ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: **Botir Elov**

Bosh muharrir o‘rinbosari: **Shahlo Hamroyeva**

Mas’ul kotib: **Oqila Abdullayeva**

Tahrir kengashi

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 32-45

NLPDA SEMANTIK TEGLASH USULLARI

Botir Elov[1]
Shahlo Hamroyeva[2]
Marjona Hamroqulova[3]

Annotatsiya. Jahon tilshunosligida kompyuter lingvistikasi muammolarini o‘rganish XX asrning 40-yillarida boshlandi, 60-yillarda mazkur jarayon jadallashdi. NLP tizimlari, avtomatik tarjima, tezaurus, elektron lug‘atlardan foydalanish imkoniyati kengaydi, ilmiy-nazariy asoslari yaratildi, amaliyotda qo‘llana boshlandi. Bu yangilanishlar axborot texnologiyalarini tilshunoslikka tatbiq etish bilan bog‘liq istiqbolli ilmiy yo‘nalishlar paydo bo‘lishiga yo‘l ochdi. Bu esa tabiiy tilni qayta ishlash(NLP)da semantik teglash usullarini o‘rganish zaruratini kun tartibiga qo‘ydi. Ushbu maqolada tabiiy tilni qayta ishlashda qo‘llaniladigan semantik teglash usullari haqida fikr yuritilgan. Ular o‘zaro qiyoslab, tahlilga tortilgan.

Kalit so‘zlar: *tabiiy tilni qayta ishlash (NLP), semantik teg, semantik teglash usullari.*

Kirish

Har bir tilning o‘ziga xos xususiyatlari mavjud bo‘lib, ularning til xususiyatlari madaniyat bilan bog‘liqdir. Shu o‘rinda ayta olamizki, millatning umumiy dunyoqarashi tilda namoyon bo‘ladi. Kompyuter lingvistikasi sohasiga zamonaviy axborot texnologiyalarini keng joriy qilish orqali mavjud vazifalarni avtomatlashtirish va qulaylashtirishga bo‘lgan urinishlar natijasida XXI asrga kelib insonlar tabiiy tillarni o‘rganish, undagi gaplarni tahlil qilishni axborot tizimlari orqali amalga oshirilmoqda. Bunday vositalar so‘nggi ax-

[1] *Elov Botir Boltayevich* – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: *elov@navoiy-uni.uz*
ORCID: 0000-0001-5032-6648
[2] *Hamroyeva Shahlo Mirdjonovna* – filologiya fanlari doktori (DSc), dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: *shaxlo.xamrayeva@navoiy-uni.uz*
ORCID: 0000-0002-5429-4708
[3] *Hamroqulova Marjona Nabijon qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti tayanch doktoranti
E-pochta: *hamroqulovamarjona@mail.ru*
ORCID: 0009-0002-2297-1492

borot-kommunikatsiya texnologiyalarining mahsuli bo‘lib, tasvirlar va nutqqa texnik ishlov berish (processing) hamda tanib olish (recognition) tizimlari, matnlarni morfologik, sintaktik analiz qilish tizimlari shular jumlasidandir.

Kompyuter lingvistikasida tabiiy tilni qayta ishlash ancha murakkab jarayon bo‘lib, unda ijtimoiy tabiatga ega tilning barcha hodisalari, umumiy va xususiy jihatlari, istisnoli holatlari, fonetik, morfonologik, leksik, semantik, grammatik va hatto orfoepik xususiyatlarini e’tiborga olish zarur. Tabiiy tilni qayta ishlash (NLP) sun’iy intellektning sohasi bo‘lib, uning maqsadi kompyuter va inson o‘rtasida o‘zaro aloqa o‘rnatish, kompyuterlarga strukturlangan matnni tushunish va undan mazmunini olishni o‘rgatishdan iborat. Semantika esa so‘zlarning, belgilarning va gap tuzilishining ma’nosi va talqinini anglatadi. Semantika, asosan, leksemaning ma’no taraqqiyoti, kengayishi, torayishi, ixtisoslashishi, uzual va okkazional ma’no, bosh va hosila ma’no, sinonimiya, antonimiya, polisemiya kabi semantik hodisalar doirasidagi masalalar tadqiqi bilan shug‘ullanuvchi lingvistik soha sanaladi. So‘nggi yillarda matnlarni semantik tahlil qilish muhim ahamiyat kasb etdi. Axborotning rivojlanishi internet resurslarini haddan ortiq ko‘p yuklanishi muammosini ko‘paytiradi. XXI asr boshlarida internetdagi sahifalar soni 4 milliarddan oshdi va har kuni 7-9 millionga ko‘paymoqda. Strukturlanmagan ma’lumotlar bilan shug‘ullanadigan foydalanuvchilar, ko‘plab tashkilotlar va shaxslar tomonidan taqdim etilgan matnli axborotlar ma’lumotlarning katta qismini tashkil qiladi.

Tabiiy tillarga ishlov berish jarayonining muammolarini ketma-ket modellashtirish zamon talabidir. Omonim so‘zlarning kontekstda aniqlanishi bevosita so‘z turkumlarini teglash muammosiga duch keladi. Ayniqsa, so‘z turkumlarini teglash qadimiy va eng mashhur muammolardan hisoblanadi. o‘zbek tilidagi omonimlarni teglash muammolari bo‘yicha ayrim mulohazalar, omonimiyani aniqlash algoritmini tuzish bo‘yicha dastlabki harakatlar amalga oshirilgan. Tadqiqotchilar D.Axmedova, Sh.Hamroyeva, O‘.Xolyorov, Sh.Gulyamova hamda M.Musayev, B.Elov, X.Axmedova, M.Sharipov, M.Abjalova, U.Salayev, H.Madatov, Sh. Bekchanov, I.Bakayev, B. Akmuradov, O‘.Xamdamov, J.Elov kabi ko‘plab lingvist va texnik tadqiqodchilarning izlanishlari bunga yaqqol misol bo‘ladi.

Semantik teglash masalasi bo‘yicha jahonda ham bir qancha ishlar amalga oshirilgan. A.M.Yelizarov, T.I.Reznikova, Y.I.Yakovchuk, O.Y.Shemanayeva, V.V.Ivanov, I.M.Boguslavskiy, V.P.Zaxarov kabi olimlar semantik razmetkalash, semantik teglarni yaratish muammolari to‘g‘risida tadqiqot olib borganlar.

Teglash tabiiy tilni qayta ishlash (NLP) dagi dastlabki ishlov berish bosqichi bo‘lib, u matndagi har bir so‘zga grammatik toifa belgilaydi. Semantik teg (rus., razmetka) – til korpuslariga kiritiladigan birliklarning muayyan semantik kategoriya yoki kichikroq semantik guruh (leksik-semantik guruh, semantik maydon)ga tegishliligini bildiradigan, ma’noni xususiylashtiruvchi belgi, izohlar majmuyi [Захаров, Богданова, 2011. 161]. Semantik tahlil murakkab matematik muammo bo‘lib, uning yechimi sun’iy intellektni yaratish jarayonida qo‘llanadi va tabiiy tilni qayta ishlash zarurati bilan murakkablashadi. Qiyinchilik shundan iboratki, kompyuter ramzlar yordamida odam uzatadigan tasvirlarni to‘g‘ri tushuntirishni bilmaydi. Sifatli semantik tahlil ma’lumotlari savdo-sotiq, tovarlarga bo‘lgan talabni tahlil qilish hamda avtomatik tarjima tizimlarida ishlatilishi mumkin [Гулямова, 2021. 32].

Semantik teglashda, boshqa razmetkalarda bo‘lganidek, yagona standart shakl bo‘lmasa ham, harf, raqam yoki faqat raqamdan iborat kodlardan foydalaniladi. Birinchi harf yoki raqam umumiy semantik ma’noni, keyingi belgi esa so‘z ma’nosini yanada maxsuslashtiruvchi kichik semantik guruhni ifodalaydi. Semantik teg nafaqat so‘z, balki ko‘plab birikmalarni ham semantik guruhlarga birlashtiradi, bunday paytda turli birikuvdagi bir ma’noni bildiruvchi birikmalar bitta belgi bilan kodlanadi. Idiomatik birlik (ibora) tarkibidagi so‘zlar miqdorini bildiruvchi axborot ham razmetkadan joy oladi. Semantik teg korpusdagi so‘z ma’nosining ixtisoslashuvi, omonimlik, sinonimlik, ma’noviy guruhga ajratish kabi muammolarni hal qiladi. V.P.Zaxarov, S.Y.Bogdanovalar ham rus tili milliy korpusini tuzishda semantik razmetkalashning o‘z variantini taklif qiladi [Захаров, Богданова, 2011. 45].

Semantik teglashda eng asosiy muammo ma’lum birlikning kontekstga qarab ikki va undan ortiq semantik guruhga mansublik holatini aniqlashdir. So‘z turkumlari ichidagi leksik birlikni tizimlashtirish shakldan ma’noga qarab emas, balki ma’nodan shaklga prinsipiga yondashuvi asosida amalga oshiriladi: turkum ichidagi har bir so‘z alohida izohlanmaydi (bunday izohlash morfologik annotatsiyalashga xos), balki semantik maydon yoki guruh aniqlanib, shu guruhga mansub so‘zlar yig‘iladi. Bunday yondashuv til birligiga tafakkur va muloqot birligi sifatida qarashdan kelib chiqadi.

Materiallar va usullar

Tadqiqot metodi sifatida ushbu maqolada tavsiflash hamda qiyoslash usullaridan foydalanildi.

Semantik tegning ham, boshqa razmetkalarda bo‘lganidek,

yagona standart shakli bo‘lmasa ham, harf, raqam yoki faqat raqamdan iborat kodlardan foydalaniladi. Birinchi harf yoki raqam umumiy semantik ma’noni, keyingi belgi esa so‘z ma’nosini yanada maxsuslashtiruvchi kichik semantik guruhni ifodalaydi. Semantik teg nafaqat so‘z, balki ko‘plab birikmalarni ham semantik guruhlarga birlashtiradi, bunday paytda turli birikuvdagi bir ma’noni bildiruvchi birikmalar bitta belgi bilan kodlanadi. Idiomatik birlik (ibora) tarkibidagi so‘zlar miqdorini bildiruvchi axborot ham razmetkadan joy oladi. Semantik teg korpusdagi so‘z ma’nosining ixtisoslashuvi, omonimlik, sinonimlik, ma’noviy guruhga ajratish kabi muammolarni hal qiladi [Zaxarov, Mengliyev, Xamroyeva, 2021. 137]. V.P.Zaxarov, S.Y.Bogdanovalar rus tili milliy korpusini tuzishda semantik razmetkalashning o‘z variantini taklif qiladi [Захаров, Богданова, 2011. 45].

Tabiiy tilni qayta ishlashda (NLP) semantik teglash matn ma’lumotlaridan ma’no chiqarish uchun mo‘ljallangan bir qator texnika va usullarni o‘z ichiga oladi. Ushbu usullar mashinalarga inson tilining nozik tomonlarini o‘rganishga va matnni to‘g‘ri talqin qilishga imkon beradi. Ular quyidagilar:

1. Qoidalarga asoslangan usul.

2. Stoxastik (yoxud statistik) usul.

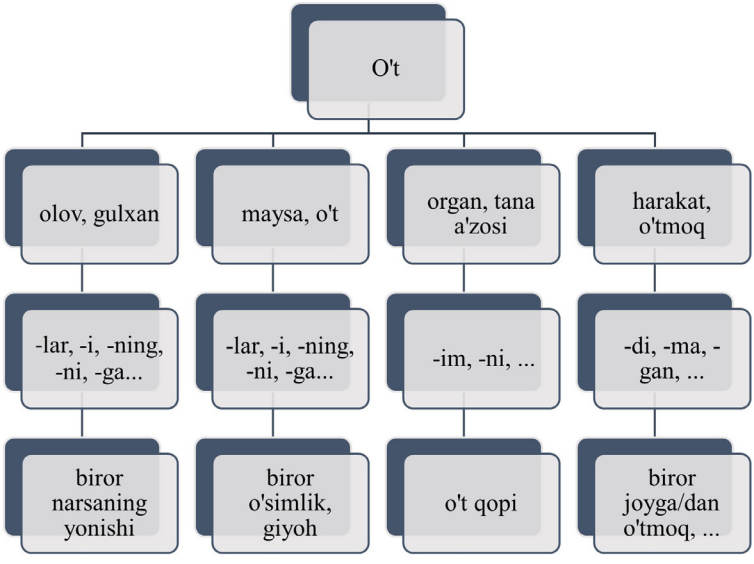
3. Neyron tarmoqlari va chuqur o‘rganish (Neural networks and Deep Learning) usuli.

4. Gibrid usul.

Qoidalarga asoslangan usulda so‘zlarni til grammatikasi qoidalari yordamida semantik tahlil qilish nazarda tutilgan. O‘zbek tilida qator qoidalar ishlab chiqilgan. Ushbu qoidalar asoslangan holda ba’zi turdagi omonim so‘zlarni farqlovchi matematik modellar hosil qilinadi. Hosil qilingan matematik modellar asosida alohida modullar yozilishi zarur bo‘lib, ular kontekstga va omonim so‘zlarning o‘ziga, ularning so‘z turkumiga yoki parametrlariga qarab faqat ma’lum hollarda o‘z isbotini topadi. Ushbu usulning mohiyati shundan iboratki, ba’zi hollarda kontekst tahlili gapning bir qismining sintaktik tuzilishini va uning yordamida so‘z shakllarini tushunishga yordam beradi. Mazkur usulda har bir so‘zni teglash uchun lug‘at yoxud leksikadan foydalaniladi.

Bu yondashuvda POS teglash grammatika qoidalari va qoidaga asoslangan tizimlar yordamida amalga oshiriladi. Ushbu algoritmlar so‘zning qo‘shimchalari va morfologik xususiyatlari kabi grammatik ma‘lumotlarni tahlil qiladi va shunga mos ravishda POS teglarini tayinlaydi. Bu usul grammatik ma‘lumotlarni chuqur tah-

lil qilishni talab qiladi. Masalan, ot va fe'l so'z turkumlari orasidagi omonimiyaga qaraydigan bo'lsak:



O't so'zi matnda ot turkumiga mansub bo'lib kelganida, otlar kabi egalik, kelishik, ko'plik qo'shimchalarini, fe'l bo'lib kelganida esa fe'llarga xos zamon, mayl, shaxs-son, fe'lning vazifa shakllari qo'shimchalarini olib kelishi mumkin.

Stoxastik usul ba'zi manbalarda statistik metod deb ham qo'llaniladi. Bu usuldan masalani yechishda so'zlarning grammatik parametrlarini tasniflash orqali foydalaniladi. Bu parametrlar turli tabiiy tillarda turlicha tanlab olinadi. Masalan, rus tilida morfologik omonimiyani aniqlashda so'z turkumi, so'zning qaysi jinsga oidligi, birlik yoki ko'plik shakli, lemma, lemma va so'z turkumi, faqat lemmasi, omonimligi kabi parametrlar ajratib olingan. Stoxastik usul ma'lumotlarga asoslangan usullarga tayanadi, ko'pincha, so'zlar orasidagi munosabatlarni aniqlash uchun katta matn korpusidan foydalanadi. Har qanday holatda chastota yoki ehtimollikni o'z ichiga olgan har qanday model to'g'ri stoxastik model deb belgilanadi. Eng oddiy stoxastik teglash so'zlarni faqatgina so'zning ma'lum bir teg bilan paydo bo'lish ehtimoli asosida ajratib turadi. Boshqacha qilib aytadigan bo'lsak, ushbu so'z bilan uchrashi mumkin bo'lgan so'zlar to'plamida eng ko'p uchraydigan teg bu so'zning noma'lum misoliga berilgan tegdir. Mazkur usul chastota yoki ehtimollik(statistika)ka asoslanadi. Shu bois ayrim manbalarda statistik yoxud ehtimolikka asoslangan usul tarzida tushuntiriladi.

Ushbu usulning asosiy vazifasi kontekst tarkibini n-grammlarga ajratish, ya'ni kontekstdagi kirish so'zining birikuvchilarini

aniqlab, baholash metodlari yordamida baholanadi. Ba'zi statistik usullarda so‘z va uning qo‘shimchalari orqali qaror qabul qilinsa, ba'zilarida so‘zning kontekstdagi semantik valentliklari yordamida xulosa qilinadi. Bundan kelib chiqadiki, statistik ma'lumotlarga asoslangan usullar qaror qabul qilish parametrlariga ko‘ra bir necha guruhlarga bo‘linadi

Stoxastik teglashda quyidagi metodlardan foydalaniladi:

1. Chastotali yondashuv. Ushbu yondashuvda stoxastik teggerlar so‘zning matnda ma'lum bir teg bilan uchrashi ehtimoli asosida grammatik noaniqliklarni bartaraf etadi. Shuni ham aytish mumkinki, o‘rganilayotgan to‘plam (matn qismi)da muayyan so‘z bilan tez-tez qo‘llaniladigan teg o‘sha so‘zning noaniqligiga ma'lumot berishga yordamchi tegdir. Masalan, qo‘llanilish darajasiga binoan tushum kelishigidagi so‘zdan so‘ng kelgan lingvistik birlik fe'l turkumiga mansub hisoblanadi: *uyni qurmoq, kosani olmoq, ko‘chani tozalamoq* kabi. Teglash jarayonidagi bunday yondashuvning asosiy muammosi, u tabiiy tilda birikuvchanligi bo‘lmagan teglar ketma-ketligini keltirib chiqarishi mumkin. Masalan, *"Men bo‘yoqlarimni rasm chizish uchun olib keldim"* misolida rasm so‘zi o‘ng tomondan "chizish uchun" birikmasi bilan birikuvchanlikka ega, ammo chapdagi so‘z bilan (bo‘yoqlarimni) na grammatik, na semantik jihatdan birika oladi. Bunday hollarda teglashtirishda birikuvchanlikka ega bo‘lmagan so‘zlar nomutanosibligiga asoslanish natijasida morfotahlil jarayonida noto‘g‘ri ma'lumot yuzaga keladi.

2. Teglarning ketma-ketligi ehtimoli yoxud n-gramma usuli. Stoxastik usulning mazkur yondashuvi tegger berilgan teglar ketma-ketligining qo‘llanilish ehtimolini hisoblaydi. Ketma-ketlik o‘lchovi, ya'ni n (bigram – ikki element ketma-ketligi, trigram – uch ketma-ket teg, 4 gram – to‘rt teg ketma-ketligi) teglarga asoslangani uchun bu yondashuv N-gramma usuli ham deyiladi. N-gramma – matnlarga avtomatik ishlov berishda keng qo‘llaniladigan matematik hisob vositasidir. O‘zbek kompyuter lingvistikasida S.Rizayev harf birikmalarini bigramm, trigramm terminlari bilan ifodalagan [Rizayev, 2006. 18].

Yashirin Markov modeli stoxastik usulda faol qo‘llaniladi. 1960- yillarda Baum L.E. va uning hamkasblari tomonidan ishlab chiqilgan [Baum, Sell, 1968. 211-227] mazkur usul statistik jarayonda yuzaga keladigan barcha variantlar ehtimolligini hisobga olishga yordam beradi. Masalan, ma'lum bir matnda ot turkumiga oid so‘zlar bog‘lovchiga nisbatan tez-tez va ko‘p uchrasa unda ayni kontekstda mavjud omonim katta ehtimollik bilan bog‘lovchi emas, ot turkumi-

ga oid so‘z bo‘ladi, keyingi ehtimollikda bog‘lovchi sifatida hisobga olinadi. Kontekstni tavsiflash uchun N-grammadan foydalaniladi. N-gramma – so‘zlar yoki teglar kabi N-identifikator elementlarning ketma-ketligini ifodalaydi.

Yashirin Markov modellari termodinamika, statistik mexanika, fizika, kimyo, iqtisodiyot, moliya, signallarni qayta ishlash, axborot nazariyasi, nutqni qayta ishlash, husnixat, imo-ishoralarni tanib olish, so‘z turkumlarini teglash va bioinformatikada keng qo‘llaniladigan statistik model hisoblanadi.

Neyron tarmoqlari va chuqur o‘rganish (Neural networks and Deep Learning) usuli to‘g‘ridan-to‘g‘ri ma‘lumotlardan semantik tasvirlarni olish uchun neyron tarmoqlardan foydalangan holda semantik tahlilni amalga oshiradi. Ushbu usul statistik usullarga qo‘shimcha bo‘lib, hozirda omonimiyani aniqlashda keng qo‘llanmoqda. Deep Learning – bu Machine Learning algoritmlarining ixtisoslashuvi, sun’iy neyron tarmog‘idir. So‘nggi paytlarda Deep Learning usullari keng qo‘llanilib, yaxshi natijalarga erishayotgani kuzatilmoqda. Arxitektura bo‘yicha qaror qabul qilishda Deep Learning texnikasi tomonidan taqdim etilgan moslashuvchanlik ushbu texnikalar muvaffaqiyatining muhim sabablaridan biridir. Deep Learning usullari tabiiy tilni qayta ishlash bo‘yicha tadqiqotlar uchun ishlatiladigan Machine Learning usullari orasida birinchi o‘rinda turadi [Sharipov, Adinayev, Raximov, 2023. 48]. Diqqatga sazovor joyi shuki, bu usul quyidagilarni o‘z ichiga oladi:

Transformerlar, BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) va T5 (Text-to-Text Transfer Transformer) [https://spotintelligence.com].

Semantic Role Labeling (SRL): jumladagi so‘zlar yoki iboralarga ularning jumladagi semantik rolini ko‘rsatadigan yorliqlarni belgilash jarayonidir, masalan, agent, maqsad yoki natija. U gapning ma‘nosini topish uchun xizmat qiladi. Buning uchun u jumlaning predikati yoki fe’li bilan bog‘liq dalillarni va ularning o‘ziga xos rollariga qanday tasniflanganligini aniqlaydi . Umumiy misol “Maryam kitobni Yusufga sotdi” jumlasidir. Agent “Maryam”, predikat “sotilgan” (aniqrog‘i, “sotish”), mavzusi “kitob” va qabul qiluvchi “Jon”. Semantik rol belgilari asosan jumlalardagi so‘zlarning rolini tushunish uchun mashinalar uchun ishlatiladi. Bu nafaqat tillardagi so‘zlarni, balki ularni turli jumlalarda qanday ishlatishni ham tushunishi kerak bo‘lgan Tabiiy tilni qayta ishlash (NLP) dasturlariga foyda keltiradi. Semantik rol yorlig‘ini yaxshiroq tushunish savollarga javob

berish, ma‘lumot olish, matnni avtomatik umumlashtirish, matn ma’lumotlarini yig‘ish va nutqni aniqlashtirishda yutuqlarga olib kelishi mumkin [Wikipedia].

Chuqur o‘rganish algoritmlari yordamida o‘qitiladigan tarmoqlar nafaqat aniqlik bo‘yicha eng yaxshi alternativ yondoshuvlardan ustun keldi, balki bir qator vazifalarda (masalan, tasvirni aniqlash, matn ma’lumotlarini tahlil qilish va boshqalar) taqdim etilgan ma’lumotlarning ma’nosini tushunishning boshlanishini ko‘rsatdi. Kompyuterda nutqni aniqlash va mashina tarjimasining eng muvaffaqiyatli, zamonaviy sanoat usullari neyron tarmoqlardan foydalanishga asoslangan va axborot kommunikatsion texnologiyalari sohasining gigantlar Apple, Google, Facebook kabi kompaniyalari neyron tarmoqlarini o‘rganuvchi laboratoriyalar guruhlarini sotib olishmoqda. So‘zni vektor sifatida ifodalash hozirda chuqur o‘rganish bo‘yicha tadqiqotning eng qiziqarli yo‘nalishlaridan biridir, garchi bu yondoshuv dastlab Bengio va boshqalar tomonidan o‘n yildan ko‘proq vaqt oldin kiritilgan bo‘lsa-da. So‘zni vektor sifatida ifodalashning eng mashhur modellari – Continuous-Bag-of Word va Skip-Gram, Mikolov tomonidan berilgan tavsif va Google kompaniyasi saytida chop etilgan Word2vec modellarini amalga oshirish, so‘nggi to‘qqiz yil ichida ko‘pchilikning e’tiborini tortdi. Ushbu modellar oddiy neyron tarmoqdan foydalangan holda ma’nosi bo‘yicha so‘zga yaqinroq so‘zlarni ko‘rsatish imkoniyati beradi. Ko‘plab xorijiy tajribalar shuni ko‘rsatadiki, neyron tarmoqlari modelidan foydalanish uchun ma’nosi oldindan aniqlangan, belgilangan matnlar kerak bo‘ladi. Bundan kelib chiqadiki, o‘zbek tili milliy korpusi ma’lumotlar bazasidagi katta hajmdagi matnlarni teglash, ma’nolarini belgilab olish dolzarb masala. Matnlarni belgilash qator vazifalarni bajarishni taqozo etadi [Axmedova, 2023. 24].

Gibrid usul. Bu metod qoidalarga asoslangan va statistik usullarning kuchli tomonlarini birlashtiradi. Bu usullarni birlashtirish, asosan, murakkab matnli ma’lumotlar bilan ishlashda aniqlikni oshiradi. Birinchidan, mumkin bo‘lgan POS teglari grammatika qoidalaridan foydalangan holda yaratiladi, so‘ngra bu teglarning ehtimolliklari statistik usullar yordamida hisoblanadi. So‘ngra eng yuqori ehtimoli bo‘lgan teg tanlanadi. Ushbu usul grammatik qoidalarning ham to‘g‘riligini ta’minlaydi [https://www.hayaletyazar.net.tr/blog-detay-852192-POS-Etiketleme-Icin-Kullanilan-Algoritmalar.html].

Jahon tajribasi shuni ko‘rsatadiki, omonim so‘zlarni semantik farqlashda gibrid yondoshuvlar ahamiyatlidir. Gibrid yondoshuv o‘z

ichiga qoidalarga asoslangan usul, statistik ma'lumotlarga asoslangan usullarni qamrab oladi. Qoidalarga asoslangan usul vazifaning ma'lum bir qismini bajarsa, statistik ma'lumotlarga asoslangan usul yana bir qismini bajaradi. Albatta, bu ikkita usul bilan omonimiya, polisemiya hamda polifunksionallik to'liq o'z yechimini topmaydi. Bunday hollarda Neyron tarmoqlari va chuqur o'rganish usuliga murojaat qilish mumkin.

POS yorlig'i algoritmlari tilning murakkabliklarini ko'rib chiqsa-da, o'zbek tili alohida qiyinchiliklarni keltirib chiqaradi. O'zbek tilining agglyutinativ tuzilishi va keng morfologik xususiyatlari POS-ni aniq belgilash uchun batafsil tahlilni talab qiladi. Shuning uchun o'zbek tilidagi POS teglash algoritmlari keng qamrovli grammatik ma'lumotlar to'plami bilan o'qitilishi kerak, bu tilning tuzilishini to'liq tushunish va keng grammatik bilim bazasiga ega bo'lish uchun mutlaqo zarurdir.

Ushbu usullardan foydalanib, NLP tizimlari inson tilini chuqurroq tushunishiga erishish mumkin, bu ularni yanada ko'p qirrali qiladi va hissiyotlarni tahlil qilishdan tortib, mashina tarjimasi va savollarga javob berishgacha bo'lgan turli vazifalarni bajartira oladi.

Natijalar va muhokama

NLPda semantik teglashni amalga oshirishda bir qancha qiyinchiliklar bor. Tabiiy tilni qayta ishlashda (NLP) semantik tahlil sezilarli yutuqlarga erishdi, ammo qiyinchiliklarga ham duch keldi. Matndan ma'no chiqarishda NLP tizimlari murakkab yechimlarni talab qiladigan ko'plab muammolarga duch keladi. Asosiy qiyinchiliklardan ba'zilari:

1. Noaniqlik va ko'p ma'nolilik.

• Tabiiy til kontekstga qarab bir necha ma'noli so'zlar va iboralardan iborat bo'lishi mumkin. Bu o'ziga xos noaniqlik semantik tahlil tizimlarini chalkashtirib yuborishi mumkin.

• Qiyinchilik: berilgan kontekstdagi so'z yoki iboraning to'g'ri ma'nosini aniqlash kontekstli tushunishni talab qiladi.

2. Madaniy va kontekstual farqlar.

• Til madaniyat va kontekstga katta ta'sir ko'rsatadi. Bir madaniyatda umumiy bilim deb hisoblangan narsa boshqasida noaniq bo'lishi mumkin va kontekst matnning umumiy ma'nosini keskin o'zgartirib yuborishi mumkin.

• Qiyinchilik: madaniy unsurlarni tushunish va matnni kontekstga xos usullarda izohlash uchun NLP modellarini moslashtirish.

3. Kamyob yoki yangi so‘zlarni qo‘llash.

• Til rivojlanib, yangi so‘zlar, jargonlar, domenga xos terminologiya yoki shu kabilar bilan boyiydi. NLP tizimlari ilgari hech qachon duch kelmagan so‘zlarga duch keladi.

• Qiyinchilik: So‘z boyligidan tashqari so‘zlarni boshqarish va o‘zgaruvchan tilga moslashish strategiyalarini ishlab chiqish.

4. Baholash ko‘rsatkichlari.

• Semantik tahlil tizimlarining ishlashini baholash oson emas. Aniqlik va to‘g‘rilik kabi an’anaviy o‘lchovlar semantik tahlilning barcha unsurlarini qamrab ololmasligi mumkin.

• Qiyinchilik: Kontekst, noaniqlik va madaniy o‘zgarishlarni hisobga olgan holda semantik tahlil sifatini baholaydigan tegishli baholash ko‘rsatkichlarini ishlab chiqish.

5. Sog‘lom fikrlashning yo‘qligi.

• NLP modellari sezilarli yutuqlarga erishgan bo‘lsa-da, ular, ko‘pincha, aql-idrok qobiliyatiga ega emaslar. Kundalik vaziyatlarni tushunish va mantiqiy xulosalar chiqarish ular uchun hali ham imkonsiz bo‘lib qolmoqda.

• Qiyinchilik: sog‘lom fikrlash va dunyo bilimlarini birlashtirish uchun NLP tizimlarini rivojlantirish.

6. Kengaytirish.

• Raqamli kontent hajmi jadallik bilan o‘sib borayotganligi sababli, katta hajmdagi ma’lumotlarni samarali qayta ishlash uchun semantik tahlil tizimlarini kengaytirish doimiy muammodir.

• Qiyinchilik: kengaytiriladigan hamda yuqori samarali NLP modellari va infratuzilmalarini ishlab chiqish.

7. Multimodal va tillararo semantika.

• Matn, tasvir, video va audioni tushunish uchun semantik tahlilni kengaytirish yangi chegarani taqozo etadi. Bundan tashqari, bir nechta tillar va til juftliklari bilan ishlash murakkab vazifadir.

• Qiyinchilik: Turli xil uslublar va tillarda samarali ishlash uchun semantik tahlilni kengaytirish.

8. Maxfiylik va axloqiy tashvishlar.

• Semantik tahlil axloqiy ta’sirga ham ega, ayniqsa AI (sun’iy intellekt) tizimlari maxfiylik va tarafkashlik bilan bog‘liq.

• Qiyinchilik: NLP modellaridagi axloqiy muammolar va noto‘g‘ri fikrlarni hal qilish, to‘g‘ri va adolatli semantik tahlilni ta’minlash.

Ushbu muammolarni hal qilish NLPda semantik teglashni rivojlantirish uchun juda muhimdir. Tadqiqotchilar va amaliyotchilar inson tilining nozik tomonlarini hal qiladigan yanada mustah-

kam, kontekstdan xabardor va madaniy jihatdan sezgir tizimlarni yaratish ustida bosh qotirishi lozim.

Semantik teglash rivojlanib borar ekan, u insonlarning mashinalar bilan o‘zaro munosabatini yuksaltirish va turli xil ilovalarda tilni tushunish kuchidan foydalanish imkoniyatini beradi.

Xulosa

Tabiiy tilni qayta ishlashda (NLP) semantik teglash inson va mashina tili tushunchasi o‘rtasidagi bo‘shliqni yopadigan dinamik va o‘zgarib turuvchi sohadir . Ushbu tadqiqotimizda ko‘rib chiqilganidek, u mashinalarga matn ichiga kiritilgan ma’no, kontekst va har xil unsurlarni shifrlash imkonini beradi, shuningdek, ko‘plab sohalar va domenlar bo‘ylab turli ilovalar uchun eshiklarni ochadi.

Semantik teglash his-tuyg‘ularni tahlil qilishdan tortib, ijtimoiy tarmoqlardagi kontentni moderatsiya qilishgacha, inson bilan mashinaning o‘zaro munosabatda bo‘lishi va matnli ma’lumotlardan qimmatli tushunchalarni ajratib olish usullarigacha dasturilamal bo‘lib xizmat qiladi. Semantik teglash tizmini to‘g‘ri yo‘lga qo‘yish orqali korxonalarga ma’lumotlarga asoslangan qarorlar qabul qilish imkoniyatini berish mumkin, bundan tashqari, insonlarga shaxsiy tajribalarni taklif qiladi va yuridik hujjatlarni ko‘rib chiqishdan tortib klinik tashxislargacha bo‘lgan ishlarda mutaxassislarni qo‘llab-quvvatlaydi.

Biroq, semantik tahlilning qiyinchiliklari, jumladan, til noaniqligi, madaniyatlararo farqlar va axloqiy masalalar haligacha o‘z yechimini topmagan. Ushbu soha rivojlanishda davom etar ekan, tadqiqotchilar va amaliyotchilar ushbu qiyinchiliklarni yengish va semantik teglashni yanada mustahkam, aniq va samarali qilish uchun harakat qilishi lozim.

Bugungi kunda ilg‘or til modellarining paydo bo‘lishi, tizimda sog‘lom fikrlashning yanada yaxshilanishi va multimodal ma’lumotlar tahlilining uzluksiz integratsiyasi kabi muammolar o‘z yechimini kutmoqda. NLPda semantik teglash rivojlanib borar ekan, uning ta’siri alohida sohalardan tashqariga chiqadi, innovatsion yechimlar orqali inson va mashinaning o‘zaro muloqotini mukammallashtiradi.

Xulosa qilib aytadigan bo‘lsak, NLPda semantik teglash texnologik innovatsiyalar orasida birinchi o‘rinda turadi, bu esa mashinaning insonlar tilini tushunish va o‘zaro munosabatlarda inqilobni amalga oshiradi. Bu bizga ko‘p narsalarni o‘zgartirish imkonini beradi, muloqotni yanada qulay, samarali va mazmunli qiladi. Muammolarni hal qilish va kelajakdagi tendensiyalarni qabul qilish bo‘yicha

davom etayotgan izlanishlar, shuningdek, semantik teglash masalasi tadqiqotchilarning diqqat markazida bo‘lib kelmoqda.

Foydalanilgan adabiyotlar

Axmedova X. O‘zbek tilidagi gaplarni semantik analiz qilish modeli, algoritmlari va axborot tizimini ishlab chiqish: Texnika fan. bo‘yicha falsafa doktori (PhD) disser. – Toshkent, 2023. – B. 24.

Baum L.E., Sell G.R. Growth transformations for functions on manifolds. Pacific Journal of Mathematics. 27 (2), 1968. – P. 211-227.

Rizayev S. O‘zbek tilshunosligida lingvostatistika asoslari. – Toshkent: Fan, 2006. – B. 18.

Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари: Филол. Фан. докт. (DSc) дисс. – Тошкент, 2021. – Б. 32.

Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с. – С. 45, 48-49.

Zaxarov V., Mengliyev B., Xamroyeva Sh. Korpus lingvistikasi: korpus tuzish va undan foydalanish: o‘quv qo‘llanma. – Toshkent, 2021. – B. 137.

Sharipov M., Adinayev X., Raximov R. O‘zbek tili uchun tabiiy tilni qayta ishlashda (NLP) Mashinali o‘rganishning o‘rni. Kompyuter lingvistikasi: muammolar, yechimlar, istiqbollar, http://compling.navoiy-uni.uz/ 2023, 1(01). 48.

https://spotintelligence.com

https://www.hayaletyazar.net.tr/blog-detay-852192-POS-Etiketleme-Icin-Kullanilan-Algoritmalar.html

https://ru.wikipedia.org/

SEMANTIC TAGGING METHODS IN NLP

Botir Elov[1]
Shahlo Hamroyeva[2]
Marjona Hamroqulova[3]

Abstract. The study of computer linguistics problems in world linguistics began in the 40s of the 20th century, and this process accelerated in the 60s. The possibility of using NLP systems, automatic translation, thesaurus, electronic dictionaries has expanded, scientific and theoretical foundations have been created, and they have been used in practice. These updates opened the way for the emergence of promising scientific directions related to the application of information technologies to linguistics. This has put the need to study semantic tagging methods in natural language processing (NLP) on the agenda. This article discusses semantic tagging techniques used in natural language processing. They were compared a nd analyzed.

Key words: *natural language processing (NLP), semantic tagging, semantic tagging methods.*

References

Axmedova X. O‘zbek tilidagi gaplarni semantik analiz qilish modeli, algoritmlari va axborot tizimini ishlab chiqish: Texnika fan. bo‘yicha falsafa doktori (PhD) disser. – Toshkent, 2023. – B. 24.

Baum L.E., Sell G.R. Growth transformations for functions on manifolds. Pacific Journal of Mathematics. 27 (2), 1968. – R. 211-227.

Rizayev S. O‘zbek tilshunosligida lingvostatistika asoslari. – Toshkent: Fan, 2006. – B. 18.

Gulyamova Sh. O‘zbek tili semantik analizatorining lingvistik asoslari: Filol. Fan. dokt. (DSc) diss. – Toshkent, 2021. – B. 32.

[1] *Elov Botir Boltayevich* – doctor of philosophy of technical sciences (PhD), associate professor. Tashkent State University of Uzbek Language and Literature named after Alisher Navo’i.
E-pochta: *elov@navoiy-uni.uz*
ORCID: 0000-0001-5032-6648
[2] *Hamroyeva Shahlo Mirdjonovna* – doctor of philological sciences, associate professor, etc. Alisher Navo’i Tashkent State University of Uzbek Language and Literature.
E-pochta: *shaxlo.xamrayeva@navoiy-uni.uz*
ORCID: 0000-0002-5429-4708
[3] *Hamroqulova Marjona Nabijon qizi* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navo’i.
E-pochta: *hamroqulovamarjona@mail.ru*
ORCID: 0009-0002-2297-1492

Zaxarov V.P., Bogdanova S.Y. Korpusnaya lingvistika: uchebnik dlya studentov gumanitarnix vuzov. – Irkutsk: IGLU, 2011. – 161 s. – S. 45, 48-49.

Zaxarov V., Mengliyev B., Xamroyeva Sh. Korpus lingvistikasi: korpus tuzish va undan foydalanish: o‘quv qo‘llanma. – Toshkent, 2021. – B. 137.

Sharipov M., Adinayev X., Raximov R. O‘zbek tili uchun tabiiy tilni qayta ishlashda (NLP) Mashinali o‘rganishning o‘rni. Komp-yuter lingvistikasi: muammolar, yechimlar, istiqbollar, http://compling.navoiy-uni.uz/ 2023, 1(01). 48.

https://spotintelligence.com

https://www.hayaletyazar.net.tr/blog-detay-852192-POS-Etike-tleme-Icin-Kullanilan-Algoritmalar.html

https://ru.wikipedia.org/

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 29.02.2024-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.