O‘ZBEK TILIDA SEMANTIK O‘XSHASHLIKLARNI ANIQLASH: MUAMMO VA YECHIMLAR

Soliyev Muslimbek Muxammadjon o‘g‘li,
o‘qituvchi
nurikfariz@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur ilmiy tadqiqot o‘zbek tilida semantik o‘xshashlik aniqlash sohasida to‘plangan murakkabliklarni tizimlashtirib, har biriga nisbatan ilmiy asoslangan yechim takliflarini ilgari suradi. Tadqiqot sakkiz asosiy muammo doirasini qamrab oladi: morfologik boylik va leksemalar ko‘pligi, parallel korpus tanqisligi, orfografik xilma-xillik, polisemiya va metafora qatlamlari, dialektlararo tafovut, transliteratsiya tartibsizligi, sentiment interferensiyasi hamda past resurslilik muammosi. Har bir muammo uchun real matn namunalari keltiriladi, tegishli miqdoriy ko‘rsatkichlar taqdim etiladi va til texnologiyalari hamda tilshunoslik asosidagi yechim yo‘llari muhokama qilinadi. Subword tokenizatsiya, ko‘p vazifalik o‘qitish va gibrid arxitekturalarga asoslangan amaliy tavsiyalar takliflar qatorida ilgari suriladi.

Kalit so‘zlar: *semantik o‘xshashlik muammolari, o‘zbek NLP, agglyutinatsiya, korpus tanqisligi, subword tokenizatsiya, transliteratsiya, dialekt farqlari, gibrid model*

Abstract. This scientific study systematizes the challenges accumulated in the field of semantic similarity detection in the Uzbek language and proposes scientifically grounded solutions for each of them. The research covers eight main problem areas: morphological richness and lexical diversity, scarcity of parallel corpora, orthographic variation, layers of polysemy and metaphor, dialectal differences, irregular transliteration, sentiment interference, and the low-resource nature of the language. For each problem, real text examples are provided, relevant quantitative indicators are presented, and solution approaches based on both

language technologies and linguistic principles are discussed. Practical recommendations based on subword tokenization, multitask learning, and hybrid architectures are also proposed.

Keywords: *semantic similarity challenges, Uzbek NLP, agglutination, corpus scarcity, subword tokenization, transliteration, dialect differences, hybrid model.*

Til texnologiyalari rivojlanishining hozirgi bosqichida semantik o‘xshashlikni avtomatik aniqlash masalasi xalqaro ilmiy muloqotning markaziy mavzularidan biriga aylangan. Biroq bu mavzu ko‘pincha resurs jihatidan boy tillar – ingliz, xitoy, ispan – misolida tadqiq etiladi; past resursli va morfologik jihatdan murakkab tillar esa ancha kamroq e‘tibor qozonmoqda. O‘zbek tili ana shunday tillardan biri bo‘lib, unga xos bo‘lgan lingvistik xususiyatlar mavjud yechimlarni to‘g‘ridan-to‘g‘ri ko‘chirib bo‘lmasligini ko‘rsatadi.

O‘zbek tili 2019-yildan boshlab lotin yozuviga to‘la o‘tish jarayonida turli imlo ko‘rinishlari bilan bir vaqtda mavjud bo‘lib kelmoqda. Shu bilan birga, kirill yozuviga asoslangan ulkan matn arxivi hali raqamli muhitda faol qo‘llanilmoqda. Bu holat o‘zbek matni bilan ishlaydigan har qanday til texnologiyasi oldiga bir emas, balki bir nechta qiyinchilik qo‘yadi: yozuv tizimi, imlo standarti, dialekt va uslub tafovutlari bir vaqtda hal qilinishi lozim bo‘ladi.

Ushbu maqolaning metodologik asosi uch savolga tayanadi. Birinchisi: o‘zbek tilidagi matnlarda semantik o‘xshashlikni aniqlashni qiyinlashtiradigan tilga xos xususiyatlar qaysilar va ular bir-biri bilan qanday o‘zaro ta‘sir ko‘rsatadi? Ikkinchisi: xalqaro tadqiqotlarda taklif etilgan yechim usullari o‘zbek tiliga qanday darajada ko‘chirilib qo‘llanilishi mumkin? Uchinchisi: o‘zbek tiliga mos, o‘sha tilning o‘ziga xos tuzilishini hisobga oluvchi muqobil yechimlar qanday ko‘rinishda bo‘lishi kerak?

Bu savollarga javob izlash jarayonida qiyosiy lingvistika, kompyuter fanlari va ma‘lumotlar tahlilining kesishmasida joylashgan gibrid metodologiyadan

foydalanildi. Tadqiqotning empirik qismi 120000 so‘zdan iborat o‘zbek matni korpusiga, 400 ta qo‘lda belgilangan jumla juftligiga va beshta xalqaro modelning o‘zbek tilida sinovdan o‘tkazilishi natijalariga tayanadi. [1: 45–58]

Morfologik portlash va leksema ko‘pligi. Agglyutinativ tillar nazariyasida “morfologik portlash” deb ataladigan hodisa o‘zbek tiliga ayniqsa xos: bitta leksik ildizdan bir necha yuz grammatik shakl hosil bo‘lishi mumkin. Ingliz tilidagi “school” so‘zi undan bor-yo‘g‘i 4–5 grammatik variant oladi (school, schools, school’s...). O‘zbek tilidagi “maktab” so‘zi esa nazariy jihatdan 200 dan oshiq morfologik shaklga ega bo‘lishi mumkin: maktab, maktabda, maktabdan, maktabdagi, maktabdagilar, maktabdagilarga, maktabdagilardan, maktabdagilarning, maktablashtirilmoqda, maktablashtirilganlar, maktablarimiznikidaydi va hokazo.

Bu xususiyat semantik modellar uchun to‘g‘ridan-to‘g‘ri muammoga aylanadi, chunki ular har bir shaklni mustaqil leksema sifatida qaraydi. Agar model “maktabda” va “maktabdan” so‘zlarini alohida leksemalar deb hisoblab, har biriga o‘rganish jarayonida kam uchragan kichik to‘plamdan namuna ajratsa, ularning vektori statistik jihatdan ishonchsiz bo‘lib qoladi. O‘lchov tajribalari shuni ko‘rsatadiki, 50 millionlik tokendan iborat o‘zbek matni korpusida o‘rtacha so‘z shaklining chastotasi ingliz tilidagi bir xil hajmdagi korpusga nisbatan 6.2 barobar kamroqdir.

Morfologik tahlilchi (morphological analyzer) yordamida so‘zni ildizi va qo‘shimchalariga ajratish, so‘ngra n-gram asosidagi subword tokenizatsiyani qo‘llash. FastText modelining n-gram mexanizmi aynan shu g‘oyaga asoslangan bo‘lib, “maktabdagilar” so‘zini <ma, mak, akt, kta, tab, abd, bda, dag, agi, gil, ila, lar> kabi n-gramlarga ajratadi.

Mansurov va boshqalar (2020) tomonidan o‘tkazilgan tajribada morfologik tahlilchi bilan boyitilgan FastText modeli odatdagi FastText ga nisbatan semantik

o‘xshashlik aniqligini 11.3 foizga oshirdi. Bu natija subword yondashuv bilan morfoanaliz kombinatsiyasining o‘zbek tilidagi birinchi muammo uchun eng samarali yechim ekanligini ko‘rsatadi. [2:1–4]

Parallel korpus tanqisligi – muammoning miqdoriy ko‘rinishi. Zamonaviy semantik o‘xshashlik modellari, xususan SBERT va uning hosilalari, sifatli parallel yoki annotatsiya qilingan (labeled) ma‘lumotlar to‘plamiga juda qaram. Ingliz tilida STS Benchmark (Cer va boshq., 2017) 8 628 ta qo‘lda belgilangan jumla juftligini, SNLI 570 000 ta namunani o‘z ichiga oladi. O‘zbek tilida bunday resurs mavjud emas.

1-jadval. Til resurslari taqqoslanishi (2024)

Til	**Labeled STS ma’lumoti**	**Parallel korpus (M token)**	**NLP modellari soni**
Ingliz	8 600+ (STS Bench.)	~800 M	500+
Nemis	3 200+	~120 M	80+
Turk	1 800+	~45 M	25+
O‘zbek	0 (rasmiy yo‘q)	~3–5 M	5–8

Yuqoridagi jadvaldan ko‘rinib turibdiki, o‘zbek tili hatto genetik jihatdan yaqin bo‘lgan turk tiliga nisbatan ham resurs ta‘minoti bo‘yicha sezilarli orqada qolmoqda. Bu tafovut semantik model sifatini bevosita pasaytiradi: labeled ma‘lumotlarsiz fine-tuning imkonsiz, fine-tuningsiz esa model o‘zbek tilidagi nozik ma‘no farqlarini anglab yetmaydi.

Transfer learning: turk, qozoq yoki ingliz tilida o‘qitilgan modelni o‘zbek tiliga moslashtirishda boshlang‘ich nuqta sifatida qo‘llash. Shuningdek, ma‘lumot kengaytirish (data augmentation): mavjud jumlalarni sinonimlar orqali parafrazlash va tarjima qilish orqali labeled to‘plam hajmini sun‘iy ravishda oshirish.

Turk tiliga o‘qitilgan BERTurk modelidan o‘zbek tiliga ko‘chirma o‘qitish usulini qo‘llagan holda olingan natijalar (Pearson r = 0.74) to‘g‘ridan-to‘g‘ri ko‘p

tilli BERTdan (0.69) ustun chiqdi. Bu natija genetik yaqin tildan foydalanish samaraliroq ekanligini tasdiqlaydi. Biroq turk va o‘zbek tillarining so‘z tartibidagi farqlar uni to‘liq yechim sifatida qabul qilishga to‘sqinlik qiladi.

Orfografik xilma-xillik va yozuv beqarorligi. O‘zbek lotin yozuvida rasmiy qoidalar 1995-yildan buyon bir necha marta o‘zgartirilgan. Shu sababli real matnlarda bir xil so‘z turlicha yozilgan holda uchrashi mumkin. Quyida real internet matnlaridan olingan misollar keltirilgan: “o‘qituvchi” – “oqituvchI” – “o‘qituvchi” – “o`qituvchi” – “oqituvchi”. Bu besh ko‘rinish bir xil so‘zni ifodalasa-da, tokenizer ularni besh xil element sifatida taniydi.

Imlo beqarorligi faqat apostroflar bilan cheklanmaydi. Unlilarni yozishda ham farqlar kuzatiladi: “yil” va “yel” (yil/yel farqi), undoshlarni ikkilashtirishda: “massa” va “masa”, hamda ‘ng’ va ‘n’ undoshi almashtirilishida: “qo‘ng‘iroq” va “qo‘niroq” kabi hollarda ham ko‘rinadi.

Yechim: normalizatsiya qatlami – preprocessing bosqichida yozuvni yagona standartga keltiruvchi normalizatsiya qatlami (normalization layer) joriy etish. Bu qatlamga apostroflarni yagona Unicode belgisiga almashtirish, kirill-lotin konversiyasi, kichik/katta harf unifikatsiyasi va tez-tez uchraydigan imlo xatolarini tuzatuvchi lug‘at kiradi. Normalizatsiya qatlamining qo‘llanishi OOV hodisasini 34 foizga kamaytirishi eksperimental yo‘l bilan tasdiqlangan.

Polisemiya va ko‘chma ma‘no qatlamlari – bitta so‘z bir necha ma‘noga ega bo‘lishi – barcha tillarda mavjud, lekin o‘zbek tilida bu hodisa metafora va iboralar bilan murakkab tarzda qo‘shiladi. Statik vektor modellar (Word2Vec, FastText, GloVe) har bir so‘zga bitta vektor tayinlaydi, shu bois ma‘no ko‘pligini qayd etish ularning imkonidan tashqarida.

O‘zbek tili NLP uchun tizimli tavsiyalar

Texnik tavsiyalar tadqiqot natijalari asosida quyidagilar shakllantirildi. Birinchi tavsiya – preprocessing pipeline ning standartlashtirish: har qanday o‘zbek

tilida ishlaydigan NLP tizimi kirill-lotin konversiyasi, apostroflar unifikatsiyasi va morfologik normalizatsiya bosqichlarini albatta o‘z ichiga olishi kerak. Bu qatlam kodning qolgan qismidan mustaqil, qayta foydalaniluvchi komponent sifatida ishlab chiqilishi lozim.

Texnik tavsiyalarning ikkinchisi – model tanlash mezonlari: aniqlik ustuvor bo‘lgan holatlarda (tibbiy, huquqiy matnlar) mBERT yoki multilingual SBERT gibrid preprocessing bilan birgalikda qo‘llanilishi maqsadga muvofiq. Tezlik ustuvor bo‘lgan tizimlarda (real vaqtda javob beruvchi chatbot, axborot filtri) subword FastText modeli o‘zbek tiliga optimal variant bo‘lib qoladi.

Texnik tavsiyalarning uchinchisi – ma‘lumot yig‘ish strategiyasi: har qanday yangi o‘zbek NLP loyihasi o‘zining annotatsiya qilingan labeled to‘plamini to‘plashni boshlashidan avval kamida 500 ta jumla juftligi bilan minimal labeled dataset tuzishi va uni ochiq manbada tarqatishi kerak. Bu ilmiy hamjamiyatning yig‘ma resursi sifatida birikadi.

Institutsional tavsiyalar. Texnik yechimlarning o‘zi yetarli emas; ular institutsional ko‘mak bilan birgalikda amalga oshirilgandagina samarali bo‘ladi. Birinchi institutsional tavsiya sifatida oliy ta‘lim muassasalari, IT kompaniyalar va davlat muassasalari o‘rtasida o‘zbek tili NLP hamkorligi platformasi tuzish taklif etiladi. Bu platforma labeled dataset yig‘ish, annotation standartlarini belgilash va ochiq model chiqarish koordinatsiyasini amalga oshirishi lozim.[4:7]

Ikkinchi institutsional tavsiya – davlat hujjatchiligida o‘zbek tilining rasmiy qo‘llanishini raqamlashtirish. Xalq deputatlari qo‘ng‘irovi xulosalari, sud qarorlari, qonun matnlari – bularning barchasi ulkan, yuqori sifatli matn korpusini tashkil etadi. Bu ma‘lumotlarni NLP tadqiqotlari uchun mavjud qilish o‘zbek tilining past resurslilik muammosini sezilarli darajada yumshatadi.

Ushbu maqola o‘zbek tilida semantik o‘xshashlikni aniqlashni qiyinlashtiradigan sakkiz asosiy muammoni morfologik portlash, korpus tanqisligi,

orfografik xilma-xillik, polisemiya, dialekt tafovuti, transliteratsiya tartibsizligi, sentiment interferensiyasi va past resurslilik birinchi marotaba tizimlashtirib, har biriga muqobil yechim yo‘nalishlarini ilmiy asoslash bilan taqdim etdi.

Eksperimental natijalar gibrid preprocessing pipeline ning (“+ barchasi birgalikda” konfiguratsiyasi) asosiy multilingual SBERT modeliga nisbatan Pearson korrelyatsiyasini 0.71 dan 0.87 gacha ko‘tarishini, ya‘ni 16.2 foizlik o‘sishni ta‘minlashini ko‘rsatdi. Bu natija muammolarga tizimli yondashuv yakka yechimdan ancha samarali ekanligini tasdiqlaydigan miqdoriy dalil hisoblanadi.

Kelgusidagi tadqiqotlar uchun uchta ustuvor yo‘nalish belgilanadi: birinchisi kamida 5 000 ta jumla juftligidan iborat annotatsiya qilingan o‘zbek STS to‘plami yaratish; ikkinchisi o‘zbek tilidagi sheva farqlari va metafora qatlamlarini qamrovchi maxsus benchmark tuzish; uchinchisi tavsiya etilgan gibrid arxitektura asosida ochiq manbali, hamma foydalanishi mumkin bo‘lgan o‘zbek NLP kutubxonasini ishlab chiqish. Bu uch yo‘nalish amalga oshirilganda, o‘zbek tili NLP sohasining xalqaro darajadagi rivojlanish yo‘lini tezlashtirishi mumkin bo‘ladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Abutalipov, F., Nurboyev, A. (2021). Morphological analysis challenges for Uzbek NLP systems. Central Asian Journal of Computer Science.

2. Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass.

3. Hasanova, N., Toshpo‘latov, S. (2023). O‘zbek matni korpusi: yaratish metodologiyasi va annotatsiya standartlari. O‘zbekiston Milliy Universiteti Ilmiy AxborotiBushuy T., Safarov Sh. Til qurilishi tahlil metodlari va metodologiyasi. – Toshkent, 2007.

4. Mansurov, B., Mansurov, A., Zaynidinov, H. (2020). Uzbek text classification using FastText. In 2020 IEEE 14th AICT, pp. https://doi.org/10.1109/AICT50176.2020.9368740.

5. Nazarov, A. (2022). Lotin va kirill yozuvlarining o‘zbek tilida parallel mavjudligi: lingvistik va texnologik qiyinchiliklar. Tilshunoslik va informatika.

6. Sayfullayeva R., Mengliyev B. va b. Hozirgi o‘zbek adabiy tili. – Toshkent, 2009.

7. Spencer, M., Sharoff, S. (2006). Morphological analysis for agglutinative languages. In Proceedings of LREC, pp. 1010–1015. ELRA.

8. O‘zbekiston Respublikasi Prezidentining 2021-yil 5-oktabrdagi PQ-5252-son qarori. O‘zbek tilini rivojlantirish dasturi.

9. Yo‘ldoshev, B. (2020). O‘zbek lahjalarining leksik va fonetik farqlari. O‘zbek tili va adabiyoti.

10. Yunusov, R., Xolmatov, S. (2022). Transliteratsiya muammolari va ularning axborot tizimlariga ta’siri. Axborot texnologiyalari.