O‘ZBEK TILI MATNLARIDA POLISEMANTIK BIRLIKLARNI AVTOMATIK TAHLIL QILISHNING STATISTIK METODLARI

Xusainova Zilola Yuldashevna,
f.f.f.d. (PhD), dotsent v.b.
xusainovazilola@navoiy-uni.uz
*Tosh*DO‘TAU
Norbekova Bahora Ibrohim qizi,
I bosqich magistrant
bahoraibrohimovna7277@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tili kabi morfologik boy va agglyutinativ tillarda uchraydigan polisemantik birliklarni (WSD) avtomatik tahlil qilishning statistik hamda gibrid metodlari tadqiq etiladi. Tadqiqot doirasida an’anaviy statistik modellar (SVM) va zamonaviy kontekstual enkoderlarning (mBERT) polisemik so‘zlarni farqlashdagi imkoniyatlari hamda kamchiliklari o‘zaro qiyoslanadi. Agglyutinativ tabiatdagi o‘zbek tilida so‘z ma’nosining kelishik qo‘shimchalari, uslubiy registrlar va idiomatik birikmalar orqali o‘zgarishi statistik modellarning aniqligiga ta’sir etuvchi asosiy omillar sifatida ko‘rsatiladi. Ochiq lingvistik modellarning integratsiyasi statistik modellarning aniqlik ko‘rsatkichini (F1) sezilarli darajada oshirishi va semantik noaniqliklarni tizimli hal etishi asosan dalillanadi.

Kalit so‘zlar: *Polisemantik birliklar, WSD, statistik metodlar, mBERT, agglyutinativ tillar, semantik tahlil*.

Abstract: This article investigates statistical and hybrid methods for automatic analysis of polysemantic units (WSD) in morphologically rich and agglutinative languages such as Uzbek. The study compares the capabilities and shortcomings of traditional statistical models (SVM) and modern contextual encoders (mBERT) in distinguishing polysemantic words. In the agglutinative Uzbek language, the changes in word meaning through declensional suffixes, stylistic registers, and idiomatic combinations are shown as the main factors

affecting the accuracy of statistical models. It is mainly argued that the integration of open linguistic signals significantly increases the accuracy index (F1) of statistical models and systematically resolves semantic uncertainties.

Keywords: *Word Sense Disambiguation (WSD), polysemantic units, statistical methods, mBERT, agglyutinative languages, semantic analysis.*

Kirish. Hozirgi kunda tilshunoslikda leksik-semantik tadqiqotlar alohida o‘rin egallaydi. Polisemiya hodisasi ya’ni bir so‘z shaklining bir nechta ma’no ifodalashi tilshunoslarning diqqat markazida bo‘lib kelmoqda. O‘zbek tili ham boy polisemantik tizimga ega bo‘lib, bu tizimning qonuniyatlarini aniqlash va tasniflash nazariy, balki amaliy jihatdan ham muhim ahamiyat kasb etadi.

Leksikografik va semantik tahlil usullari katta hajmdagi matn ma’lumotlari bilan ishlashda samaradorligi jihatidan cheklangan imkoniyatlarga ega. Zamonaviy lingvistika va sun’iy intellekt texnologiyalarining rivojlanishi polisemantik birliklarni tahlil qilishda sifat jihatidan yangi yondashuvlarni qo‘llash imkonini bermoqda. Xususan, statistik va matematik metodlar yordamida so‘zning turli kontekstlardagi qo‘llanilish chastotasi, taqsimlanish xususiyatlari va semantik o‘zgarishlarini miqdoriy jihatdan o‘lchash mumkin bo‘lmoqda.

O‘zbek tili matnlarida polisemantik birliklarni avtomatik tahlil qilish masalasi dolzarb hisoblanadi. Jumladan, o‘zbek tilining agglutinativ morfologik tuzilishi so‘z shakllarining turli-tumanligini keltirib chiqaradi, bu esa kompyuter tahlilini murakkablashtiradi. O‘zbek tili leksikasida arab, fors, rus tillari va o‘zlashtirilgan so‘zlarning mavjudligi leksik-semantik tahlilni yanada qiyinlashtiradi. Shuningdek, raqamli korporuslarning to‘liq shakllanmaganligi mavjud metodologik bazani boyitish zaruriyati yanada ko‘proq seziladi. O‘zbek tili matnlarida polisemantik leksik birliklarni avtomatik aniqlash, ularning semantik variantlarini farqlash va tasniflash uchun statistik metodlarga asoslangan kompleks tizim ishlab chiqish muhim hisoblanadi.

So‘z ma’nosini farqlash (WSD) sohasidagi tadqiqotlar bir necha metodologik bosqichlarni bosib o‘tdi. Bugungi kunda WSD yo‘nalishida bir necha metodologik bosqichlar shakllangan:

– knowledge-based yondashuvlar
– statistik modellar
– mashinali o‘qitish algoritmlari
– transformer asosidagi chuqur neyron modellar.

Statistik metodlar

O‘zbek tili kabi morfologik boy tillarda so‘z ma’nosini aniqlash faqatgina so‘zning shakliga emas, balki uning atrofidagi kontekst hamda ichki morfologik tuzilishiga bevosita bog‘liq. Ushbu maqolada quyidagi uchta asosiy yondashuv o‘zaro qiyoslanadi:

Chiziqli SVM (Tayanch model). Ushbu an’anaviy statistik usul maqsadli so‘z atrofidagi 10 ta so‘zdan iborat kontekst oynasiga tayanadi. Bunda TF-IDF ko‘rsatkichlari, unigram va bigram xususiyatlari asosiy model vazifasini o‘taydi. Biroq, ushbu model o‘zbek tilidagi murakkab morfologik o‘zgarishlarga (agglyutinatsiyaga) nisbatan past sezgirlik ko‘rsatadi. Masalan, *o‘rtoq* so‘zining quyidagi ma’nolari mavjud:

1. Yoshi teng yoki yaqin bo‘lib, o‘zaro yaxshi aloqada bo‘lgan tengdoshlar (bir-biriga nisbatan); birga o‘sgan; tengdosh, do‘st, oshna;
2. Dunyoqarashi, faoliyati, yashash sharoiti va shu kabilar jihatidan bir, g‘oyaviy jihatdan yaqin kishi; hamfikr, hamg‘oya;
3. Kishilarga murojaatda, odatda ularning familiyasi, kasbi, unvoni va shu kabilar bildiruvchi so‘zga qo‘shib ishlatiladi.

Model *“o‘rtoq”* so‘zi atrofida *“maktab”, “sinf”, “qalin”* kabi n-gramlarni ko‘rsa, uni “tengdosh/do‘st” (1-ma’no) deb klassifikatsiya qiladi. Biroq, agar gapda “o‘rtoq direktor” yoki “o‘rtoq bemor” kabi rasmiy murojaat shakllari kelsa, SVM

modeli morfologik va uslubiy belgilarni inobatga olmagani sababli, murojaat (3-ma’no) va do‘stlik ma’nosini ajratishda past sezgirlik ko‘rsatadi.

Kontekstual enkoderlar (mBERT). Ko‘p tilli BERT modeli gapdagi so‘zlararo munosabatlarni chuqur neyron tarmoqlari orqali vektor ko‘rinishida (embeddings) ifodalaydi [5:4171-4186] U kontekstni yaxshi anglasa-da, o‘zbek tilidagi formal shakllarning (masalan, kelishik qo‘shimchalarining) semantik qiymatini ba’zan e’tibordan chetda qoldiradi. Buni *“zar”* so‘zi misolida ko‘rishimiz mumkin. *Zar* so‘zining ma’nolari:

1. Oltin, tilla;
2. Oltin bilan teng, oltinga o‘xshash narsa haqida;
3. Tilla, pul, umuman boylik;
4. Zardo‘zlikda zarbof to‘n, zar do‘ppi va boshqalar tayyorlashda ishlatiladigan, oltin yoki kumush suvi yuritilgan, zarlangan ipak yoki sim; shunday ipak yoki simdan tikilgan kiyim;
5. Bezak yoki o‘rov uchun ishlatiladigan yupqa yaltiroq metall varag‘i yoki zarhallangan qog‘oz.

“Zar qog‘oz” yoki *“zarga o‘ralgan popuk”* birikmalarida mBERT kontekstual bog‘liqlik orqali bu yerdagi *“zar”* so‘zini “bezak/yaltiroq qog‘oz” (5-ma’no) ekanligini yaxshi anglaydi. Ammo ***“Zar qadrini zargar biladi”*** (maqol) va ***“xazinamda har qancha zar bo‘lsa”*** (tarixiy/badiiy matn) kabi holatlarda, mBERT formal shakllarning semantik qiymatini (masalan, tarixiy terminologiya) ba’zan e’tibordan chetda qoldirib, oddiy “pul” yoki “metal” ma’nolari bilan adashtirishi mumkin.

Gibrid Morph-mBERT. Bu yondashuv mBERTning kontekstual imkoniyatlarini simvolik morfologiya bilan birlashtiradi. Modelda kontekstual oqim va FST analizatori [3:57-62] orqali olingan simvolik oqim bitta umumiy vektorga

birlashtiriladi. *“o‘zga”* polisemik so‘zi yordamida ko‘rishimiz mumkin. *“O‘zga”* so‘zining ma’nolari:

1. Qarindoshlik, tug‘ishganlik va shu kabilar jihatdan aloqasi yo‘q odam; begona;
2. So‘zlovchi yoki so‘zlovchi va tinglovchidan boshqa odam;
3. O‘zi yoki o‘ziga mansub bo‘lmagan; notanish, yot;
4. Chiqish kelishigidagi so‘z bilan (O‘zi bog‘lanib kelgan so‘z bildirgan shaxs yoki narsadan boshqasiga ishora qiladi; -dan boshqa);
5. Odatdagidan boshqacha, o‘zgacha.

Bunda mexanizm quyidagicha bo‘ladi. *“Otabekdan o‘zga farzandimiz yo‘q”* gapida *“o‘zga”* so‘zi chiqish kelishigi (-dan) bilan bog‘lanib kelmoqda. Morph-mBERT modeli FST analizatori orqali kelishik tegi va uslubiy ko‘rsatkichni ajratib oladi [2:1–6;4:380-384]. Kontekstual oqim va simvolik oqim birlashtirilishi natijasida model bu yerdagi ma’noni aniq “bo‘lak/boshqa” (4-ma’no) deb belgilaydi.

1-jadval. Modellarning semantik qamrovi

Metod	So‘z	Qo‘llaniladigan asosiy model	Natija
SVM	O‘rtoq	n-gramlar (TF-IDF)	Kontekst bo‘lsa barqaror.
mBERT	Zar	Kontekstual embeddinglar	Kontekstni anglaydi, sirtqi shaklda adashadi.
Gibrid	O‘zga	Kontekst + Kelishik (-dan) + Uslub	Kelishikka bog‘liq ma’nolarda eng yuqori aniqlik.

Modellarning samaradorligi o‘zbek tiliga xos noaniqlik turlari bo‘yicha quyidagi jadvalda ko‘rsatilgan:

2-jadval. Statistik metodlarning funksional tahlili.

Metod turi	Ishlatiladigan asosiy modellar	Semantik samaradorligi
SVM	n-gramlar, TF-IDF va maqsadli lemma	Kam resursli holatlarda barqaror, ammo morfologik jihatdan zaif.
mBERT	Kontekstual embeddinglar	Kontekstni chuqur anglaydi, biroq sirtqi shakl o‘zgarishlarida adashadi.

Gibrid (Morph-mBERT)	Kontekst + Kelishik + Uslub + Idioma	Kelishikka bog‘liq ma’nolarda va idiomalarda eng yuqori aniqlikni ta’minlaydi.

Tadqiqotlar shuni ko‘rsatadiki, kelishik teglari, uslubiy teglar (scientific, fiction, publicistic, formal) va idioma teglarini integratsiya qilish mBERT modelining aniqligini o‘zbek tili uchun qo‘shimcha 4.6 ballga oshiradi.

Xulosa

O‘zbek tili matnlarida polisemantik birliklarni avtomatik tahlil qilish bo‘yicha o‘tkazilgan tadqiqotlar quyidagi fundamental xulosalarni shakllantirish imkonini beradi. O‘zbek tilida so‘zning semantik qiymatini aniqlashda faqatgina kontekstual qism-so‘z (subword) segmentatsiyasiga tayanib qolish yetarli emas. Kelishik qo‘shimchalari va so‘z yasovchi elementlar ma’no farqlashda statistik modellar uchun asosiy model vazifasini o‘taydi. Standart mBERT modeli kontekstni chuqur anglasa-da, morfologik o‘zgarishlarda adashishga moyil. Kontekstual kiritmalarni FST orqali olingan ochiq morfologik modellar (Morph-mBERT) bilan birlashtirish o‘zbek tili uchun model aniqligini qo‘shimcha 4.6 ballga oshirish imkonini berdi. polisemantik so‘zlarning ma’no ko‘lamini aniqlashda nutq uslubi (registr) va idiomatik teglarning (idiom flag) integratsiyasi sohalararo semantik chalkashliklarni bartaraf etuvchi eng samarali statistik vosita hisoblanadi. Taklif etilgan statistik va gibrid metodologiya mashina tarjimasi, semantik qidiruv va NLP tizimlarining quyi bosqich komponentlarida ma’no noaniqliklarini kamaytirishda muhim amaliy ahamiyatga ega.

Foydalanilgan adabiyotlar ro‘yxati

1. Agirre E., Edmonds P. Word Sense Disambiguation: Algorithms and Applications // Elektron resurs. – URL: https://arxiv.org/abs/2205.06072

2. Akhundjanova A., Talamo L. Universal Dependencies Treebank for Uzbek // Proceedings of the Third Workshop on Resources for Under-Resourced Languages. – 2025. – P. 1–6.

3. Boltayevich E.B., va boshq. The Problem of POS Tagging and Stemming for Agglutinative Languages // 2023 8th International Conference on Computer Science and Engineering (UBMK). – 2023. – P. 57–62.

4. Boltayevich E.B., va boshq. A Morphological Tagging Model of the Uzbek Language in the Universal Dependencies Format // 2025 10th International Conference on Computer Science and Engineering (UBMK). – 2025. – P. 380–384.

5. Devlin J., va boshq. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.

6. Elov B.B., va boshq. The pipeline processing of NLP // E3S Web of Conferences. – 2023. – Vol. 413. – Art. 03011