

Matnlarda NER obyektlarini aniqlashdagi muammolar

Madina Samatboyeva¹

Abstrakt:

Nomlangan obyekttni tanib olish axborotni boshqarishning ko'plab boshqa sohalari uchun asos bo'lib xizmat qiladi. Tabiiy tilni qayta ishlashda nomli obyektlarni tanib olish (NER) matndagi obyektlarning muayyan turlarini tanib olish va ajratib olish muammosidir. Haqiqatan ham, ismga ega bo'lgan har qanday aniq "narsa" NER obyekti sifatida ajratiladi. Ushbu maqolada mashinaviy o'qitish yordamida NER obyektlarini aniqlashda yuzaga keladigan muammolar ingliz tili va o'zbek tili matnlari tarkibida solishtirildi va tadqiq qilindi.

Kalit so'zlar: *tabiiy tilni qayta ishlash, NLP, mashinaviy o'qitish, nomlangan obyekt, NER, matn.*

Kirish

Axborotni olish (IE) kompyuter lingvistikasining asosiy kichik vazifalaridan biri bo'lib, tizimli ma'lumotlarni tizimli bo'lmagan yoki yarim tuzilmali mashina tomonidan o'qilishi mumkin bo'lgan hujjatlardan avtomatik ravishda ajratib olishga qaratilgan. Umatndagi barcha kompaniya nomlarini topishdan tortib, voqeaning barcha obyektlarini topishgacha bo'lgan keng ko'lamli kichik vazifalarni o'z ichiga oladi. Masalan, kim qaysi kompaniyaga ishga kirgan yoki kim qaysi bo'limda ishlayapti. Bunday imkoniyatlar foydalanuvchiga kerakli ma'lumotlarni topish uchun katta hajmdagi onlayn matnlarni saralash uchun tobora muhim ahamiyat kasb etmoqda.

Nomlangan obyektlar (NER)ni tanib olish va uning turlarini avtomatik identifikatsiya qilish vazifasi axborot olishning eng intensiv o'rganilgan usullaridan biridir. Til tahlilining taqdimoti odatda lug'atdagi so'zlarni qidirish va ularni ot, fe'l, sifat va boshqalar sifatida aniqlashdan boshlanadi. Lekin ko'pchilik matnlar juda ko'p

¹*Samatboyeva Madina To'liqinjon qizi* - Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxassisligi 1-kurs magistranti.

E-pochta: samatboyevamadina@navoiy-uni.uz

ORCID: 0000-0002-5146-4484

Iqtibos uchun: Samatboyeva M. 2022. "Matnlarda NER obyektlarini aniqlashdagi muammolar". *O'zbekiston: til va madaniyat. Amaliy filologiya*. 2 (5): 110-118.

ismlarni o'z ichiga oladi. Agar tizim ularni lug'atdan topa olmasa, ularni aniqlay olmaydi. Bu o'rinda matnni lingvistik tahlil qilish qiyin. Shunday qilib, NER ko'plab tabiiy tillarni qayta ishlash (NLP) vazifalarida (masalan, ma'lumot olish (IR) yoki mashina tarjimasini (MT)) muhim ahamiyatga ega. [Simon, 2013]

Asosiy qism

Matn tarkibidan NER obyektlarini avtomatik aniqlashda bir nechta muammolar mavjud. Bu muammolar matn tarkibidagi NER obyektlarini avtomatik tanib olish, ularga mos modellar yaratish, teglash, obyekt turini hamda uning matn tarkibidagi chegarasini belgilashda yuzaga keladi. Dastlab, biz ikki tushunchani o'rganib olishimiz zarur. Bular: *NER* va *nomlangan obyekt*.

Nomlangan obyektни tanib olish (NER) - *obyektni identifikatsiyalash* yoki *obyektni ajratib olish* deb ham ataladi - matndagi nomlangan obyektlarni avtomatik ravishda aniqlaydigan va ularni oldindan belgilangan toifalarga tasniflaydigan tabiiy tilni qayta ishlash (NLP) usuli. Matndagi NER obyektlari shaxslar, odamlar, tashkilotlar, joylar, vaqtlar, miqdorlar, pul qiymatlari, foizlar va boshqalarning nomlari bo'lishi mumkin. [<https://monkeylearn.com>]

Nomli obyekt - tegishli nom bilan ko'rsatilgan yoki ko'rsatilishi mumkin bo'lgan haqiqiy yoki xayoliy har qanday narsa. Nomlangan obyekt nima ekanligini bilish uchun NER modeli obyektни tashkil etuvchi so'z yoki so'zlar qatorini (masalan, London shahri) aniqlay olishi va u qaysi obyekt toifasiga tegishli ekanligini (London – Shahar (city)) bilishi kerak. *Nomlangan obyektlarni tanib olish* ularning *havolalarini topish! nazarda tutilgan obyekt yoki shaxsni aniq ko'rsatish! vazifasi* nafaqat matnda, shuningdek, *unga toifa berishni!* o'z ichiga oladi.

(NER), ehtimol, ma'lumot olish yo'lidagi birinchi qadam bo'lib, matndagi nomli obyektlarni oldindan belgilangan toifalarga, masalan, shaxslar, tashkilotlar, joylar, vaqt ifodalari, miqdorlar, pul qiymatlari, foizlar va boshqalarni aniqlash va tasniflashga intiladi. [Susan Li. 2018] SpaCy kutubxonasidan foydalangan holatda ajratib olingan NER obyektlari bilan tanishamiz [<https://towardsdatascience.com>]:

№	TYPE	DESCRIPTION
1	PERSON	People, including fictional.
2	NORP	Nationalities or religious or political groups.

3	FAC	Buildings, airports, highways, bridges, etc.
4	ORG	Companies, agencies, Institutions, etc.
5	GPE	Countries, cities, states.
6	LOC	Non-GPE locations, mountain ranges, bodies of water.
7	PRODUCT	Objects, vehicles, foods, etc. (Not services)
8	EVENT	Named hurricanes, battles, wars, sports events, etc.
9	WORK OF ART	Titles of books, songs, etc.
10	LAW	Named documents made into laws.
11	LANGUAGE	Any named language.
12	DATE	Absolute or relative dates or periods.
13	TIME	Times smaller than a day.
14	PERCENT	Percentage, including «%»
15	MONEY	Monetary values, including unit.
16	QUANTITY	Measurements, as of weight or distance.
17	ORDINAL	«first», «second», etc.
18	CARDINAL	Numerals that do not fall under another type.

1-jadval. SpaCy kutubxonasidagi NER obyektlari

No	Type	Tag	Sample Categories	Example sentences
1	Person	PER	people, characters	Turing is a giant of computer science.
2	Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
3	Location	LOC	regions, mountains, seas	The Mt. Sanitas loop is in Sunshine Canyon.
4	Geo-Political Entity	GPE	countries, states, provinces	Palo Alto is raising the fees for parking.
5	Facility	FAC	bridges, buildings, airports	Consider the Golden Gate Bridge.
6	Vehicles	VEH	planes, trains, automobiles	It was a classic Ford Falcon.

2-jadval. NER obyektlarining teglgan va kategoriyalangan ko'inishi [<https://medium.com>]

Shunday qilib, birinchi navbatda, biz nom, joylashuv, voqea, tashkilot va boshqalar kabi obyektlar toifalarini yaratishimiz va NER modeliga tegishli ma'lumotlarni ta'minlashimiz kerak. Keyin, ba'zi so'z va ibora namunalarini tegishli obyektlar bilan belgilash orqali biz NER modeliga obyektlarning o'zini qanday aniqlashni o'rgatamiz. [<https://monkeylearn.com>]

Endi esa ushbu NER obyektlari aniqlangan dasturdan misol taqdim etamiz [<https://towardsdatascience.com>]:



3-rasm. NER obyektlarini aniglovchi dastur ko'rinishi

Ushbu 3-rasm-jadvalga e'tibor qaratsak, uning tarkibida barcha NER obyektlari maxsus ranglar bilan ajratilgan va obyektlarning barchasiga toifa (obyekt turi nomi) belgilangan. Bundan tashqari, NER hujjat tahlilining bir turi hisoblanadi. U matnda qaysi obyektlar - shaxslar, joylar, tashkilotlar, sanalar, manzillar va hokazolar eslatib o'tilganligini va semantik tahlil orqali xulosa chiqarish mumkin bo'lgan shaxsning atributlarini belgilaydi. Masalan, shaxslar olmoshlarda va nomli eslatmalarga ishora qiluvchi qisqaroq shakllarda ham belgilanadi. Ushbu usul bilan nomni tanib olishning bunday turi anaforikdir, chunki obyektlar anaforalar orqali tan olinadi. Quyidagi misolga e'tibor qarating:

Michael Jordan was one of the best basketball players of all time. Scoring was Jordan's stand-out skill, but he still holds a defensive NBA record, with eight steals in a half.

4-rasm. <https://docs.expert.ai/nlapi/latest/guide/entity-recognition/>

Ushbu misolda Maykl Jordan haqida ikkita eslatma mavjud:

To'liq ismi: Michael Jordan

Anoforalar: Jordan (Jordan's), u – olmoshi (he – pronoun)

Ushbu anaforik usul NER obyektlarini matn tarkibidan aniq ajratib olish uchun to'liq samarali usul bo'la olmaydi. Matnni qayta ishlab, uning tarkibidagi obyektlarni avtomatik aniqlash jarayonida ham bir qancha muammolar yuzaga keladi. Quyida ushbu NER obyektlarini aniqlashdagi muammolar bilan tanishamiz.

Segmentatsiyaning noaniqligi. Bunda soʻzlar semantik jihatdan tahlil qilinib, ular orasidagi chegara toʻgʻri ajratilishi talab

etiladi. Masalan: Buyuk Britaniya (yoki: O'zbekiston Respublikasi) – bunda so'zlar qismi 2 ta, ammo ma'no bitta. Demak ular bitta obyekt, bitta NER.

Nyu York – bunda esa qismlar bitta ma'noga egami yoki ular alohida ma'noga ega so'zlarmi? Ya'ni birikmali so'zlar kelganda ular bitta obyekt deb qaraladimi yoki ular alohida semaga ega ikkita obyektmi? Mana shunday muammoli vaziyatlar bizda dastlabki yechimlarni taqdim etishni talab etadi.

Teglarni belgilashdagi noaniqlik. NER obyektlarini avtomatik aniqlashda, avvalo, teglash va modellashtirish jarayoni amalga oshiriladi. Teglash jarayoni esa tildagi har bir sintaktik bo'lak uchun mos teglarni talab etadi. Teglash jarayonida teglanayotgan obyekt bitta tegga bo'ysunsa yaxshi, ammo teglanish uchun tanlangan obyekt bir nom ostida bir nechta ma'noni hosil qilsa, bu holatda teglash jarayonida ham noaniqliklar yuzaga keladi. Masalan, quyidagi matnga e'tibor qaratamiz [<https://medium.com>]:

No	Name	Possible Categories
1	Washigton	Person, Location, Political Entity, Organization, Vehicle
2	Downing St.	Location, Organization
3	IRA	Person, Organization, Monetary Instrument
4	Louis Vuitton	Person, Organization, Commercial Product

5-jadval. Bitta NER nomi ostida turli obyektlarning ko'rinishi
[PER Washigton] was born into slavery on the farm of James Burroughs.

[ORG Washington] went up 2 games to 1 in the four-game series.

Blair arrived in [LOC Washington] for what may well be his last state visit. In June, [GPE Washington] passed a primary seatbelt law.

The [VEH Washington] had proved to be a leaky ship, every passage I made... [<https://medium.com>]

Yuqoridagi ingliz tilidagi matn tarkibida bitta "Vashington" obyekt ostida bir nechta teglarni talab etuvchi NER turlari (inson ismi (PER), tashkilot (ORG), yo'nalish (LOC), siyosiy tuzilma (GPE), transport (VEH)) keltirilgan. Bu o'rinda matn qayta ishlanganda dastur NER obyekt qaysi obyekt turi aynan shu matn tarkibida kelayotganini to'g'ri ajrata olishi kerak. Bunday muammo, nafaqat ingliz tili matnlari uchun, balki o'zbek tilidagi matnlarni qayta ishlashda ham mavjuddir. Quyida bitta "Alisher Navoiy" NER obyektiga e'tibor qaratamiz:

1. Alisher Navoiy – buyuk shoir. (Shaxs nomi)
2. Men “Alisher Navoiy” mahallasida yashayman. (Mahalla nomi)
3. U kecha “Alisher Navoiy” romanini o‘qidi. (Roman)
4. “Alisher Navoiy” fondi matnshunoslik bilan shug‘ullanadi. (Fond nomi)
5. Shahrimiz markazida “Alisher Navoiy” kutubxonasi joylashgan. (Kutubxona)
6. “Alisher Navoiy” bog‘i yurtimiz ko‘rkiga ko‘rk qo‘shmoqda. (Bog‘)
7. Davron Navoiy viloyatida yashaydi. (Viloyat)
8. Alisher Navoiy nomidagi davlat mukofoti ta‘sis etildi. (Davlat mukofoti)
9. Kecha televizorda “Alisher Navoiy” filmi namoyish etildi. (Film nomi)
10. O‘zRFA Alisher Navoiy nomidagi Til va adabiyot instituti (Institut nomi)
11. Alisher Navoiy nomidagi opera va balet akademik teatri ochildi. (Teatr nomi)
12. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti (Universitet)
13. Alisher Navoiy qishloq vrachlik punkti (Samarqand, Bulung‘ur tumanidagi qishloq vrachlik punkti)
14. Alisher Navoiy QFY (Samarqand, Bulung‘ur tumanidagi QFY)
15. Alisher Navoiy nomidagi fermer xo‘jaligi (Samarqand, Narpay tumanidagi fermer xo‘jaligi)
16. Alisher Navoiy haykallari (Toshkent, Navoiy, Samarqand, Boku, Tokio, Moskva, O‘sh, Shanxayda mavjud haykallari)
17. Alisher Navoiy nomidagi Merkuriydagi krater (Astronomik krater)

Chegaralarning noaniqligi. Ushbu vaziyatda matn tarkibidagi NER obyektlari chegarasi aniq bo‘linishni talab etadi. Quyida ikki misolga e‘tibor qaratamiz:

1. Bog‘da *qizil* va *oq* gullar ochilib turibdi.
 2. Tarixda Angliyada “*Qizil va oq gullar*” qo‘zg‘oloni bo‘lib o‘tgan.
- 1-gapda ikkita NER obyekti mavjud, ya‘ni “qizil” va “oq” (**COL**).
2-gap tarkibida esa bitta NER obyekti mavjud – “Qizil va oq gullar” (**voqea nomi – VEH**)

Ushbu vaziyatda dastur matnni qayta ishlashda obyekt chegarasini to'g'ri aniqlay olishi zarur. Bunda, albatta, kompyuter uchun biz taqdim etgan modellar ishlaydi. Ya'ni bunda *B – boshi (begin)*, *E – oxiri (end)*, *I – oraliq (intermediate)*, *S – bitta so'z (singular)* modellaridan foydalanishimiz mumkin.

Bosh harflar bilan bog'liq noaniqlik. Juda ko'p hollarda, nomlangan obyektlar matn tarkibida bosh harf bilan beriladi. Bu ularning "nomlanganlik" belgisi bilan bog'liq. Qayta nomlangan obyekt, o'z sifatiga, ko'rinishiga, holatiga qarab boshqa nom bilan ataladi. Bu o'rinda atab qo'yilgan nom matn tarkibida (gapning qaysi o'rnida kelishidan qat'iy nazar) bosh harf bilan beriladi. Ammo NER obyektlarini aniqlashda har doim ham bosh harflar qoidasiga tayanish biz kutgan ijobiy samarani bermaydi. Masalan:

"Davlat rahbarining topshirig'iga asosan, davlat oliy ta'lim muassasalariga o'qishga qabul qilish jarayonlarini muvofiqlashtirish bo'yicha komissiyaning tegishli bayoni bilan Qrim va Sevastopol hududlaridagi OTMda tahsil olayotgan O'zbekiston fuqarolarining ham o'qishini O'zbekistondagi davlat OTMga istisno tariqasida ko'chirishga ruxsat berildi. Bu qonun ko'pgina talabalarga yengillik taqdim etdi" [<https://kun.uz/uz/>].

Ushbu gapda qizil bilan belgilab ko'rsatilgan obyektlar NER obyektlaridir va ular gap tarkibida bosh harfda kelmoqda. Ammo nuqtadan so'ng punktuatsion qoidaga bo'ysungan holatda bosh harf bilan yoziluvchi so'zlar (davlat, bu) NER obyekti emas.

Boshqa tillar doirasida tadqiq qilganimizda ham, masalan, ingliz va rus tillarida nomlar bosh harfda yozilishi ham yozilmasligi ham mumkin. Bunga qarama-qarshi tarzda, nemis tilida esa barcha nomlar to'liq bosh harfda ifodalanadi. Ammo arab va xitoy tilida esa bosh harf tushunchasi mavjud emas. Bundan ko'rinadiki, NER obyektlarini avtomatik aniqlashda bosh harflar har doim biz kutgan natijani bermaydi.

Sonlar va yillar bilan bog'liq noaniqliklar. Bunda raqamlar bilan ifodalangan yillar aniqlikni beradi. Masalan, *"Biz 2010-yil Toshkentga borgan edik"* – ushbu gapda 2010-yil birligini NER obyekti sifatida qabul qilish mumkin. Ammo *"Biz may oyida Toshkentga borgan edik"* – gapida may oyi birligini NER obyekti sifatida qabul qilib bo'lmaydi. Chunki bunda vaqtning aniqligi yo'q. Ya'ni may oyi qaysi yilga tegishli ekanligi noma'lum. Bundan ko'rinadiki, vaqt bu - NER degan tushunchaga kelmasligimiz kerak.

Xulosa

Axborot yaratish tezligi yillar davomida o'sib bormoqda,

ayniqsa hujjatlar va hisobotlar kabi matnga asoslangan ma'lumotlar juda ko'p. Ushbu ma'lumotlardan foydalanib, biz kuchli algoritmlar yarata olamiz. Biroq, bu algoritmlar yoki kompyuter dasturlarini noldan yaratish, ayniqsa, matnga asoslangan ma'lumotlardan axborot olish uchun keng ko'lamli tajribalarni talab qiladi. Shuning uchun biz matnga asoslangan hujjatlardan muhim obyektlarni aniqlash va chiqarish uchun NLP (Tabiiy tilni qayta ishlash) usullariga tayanamiz.

Nomlangan obyektlarni tanib olish (NER) obyektlarini aniqlash matnni avtomatik qayta ishlashga mo'ljallangan NLP dasturlari uchun muhim qism hisoblanadi. Maqolada, nomlangan obyekttni tanib olish asoslari va undan foydalanish holatlari yoritildi. NER obyektlarni aniqlashdagi muammoli vaziyatlar tahlil qilindi va o'rganildi. Shuni alohida aytib o'tishimiz mumkinki, SpaCy - maxsus NER modellarini yaratish va ulardan foydalanish uchun yaxshi kutubxona. Demak, ushbu kutubxona yordamida yuqorida biz taqdim etgan muammoli vaziyatlarga ham yechim topa olishimiz mumkin.

Adabiyotlar:

- Simon 2013 Simon, Eszter. 2013. Approaches to Hungarian Named Entity Recognition. PhD. Thesis. http://www.cogsci.bme.hu/~ktkuser/PHD_iskola/dissertations/20131011_Simon_Eszter/ertekezes.pdf
- [https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20\(NER\)%20helps,to%20deal%20with%20large%20datasets](https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20(NER)%20helps,to%20deal%20with%20large%20datasets)
- Susan Li. 2018 Changing the world, one post at a time. Sr Data Scientist, Toronto Canada. <https://www.linkedin.com/in/susanli/>.
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- [https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20\(NER\)%20helps,to%20deal%20with%20large%20datasets](https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20(NER)%20helps,to%20deal%20with%20large%20datasets)
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- Mehal Gupta. 2020. <https://medium.com/data-science-in-your-pocket/named-entity-recognition-ner-using-conditional-random-fields-in-nlp-3660df22e95c>
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- <https://kun.uz/uz/news/2022/11/09/grim-va-sevastopol-hududlaridagi-talabalar-ham-istisno-tariqasida-oqishini-kochirishi-mumkin>

Problems of NER object identification in texts

Madina Samatboeva¹

Abstract:

Named entity recognition serves as the basis for many other areas of information management. In natural language processing, named entity recognition (NER) is the problem of recognizing and extracting certain types of entities in text. In fact, any particular “thing” that has a name stands out as a NER object. This article compared and studied the problems that arise when identifying NER objects using machine learning, as part of English and Uzbek texts.

Key words: *natural language processing, NLP, machine learning, named object, NER, text.*

References:

- Simon 2013 Simon, Eszter. 2013. Approaches to Hungarian Named Entity Recognition. PhD. Thesis. http://www.cogsci.bme.hu/~ktkuser/PHD_iskola/dissertations/20131011_Simon_Eszter/ertekezes.pdf
- [https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20\(NER\)%20helps,to%20deal%20with%20large%20datasets](https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20(NER)%20helps,to%20deal%20with%20large%20datasets)
- Susan Li. 2018 Changing the world, one post at a time. Sr Data Scientist, Toronto Canada. <https://www.linkedin.com/in/susanli/>.
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- [https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20\(NER\)%20helps,to%20deal%20with%20large%20datasets](https://monkeylearn.com/blog/named-entity-recognition/#:~:text=Named%20entity%20recognition%20(NER)%20helps,to%20deal%20with%20large%20datasets)
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- Mehal Gupta. 2020. <https://medium.com/data-science-in-your-pocket/named-entity-recognition-ner-using-conditional-random-fields-in-nlp-3660df22e95c>
- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
- <https://kun.uz/uz/news/2022/11/09/qrim-va-sevastopol-hududlaridagi-talabalar-ham-istisno-tariqasida-oqishini-kochirishi-mumkin>

¹Samatboyeva Madina To'liqinjon qizi - 1st year master's degree in Computer Linguistics at Alisher Navoi Tashkent State University of Uzbek Language and Literature.

E-mail: samatboyevamadina@navoiy-uni.uz

ORCID: 0000-0002-5146-4484

For reference: Samatboyeva M. 2022. “Problems of NER object identification in texts”. *Uzbekistan: language and culture. Applied philology.* 2 (5): 110-118.