KNN-ALGORITMI YORDAMIDA SO‘Z MA’NOSINI ANIQLASH (sifat so‘z turkumi misolida)

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

Xushmuratova Madina Baxtiyor qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
xushmurodovamadina209@gmail.com
ToshDO‘TAU

Annotatsiya: Kompyuter lingvistikasi sohasining muhim vazifalaridan biri so‘z ma’nosini aniqlashdir. So‘z ma’nosini aniqlash atamasi negizida semantik tahlil elementlari aniqlash masalasi yetakchilik qiladi. Bu elementlar omonim, polisemantik, polifunksional so‘zlar, nomlangan obyektlarni tanib olish (NER) meronim, holonim so‘zlardir. Avtomatik semantik tahlilni amalga oshirish turlicha yondashuvlarni talab qiladi. Zamonaviy yondashuvlardan foydalanish yuqori aniqlikka erishishni ta’minlaydi. Ana shunday yondashuvlardan biri Mashinali o‘qitish yondashuvi bo‘lib, bu yondashuv algoritmlari uchta guruhga bo‘linadi: Nazorat ostida o‘qitish, nazoratsiz o‘qitish, yarim nazorat ostida o‘qitish algoritmlari. Ushbu maqolada nazorat ostida o‘qitish algoritmlaridan bir KNN algoritmi yordamida omonim so‘zlarni semantik farqlash ketma-ketligi keltirilgan.

Annotation: One of the key tasks in the field of computational linguistics is word sense disambiguation. The term “word sense disambiguation” is primarily associated with identifying elements of semantic analysis. These elements include homonyms, polysemantic and polyfunctional words, named entity recognition (NER), meronyms, and homonyms. Performing automatic semantic analysis requires various approaches. Using modern methods ensures high accuracy. One such method is the Machine Learning approach, which is divided into three groups: supervised learning, unsupervised learning, and semi-supervised learning algorithms. This article presents a sequence of semantic disambiguation of homonymous words using one of the supervised learning algorithms-the KNN algorithm.

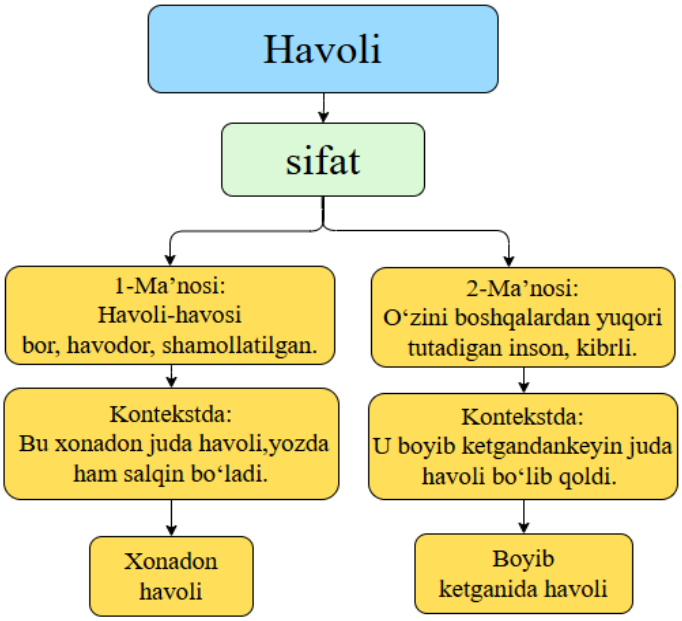
Аннотация: Одной из важных задач в области компьютерной лингвистики является определение значения слова. Термин «определение значения слова» подразумевает выявление элементов семантического анализа. К таким элементам относятся омонимы, полисемантические и полифункциональные слова, распознавание именованных сущностей (NER), меронимы и холонимы. Автоматический семантический анализ требует различных подходов. Применение современных методов обеспечивает

высокую точность. Одним из таких подходов является машинное обучение, которое подразделяется на три группы: обучение с учителем, обучение без учителя и частично контролируемое обучение. В данной статье представлена последовательность семантического различения омонимов с использованием одного из алгоритмов обучения с учителем — алгоритма KNN.

Kalit so‘zlar: *So‘z ma’nosini aniqlash, tabiiy tilga ishlov berish, omonim, mashinali o‘qitish, KNN*

Kirish

So‘z ma’nolarini aniqlash (WSD) tabiiy tilga ishlov berish (Natural Language Processing, NLP) sohasidagi muhim muammolardan biri bo‘lib, axborot izlash, mashina tarjimasi, nutqni tanib olish kabi ko‘plab ilovalar uchun zarurdir. WSD muammosi shundan iboratki, ko‘p ma’noli so‘z (polisemiya), vazifadoshlik (polifunksional), omonim so‘zlar turli kontekstlarda ma’nolarni ifodalaydi va bu so‘zning joriy kontekstdagi ma’nosini aniqlash muhim vazifa hisoblanadi [1]. Mashina odam singari so‘zning ma’nosini to‘g‘ri tushunishi uchun katta hajmdagi ma’lumotlar talab etiladi. Shu sababli, ko‘p hollarda statistik va ehtimollik modellari asosida yondashuvlar ishlab chiqilgan. Buni quyidagi “havoli” so‘zi misolida ko‘rib chiqildi. Bu so‘z bir so‘z turkumida ya’ni sifat so‘z turkumida omonimlikni tashkil qiluvchi so‘z hisoblanadi. Bu omonim so‘z konteksda kelganda qaysi ma’noda kelganini aniqlash uchun avval uning yonidagi qo‘shni so‘zlarini topib, eng ko‘p qaysi so‘z bilan kelishi aniqlanib olib, shu orqali ma’nosi aniqlanishi mumkin:



1-rasm. Havoli so‘zining ma’nolari

WSD muammosini hal qilish uchun turli yondashuvlar mavjud bo‘lib, shulardan eng ko‘p qo‘llaniladigan yondashuv mashinali o‘qitish yondashuvidir.

Mashinali o‘qitish yondashuvi o‘qitiladigan ma’lumotlarning turiga qarab uchta guruhga bo‘linadi:

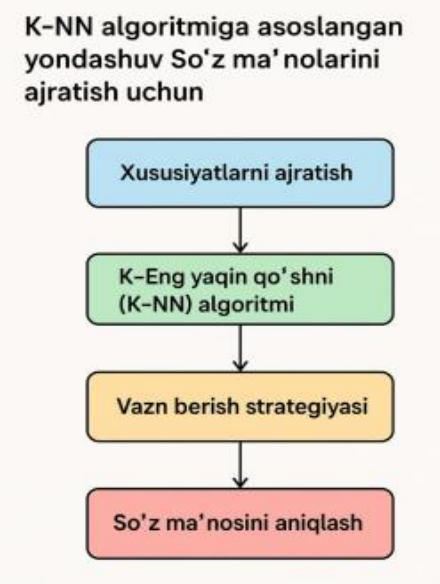
Nazorat ostidagi algoritmlar – annotatsiyalangan (sense-tagged) korpuslardan foydalanib mashinani o‘rgatish [2].

Nazoratsiz algortimlar – belgilanmagan yoki teglanmagan (unlabeled) korpuslar asosida klasterlash usullari orqali ma’nolarni ajratish [2].

Yarim nazorat ostidagi algoritmlar – yuqoridagi ikkala yondashuvni birlashtirgan usullar [2].

Ushbu maqolada nazorat ostidagi algoritmlar guruhiga mansub bo‘lgan KNN algoritmi yordamida so‘z ma’nosini aniqlashni ko‘rib chiqiladi.

K-NN algoritmi. K-NN – nazorat ostidagi mashinali o‘qitish algoritmi bo‘lib, semantik o‘xshashlikka asoslangan klassifikatsiya usulini qo‘llaydi [3]. Noaniq so‘z uchun eng yaqin k ta o‘xshash namunalar topilib, ular asosida ushbu so‘zning ma’nosi aniqlanadi. Bu jarayon ketma-ketligini 2-rasmda ko‘rish mumkin:



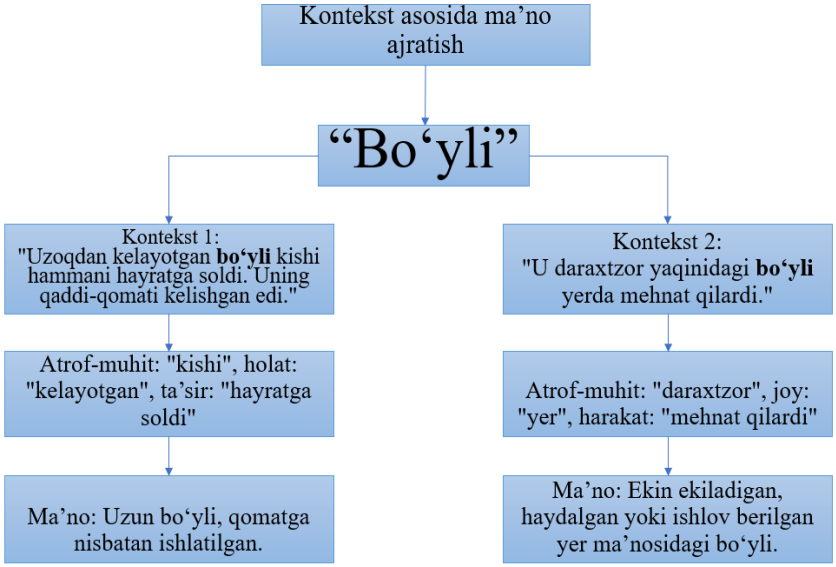
2-rasm. K-NN algoritmi ketma-ketligi

2-rasmdagi jarayonlarni ta’riflab o‘tsak.

Xususiyatlarni ajratish. Bu atama ingliz tilida Feature Extraction ya’ni xususiyatlarni ajratish deb nomlanadi [3]. WSD tizimini ishlab chiqishda xususiyatlarni to‘g‘ri ajratish juda muhim. Xususiyatlar to‘plami quyidagi turlardan iborat:

Ko‘p uchraydigan so‘zlar to‘plami (TF - Term Frequency). Bu usulda, matndagi eng ko‘p uchraydigan so‘zlar tanlanadi. TF asosida noaniq so‘zning biror ma’nosini ifodalaydigan kontekstda eng ko‘p uchraydigan so‘zlar bo‘lib, ular noaniq

so‘zning ma’nosini aniqlashda yordam beradi [4]. Bu jarayonni quyidagi 3-rasmdagi 1 va 2-kontekstlarda uchragan *bo‘yli* omonim so‘zi misolida ko‘rib chiqiladi, bu so‘z sifat so‘z turkumi doirasida omonimlikni hosil qilib, turli kontekstlarda turli ma’nolarni ifodalashi mumkin:



3-rasm. Konteks asosida ma’no ajratish (bo‘yli so‘zi misolida)

“Bo‘yli” so‘zi 1- kontekstda “kishi” so‘zi bilan, 2-kontekstda esa “daraxtzor” so‘zlari atrofida paydo bo‘lgan. Omonim so‘zning eng ko‘p uchraydigan birikuvchilarini hisobga olish orqali kontekst ma’nosi aniqlanadi.

Atrofdagi so‘zlar to‘plami. Noaniq so‘zning atrofidagi so‘zlar hisobga olinadi, masalan, Word2Vec yoki boshqa embedding usullari yordamida [5]. 3-rasmdagi 2 ta kontekstda ham mavjud bo‘lgan “bo‘yli” so‘zining atrofidagi bu so‘z bilan kelgan so‘zlari turlicha. Birinchi jumlada “kishi” va “kelayotgan” so‘zlari bilan, ikkinchi jumlada esa “daraxtzor” va “yer” so‘zlari bilan birikib kelgan. Omonim so‘zning har bir ma’nosi bo‘yicha jamlangan gaplarning soni ortishi bilan ularning birikuvchilari soni ham oshadi, bu esa so‘zning konteks ma’nosi aniqlanishida yordam beradi. Omonim so‘zning kontekst ma’nosini uning birikuvchilari yordamida aniqlashda turli algoritmlardan foydalanish mumkin. Shunday algoritmlardan biri KNN algoritmi.

*K-eng yaqin qo‘shni (K-NN) algoritmi*ning ishlash prinsipi:

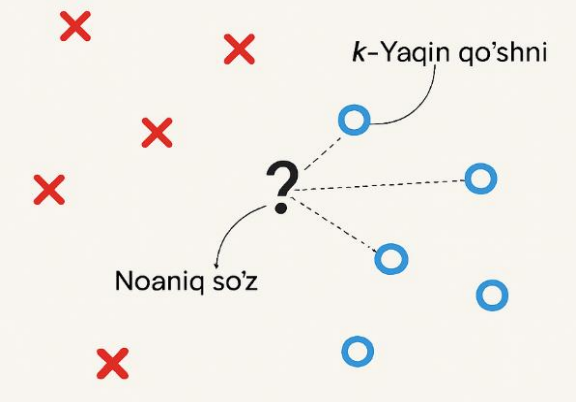
Ma’lumotlarni tayyorlash: Boshqa so‘zlar bilan ajratilgan ma’lumotlar (masalan, matndagi so‘zlar) mavjud bo‘ladi.

Masofa o‘lchovi: So‘zlar orasidagi o‘xshashlikni masofa yordamida o‘lchash mumkin. K-NN algoritmida masofa o‘lchovlari sifatida Evklid masofasi ko‘pincha ishlatiladi.

Eng yaqin k ta qo‘shnini topish: Yangi so‘z uchun eng yaqin k ta qo‘shni so‘zlar tanlanadi.

O‘rganish va klassifikatsiya: Bu k ta qo‘shnilar asosida yangi so‘zning ma’nosi aniqlanadi. Ya’ni, yangi kiritilgan gapdagi noaniq so‘zning ma’nosi, uning atrofidagi k ta qo‘shni so‘zlarning umumiy ma’nosi bilan bog‘lanadi [6].

Sifat so‘z turkumi doirasidagi omonim so‘zlarni knn algoritmi yordamida semantik farqlash jarayonini ko‘rib chiqamiz. O‘zbek tilida faqat sifat so‘z turkumi doirasida omonimlik hosil qiluvchi omonim so‘zlarni “O‘zbek tili izohli lug‘ati” dan yig‘ilganida 82 tani tashkil etib, ular jami 175 ma’noni qamrab oladi. Ushbu omonim so‘zlarning kontekst ma’nosini KNN algoritmi yordamida aniqlashdagi dastlabki jarayon ma’lumotlarni jamlash bo‘lib, omonim so‘zlarning har bir ma’nosini anglatuvchi 100 tacha gaplar jamlandi. Jamlangan gaplar asosida sifat omonimlarni aniqlash uchun KNN algoritmidan foydalanildi. K-NN algoritmi uchun k koeffitsentini tanlash muhim sanaladi. Jahon tajribasiga ko‘ra k=3, 5 kabi belgilash mumkin ekan, asosa k=3 kabi qo‘llanilgan.



4-rasm. Qo‘shni so‘zlarni aniqlash

Vazn berish strategiyasi (Weighting Strategy) so‘zlarning ma’nolarini aniqlashda, ularning matndagi joylashuvi va chastotasi muhim rol o‘ynaydi. Har bir so‘zning vaznini quyidagi formulaga asoslanib hisoblash mumkin:

$$w(k, f_i) = \frac{C(k, f_i)}{C(k)}$$

Bu yerda:

N (k, f-i) – k-ma’nodagi so‘z bilan birga uchraydigan f-i xususiyatli so‘zlar soni,

N (k) – k-ma’nodagi umumiy namunalar soni.

1-jadval.Omonin so‘zlar statistikasi

So‘z	Atrofdagi so‘zlar bilan (%)	Tez-tez uchraydigan so‘zlar bilan (%)	Ikkisini birga ishlatganda (%)
bo‘ydor	90.7	89.7	90.7
bo‘sh	76.8	75.8	76.8
bo‘yli	75.6	72.6	75.6
bog‘li	78.6	82.1	78.6
sirli	61.2	56.9	63.8
belli	70.1	65.7	71.1
O‘rtacha:	75.5	73.8	76.1

O‘rtacha aniqlik: 76.1%

Taklif etilgan metod ilgari ishlab chiqilgan yondashuvlardan yaxshiroq natija ko‘rsatildi.

Xulosa

Mazkur maqolada tabiiy tilni qayta ishlashda muhim hisoblangan so‘z ma’nolarini aniqlash muammosi amaliy jihatdan yoritildi. K-NN algoritmiga asoslangan nazorat qilinuvchi yondashuv yordamida sifat so‘z turkumidagi ko‘p ma’noli so‘zlar kontekst asosida tahlil qilindi. Eksperimentlar natijalari shuni ko‘rsatdiki, kontekst oynasi (Context Window) va so‘z chastotasi asosida xususiyatlar ajratish, shuningdek, vazn berish strategiyalarining qo‘llanishi model aniqligini sezilarli darajada oshiradi. To‘plangan gaplar test qilish natijasida o‘rtacha 76.1% aniqlik kuzatildi, bu esa ilgari mavjud yondashuvlarga nisbatan yuqori ko‘rsatkich hisoblanadi.

Kelajakdagi tadqiqotlarda kontekstga asoslangan chuqur o‘rganish (deep learning) modellarini, jumladan BERT kabi neyron tarmoqlarni qo‘llash, boy va ko‘p tilli korpuslar asosida ma’no ajratish tizimlarini ishlab chiqish, hamda cross-lingual yondashuvlar orqali turli tillarda WSD tizimlarini taqqoslab o‘rganish rejalashtirilgan. Bu esa tilshunoslik va sun’iy intellekt kesishgan sohaga katta ilmiy va amaliy hissa bo‘ladi.

Foydalanilgan adabiyotlar:

1. Boulder, Colorado. Statistical Post-Editing of a Rule-Based Machine Translation System, Proceedings of NAACL HLT 2009: Association for Computational Linguistics, pp 217-224

2. Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes, 8th EAMT conference, 22-23 September 2008, Hamburg, Germany, pp. 6-11.

3. Abdul-Rauf and Holger Schwenk. On the use of Comparable Corpora to improve SMT performance Sadaf. 12th Conference of the European Chapter of the ACL, pages 16–23, Athens, Greece, 30 March – 3 April 2009.pp 16-23

4. Yarowsky.: 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. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (2013).

6. Gale, K. Church, and D. Yarowsky.: A Method for Disambiguating Word Senses in a Large Corpus. Computers and Humanities, vol. 26, pp. 415-439 (1992).

7. Baker, P., McEnery, T., Hardie, A. . A glossary of corpus linguistics. Edinburgh: Edinburgh University Press. (2006).