun). U anchagina katta ma’lumotlar to‘plami bo‘lish bilan bir paytda, multilingual hisoblanadi. Chunki yuqorida aytilganidek, uning data-ma’lumotlari 2 til uchun birdek ishlaydi[8:142–143]. Ushbu datasetdagi modellar annotatsiyaga muvofiq, “shaxs” (Person), “joy” (Location), “tashkilot” (Organization) va “boshqa”

(Miscellaneous) kabi obyekt turlarini aniqlashga o‘rgatiladi. Yana bir WNUT-2017 ma’lumotlar to‘plami esa asosan foydalanuvchilar tomonidan yaratilgan kontentlardan – masalan, Reddit, YouTube, Twitter yoki StackExchange platformalaridagi izohlar va blog yozuvlaridan tashkil topgan[2:141–142]. Ushbu dataset shaxslar, joylar, kompaniyalar, mahsulotlar, ijodiy asarlar hamda guruhlarni aniqlash vazifasiga yo‘naltirilgan.

NLP sohasida dastlabki ilmiy tadqiqotlar NER doirasida faqatgina qoidaga asoslangan yondashuvlar bilan cheklangan. Dastlabki NER modellari ham qo‘lda ishlab chiqilgan, til qoidalariga asolangan grammatik asosida qurilgan. Bunday tizimlarda nomlangan obyektlarni ajratib olish qo‘lda amalga oshiriladi. Nomlangan obyektlarni ajratib olish qoidalariga katta-kichik harflar tizimi, punktuatsion qoidalar, prefiks va suffiks qoidalari, semantik qoliplar, gap qurilishi kabilar kiritilgan (Qarang:1-rasm).



1-rasm. NER aniqlash yondashuvlari algoritmi

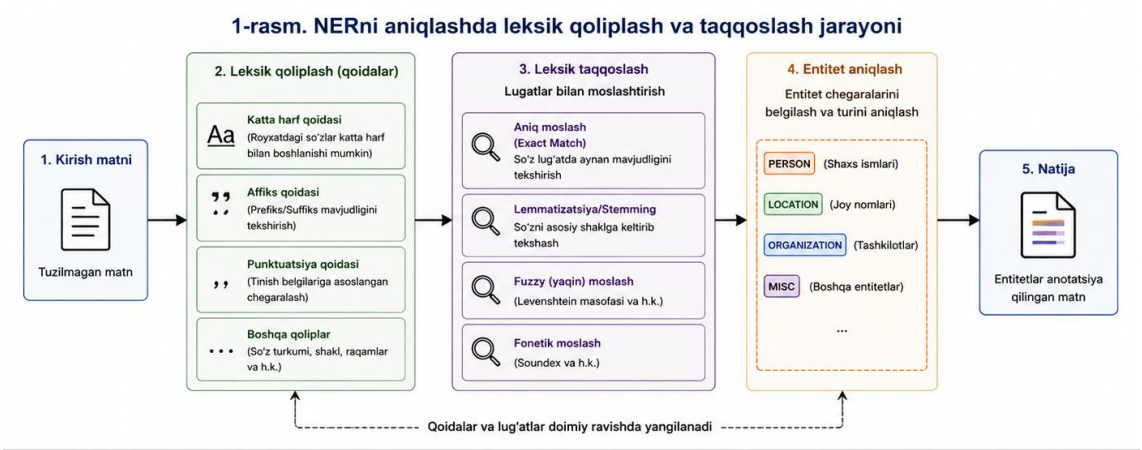
Nedau va Shekine (2007) fikricha, qoidalarga asoslangan yondashuvlar bilan bir vaqtda lug‘atlar va leksik moslashtirish yondashuvlari ham kiradi. Bu yerda

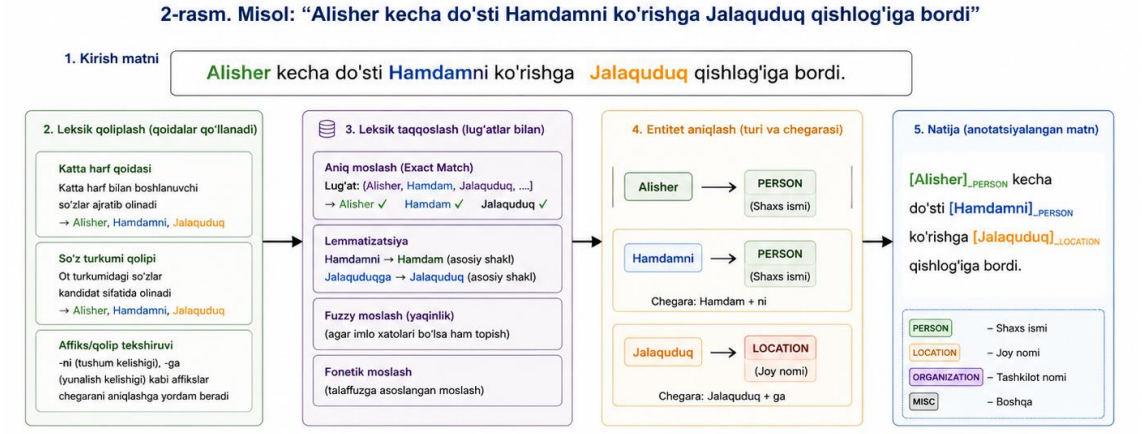
tayyor ro‘yxatlar deganda solishtirish uchun ishlatiladigan lug‘atlar yoki leksikonlar tushuniladi. Asosan uch turdagi ro‘yxatlar farqlanadi: umumiy ro‘yxatlar (turli aralash lug‘atlar), nomlangan obyektlar (inson ismlari, joy nomlari…) ro‘yxatlari va nomlangan obyektlariga ishora qiluvchi ro‘yxatlar. Bunda, Umumiy ro‘yxat lug‘atlari domenlarga bog‘liq bo‘lmagan so‘zlarni, nomuhim so‘zlar (stop-words), katta harf bilan yoziladigan otlarni yoki keng qo‘llaniladigan qisqartmalarni o‘z ichiga oladi va ular nomlangan obyekt bo‘lmagan birliklarni aniqlashda qo‘llanadi. Ya’ni bu yerda teskari qoida ishlaydi. Bu jarayon datalarni teglashda va bu ish avtomatik amalga oshirilganda NER bo‘lmagan birliklarni ajratib olish uchun juda muhim[6:3-26].

Nomlangan obyektlar (NER) ro‘yxati lug‘atlari esa tilda mavjud barcha atoqli otlar lug‘atlaridir. Bular ichida inson ismlari lug‘ati, joy nomlari lug‘atlari bo‘lishi mumkin. Nomlangan obyektlarga ishora qiluvchi lug‘atlar esa muayyan nomlangan obyekt turlarining mavjudligini bildiruvchi tipik hamroh so‘zlar (indicators words) o‘z ichiga oladi. Masalan, tashkilot turlarini bildiruvchi agentlik, muassasa, korxona, firma, kompaniya, yoki ko‘cha, viloyat, qishloq, mahalla kabi geografik yo‘nalishlardir[9:12].

Leksik moslashtirish turli usullar orqali amalga oshirilishi mumkin: aniq moslik (exact matching), lemmatizatsiya va asos shaklga keltirish, taxminiy moslik (fuzzy matching) yoki fonetik xususiyatlarga asoslangan usullar (masalan, Soundex) orqali. Quyidagi jarayonga e’tibor qaratsak, leksik qoidalar va lug‘atlar asosida matndan nomlangan obyektlarni ajratib olish jarayonini bosqichma-bosqich tushuntiriladi. Birinchi diagrammada kirish sifatida tuzilmagan matn olinadi va u ketma-ket bir necha bosqichdan o‘tadi: leksik qoidalar, leksik taqqoslash, obyektlarni aniqlash va yakuniy natija. Leksik qoidalar bosqichida katta harf, affikslar, punktuatsiya kabi qoidalar orqali so‘zlar filtrlanadi. Keyingi bosqichda so‘zlar lug‘atlar bilan taqqoslanadi: aniq moslik, lemmatizatsiya, fuzzy va fonetik

moslash usullari qo‘llaniladi. So‘ngra matndan *PERSON, LOCATION, ORGANIZATION* va *MISC* kabi entitet turlari aniqlanadi va natijada annotatsiyalangan matn hosil bo‘ladi. Ushbu jarayonni o‘zbek tilidagi misol orqali ko‘rib chiqamiz: *“Alisher kecha do‘sti Hamdamni ko‘rishga Jalaquduq qishlog‘iga bordi.”* Ushbu gap NER tizimiga kirish sifatida beriladi. Tizim “Alisher” va “Hamdamni” so‘zlarini *shaxs (PERSON)*, “Jalaquduq”ni esa joy nomi *(LOCATION)* sifatida aniqlaydi. Natijada matn ichidagi muhim obyektlar ajratilib, teglar bilan belgilangan ko‘rinishda qayta ishlangan annotatsiyalangan matn hosil bo‘ladi. Bu jarayon tuzilmagan matndan *strukturaviy ma’lumot olish* imkonini beradi (Qarang:2-rasm).





2-rasm. Leksik taqqoslash

Xulosa qilib shuni aytish mumkinki, NLP inson tilini kompyuter tizimlari orqali qayta ishlash imkonini beruvchi muhim ilmiy soha sifatida ko‘rib chiqildi.

Ayniqsa, axborot ekstraksiyasi va Named Entity Recognition (NER) texnologiyalari zamonaviy ma’lumotlarni tahlil qilish jarayonlarida markaziy o‘rin tutishi aniqlandi. Tadqiqot davomida NER tizimlarining tarixiy rivojlanishi, dastlabki qoidaga asoslangan yondashuvlardan boshlab, leksik moslashtirish va ma’lumotlarga asoslangan yondashuvlargacha bo‘lgan evolyutsiyasi yoritildi. CoNLL-2003 va WNUT-2017 kabi datasetlarning NER modellarini o‘qitishdagi ahamiyati ham tahlil qilindi. Shuningdek, axborot ekstraksiyasi tizimlarining ko‘p bosqichli transformatsion arxitekturasi va uning har bir bosqichida ma’lumotlarni qayta ishlash jarayoni izohlandi. Tadqiqot natijalari shuni ko‘rsatadiki, NER tizimlari tuzilmagan matndan strukturaviy axborotni ajratib olishda muhim vosita bo‘lib, zamonaviy NLP tizimlarining ajralmas qismi hisoblanadi. Kelgusida ushbu soha chuqur o‘rganilishi va yanada aniqroq modellar ishlab chiqilishi zarur.

Foydalanilgan adabiyotlar ro‘yxati

1. Chowdhary K. R. Fundamentals of Artificial Intelligence, 2020. Springer India. https://doi.org/10.1007/978-81-322-3972-7

2. Derczynski L., Nichols E., Van Erp M., et al. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. Proceedings of the 3rd Workshop on Noisy User-Generated Text, 2017. Pp. 140-147. https://doi.org/10.18653/v1/W17-4418

3. Holdsworth J., Cole S. What Is NLP (Natural Language Processing)? IBM, 2024, September 11. https://www.ibm.com/topics/natural-language-processing.

4. Jehangir B., Radhakrishnan S., Agarwal R. A survey on Named Entity Recognition – datasets, tools, and methodologies. Natural Language Processing Journal, 3, 2023.100017. https://doi.org/10.1016/j.nlp.2023.100017

5. Klontzas M. E., Fanni S. C., Neri E. Introduction to Artificial Intelligence. Springer International Publishing, 2023. https://doi.org/10.1007/978-3-031-25928-9

6. Nadeau D., Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 2007. Pp. 3–26. https://doi.org/10.1075/li.30.1.03nad

7. Nasar Z., Jaffry S. W., Malik M. K. Named Entity Recognition and Relation Extraction: State-ofthe-Art. ACM Computing Surveys, 54(1), 2022. Pp. 1–39. https://doi.org/10.1145/3445965

8. Tjong Kim Sang E. F., De Meulder F. Introduction to the CoNLL-2003 Shared Task: LanguageIndependent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Pp. 142–147. https://aclanthology.org/W03-0419/

9. Vili Lucic. Extraktion von Technologien aus Stellenausschreibungen: Eine Analyse verschiedener NER-Ansätze. Masterarbeit. Dornbirn, 04. Juli 2025. Pp. 12.