

TABIIY TILGA ISHLOV BERISH SOHASIDA SO‘Z MA’NOSINI ANIQLASH

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

Hasanova Gulshoda Ikrom qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
ToshDO‘TAU

Annotatsiya. Tabiiy tilga ishlov berish sohasi matnlarni avtomatik tahlil qilish masalasini yechish bilan shug‘ullanadi. Avtomatik tahlil turlaridan biri so‘z semantic tahlil bo‘lib muhim ahamiyatga ega. Avtomatik semantik tahlil amalga oshirishda uni bir nechta elementlarga ajratgan holda amalga oshiriladi. Ushbu aniqlash, matnni tahlil qilish va so‘zlar orasidagi bog‘lanishni tushunish asosiy vazifasi sanaladi.

Annotation. The field of natural language processing deals with solving the problem of automatic text analysis. One of the types of automatic analysis is word semantic analysis, which is of great importance. When performing automatic semantic analysis, it is carried out by dividing it into several elements. This identification, text analysis, and understanding of the relationships between words are considered the main tasks.

Аннотация. Область обработки естественного языка занимается решением задачи автоматического анализа текстов. Одним из видов автоматического анализа является семантический анализ слов, который имеет важное значение. При выполнении автоматического семантического анализа он осуществляется с разделением на несколько элементов. Это определение, анализ текста и понимание связей между словами считаются основными задачами.

Kalit so‘zlar: *So‘z ma’nosini aniqlash, NLP, noaniqlik, leksema, leksik noainqlik, struktural noaniqlik, referensial noainqlik*

Kirish

Kompyuter lingvistikasi sohasida matnlarni avtomatik semantik tahlil qilish masalasining yechilishi kunning dolzarb masalalaridan biri. Semantik tahlil matn yoki nutqning ma’nosini, uning tarkibini va mazmunini o‘rganadi. Semantik tahlil orqali so‘zlar, iboralar, jumlar yoki butun matnning semantik (ya’ni ma’noga oid) tuzilishi tahlil qilinadi. Barcha tabiiy tillarida noaniq so‘zlar mavjud, bu so‘zlar o‘z ma’nosini kontekstdan kelib chiqqan holda o‘zgartiradi. Biror so‘zning to‘g‘ri

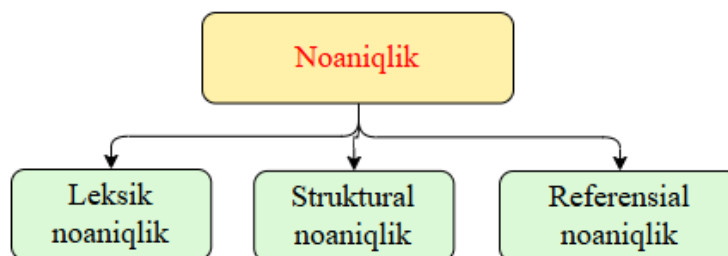
ma'nosini kontekstga qarab aniqlash jarayoni So'z Ma'nosini Aniqlash (WSD) deb ataladi. Bu jarayon, bir so'z jumlada ishlatilganda uning maqsad qilingan ma'nosini avtomatik tarzda aniqlaydi. Noaniq so'zlarning ba'zi misollari quyidagilar:

1-jadva: Noaniq so'zlar va ularning ma'nolari

t/r	So'zlar	Ma'nolari
1	ich	1.suyuqlikni ichga yutmoq 2.ichkaridagi joy
2	yara	1. chiqiq, jarohat 2. kerak bo'lib qolmoq
3	qazi	1. otning to'sh va peshnop go'shti shulardan yo'g'on ichakka tiqib tayyorlanadigan mahsulot 2. yerni o'yib, chuqur hosil qilmoq, kovlab olmoq

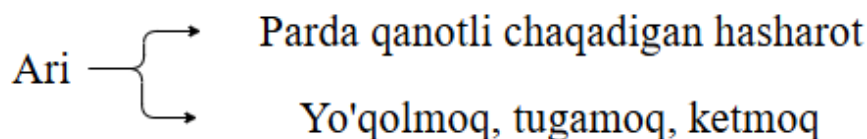
So'z ma'nosini aniqlash (WSD) turli sohalarda, masalan, mashina tarjimasi (Machine translation, MT), axborotni qayta ishlash (Information reworking, IR), nutqni tanib olish (Automatic Speech Recognition, ASR), matnni qayta ishlash kabi ko'plab sohalarda qo'llaniladi. WSD boshqa NPL axborot tizimlarining natijasining aniqligini oshirishda juda muhim rol o'ynaydi. U so'zning ma'nosini kontekstga nisbatan saqlaydi.

Noaniqlik (aniqmaslik). Agar bir so'z, ibora yoki jumla bir nechta ma'noga ega bo'lsa, u noaniqlik deyiladi. Quyida turli noaniqlik turlari keltirilgan:



5-rasm: Noaniqliklarning turlari

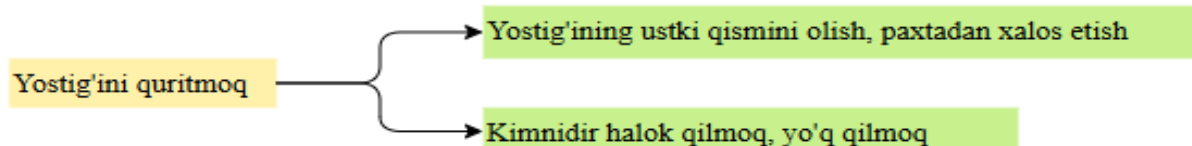
Leksik noaniqlik. Bir nechta ma'noga ega bo'lgan so'zlar omonim yoki polisemantik bo'lishi mumkin. Agar bir so'zning ikki ma'nosi butunlay farqli bo'lsa, ammo yozilishi bir xil bo'lsa, ular omonimlar deb ataladi. Masalan, "ari" so'zining ikki ma'nosi quyidagicha ishlatilishi mumkin:



2-rasm: Ari so'zining ma'nolari

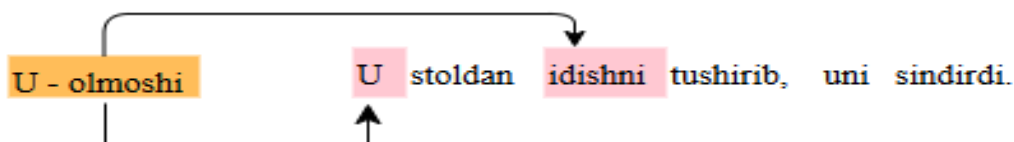
Bir so'zning ma'nosi bir xil asosiy ma'noning turli xil tuslarini ifodalasa, ya'ni u turli lekin bir-biriga bog'liq ma'nolarga ega bo'lsa, u polisemantik deb ataladi. Masalan, “moya” so'zi polisemantik hisoblanadi, chunki uning ikki ma'nosi – urg'ochi tuya va hamirturish, achitqi – bir-biriga aloqador bo'lgan ma'nolardir.

Struktural (Sintaktik) noaniqlik. Agar noaniqlik jumla yoki ibora ko'rinishida bo'lsa, bu struktural noaniqlik deb ataladi. Bu holat, biror ibora yoki jumla bir nechta asosiy strukturalarga ega bo'lganda yuzaga keladi. Masalan:



3-rasm: Struktural noaniqlikka misol

Referensial noaniqlik. Agar gapdagi olmoshlar ma'lum bir so'zlarga ishora qilsa, lekin qaysi so'zga ishora qilayotganini aniqlash qiyin bo'lsa, bu referensial noaniqlik deb ataladi. Misol uchun:



4-rasm: Referensial noaniqlikka misol

Bu yerda noaniqlik shundaki, “u” olmoshi qaysi otga ishora qilmoqda: *stolga* yoki *idishga*.

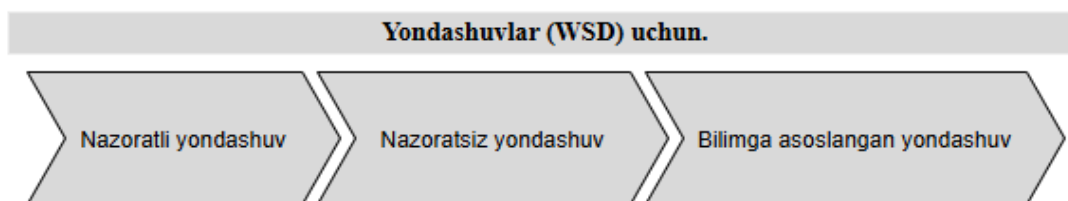
Bunday ishoralar **anforik ishoralar** yoki **anafora** deb ataladi.

Ko'p so'zli konstruktsiyalar: Ko'p so'zli konstruktsiyalar, masalan, idiomalar va frazali fe'llar tufayli yuzaga kelgan noaniqlik. Misol uchun:

t/r	Matn	Ma'nosi
1	Qishloq bug'doy ekdi.	Qishloq – butun odamlar.
2	Universitet bugun e'lon chiqardi.	Universitet – rahbariyat yoki matbuot xizmati.

5-rasm: Ko'p so'zli konstruktsiyalarga misol

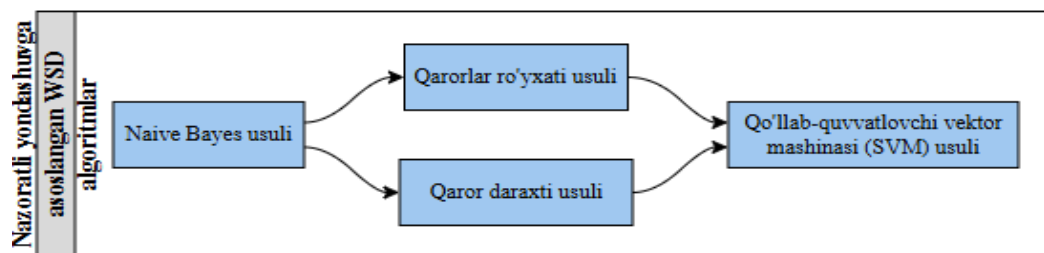
Idiomning haqiqiy ma'nosini uning tarkibiy so'zlaridan aniqlash qiyin. Bu noaniqliklar semantik tahlilning asosiy masalasidir. Jahon tajribasiga ko'ra noaniqliklarni yechish so'z ma'nosini aniqlash (WSD- Word Sense Disambiguation) deb nomlanadi. Noaniqliklarni aniqlashda turli yondashuvlardan foyalanish mumkin. **WSD uchun yondashuvlar.** So'z ma'nosini aniqlash (WSD) yondashuvlari uchta asosiy kategoriya bo'yicha tasniflanadi:



6-rasm: WSD yondashuvlar turlari

Nazoratli yondashuv. Nazoratli yondashuv semantic teglangan ma'lumotlarga yoki korpusga asoslanadi. Nazoratli yondashuv, tasniflagichlarni qurish uchun o'qitilgan ma'no teglangan korpusga asoslanadi. Dastlab, tasniflagichni qurish uchun teglangan korpus talab qilinadi. Tasniflagich ko'p ma'noli so'zning ma'nosini kontekstga asoslanib aniqlash uchun ishlatiladi. Ushbu yondashuvning asosiy kamchiligi — katta hajmdagi semantic teglangan korpusni talab qilishi.

Tasniflagichning aniqligi korpusga kiritilgan ma'lumotlar hajmi va xilma-xilligiga bog'liq. Korpus kattaroq bo'lsa, natijalar yaxshiroq bo'ladi. Korpus yaratish o'zi vaqt va mablag' jihatidan muammo bo'lishi mumkin. Nazoratli yondashuvga asoslangan ba'zi muhim WSD algoritmlari quyidagilardir:



7-rasm: Nazoratli yondashuvga asoslangan WSD algoritmlar

Naive Bayes usuli – omonimlarni farqlashda va guruhlariga ajratishda foydaniladi.

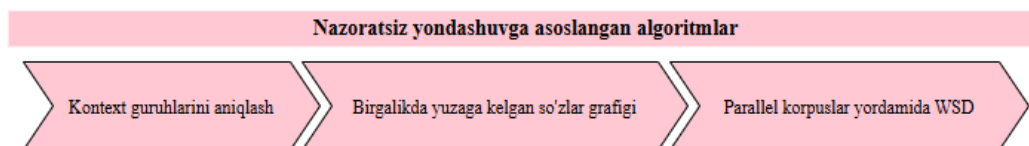
Qarorlar ro'yxati usuli – so'zning turli ma'nolarda kelgan matnlarni korpusga to'playdi va har bir misolni tegishli ma'nosini belgilab oladi. Lekin katta hajmdagi ma'lumotlar va matn murakkabligida ishlamaydi.

Qarorlar daraxti usuli – bu usul so'zning ma'nosini aniqlash uchun daraxt shaklida tuzilgan bo'lib, unda har bir tugun matnni tekshiradi, shoxlar natija va burglar esa so'zning ma'nosini bildiradi.

Qo'llab-quvvatlovchi vector mashinasi (SVM) usuli – so'zlarning ma'nolarini o'rganadi, SVM ma'lumotlarini aniqlash uchun chegara (gipertekislik) hosil qiladi.

Nazoratsiz yondashuv. Nazoratsiz yondashuv teglanmagan korpusga asoslanadi. Bu yondashuv so'zlarni guruhlashga asoslanadi. U noaniq so'zning ma'nosi qo'shni so'zlardan hosil bo'lishi mumkin deb hisoblaydi. Ushbu yondashuv

annotatsiya qilinmagan ma'lumotlarga asoslangani uchun, o'qitish korpusida so'zning ko'plab ma'nolari mavjud bo'lishi juda muhimdir. Ushbu yondashuvga asoslangan ba'zi algoritmlarni keltiramiz:



8-rasm: Nazoratsiz yondashuvga asoslangan agoritmlar

Kontext guruhlarini aniqlash - yangi tillarni o'rganadi, so'zlarni qatnashgan konteksti bo'yicha guruhlarga ajratadi. Katta hajmdagi matnlarni avtomatik tahlil qilishda aniq tahlil qiladi.

Birgalikda uchragan so'zlar grafigi - so'zlarning bir-biri bilan qanday bog'liqligini grafik shaklda ko'rsatadi. Matnning muhim so'zlarini ajratadi.

Parallel korpuslar yordamida WSD - ikki yoki undan ortiq tillarni tarjima qilib, so'zlarni matnga qarab ma'nolarini aniqlaydi.

Bilinga asoslangan yondashuv. Ushbu yondashuv so'z ma'nosini aniqlash uchun bilim bazalarini ishlatishga asoslanadi. Foydalanilishi mumkin bo'lgan bilim bazalari quyidagilar: WordNet, mashinaga o'qitish mumkin bo'lgan lug'atlar, tezauruslar va hokazo. Ushbu yondashuv, ayniqsa, ma'lum bir soha uchun so'z ma'nosini aniqlashda samarali bo'ladi. Bilinga asoslangan yondashuvda foydalaniladigan mashhur algoritmlardan biri 1986-yilda Maykl Lesk tomonidan taklif qilingan va **Lesk algoritmi** deb nomlangan algoritmdir.

Lesk Algoritmi Maykl Lesk tomonidan 1986-yilda ishlab chiqilgan [5]. Bu polisemantik so'zlarning ma'nolarini aniqlashda ishlatiladi. U Oxford Advanced Learner's Dictionary of Current English (OALD) lug'atidagi so'z ta'riflarining o'z va ko'chma ma'nolarini ajratib olgan. Hosil bo'lgan kontekst yordamida so'z ma'nolarini aniqlagan. O'zbek tilidagi omonimiyani aniqlashda ham ushbu algoritmda foydalanilgan[15]. **Banerjee va Pedersen** WordNetni bilim manbasini Lesk algoritmi uchun dataset sifatida ishlatganlar. Ishlab chiqilgan axborot tizimining umumiy aniqligi 32% deb baholangan [7].

Sinha M. va boshqalar tomonidan hind tili matnlari uchun avtomatik WSD tizimini ishlab chiqilgan. Ushbu tizimni ishlab chiqishda **Hindi WordNet**dan foydalanilgan [8]. Olimlar so'z ma'nolarini aniqlashda statistik usullardan foydalanganlar. Ishlab chiqilgan axborot tizim matn tarkibidagi faqat ot so'z turkumidagi noaniqlikdan bartaraf qiladi. Ular polisemantik so'z joylashgan kontekstni WordNetdan olingan kontekstlar bilan solishtirganlar. Ular tizimni Hindistonning Markaziy Til institutidan (CIIL) taqdim etilgan hind korpuslari yordamida baholashgan. Ular algoritmnining aniqligini 40% dan 70% gacha bo'lgan oraliqda aniqlagan.

Shrestha N. va boshqalar[9] Lesk algoritmi asosida Nepali noaniq soʻzlarini noaniqlikdan bartaraf qilish uchun WSD moduli ishlab chiqishgan. Ular bilim resursi sifatida *synset*, *gloss*, *misollar* va *hypernym*dan foydalanib, soʻzlarni toʻplashgan. Har bir maʼno uchun maqsadli soʻzning maʼlumotlar bazasida faqat bitta misol mavjud edi. Har bir matndagi soʻzi, tuzilgan maʼlumotlar bazasida mavjud boʻlgan soʻzlar toʻplamidagi har bir soʻz bilan taqqoslanadi.

Dhungana va Shakya [12] Lesk algoritmi yordamida Nepali tilidagi polisemantik soʻzlarni noaniqlikdan xoli qilish ustida ishlashgan. 348 ta soʻz (shu jumladan, 59 ta polisemantik soʻzni va matnning soʻzlarini oʻz ichiga olgan) boʻyicha bajarilgan tajribalar va 201 ta Nepali jumlasini oʻz ichiga olgan test maʼlumotlari ularning tizimining aniqligini 88,05% deb koʻrsatgan.

Richard Singh va K. Ghosh 2013-yilda Manipuri tili uchun arxitektura taklif qilishgan. Tizim WSDni ikki bosqichda amalga oshiradi: oʻqitish bosqichi va test bosqichi. Manipuri tilidagi soʻz maʼnosini aniqlash tizimi quyidagi bosqichlardan iborat: (I) oldindan qayta ishlash, (II) xususiyatlarni tanlash va yaratish, (III) oʻqitish, (IV) test qilish va (V) ishlash samaradorligini baholash [14].

Haroon, R.P. Malayalam tilida avtomatik WSD ning birinchi urinishini taqdim etgan. U bilimga asoslangan yondashuvni ishlatgan [10]. **Rakesh va Ravinder** Punjabi matn hujjatidan noaniqlikni olib tashlash uchun WSD algoritmini taklif qilishgan. Mualliflar WSD uchun *Modified Lesk Algorithm*ni ishlatishgan [16].

Ayan Das va Sudeshna Sarkar Bengali tili uchun nazoratsiz grafikka asoslangan guruhlash yondashuviga asoslangan WSD tizimini ishlab chiqqanlar, bu tizim Bengali-Hindi mashina tarjimasini uchun qoʻllaniladi [10].

Xulosa. Soʻz maʼnosini aniqlash (WSD) hisoblash lingvistikasi sohasida juda faol tadqiqot mavzusi boʻlib kelmoqda. U tizimlar tomonidan yaratilgan natijalarni yanada aniqroq qilishda muhim rol oʻynaydi. WSD tadqiqotlarini cheklagan omillardan biri — lingvistik resurslarning mavjud emasligidir. Hozirgi kunda koʻpgina tadqiqotchilar asosan *Lesk algoritmi*ga asoslangan maʼno aniqlashning bilimga asoslangan yondashuvini ishlatishmoqda. Korpusning yetishmasligi sababli, tadqiqotchilar noaniqlikni bartaraf etish uchun *WordNet*dan foydalanmoqdalar. Shuningdek, bilimga asoslangan yondashuvlarning noaniq soʻzlarning maʼnolarini hal qilishda samarali ekanligi oʻrganilgan, ammo tabiiy tilni qayta ishlash vositalarining samaradorligini oshirish uchun leksik resurslarni ishlab chiqish ehtiyoji katta.

Foydalanilgan adabiyotlar:

1. Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, “IndoWordNet va Word Sense Disambiguation uchun Asboblari

Tanishtiruv”,

http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/18_Tools%20for%20IndoWordNet%20&%20WSD.pdf

2. C. Fellbaum, “WordNet: An Electronic Lexical Database”, MIT Press, 1998. Lokesh Nandanwar, Kalyani Mamulkar, “Supervised, Semi-Supervised and Unsupervised WSD Approaches: An Overview”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

3. Kuhoo Gupta, Manish Shrivastava, Smriti Singh va Pushpak Bhattacharyya, “Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi”, COLING/ACL-2006, Sidney, Avstraliya, iyul, 2006.

<https://pdfs.semanticscholar.org/1daa/370563deb1e9f5772d6ada1edfb4ec4cd7b4.pdf>

4. M. Lesk “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone”. In Proceedings of the 5th annual international conference on Systems documentation, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

5. Udaya Raj va boshqalar, “Word Sense Disambiguation using WSD specific Wordnet of polysemy Words”,
<https://arxiv.org/ftp/arxiv/papers/1409/1409.3512.pdf>

6. Banerjee va Pederson, “An adapted Lesk algorithm for Word Sense Disambiguity using Word Net”,
<http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.pdf>

7. Sinha M. va boshqalar, “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”,
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.3892&rep=rep1&type=pdf>

8. Shrestha N va boshqalar, “Resources for Nepali WSD” in proceedings of IEEE conference on Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. 2008.

9. Alok Ranjan Pal va Diganta Saha, “WORD SENSE DISAMBIGUATION: A SURVEY”, International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, iyul 2015

10. Z.Z. Merhbene Laroussi, “Unsupervised system for Lexical Disambiguation of Arabic Language using a vote procedure,” IEEE-ICoAC, 2011

11. Raj Dhungana, Subarna Shakya va boshqalar, “Word Sense Disambiguation Using WSD Specific Wordnet of Polysemy Words”, Udaya International Journal on Natural Language Computing (IJNLC) Vol. 3, No.4, avgust 2014



12. Shallu va boshqalar, “A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 5, NO. 4, Noyabr 2013
13. E. B. Boltayevich, A. X. Ilxomovna, P. M. Hakim Qizi and K. N. Uktambay O'g'li, “Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm,” 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 137-140, doi: 10.1109/UBMK59864.2023.10286666.