KORPUSGA ASOSLANGAN YONDASHUV: NERNI N-GRAMM METODIDA ANIQLASH

Elov Botir Boltayevich,
Texnika fanlari doktori (DSc), dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Samatboyeva Madina To‘lqinjon qizi
tayanch doktorant
msamatboyeva@gmail.com
ToshDO‘TAU

Annotatsiya: Mazkur maqolada korpusga asoslangan yondashuv doirasida nomlangan obyektlarni aniqlash (Named Entity Recognition – NER) masalasi n-gramm metodlari yordamida o‘rganiladi. Tadqiqotda o‘zbek tili matnlari asosida shaxs nomlari, geografik nomlar, tashkilot nomlari va boshqa nomlangan birliklarni aniqlashda unigram, bigram va trigram modellarining samaradorligi tahlil qilinadi. O‘zbek tilining agglutinativ xususiyati, so‘z shakllarining ko‘pligi hamda lotin va kirill yozuvlari o‘rtasidagi tafovutlar NER tizimiga ta’sir qiluvchi omillar sifatida ko‘rib chiqiladi. Tadqiqot natijalariga ko‘ra, trigram model asosida korpus chastotalariga tayangan holda nomlangan birliklarni aniqlash yuqoriroq aniqlik ko‘rsatgan.

Kalit so‘zlar: *korpus lingvistikasi, NER, n-gramm, o‘zbek tili, nomlangan obyekt, NLP, trigram.*

Abstract: This article examines the problem of Named Entity Recognition (NER) within the framework of a corpus-based approach using n-gram methods. The study analyzes the effectiveness of unigram, bigram, and trigram models in identifying personal names, geographical names, organization names, and other named entities based on Uzbek language texts. The agglutinative nature of the Uzbek language, the abundance of word forms, and the differences between Latin and Cyrillic scripts are considered as factors influencing the performance of NER

systems. According to the research results, the trigram model, relying on corpus frequency data, demonstrated higher accuracy in recognizing named entities.

Keywords: *corpus linguistics, NER, n-gram, Uzbek language, named entity, NLP, trigram.*

Tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasida nomlangan obyektlarni aniqlash (NER) muhim vazifalardan biri hisoblanadi. NER matndan shaxs nomi, joy nomi, tashkilot nomi, sana, vaqt, pul birligi kabi semantik birliklarni avtomatik ajratib olish jarayonidir. Ushbu texnologiya axborot izlash, mashina tarjimasi, savol-javob tizimlari, chatbotlar va matn tahlili kabi yo‘nalishlarda keng qo‘llaniladi.

O‘zbek tilida NER bo‘yicha tadqiqotlar hali rivojlanish bosqichida bo‘lib, ayniqsa korpusga asoslangan statistik metodlar yetarlicha o‘rganilmagan. Shu sababli ushbu maqolada n-gramm metodlari yordamida o‘zbek tili matnlarida nomlangan obyektlarni aniqlash masalasi yoritiladi.

NER termini dastlab Lisa F. Rau (1991) [6:25] tomonidan ishlatilgan. Keyinchalik David Nadeau va Satoshi Sekine NER bo‘yicha statistik va qoidaga asolangan yondashuvlarni tasniflagan [5:64]

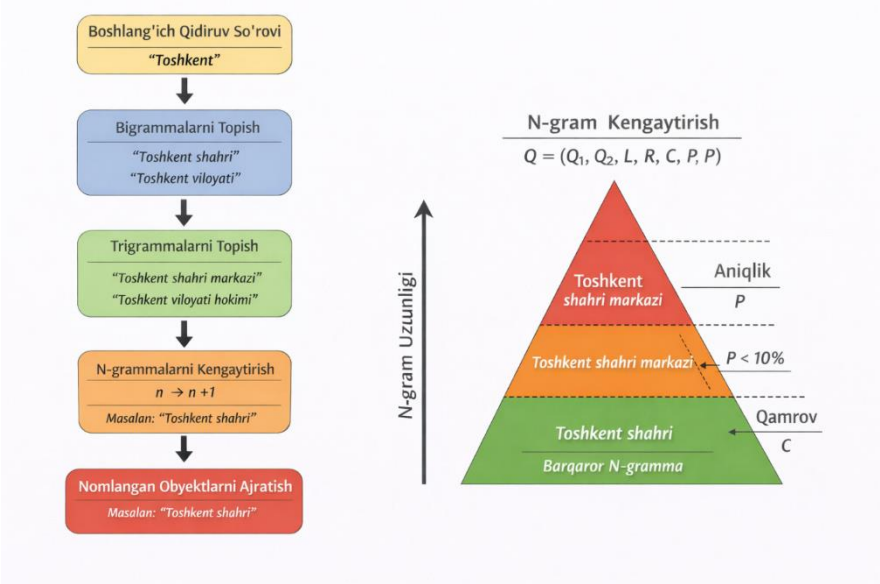
Jurafsky va James H. Martin o‘z asarlarida n-gramm modellarini til birliklari ehtimolligini aniqlashda samarali vosita sifatida baholaydilar [2:180].

So‘nggi o‘n yil ichida strukturlanmagan matnli ma’lumotlardan zarur tushunchalarni ajratib olishga qaratilgan ko‘plab NLP ilovalarini ishlab chiqish uchun asos sifatida til modellari shakllantirildi. Tilni modellashtirish – gapdagi so‘zning ehtimolini bashorat qilishni amalga oshirishga asoslangan tabiiy tilni qayta ishlashning asosiy vazifalaridan biri. Tilni modellashtirish amaliyotda avtoto‘ldirish (autocomplete), imloni tuzatish (spelling correction) yoki matnni generatsiya qilish (text generation) kabi ko‘plab NLP ilovalarida qo‘llaniladi. Til modeli oldingi yozuvlar asosida gapdagi keyingi so‘zni bashorat qilish uchun ishlatiladigan so‘zlar

bo‘yicha ehtimollik taqsimoti vositasida mashinali o‘qitishdan foydalanadi. Til modellari katta hajmdagi matnlarni o‘rganadi va matnni yaratish, matndagi keyingi so‘zni bashorat qilish, nutqni aniqlash, belgilarni optik aniqlash va qo‘l yozuvini aniqlash uchun ishlatilishi mumkin. [1:130-144]

N-gramm tahlili tilni qayta ishlashning muhim metodi (usuli) bo‘lib, u til tuzilishini tushunishga va gapda keyin nima kelishini bashorat qilishga yordam beradi. N-gramm tahlilini amalga oshirish matnni n-grammga ajratish orqali, har bir n-gramm chastotasini hisoblash va matn ma'lumotlari haqida tushunchaga ega bo'lish uchun n-gramm chastotasi va ularning tarqalishini tahlil qilishni o'z ichiga oladi. Bu jarayon NLP dasturi yordamida avtomatlashtirilishi yoki matndagi n-gramm chastotasini hisoblash orqali qo‘lda bajarilishi mumkin. N-grammlarni tahlil qilishdan matnni tasniflash, matn yaratish, imloni tuzatish va his-tuyg‘ularni (sentiment) tahlil qilish kabi murakkab NLP vazifalarni mashinali o'qitish modellarini o'rgatishda muhim rol o'ynaydi. [4:145-151]

Bugungi kunda kompyuterlar va Internetdan tobora keng foydalanilishi natijasida tabiiy tillaridagi katta hajmdagi ma’lumotlar hosil bo‘lmoqda. Ushbu avtomatik ma’lumotlarni qayta ishlash va qidirish qizimlarini (information retrieval, IR) optimallashtirish dolzarb vazifa hisoblanadi. Jumladan, P.Majumder, M.Mitra va B.B.Chaudhuri ko‘p tilli mamlakat hisoblangan Hindiston kontekstidagi IRda n-grammlardan muvaffaqiyatli foydalanilgan [3:20].



1-rasm. Nomlangan obyektlarni n-gramm metodida aniqlash bosqichlari

Quyidagi model nomlangan obyektlarni ajratib olishda **iterativ n-gramm kengaytirish va statistik barqarorlikka asoslangan filtratsiya** prinsipiga tayangan murakkab yondashuv sifatida tasniflanadi. Ushbu yondashuvda korpus ichidagi lokal kontekstlar global chastota bilan integratsiyalashadi.

Boshlang‘ich bosqichda qidiruv quyidagi ko‘rinishda beriladi:

$$Q = (Q_1, Q_2, L, R, C, P)$$

Bu yerda Q_1, Q_2 – lingvistik so‘rov (lemma yoki forma), L,RL, RL,R – chap va o‘ng kontekst bo‘yicha kengaytirish chegaralari, CCC – qamrov koeffitsienti, PPP esa moslik aniqligi.

Algoritmning asosiy dinamikasi quyidagicha ifodalanadi:

$$G_n = w_{i-n+1}, \ldots, w_i$$

$$G_{n+1} = w_{i-n+1}, \ldots, w_i, w_{i+1}$$

ya’ni har bir iteratsiyada n-gramma uzunligi rekursiv tarzda kengaytiriladi. Diagrammatik jihatdan bu jarayon **ketma-ket chuqurlashuvchi daraxt yoki piramida strukturasini** hosil qiladi: pastki qatlamda yuqori chastotali qisqa birliklar, yuqorida esa kamroq, ammo semantik jihatdan boyroq uzun birliklar joylashadi.

Filtrlash bosqichi quyidagi nisbat orqali amalga oshiriladi:

$$\frac{f(G_n)}{f(G_{n+1})} < P$$

Agar ushbu shart bajarilsa, G_n barqaror segment sifatida ajratib olinadi. Bu yerda muhim jihat – chastotalar o‘rtasidagi nisbat **lokal maksimal barqarorlik nuqtasini** aniqlaydi. Diagrammada bu holat piramidaning ma’lum bir qatlamida “to‘xtash zonasi” sifatida talqin qilinadi.

Mazkur yondashuvni konseptual jihatdan uch bosqichga ajratish mumkin:

1. **Ekstraksiya (Extraction)** – boshlang‘ich n-grammalarni hosil qilish
2. **Ekspansiya (Expansion)** – n → n+1 o‘tish orqali kontekstni kengaytirish
3. **Stabilizatsiya (Stabilization)** – chastota nisbatlari orqali optimal uzunlikni aniqlash

Natijada tizim **eng ko‘p takrorlanadigan emas, balki eng barqaror kontekstga ega birliklarni** aniqlaydi. Shu sababli bu model oddiy chastotaga asoslangan yondashuvlardan farqli ravishda, yashirin nomlangan obyektlarni ham aniqlash imkonini beradi.

Xulosa

Korpusga asoslangan yondashuv o‘zbek tilida nomlangan obyektlarni aniqlash (NER) tizimlarini yaratishda samarali va istiqbolli metodlardan biri ekanligi tadqiqot davomida o‘z tasdig‘ini topdi. Matnlar korpusi asosida shakllantirilgan statistik ma’lumotlar til birliklarining real qo‘llanish chastotasi, kontekstdagi o‘rni hamda o‘zaro bog‘liqligini aniqlash imkonini beradi. Shu jihatdan korpus resurslari NER tizimlarining aniqligi va barqarorligini oshirishda muhim manba vazifasini bajaradi. Ayniqsa, o‘zbek tilining agglutinativ xususiyati, so‘z shakllarining ko‘pligi hamda grammatik qo‘shimchalarning serqatlamligi kabi murakkab lingvistik jihatlar korpusga tayangan holda samaraliroq tahlil qilinadi.

Tadqiqot natijalari shuni ko‘rsatdiki, n-gramm modellar orasida trigram model eng yuqori samaradorlikni namoyish etdi. Sababi trigram model uchta ketma-ket

birlik asosida kontekstni chuqurroq hisobga oladi va nomlangan obyektlarni aniqlashda aniqlik darajasini oshiradi. Unigram model faqat alohida so‘zga tayanishi, bigram esa cheklangan kontekstni qamrab olishi sababli ayrim hollarda xatoliklarga yo‘l qo‘yadi. Trigram model esa shaxs nomlari, tashkilot nomlari va geografik obyektlarni aniqlashda yanada ishonchli natijalar berdi.

Foydalanilgan adabiyotlar ro‘yxati

1. Botir Elov, Nizomaddin Xudayberganov, Mastura Primova. O‘zbek tili matnlari uchun unigram til modelini ishlab chiqish: muammo va yechimlar. Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali Volume 2, ISSUE 5, OCTOBER 2024 ISSN: 3030-3346 . 130-144
2. Jurafsky, D., Martin, J. H. (2023). Speech and Language Processing. P-180
3. Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002, November). N-gram: a language independent approach to IR and NLP. In International conference on universal knowledge and language (Vol. 2)
4. Mastura Primova. O‘zbek tili matnlari uchun unigram til modelini ishlab chiqish: muammo va yechimlar. Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali Volume 2, ISSUE 5, OCTOBER 2024 ISSN: 3030-3346 . 145-151
5. Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. P-64
6. Rau, L. F. (1991). Extracting company names from text. P-25