BIR SO‘Z TURKUMIDAGI OMONIM SO‘ZLARNI ANIQLASH USULLARI

Malikov Elbek
Kompyuter lingvistikasi yo’nalishi magistaranti
Toshkent davlat o’zbek tili va adabiyoti universiteti.

Annotatsiya: Tabiiy tilga ishlov berish sohasidagi muhim vazifalardan biri bu omonimiya masalasidir. Bu masala turli tabiiy tillarda mavjud bo‘lib, har bir tabiiy tilda turlicha usullar taklif qilinib kelinmoqda. Omonim so‘zlarni semantik farqlashda turli usullardan foydalanish mumkin. O‘zbek tilidagi omonim so‘zlarni semantic farqlashda ularni turli so‘z turkumlari va bir so‘z turkumi doirasida uchrashiga ko‘ra turli usullardan foydalanish mumkin. Ushbu maqolada aynan bir so‘z turkumi doirasidagi omonim so‘zlarni semantic farqlash usullari haqida so‘z olib boriladi. Bir so‘z turkumi doirasidagi omonimiyani aniqlashda Mashinali o‘qitish yondashuvining usullaridan foydalanishning jahon tajribalari o‘rganildi.

Kalit so‘zlar: tabiiy tilga ishlov berish (NLP natural language processing), omonimiya, Mashinali o‘qitish yondoshuvi, kontekst, so‘z ma’nosini aniqlash (WSD-word sense disambiguation)

Annotation: One of the important tasks in the field of natural language processing is the issue of homonymy. This problem exists in different natural languages, and different methods are proposed in each natural language. Different methods can be used for semantic differentiation of homonyms. In the semantic differentiation of homonymous words in the Uzbek language, different methods can be used depending on whether they meet in different word groups and within the same word group. This article discusses the methods of semantic differentiation of homonyms within the same word group. The world experiences of using the methods of the Machine Learning approach in determining homonymy within a word group were studied.

Keywords: NLP (natural language processing), homonymy, Machine learning approach, context, word sense disambiguation (WSD-word sense disambiguation)

Аннотация: Омонимия имеет важное значение в обработке естественного языка. Этот вопрос существует в разных естественных языках. Для семантической дифференциации омонимов можно использовать различные способы. При семантической дифференциации омонимов в узбекском языке могут применяться различные методы в зависимости от того, встречаются ли они в разных частях речи и в рамках одной части речи. В данной статье рассмотрены способы семантической дифференциации омонимов внутри одной части речи. Изучен мировой опыт использования методов подхода машинного обучения при определении омонимии внутри одной части речи.

Ключевые слова: обработка естественного языка (NLP), омонимия, подход машинного обучения, контекст, определение смысловой неоднозначности

Tabiiy tilda o‘qilishi bir xil, lekin turli kontesktlarda ma’nosi har xil bo‘lgan so‘zlar uchraydi. Bu so‘zlar omonim so‘zlar hisoblanadi. Inson omonim so‘zni oson farqlab, ma’nosini tushunadi. Lekin bu mashina uchun qiyin vazifa, uzoq yillardan buyon odamlarga o‘xshab

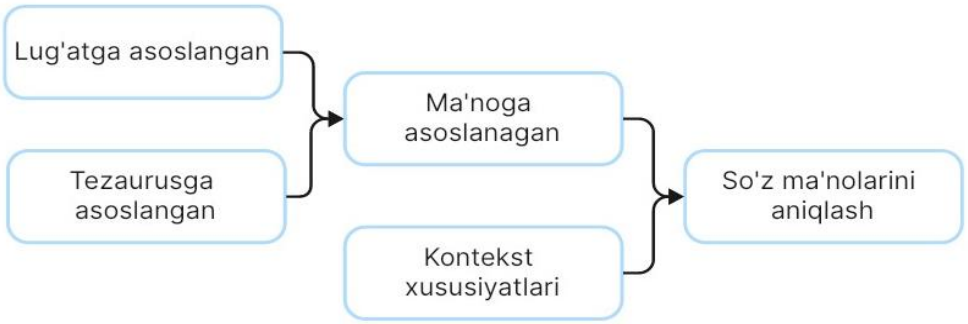
so‘zning ma’nosini avtomatik aniqlaydigan tizim ishlab chiqish masalasi o‘zining dolzarbligini saqlab kelmoqda.

So‘z manolarini aniqlash usullari omonim so‘zning to‘g‘ri manosini topish va uni to‘g‘ri o‘ringa avtomatik joylashtirishdir vazifasidir. So‘z manolarini aniqlash tabiiy tilni qayta ishlash sohasining muhim, murakkab masalasidir. Bu mashina tarjimasi, semantik xaritalash, semantik izohlash va ontologiyani o‘rganish kabi ko‘plab real dunyo ilovalari uchun zarur. Shuningdek, u ko‘plab axborot qidirish, ma'lumot olish va nutqni aniqlash kabi ilovalarning ish faoliyatini yaxshilashda foydali bo‘ladi. Ingliz [8], hind [9], fransuz [10], ispan [11], xitoy [12] va boshqa ko‘plab tabiiy tillar omonim so‘zlarga ega tillar hisoblanadi. O‘zbek tilidagi omonim so‘zlarni semantik farqlashda qator ilmiy izlanishlar olib borilmoqda. Omonimiyani aniqlashda ularni ikkita: *bir so‘z turkumi va turli so‘z turkumlari* doirasidagi omonim so‘zlarga ajratilgan holda aniqlash bo‘yicha ilmiy izlanishlar mavjud. Jumladan tadqiqotchi X.Axmedova hamda texnika fanlari falsafa doktori B.B.Elov tomonidan turli so‘z turkumlari doirasidagi omonim so‘zlarni so‘z turkumlari doirasida uchrashiga ko‘ra *ikki so‘z turkumi, uch so‘z turkumi va to‘rtta so‘z turkumi doirasidagi omonimlar* kabi guruhlarga ajratgan holda semantik farqlash bo‘yicha amaliy natijalarga erishgan [1], [2], [3], [4]. Turli so‘z turkumlari orasidagi omonimiyani aniqlash masalasi jumlalarni POS (Part of speechs tagging) teglash jarayonida o‘z yechimini topadi. So‘z turkumlarini aniqlashda Yashirin Markov modelidan foydalanish bo‘yicha ko‘plab ilmiy maqolalarni ko‘rish mumkin [5]. O‘zbek tilida shunday omonim so‘zlar borki, turli so‘z turkumlari ichida ham bir so‘z tukumiga oid bir nechta ma’nolarni anglatishi mumkin.

1- Jadval: Bir so‘z turkumi doirasidagi omonim so‘zlar

So‘z	**So‘z turkumi**	**Ma’nosi**
O‘t	Fe’l	O‘tmoq fe’li
	O‘t	Maysa, o‘t-o‘lan
	O‘t	Inson o‘rgani
	O‘t	Olov
Oz	Ravish	Kam, miqdori nisbatan ko‘p bo‘lmagan.
	Fe'l	Oriqlamoq, etidan yo‘qotmoq.
	Fe'l	Noxush bo‘lmoq, kuchsizlanmoq, holsizlanmoq.
	Fe'l	Adashmoq, to‘g‘ri yo‘ldan chetga chiqmoq
Bo‘y	Ot	Uzunlik o‘lchovi.
	Ot	Hid, is
…	…	…

1-jadvada keltirilgan *o‘t* omonim so‘zi turli so‘z turkumlari doirasida o‘ziga xos ma’nolarni anglatadi. Shuni alohida ta’kidlash kerakki, shu so‘z turkumlarining biri doirasida bir qancha ma’nolarni anglatuvchi so‘zlar ham talaygina. Bunday omonim so‘zlarni POS tegging masalasini yechish bilan aniqlab bo‘lmaydi. Buning uchun alohida usullar, modellar zatur. Faraz qilaylik, T matn W_1, W_2, W_3, ..., W_n so‘zlar ketma -ketligidan iborat. So‘z manolarini aniqlash usullari matndagi barcha omonim so‘zlar uchun to‘g‘ri ma'noni belgilash vazifasidir. So‘z manolarini aniqlash usullarining nazariy modeli 1-rasmda ko‘rsatilgan.



1-rasm. So‘z manolarini aniqlash usullarining nazariy modeli

So‘z manosini aniqlashda ishlatiladigan ikkita asosiy yondashuv - bu *Lug‘atga asoslangan yondashuv* va *Mashinali o‘qitishga asoslangan yondashuv*dir.
Ushbu yondashuvlar ham o‘z navbatida bir qator usullarni qamrab oladi.

Lug‘atga asoslangan yondashuvda target so‘zning barcha ma'nolari lug‘atdan olinadi. Keyin bu ma’nolar kontekstda qolgan barcha so‘zlarning lug‘at ta'riflari bilan taqqoslanadi. Biz So‘z manolarini aniqlash usullari yondashuvlarini so‘z ma'nosini ajratish uchun chuqur va sayoz yondashuv sifatida tasniflashimiz mumkin. Bilimga asoslangan yondashuv kompyuter o‘qishi mumkin bo‘lgan lug‘atlar, masalan, korpus WorldNet asosida ishlaydi. So‘z manosini aniqlash uchun grammatik qoidalardan ham foydalanishi mumkin.Bilimga asoslangan yondashuvning maqsadi (lug‘atga asoslangan yondashuv) kontekstdagi so‘zlarning ma'nosini aniqlash uchun bilim resurslaridan foydalanishdir. Bilim resurslari lug‘atlar, tezauriyalar, ontologiyalar, birikmalar va boshqalardir. Yuqoridagi usullar boshqariladigan muqobil usullarga qaraganda unumdorligi pastroq, lekin ular kengroq diapazondagi afzalliklarga ega**.**

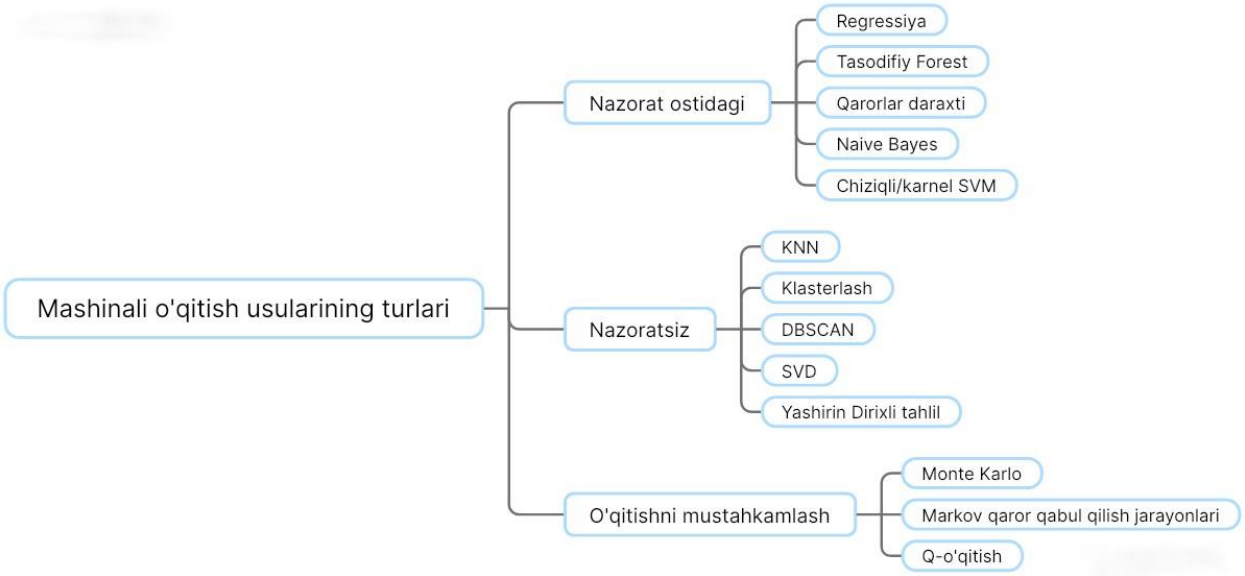
Bir-biriga o‘xshash yondashuvlar: Ushbu yondashuv mashinada o‘qiladigan lug‘at (MDR) qoidalarini talab qiladi. U kontekstdagi so‘zlarning xususiyatlari bilan bir qatorda noaniq bo‘lgan so‘zlarning ma’nolarining turli xususiyatlarini aniqlashni o‘z ichiga oladi.

Lesk algoritmi: Bir-biriga o‘xshash yondashuvga asoslangan Lesk algoritmini W so‘z ma'noni anglatuvchi so‘z, C - atrofdagi kontekstdagi so‘zlar to‘plami, S - W uchun ma’noni farqlaydi[7].

Tanlash afzalliklari: Tanlov afzalliklari so‘z turlarining mumkin bo‘lgan munosabatlari haqida ma'lumot topadi va bilim manbasidan foydalangan holda umumiy fikrni bildiradi. Masalan Son-yuz, inson a’zosi-yuz so‘zlari semantik munosabatga ega. Bunday yondashuvda noto‘g‘ri so‘z ma'nolari chiqarib tashlanadi va faqat umumiy ma'no qoidalariga mos keladigan hislar tanlanadi. Ushbu yondashuvning asosiy g‘oyasi sintaktik munosabatga ega bo‘lgan korpusda bunday so‘z juftligi necha marta sodir bo‘lishini hisoblashdir.

Mashinali o‘qitishga asoslangan yondashuvda tizimlar So‘z manolarini aniqlash usullari vazifasini bajarish uchun o‘qitiladi. Klassifikator kiruvchi gap tarkibidagi omonim so‘z(so‘zlar)ning joriy ma’nosi va xususiyatlarni aniqlash uchun foydalaniladi. Ushbu yondashuvda tizimlar *so‘z manolarini aniqlash usullari* vazifasini bajarish uchun o‘qitiladi. Klassifikatorla kiruvchi ma’lumotlarni ma’nosi va xususiyatlarini aniqlash uchun foydalaniladi. Xususiyatning qiymati so‘zning maqsadli so‘z atrofida necha marta sodir bo‘lishidir. Kontekst markazi sifatida tanlangan so‘z bilan doim birga paydo bo‘ladi. Mashinali o‘qitishga asoslangan

yondashuv usullari uch turga bo‘linadi: *nazorat qilinadigan usullar, nazoratsiz usullar va yarim nazorat qilinadigan usullar.*



2-rasm: Mashinali o‘qitish yondashuvi usullarining iyerarxiyasi

Nazorat ostidagi usullar. Bu usullarning mohiyati shundan iboratki, unda inson omili bilan semantik teglangan ma'lumotlar to‘plami asosida joriy so‘zning ma’nosini aniqlanadi.

Tasniflagichni o‘rganish uchun foydalaniladigan o‘quv majmuasi odatda ma'lum bir maqsadli so‘z va unga mos yozuvlar lug‘atining ma'no inventarizatsiyasidan olingan ma'no bilan qo‘lda teglangan misollar to‘plamini o‘z ichiga oladi. Asosan, So‘z manolarini aniqlash algoritmlari boshqa yondashuvlarga qaraganada yaxshiroq natija beradi. Quyida ushbu yondashuv usullarini ko‘rib chiqamiz.

Qaror ro‘yxatlari: ushbu usulning mohiyati shundan iboratki, test misollarini toifalarga ajratish uchun tartiblangan qoidalar to‘plamidan foydalaniladi. Buni *vaznli* [if-then-else] *qoidalari* ro‘yxati sifatida ko‘rish mumkin. O‘quv majmuasi xususiyatlar to‘plamini induktsiya qilish uchun ishlatiladi. Har qanday so‘z ko‘rib chiqilayotganda, birinchi navbatda uning paydo bo‘lishi hisoblab chiqiladi va qarorlar ro‘yxatini yaratish uchun xususiyatlar vektori nuqtai nazaridan ifodalanadi, ball hisoblanadi. Vektor uchun maksimal ball ma'noni ifodalaydi.

Qaror daraxti: Qaror daraxti ma'lumotlarini rekursiv tarzda ajratadi va daraxt tuzilishidagi tasniflash qoidalarini ifodalaydi. Ichki tomirlar xususiyatlar bo‘yicha sinovni ifodalaydi va har bir shox qaror qanday qabul qilinayotganini ko‘rsatadi va barg tomirlari natija yoki bashoratga ishora qiladi.

Naïve Bayes: Naive Bayes klassifikatori Bayes teoremasini qo‘llashga asoslangan oddiy ehtimolli klassifikatordir. U kontekstdagi fj xususiyatlarini hisobga olgan holda w so‘zning har bir Si ma'nosining shartli ehtimolini hisoblashga tayanadi. Quyidagi formulani maksimal darajaga keltiradigan S ma'nosi kontekstda eng mos ma'no sifatida tanlanadi.

$$\hat{S} = \underbrace{argmax}_{S_i \in manolar_D(w)} P(S_i|f_1, f_2, \dots, f_m) = \underbrace{argmax}_{S_i \in manolar_D(w)} \frac{P(f_1, \dots, f_m|S_i)\,P(S_i)}{P(f_1,\dots,f_m)} =$$

$$= \underbrace{argmax}_{S_i \in manolar_D(w)} = P(S_i) \neq P(f_j|S_i) \qquad (1)$$

Naïve Bayes klassifikatoridan foydalanishdan oldin target so‘zning farqlanuvchi xususiyatlari aniqlab olinadi. O‘zbek tilidagi turli so‘z turkumlari orasidagi omonimiyani aniqlashda ushbu usuldan foydalanilganiga guvoh bo‘ldik [6].
Xuddi shunday, boshqa usullarni ham tasniflash mumkin.

Nazoratsiz usul. Nazoratsiz usulning nazorat qilinadigan usuldan farqi shundaki, unda inson omili yordamida aniqlangan ma’lumotlar to‘plami kerak emas. Bu omonim so‘zlar o‘xshash kontekstlarda kelishiga asoslanadi. So‘z ma'nolari so‘zlarning voqelik klasterlarini shakllantirish orqali hosil bo‘ladi.

Kontekst klasterlash: Bu usulda birinchi kontekst vektorlari yaratiladigan, so‘ngra so‘zning ma'nosini aniqlash uchun klasterlarga guruhlanadigan klasterlash usullariga asoslanadi. Bu usuldan foydalanilganda so‘z fazosi sifatidagi vektor maydoni va uning o‘lchamlari faqat so‘zlardir. Shuningdek, bu usulda korpusda bo‘lgan so‘z vektor sifatida belgilanadi va uning kontekstda necha marta sodir bo‘lganligi hisoblanadi. Keyin birgalikda sodir bo‘lish matritsasi yaratiladi va o‘xshashlik o‘lchovlari qo‘llaniladi. Keyin farqlash har qanday klasterlash texnikasi yordamida amalga oshiriladi.

So‘z klasterlash: Ushbu usulda omonim so‘zlar bir xil klasterga biriktiriladi. So‘zlar orasidagi o‘xshashlik sintaktik bog‘liqlik orqali beriladi. Agar W w_m ga o‘xshash so‘zlardan iborat bo‘lsa, daraxt dastlab faqat bitta w_m shox bilan hosil bo‘ladi va w_1, …, $w_m$ga eng o‘xshash ma'noli so‘z ekanligi aniqlanganda w_1 shoxda w_m tugunga ega bo‘ladi. Guruh algoritmi bo‘yicha klasterlash deb ataladigan yana bir yondashuv har bir so‘zni xususiyat vektori sifatida ifodalaydi. Belgilangan so‘zlarga duch kelganda, har bir elementi w_m va w_n ikkita so‘z o‘rtasidagi o‘xshashlik bo‘lgan S_{mn}o‘xshashlik matritsasi deb ataladigan matritsa tuziladi. Ushbu algoritmning keyingi bosqichida W so‘zlar to‘plami uchun rekursiv usulda guruhlar tuziladi. Keyin klasterlash algoritmi biron bir guruhning so‘zlariga o‘xshamaydigan so‘zlarni topishga harakat qiladi. Hech qanday guruh tarkibiga kirmaydigan so‘zlar yana qo‘shimcha guruhlar tuzish uchun ishlatiladi.Yakuniy bosqichda W ga tegishli har bir belgilangan so‘z guruhning markaziy qismiga o‘xshashligiga qarab guruh a'zosi bo‘ladi.

Birgalikda sodir bo‘lgan grafiklar: Bu usulda V cho‘qqisi va E cheti bo‘lgan grafik, bu yerda V –matndagi so‘zlarni ifodalaydi va agar so‘zlar bir xil paragraf yoki matndagi sintaksisga muvofiq munosabatda birga bo‘lsa, E qo‘shiladi. Berilgan maqsadli so‘z uchun birinchi navbatda grafik tuziladi va grafik uchun qo‘shnilik matritsasi yaratiladi. Shundan so‘ng so‘zning ma'nosini topish uchun Markov klasterlash usuli qo‘llaniladi. Grafikning har bir chetiga so‘zlarning birgalikdagi chastotasi bo‘lgan og‘irlik beriladi. {m,n} chetining og‘irligi quyidagi formula bo‘yicha aniqlanadi:

$$w_{mn} = 1 - \max\{P(w_m|w_n), P(w_n|w_m)\} \qquad (2)$$

Bu yerda, $P(w_m|w_n)$–w_m va w_n so‘zlarning ketma-ket uchrash ehtimolligi. Yuqori chastotali so‘zlarga og‘irlik 0, kamdan-kam uchraydigan so‘zlarga esa 1 og‘irlik beriladi. Og‘irligi

ma'lum chegaradan oshib ketadigan chekkalar o‘tkazib yuboriladi. Keyin grafikga iterativ algoritm qo‘llaniladi va eng yuqori nisbiy darajaga ega bo‘lgan tugun markaz sifatida tanlanadi. Algoritm so‘zning o‘z markaziga chastotasi chegaradan pastroqqa yetganda tugaydi. Nihoyat, butun markaz berilgan maqsadli so‘zning ma'nosi sifatida belgilanadi. Nol og‘irlikka ega bo‘lgan maqsadli so‘zning markazlari bog‘lanadi va grafikdan minimal chegara daraxti yaratiladi. Ushbu kengaytmali daraxt maqsadli so‘zning haqiqiy ma'nosini aniqlash uchun ishlatiladi.

Yarim nazorat ostidagi usullar: Yarim nazorat ostidagi o‘qitish usullarida ma'lumotlar nazorat ostidagi kabi mavjud, ammo kamroq ma'lumot berilishi mumkin. Bu yerda aniq ma'lumot emas, faqat kritik ma'lumotlar mavjud. Misol uchun, tizim faqat maqsadli mahsulotning ma'lum bir qismi to‘g‘ri ekanligini va shunga o‘xshashligini aytishi mumkin. Yarim nazorat qilinadigan yoki minimal nazorat qilinadigan usullar kichik miqdordagi izohli ma'lumotnomalar bilan ishlash qobiliyati tufayli mashhurlikka erishmoqda va ko‘pincha katta ma'lumotlar to‘plamlarida mutlaqo nazoratsiz usullardan ustun turadi. Yordamchi ma'lumotlardan muhim xususiyatlarni o‘rganadigan va olingan ma'lumotlardan foydalangan holda ma'lumotlarni guruhlaydigan imkoniyatlar mavjud.

Xulosa.

Ushbu maqolada So‘z manolarini aniqlash usullari uchun qo‘llaniladigan turli yondashuvlarni umumlashtirildi va mavjud So‘z manolarini aniqlash usullari algoritmlarini ularning texnikasiga ko‘ra tasniflandi. Ushbu maqolada biz so‘z ma'nosini ajratib ko‘rsatishda mavjud bo‘lgan turli yondashuvlarni taqqoslash bo‘yicha so‘rovni taqdim etildi, bunda birinchi navbatda Mashinali o‘qitish yondashuvlari va lug‘atga asoslangan bilimlarga asoslangan yondashuvlarga e'tibor qaratildi. Biz nazorat ostidagi yondashuv yaxshiroq ishlaydi degan xulosaga keldik, ammo uning kamchiliklaridan biri bu katta korpusga bo‘lgan talab bo‘lib, ularsiz o‘qitish mumkin emas, uni nazoratsiz yondashuvda yengib o‘tish mumkin, chunki u noaniqlik uchun bunday keng ko‘lamli manbaga tayanmaydi. Boshqa tomondan, bilimga asoslangan yondashuv ma'lum bir kontekstda so‘zlarning ma'nosini aniqlash uchun ma'lumot manbalaridan foydalanadi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Axmedova X.I. Uchta so‘z turkumi doirasidagi omonimiyani farqlovchi biznes jarayonni modellashtirish// O‘zbekiston respublikasi innovatsion rivojlanish vazirligining, Ilm-fan va innovasion rivojlanish ilmiy jurnal 2022 / 1, 150-162-b.
2. Axmedova X. I. Turli so‘z turkumlari orasidagi omonimiyani aniqlovchi matematik modellar// Science and innovation international scientific journal volume 1 issue 7 uif-2022: 8.2 | issn: 2181-3337. https://doi.org/10.5281/zenodo.7238546
3. Axmedova X.I. Chastotali usul yordamida omonimiyani aniqlash// “O‘ZBEK AMALIY FILOLOGIYASI ISTIQBOLLARI” Respublika ilmiy-amaliy konferensiyasi Toshkent: 2022.-164-170 b.
4. Elov B.B., Axmedova X.I. Determining homonymy using statistical methods. // “Hisoblash modellari va texnologiyalari (HMT 2022)” O‘zbekiston-Malayziya ikkinchi xalqaro konferensiyasi materiallari- Toshkent, 2022 16-17 sentabr,-106 b.
5. Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N., Yodgorov U., Yuldashev A. “POS TAGGING OF UZBEK TEXTS USING HIDDEN MARKOV MODELS (HMM) AND VITERBI ALGORITHM”. “O‘zbekiston Milliy universitetining ilm-fan rivoji va

jamiyat taraqqiyotida tutgan o‘rni” mavzusidagi xalqaro ilmiy-amaliy konferensiya, 2023 yil, 12 may, 104-115-b.

6. Elov B.B., Axmedova X.I. Homonymy detection using a Naïve Bayes classifier. Journal of Computer Science Engineering and Information Technology Research (JCSEITR) ISSN(P): 2250-2416; ISSN(E): Applied Vol. 13, Issue 1, March 2023, 5–20 © TJPRC Pvt. Ltd, Hindiston
7. Elov B.B., Axmedova X.I. Primova M.H., Khudayberganov N.U. Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm. *UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering*, 2023, страницы 137–140. https://doi.org/10.1109/UBMK59864.2023.10286666
8. Eneko Agirre and Philip Edmonds "Word Sense Disambiguation: Algorithms and Applications" Springer, 2007, ISBN: 978-1-84628-855-2
9. Pushpak Bhattacharyya and Josef van Genabith, "Word Sense Disambiguation in Indian Languages", 2017, Springer, ISBN: 978-981-10-4055-4
10. "La désambiguïsation lexicale et ses applications" (Lexical Disambiguation and Its Applications) by Isabelle Tellier.
11. M. Victoria Marrero Aguiar, "Desambiguación léxica y sentido en el español" (Lexical Disambiguation and Sense in Spanish) Publicacions i Edicions de la Universitat de Barcelona, 2008, ISBN: 978-84-475-3275-7
12. Huang Xiaorong, Semantic Disambiguation Techniques for Chinese Words Science Press, 2009, ISBN: 978-7-03-024173-1