

QAROR RO'YXATI ALGORITMI YORDAMIDA SO'Z MA'NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO'TAU

Hayotova Gulmira Hakim qizi,
Kompyuter lingvistikasi yo'nalishi talabasi
hayotovagulmira208@gmail.com
ToshDO'TAU

Annotatsiya: Ba'zi so'zlar bir xil tarzda yozilib, turli kontekstda turlicha ma'nolarga ega. Insonlar intellektga ega bo'lib, ular turli kontekstlarda ko'p ma'noli so'zlarning to'g'ri ma'nosini osonlik bilan tushunadilar. Buni atrofdagi so'zlar, ya'ni lokal yoki global kontekst yordamida amalga oshiradilar. Lokal kontekst target so'zga yaqin bir nechta so'zlarni anglatadi. Global kontekst esa noaniq so'zga yaqin bo'lgan satrlar yoki paragraflarni ifodalaydi. So'z ma'nosini aniqlashda kontekst muhim ahamiyat kasb etadi. Mashinali o'qitish yondashuvi asosida so'z ma'nosini aniqlash kontekstga asoslanadi. Ushbu maqolada qaror ro'yxatlari algoritmi yordamida so'z ma'nosini aniqlash haqida batafsil ma'lumot berilgan.

Annotation: Some words are spelled the same but have different meanings in different contexts. Humans are intelligent enough to easily understand the correct meaning of words with multiple meanings in different contexts. They do this by using the surrounding words, i.e., the local or global context. The local context refers to several words close to the target word. Global context refers to lines or paragraphs that are close to the ambiguous word. Context plays an important role in determining the meaning of a word. Machine learning approaches to determining the meaning of a word are based on context. This article provides detailed information about determining the meaning of a word using a decision list algorithm.

Аннотация: Некоторые слова пишутся одинаково, но имеют разное значение в разных контекстах. Люди разумны и могут легко понимать правильное значение слов, имеющих несколько значений в разных контекстах. Они делают это, используя окружающие слова, то есть локальный или глобальный контекст. Локальный контекст относится к нескольким словам, близким к целевому слову. Глобальный контекст представляет собой строки или абзацы, которые находятся близко к неоднозначному слову. Контекст важен при определении значения слова. Подход машинного обучения к определению значения слова основан на контексте. В данной статье представлена подробная информация об определении значения слова с использованием алгоритма списка решений.

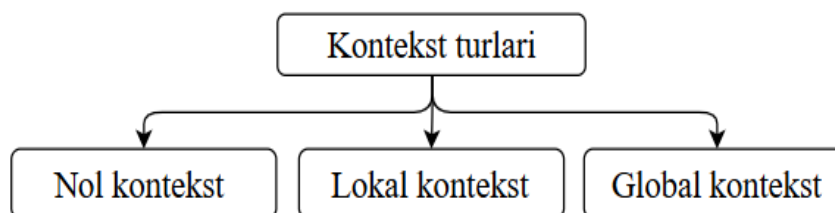
Kalit soʻzlar. *Mashinali oʻqitish yondashuvi, nazorat ostida oʻqitish algoritmlar, soʻz maʼnosini aniqlash, kontekst, vazn.*

Kirish

Dunyoning deyarli barcha tabiiy tillarida turli kontekstlarda turlicha maʼnoga ega boʻlgan soʻzlar mavjud. Bunday soʻzlar polisemantik yoki koʻp maʼnoli soʻzlar deb ataladi, ular bir xil yoziladi, lekin turli kontekstlarda alohida maʼnoga ega. Yaʼni ularning kontekstdagi maʼnosi noaniq. Noaniq soʻzning toʻgʻri maʼnosini aniqlash va haqiqiy maʼnoni belgilash jarayoni soʻz maʼnosini aniqlash (Word Sense Disambiguation, WSD) deb ataladi. Bu tabiiy tilga ishlov berish (NLP) sohasi uchun muhim yondashuvdir. Soʻz maʼnosini aniqlash mashina tarjimasini (MT), semantik xaritalash (SM), ontologiyalarni oʻrganish (OL), semantik annotatsiya (SA) kabi jarayonlarda muhim ahamiyatga ega.

Insonlar intellektga ega boʻlib, ular turli kontekstlarda koʻp maʼnoli soʻzlarning toʻgʻri maʼnosini osonlik bilan tushunadilar. Buni atrofidagi soʻzlar, yaʼni lokal yoki global kontekst yordamida amalga oshiradilar. Lekin bu jarayonni tizim orqali amalga oshirilsa, tizimga baʼzi qoida va yoʻriqnomalarni kiritish lozim. Shundan soʻng tizim kiruvchi kontekstdagi noaniq soʻzning toʻgʻri maʼnosini aniqlay oladi. Soʻzning maʼnosi – bu til jamoasi tomonidan soʻz uchun qabul qilingan maʼno. Noaniq soʻzning toʻgʻri maʼnosini topish uchun lugʻatga murojaat qilish mumkin. Yaʼni, noaniq soʻzning maʼnosini avtomatik aniqlash uchun resurslar zarur. Masalan ingliz tilida aniqmasliklarni bartaraf etish uchun WORDNET manba sifatida xizmat qiladi [1,2]. Oʻzbek tilining WORDNET tizimi ishga tushirilmaganligi sababli, boshqa manba hosil qilishga toʻgʻri keladi. Garchi oʻzbek tilining WordNet tizimi ishga tushirilmagan boʻlsada, noaniqliklarni aniqlash boʻyicha qator izlanish olib borilmada [7-9]. Yaʼni aniqmasliklarni turli maʼnolarini ifodalovchi gaplar toʻplamini jamlash zarur.

Gaplar shaxs tomonidan ifodalanganligi sababli, faqatgina gapga eʼtibor beriladi, his-tuygʻularga emas. Inson omili yordamida gaplardagi his-tuygʻuni aniqlash oson, ammo Mashinali oʻqitish yondashuvlari yordamida faqatgina gapga ishlov berish mumkin va koʻp maʼnoli soʻz uchrasa, uning atrofidagi soʻzlar tahlil qilinadi. Koʻp maʼnoli soʻzning birikuvchilari esa kontekstning oʻlchamiga bogʻliq. Kontekstlar quyidagicha boʻlishi mumkin:



1-rasm: Kontekstlarning turlari

Nol kontekst – soʻzning oʻzi.

Lokal kontekst (Mahalliy) – ibora va qoʻshma jumla.

Global kontekst – gap, muloqot, mavzu va soha.

Kontekstda noaniq soʻzning toʻgʻri maʼnosini aniqlash murakkab vazifadir. Kontekst *mahalliy* yoki *global* boʻlishi mumkin. Mahalliy kontekst noaniq soʻz atrofdagi bir nechta soʻzlarni qamrab oladi. Global kontekst esa bir jumladan koʻproq soʻzlar demakdir [1].

Kontekst turlari soʻz maʼnosini aniqlashda muhim ahamiyatga ega. Soʻz maʼnosini aniqlashda turli yondashuvlardan foydalanish mumkin.

Asosan, ikki xil soʻz maʼnosini aniqlash yondashuvi mavjud:

- 1) Mashinali oʻqitish yondashuvi.
- 2) Lugʻatga asoslangan yondashuv.

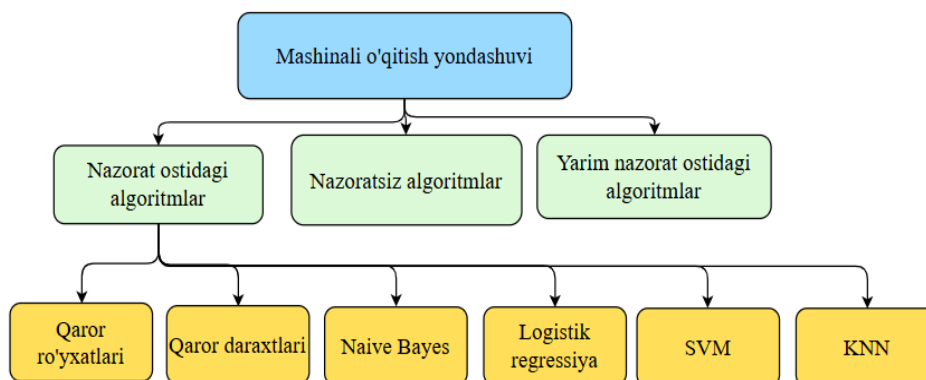
Mashinali oʻqitish yondashuvi algoritmlari 3 ta turga boʻlinadi:

- nazorat ostida oʻqitish algoritmlar,
- nazoratsiz oʻqitish algoritmlar,
- yarim nazorat ostida oʻqitish algoritmlar

Ushbu yondashuvda tizim toʻgʻri maʼnoni aniqlash uchun teglangan dataset oʻqitiladi. Klassifikator xususiyatlarni oʻrganish va yangi kiritilgan gaplardagi noaniq soʻzning maʼnolarni belgilash uchun qoʻllaniladi. Ushbu yondashuvda, noaniq soʻz va nomaʼlum soʻzning matndagi birikuvchilari tizimga kirish sifatida beriladi.

Jahon tajribasiga koʻra WSD tizimlarini ishlab chiqishda asosan mashinali oʻqitish yondashuvidan foydalanilgan. Mashinali oʻqitish yondashuvida inson omili yordamida yaratilgan, maʼnosi annotatsiyalangan maʼlumotlardan foydalaniladi. Klassifikator noaniq soʻzga tegishli misollarni ham oʻz ichiga olgan maʼlumotlar toʻplami asosida oʻqitiladi. Klassifikator joriy noaniq soʻzni kirish sifatida qabul qiladi va toʻgʻri maʼno uchun qoʻlda yaratilgan maʼno maʼlumotlar toʻplamini qidiradi. Oʻqitiladigan maʼlumotlar toʻplami klassifikatorni oʻqitish uchun ishlatiladi, bu ham noaniq soʻzni oʻz ichiga oladi.

Soʻz maʼnosini aniqlashda Nazorat ostida oʻqitish algoritmlardan foydalanish.



2-rasm: Nazorat ostida o'qitish algoritmlar

2-rasmdan ifodalanganidek nazorat ostida o'qitish algoritmlar sinfi ham bir nechta algoritmlarni qamrab oladi. Ushbu algoritmlar yordamida so'z ma'nosini aniqlashda teglangan ma'lumotlarga zarur bo'ladi. Oldingi maqolalarimizda Naïve Bayes algoritmi yordamida omonimiyani aniqlash haqida ma'lumot bergan edik [7].

Qaror ro'yxatlari. Qaror ro'yxatlari "agar - unda-aks holda" qoidalar to'plamidir. So'zning xususiyatlarini olish uchun o'qitish to'plami belgilangan ro'yxatda ishlatiladi. Ushbu qoidalar yordamida natijalar ro'yxati yaratiladi. Yarowsky ta'riflaganidek “Qaror ro'yxatlari (Decision Lists, DL) — noaniqlik muammolarini hal qilish uchun ishlatiladigan sodda vositalardan biridir “[4]. Ushbu yondashuv urg'uni tiklash, so'z ma'nosini aniqlash va omofonlarni ajratish kabi vazifalarda muvaffaqiyatli qo'llanilgan. Bu usul Senseval so'z ma'nosini aniqlash eng muvaffaqiyatli tizimlardan biri bo'lgan [5].

O'qitish ma'lumotlari xususiyatlarni (features) ajratib olish uchun qayta ishlanadi, bu xususiyatlar logarifmik ehtimollik (log-likelihood) o'lchovi yordamida baholanadi. Logarifmik ehtimollik qiymatlariga ko'ra tartiblangan barcha xususiyatlarning ro'yxati qaror ro'yxatini hosil qiladi. Olimlar formulani ikkitadan ortiq noaniqlik holatlarini hisobga olish imkonini berish uchun moslashtirganlar.

$$weight(s_i|f_k) = \log\left(\frac{P(s_i|f_k)}{\sum_{j \neq i} P(s_j|f_k)}\right) \quad (1)$$

Bu yerda, s_i — kontekstdagi noaniq so'zning i - ma'nosi,

f_k — k -xususiyat,

$weight(s_i|f_k)$ - s_i so'z ma'nosining f_k xususiyat bo'yicha vazni.

Qaror ro'yxatlari algoritmidagi so'z ma'nosining turli xususiyatlari bo'yicha vaznlarini hisoblash orqali so'z ma'nosini aniqlaydi. Bunda biror xususiyat bo'yicha so'zning har bir ma'nosi vaznlanadi va yuqori vaznga ega bo'lgan so'z ma'nosi natijasi sifatida qabul qilinadi. Xususiyatni aniqlashni turli usullari mavjud. Eng ko'p foydalaniladigan xususiyat turlaridan biri *Bag-of-Words (BOW)* lug'atidir.

Bunda teglangan datasetdan joriy ko'p ma'noli so'zning atrofidagi so'zlar chastotasi aniqlanadi. BOW ni hosil qilishda odatda joriy so'z atrofidagi 5 ta so'z olinadi.

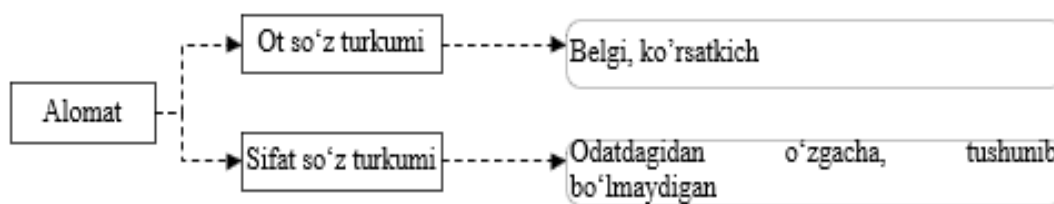
Xususiyatlarni quyidagicha tanlab olish mumkin:

- Bigrams/trigrams – datasetdagi joriy so'z ishtirokidagi bigram/trigrammalar asosida;
- Lemmalar asosida xususiyatlar-so'zning lug'aviy shakllari olinadi;
- Position xususiyatlar-so'zning oldin yoki keyin uchraydigan so'zlar;
- Semantik maydon asosida-tabiiy tilning WordNet tizimi

Kabi xususiyatlarni tanlab olish mumkin.

Osmondagi qora bulutlar shiddatli yomg'ir alomati demakdir.

Gapdagi *alomat* so'zi omonim so'z bo'lib, u turli kontekstlarda quyidagicha ma'nolarni anglatishi mumkin (3-rasm).



3-rasm: Alomat so'zining ma'nolari

Demak, ushbu aniqmaslikni qaror ro'yxatlari algoritmi yordamida aniqlashni ko'rib chiqamiz. (1) formuladag parametrlarning qiymatlarini aniqlab olamiz.

$$s = \{Belgi, ko'rsatkich; Odatdagidan o'zgacha, tushunib bo'lmaydigan\}$$

$$f = \{so'z, bigram, trigram, lemma, position, semantik maydon, win, morfologik\}$$

Yuqoridagi gapda keltirilgan *alomat* so'zi xususiyatlar to'plami bo'yicha qiymatlari 1-jadvalda keltirilgan.

1-jadval: Gapdagi alomat so'zining xususiyatlar bo'yicha qiymatlari

Xususiyat turi	Namuna
So'z	Yomg'ir / demakdir
Bigram	Yomg'ir alomati, alomati demakdir
Trigram	Yomg'ir alomati demakdi
Lemma	Lemma(shiddatli)=shiddat
position	So'zdan oldingi so'z=yomg'ir
Semantik maydon	Xususiyatlar maydoniga tegishli
Morfologik	Alomat so'zi ot sifatida foydalanilgan.

Alomat soʻzining s_i maʼnosi boʻyicha jamlangan gaplar toʻplamiga namuna keltiramiz.

2-jadval: Teglgan datasetdan namunalar

1-maʼnosi	2-maʼnosi
Bu kasallikning ilk alomati – tana haroratining koʻtarilishidir.	Bu alomat koʻz qarashlari bilan u nimadir demoqchi edi.
Osmonda qora bulutlar toʻplangani yomgʻirning yaqin alomati edi.	Alomat jimlik xonadagi muhitni yanada gʻalati qildi.
Donishmandlar aytadiki, odamning soʻzlari va xatti-harakati uning ichki dunyosining alomatidir.	Oʻsha kechadagi alomat tush hali-hanuz yodimdan chiqmaydi.
Uning koʻzidagi mung va yuzidagi charchoq yashirin gʻamning alomatlari edi.	Ular alomat harakatlari bilan atrofdagilarni bezovta qilishardi.
Til oʻrganishda muvaffaqiyatli boʻlishning eng yaxshi alomati – muntazam mashq qilishdir.	Alomat soʻzlar bilan boshlangan suhbat kutilmagan tomonga burildi
...	...

Xuddi shu kabi *alomat* soʻzining har bir maʼnosi boʻyicha 100 tadan gap jamlandi. Alomat soʻzining har bir maʼnosi boʻyicha vaznini hisoblashda biz *position* xususiyatidan foydalandik. *Position* xususiyati boʻyicha statistik koʻrsatkichlar oldik. Dastlab alomat soʻzidan oldin yomgʻir soʻzi uchrash ehtimolligini hisobladi. Quyidagi natijalar olindi.

$$P(s_1|yomg'ir) = \frac{C(s_1|yomg'ir)}{C(s_1)} = \frac{12}{100}$$

$$\sum_{j \neq i} P(s_j|f_k) = \frac{13}{200}$$

$$weight(s_1|yomg'ir) = \log\left(\frac{0.12}{0.065}\right) = \log(1.85) = 0.889$$

Koʻrinib turganidek, *alomat* soʻzidan oldin *yomgʻir* soʻzi uchraganda uning s_1 maʼnoni anglatishdagi vazni 0.889 ga teng. Endi s_2 maʼnosi boʻyicha vaznini hisoblaymiz.

$$P(s_2|yomg'ir) = \frac{C(s_2|yomg'ir)}{C(s_2)} = \frac{1}{100}$$

$$\sum_{j \neq i} P(s_j|f_k) = \frac{13}{200}$$

$$weight(s_2|yomg'ir) = \log\left(\frac{0.01}{0.065}\right) = \log(0.15) = -0.81$$

s_1 ma'nosi bo'yicha vazni -0.81 ga teng ekanligini ko'rish mumkin. Ma'lumki noaniq so'zning vazni 0 yoki manfiy qiymatga teng bo'lsa, demak ushbu xususiyat bo'yicha so'zning joriy ma'nosi ahamiyatga ega emas.

Osmondagi qora bulutlar shiddatli yomg'ir alomati demakdir.

Gapidagi *alomat* omonim so'zining ma'nosini aniqlashda Qaror ro'yxatlari algoritmidan foydalanilganda quyidagicha vaznlar aniqlandi.

$$weight(belgi, xususiyat|yomg'ir) = 0.889$$

$$weight(Odatdagidan o'zgacha|yomg'ir) = -0.81$$

Aniqlangan vaznlardan quyidagicha qaror qabul qilish mumkin, berilgan gapdagi noaniq so'z "*belgi, xususiyat*" ma'nosini anglatgan.

Qaror ro'yxatlari (Decision lists) algoritmidan foydalanish vazn hisoblash xususiyatlarining ko'pligi, uni tanlashning ixtiyoriyligi yoki mavjud barcha xususiyatlardan foydalangan holda ham vaznlashtirish mumkinligi bilan ushbu masalani yechishda qulaylik yaratib beradi. Agar omonim so'z turli so'z turkumlari doirasida turli ma'nolarga ega bo'lsa, u holda *morfologik* xususiyatdan foydalanish maqsadga muvofiq. Chunki omonim so'z mumkin bo'lgan har bir so'z turkumi doirasida faqat bitta ma'noga ega bo'ladi. Teglangan datasetda ham bu aks ettiriladi albatta, u holda statistik ko'rsatkichlar nisbatan kichikroq dataset ustida hisoblanadi hamda aniqroq natijalar olinadi.

Xulosa o'rnida shuni aytish mumkinki, Nazorat ostida o'qitish algoritmlar inson omili tomonidan teglangan, tekshirilgan ma'lumotlar ustida ish olib boradi. Qisqacha qilib aytganda namunalar asosida o'rganadi. Ushbu o'qitish yondashuvi namunalardan klassifikatsiya modelini ishlab chiqadi. Model har safar misol kelganda xususiyatlar fazosida nuqtalarni saqlaydi, bu yondashuv yangi misol va xususiyatlar fazosidagi ma'lumotlar o'rtasidagi masofani hisoblaydi. Oldindan saqlangan ma'lum bir sinfga yaqinroq bo'lsa, o'sha sinfga qabul qilish imkoniyati yuqori bo'ladi. Ushbu maqolada so'z ma'nosini aniqlash uchun qaror ro'yxatlari algoritmidan foydalanishning batafsil tahlili keltirildi. Asosan, mashinali o'qitish usulida bu jarayonni amalga oshirish uchun katta korpusdan foydalaniladi va ushbu usulda aniqlash jarayoni uchun tizimni o'qitish zarur. Ushbu mashinali o'qitish algoritmlari yaxshi natijalar ko'rsatadi, ammo asosiy kamchiliklari shundaki, ular katta korpusga ehtiyoj sezadi.

Foydalanilgan adabiyotlar:

1. Kilgarrieff, A., Evans, M. (eds): Special issue on SENSEVAL. Computer and the Humanities, 34(1-2) (2000)
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. Special Issue of International Journal of Lexicography, 3(4), (1990)

3. S.Ramana, M.V.Ramana Murthy, & N.Bhaskar. (2017). Ensuring data integrity in cloud storage using ECC technique, International Journal of Advanced Research in Science and Engineering, BVC NS CS 2017, 06(01), 170–174. ISSN Number: 2319-8346
4. Yarowsky, D. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French', in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88--95. 1994.
5. Kilgarriff, A. and M. Palmer. (eds). Special issue on SENSEVAL. Computer and the Humanities, 34 (1-2). 2000
6. R Alugubelli. (2016). Exploratory Study of Artificial Intelligence in Healthcare. International Journal of Innovations in Engineering Research and Technology, 3(1), 1–10.
7. B.Elov, X.Axmedova. Naïve Bayes klassifikatori yordamida omonimiyani aniqlash. Muhammad al-Xorazmiy avlodlari ilmiy-amaliy va axborot tahliliy jurnal 4(26)/2023 ISSN -2181-9211, 194-201-b.
8. E. B. Boltayevich and A. X. Ilxomovna, "Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speechs in The UzbekLanguage," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 278-283, doi:10.1109/UBMK55850.2022.9919453
9. Axmedova X. I. Turli soʻz turkumlari orasidagi omonimiyani aniqlovchi matematik modellar// Scienceand innovation international scientific journal volume 1 issue 7 uif-2022: 8.2 | issn: 2181-3337.https://doi.org/10.5281/zenodo.7238546[12]
10. Axmedova X.I. Chastotali usul yordamida omonimiyani aniqlash// “Oʻzbek amaliy filologiyasi istiqbollari” Respublika ilmiy-amaliy konferensiyasi Toshkent: 2022.-164-170 b.