I SHO‘BA. TABIIY TILNI QAYTA ISHLASH (NLP)

SO‘ZLARNI JOYLASHTIRISH MODELLARI YORDAMIDA O‘ZBEK TILI LEKSIK SINONIMLARINI AJRATIB OLISH: LEMMA VA TOKEN YONDOSHUVLARI TAHLILI

Madatov Xabibulla Axmedovich,
Fizika-matematika fanlari nomzodi, dotsent,
habi1972@mail.ru
Abu Rayhon Beruniy nomidagi Urganch davlat Universiteti

Khajibaeva Surayyo Maxmudjonovna,
o‘qituvchi
surayyo.khajiboyeva@gmail.com
Abu Rayhon Beruniy nomidagi Urganch davlat Universiteti

Annotatsiya. Tabiiy tilni qayta ishlash sohasida so‘zni joylashtirish modellari yordamida leksik sinonimlarni ajratib olish masalasi haligacha dolzarbligicha qolmoqda. Bu masala lug‘atga asoslangan usuldan ko‘ra afzalroqdir. Shu sababdan, ushbu maqolada so‘zlarni joylashtirish modellaridan Word2Vec va FastText yordamida nomzodlar k=10, k=50 bo‘lgan holatlarda sinonimlarni ajratib olish jarayoni ko‘rib chiqildi. Har bir maqsad so‘z uchun nomzodlar orasidan UzWordNet sinsetlariga asoslangan holatda leksik sinonimlar ajratib olindi. Xuddi shu ajratib olish va sintezga asoslangan filtrlash protsedurasi ikkita normallashtirish sozlamalarida – token asosidagi (sirt shakllari) va lemma asosidagi (kanonik shakllar) – ko‘rib chiqildi. Olingan natijalar shuni ko‘rsatadiki, agglyutinativ xususiyatga ega o‘zbek tili uchun Word2Vec kichik sinonim nomzodlari k=10 bo‘lganda, n gramlarga asoslangan FastText modeli esa yuqori sinonim nomzodlari olinganda, ya’ni k=50 va undan katta holatlarda ko‘proq sinonim nomzodlarini o‘z ichiga oladi.

Kalit so‘zlar: *leksik sinonimlar, sinset, kosinus o‘xshashlik, k-ta yaqin qo‘shnilar, maqsad so‘z.*

Abstract. The problem of extracting lexical synonyms using word embedding models in the field of natural language processing is still relevant. This problem is

preferable to the dictionary-based method. For this reason, in this article, the process of extracting synonyms from word embedding models using Word2Vec and FastText in the cases where the candidates are k=10, k=50 was considered. Lexical synonyms were extracted from the candidates for each target word based on UzWordNet synsets. The same extraction and synthesis-based filtering procedure was considered in two normalization settings – token-based (surface forms) and lemma-based (canonical forms). The obtained results show that for the agglutinative Uzbek language, the Word2Vec model contains more synonym candidates when k=10, while the FastText model based on n-grams contains more synonym candidates when k=50 and larger when high synonym candidates are obtained.

Keywords: lexical synonyms, synset, cosine similarity, k-nearest neighbours, target word.

1. Kirish

Leksik sinonimlar – talaffuzi va yozilishi har xil, ammo ma’nosi bir xil yoki bir-biriga juda yaqin bo‘lgan, nutqni boyitish va jozibadorligini oshirishga xizmat qiladigan so‘zlardir[1]. Ular qidiruv tizimlar, savol-javob tizimlari va matnni soddalashtirish kabi ko‘plab NLP vazifalari uchun amaliy talabdir. Lug‘atga asoslangan sinonimlarni ajratib olishning o‘zi yetarli bo‘lmasligi mumkin, chunki sinonim sifatida ko‘rsatilgan so‘zlar har doim ham bir xil kontekstga mos kelmaydi yoki haqiqiy matnda qayd etilmaydi. Shuning uchun biz taqsimot semantikasidan foydalangan holda ma’lumotlarga asoslangan yondashuvni qo‘llaymiz. Ma’lumki, o‘zbek tili agglyutinativ til bo‘lgani uchun bitta lemmaning ko‘plab token shakllari mavjud. Bu esa tildagi so‘zlarni vektor fazoda tasvirlash jarayonida sun’iy ishishga olib keladi. Shu muammoni oldini olish maqsadida biz korpus ustida tadqiqotni ikkita turga: token va lemma holatlarida ko‘rib chiqdiq.

2. Ma’lumotlar to‘plami va usul.

O‘zbek tili kam resursli tillar qatoriga kirgani bois, ochiq manbada yetarlicha katta miqdordagi korpuslar mavjud emas. Ushbu jarayonda maktab korpusi ustida bazis so‘zlarni ajratib olish [2] va o‘rta maktab o‘quvchilari uchun tegishli adabiyotlarni tanlash kabi ko‘plab tadqiqotlar o‘z samarasini berayotgani bois undan foydalandik. Lekin bizga ma’lumki, so‘zlarni joylashtirish(word embedding) modellari juda katta korpus bilan ishlashda yaxshi natijani ko‘rsatadi. Ushbu modellar har bir so‘z uchun “atrofidagi so‘zlar”(window) dan foydalanadi. Katta korpusda bir so‘z turli mavzu va uslublarda minglab marta uchraydi, natijada model uning ma’nosini barqarorroq tiklaydi. Shu sababdan, BertBEK modelini [4] testlash uchun ishlatilgan Wikipedia, daryo.uz va boshqa manbalardan iborat korpusdan foydalanishni maqsad qildik. Quyidagi 1-jadvalda bizning tadqiqotimizda ishlatgan ma’lumotlar omborlari haqida aks ettirilgan:

Jadval 1. Har bir ma’lumotlar omboridagi tokenlar va gaplar soni

Ma’lumotlar ombori nomi	Tokenlar soni	Gaplar soni
Vikipediya manbasi	27711575	2481232
Veb manbalar	33169827	2389912
Maktab korpusi	1408830	154239

Taqsimotga asoslangan semantik modellar, ayniqsa Word2Vec va FastText, matndan semantik bog‘liqlikni aniqlash uchun standart vositalarga aylandi. Zich vektor tasvirlarini (joylashtirish) o‘rganish orqali ushbu modellar o‘xshash kontekstlarga ega so‘zlarni bir-biriga yaqin umumiy vektor maydoniga joylashtiradi va bu eng yaqin qo‘shni qidiruvi orqali sinonimlarni topish imkonini beradi. Biroq, joylashtirish maydonidagi eng yaqin qo‘shnilar har doim ham haqiqiy sinonimlarga mos kelmaydi: ular ko‘pincha bitta to‘plamga tegishli so‘zlar, gipernimlar, antonimlar va morfologik jihatdan olingan variantlarni o‘z ichiga oladi. Bu qiyinchilik, ayniqsa, agglyutinativ tillar uchun yaqqol ko‘rinib turibdi, bu yerda boy morfologiya ko‘plab sirt shakllarini hosil qiladi va qo‘shnilik ro‘yxatlarini haqiqiy leksik o‘rinbosarlar o‘rniga flektsion yoki derivativ shakllar bilan to‘ldirishi

mumkin. Shuning uchun, o‘zbek tilida sinonim induksiyasi sinonimiyani kengroq semantik bog‘liqlikdan ajratish uchun ham kuchli taqsimot modellashtirishni, ham printsipial filtrlash strategiyasini talab qiladi.

Ushbu muammoni hal qilish uchun ushbu ish joylashtirishga asoslangan nomzod yaratishni sinsetga asoslangan validatsiya bilan birlashtiradi. So‘zlarni joylashtirish Vikipediya, Daryo.uz va Maktab korpusidan (1–11-sinflar) to‘plangan katta va stilistik jihatdan xilma-xil o‘zbek korpusida o‘rgatiladi, bu esa faqat maktab darsliklariga qaraganda kengroq leksik qamrovni ta’minlaydi. Har bir maqsadli so‘z uchun biz Word2Vec va FastText modellaridan Top-k eng yaqin qo‘shnilarini (k = 10 va 50) olamiz. Keyin nomzod qo‘shnilar o‘zbek WordNet[5] yordamida filtrlanadi: faqat maqsad bilan WordNet sintezini baham ko‘radigan so‘zlar saqlanadi va sinonim sifatida ko‘rib chiqiladi. Bu qadam ajratish jarayonini aniq leksik-semantik resursga asoslaydi va taqsimot o‘xshashligi bilan kiritilgan noto‘g‘ri ijobiylarni kamaytiradi. Turkiy tillar uchun tillararo so‘zlarni joylashtirish(word embedding)lar yaratib, kam-resursli tillarda semantik moslikni kuchaytirish va leksik resurslarni transfer qilish[6] imkonini ko‘rsatish kabi ishlar FastText modeli yordamida ko‘rib chiqilgan. Chunki, FastText modeli n-gram asosida so‘zni qism so‘zlarga bo‘lib o‘rganishi, ayniqsa o‘zbek tili uchun juda qulay hisoblanadi.

Tadqiqotning asosiy metodologik o‘lchovi – bu token asosidagi va lemma asosidagi tasvirlarni taqqoslash. Token sharoitida joylashtirish va qo‘shnilar sirt so‘z shakllari ustida hisoblanadi, bu haqiqiy korpusdan foydalanishni aks ettiradi, shuningdek, morfologik o‘zgarishlarni ham qayd etadi. Lemma sharoitida matn kanonik shakllarga normallashtiriladi, bu esa morfologik shovqinni kamaytirishi va leksik jihatdan mazmunli o‘rinbosarlarni olishni yaxshilashi mumkin. Tadqiqot morfologik normallashtirishning o‘zbek tili uchun sinonimlarni topishga qanday ta’sir qilishini nazorat ostida tahlil qiladi.

Natijada, olingan ma’lumotlar to‘plami to‘rtta asosiy shart ostida WordNet sinonimlari bilan moslashtirilgan joylashtirishdan olingan sinonim nomzodlaridan iborat: Word2Vec token, Word2Vec lemma, FastText token va FastText lemma, ularning har biri k = 10 va k = 50 uchun hisobot qilingan. Ushbu resurs o‘zbek leksik semantikasi bo‘yicha takrorlanadigan tajribalarni qo‘llab-quvvatlaydi, sinonimlarni topish uchun joylashtirish konfiguratsiyalarini tizimli baholash imkonini beradi va ishonchli sinonim bilimlarini talab qiladigan o‘zbek tiliga xos NLP vositalarini ishlab chiqishga hissa qo‘shadi.

Olingan natijalar

Mazkur jarayon token va lemma ko‘rinishlari uchun alohida bajarilib, morfologik vakillanishning sinonim topishga ta’siri taqqoslandi. Quyida 2-jadvalda lemma va token shaklida Word2Vec va FastText modellari yordamida 10, 50 ta yaqin qo‘shnilar asosida sinonim nomzodlari orasidan WordNet asosida leksik sinonimlar qiyosiy baholash jadvali keltirilgan.

Jadval 2. Word2Vec&FastText modellari qiyosiy tahlili

So‘z turi shakli	Model nomi	Baholash (%)	OOV(%)
Token	Word2Vec(10 ta yaqin qo‘shni)	5,50	67,98
	Word2Vec(50 ta yaqin qo‘shni)	8,58	55,28
	FastText(10 ta yaqin qo‘shni)	4,19	75,07
	FastText(50 ta yaqin qo‘shni)	10,23	48,81
Lemma	Word2Vec(10 ta yaqin qo‘shni)	5,24	68,93
	Word2Vec(50 ta yaqin qo‘shni)	7,69	59,22
	FastText(10 ta yaqin qo‘shni)	5,42	67,98
	FastText(50 ta yaqin qo‘shni)	11,08	44,08

Token sozlamalarida Word2Vec 5.50% (Top-10) va 8.58% (Top-50) bahosiga erishadi, OOV ko‘rsatkichlari mos ravishda 67.98% va 55.28% ni tashkil qiladi. Xuddi shu token sozlamalarida FastText 4.19% ga (Top-10) yetadi va 10.23% ga (Top-50) yaxshilanadi, OOV esa 75.07% dan 48.81% gacha pasayadi. Lemma sozlamalarida Word2Vec 5.24% (Top-10) va 7.69% (Top-50) ga ega bo‘ladi, OOV

qiymatlari esa 68.93% va 59.22% ni tashkil qiladi. Lemmaga asoslangan FastText 5.42% (Top-10) va eng yaxshi umumiy ball 11.08% (Top-50) ni beradi, eng past OOV esa Top-50 da 44.08% ni tashkil qiladi. Ikkala shaklda ham k ta yaqin qo‘shnilar hajmini 10 dan 50 gacha oshirish ikkala model uchun ham baholash ballarini doimiy ravishda yaxshilaydi. Yaxshilanish FastText uchun, ayniqsa Top-50 da eng kuchli bo‘lib, bu n-gramga asoslangan ko‘proq nomzodlar ko‘rib chiqilganda yuqori WordNetga moslashtirilgan sinonimlarni ajratib olishni ko‘rsatadi.

3. Xulosa

Xulosa qilib aytganda, natijalar shuni ko‘rsatadiki, lemma asosidagi ishlov berish va Top-50 qo‘shnilariga ega FastText modeli sinovdan o‘tganlar orasida eng samarali natijalarni ko‘rsatdi. Bu FastText modeli agglyutinativ tillar uchun samarali ekanligini yana bir bor ko‘rsatadi. OOV muammosini kamaytirishda Word2Vec modeliga nisbatan FastText modeli ustun bo‘lib, n-gramlarga asoslangani tufayli kam uchraydigan va morfologik variantlarga boy so‘zlar uchun ham vektor hosil qila oladi. Shu sababdan, FastText so’zlarni joylashtirish modeli yordamida leksik sinonimlarni ajratib olish Word2Vec modeliga qaraganda ancha samarali hisoblanadi.

Foydalanilgan adabiyotlar:

1. M. Abjalova, “O‘zbek tilidagi sinonimlar tasnifi,” Respublika ko‘p tarmoqli ilmiy-amaliy konferensiyasi to‘plami, vol. 1, no. 2, pp. 352–361, Oct. 2023, doi: 10.5281/zenodo.10206034.

2. S. Khajibaeva, “Maktab korpusi asosida bazis so‘zlarni ajratib olish va tahlil qilish,” Xalqaro ilmiy-amaliy konferensiyalar, vol. 1, no. 3, pp. 300–302, Oct. 2025.

3. K. Madatov and S. Sattarova, “Neural Network-Based Approach to Literary Selection for Grades 5–9,” in Proc. 2025 IEEE 26th Int. Conf. of Young

Professionals in Electron Devices and Materials (EDM), Jun. 2025, pp. 2180–2184, doi: 10.1109/EDM65517.2025.11096885.

4. E. Kuriyozov, D. Vilares, and C. Gómez-Rodríguez, “BERTbek: A Pretrained Language Model for Uzbek,” in Proc. 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL) @ LREC-COLING 2024, May 2024, pp. 33–44.

5. N. Abdurahmonova, “UzWordNet: A Lexical-Semantic Database for the Uzbek Language,” in Proceedings of the 10th Global WordNet Conference (GWC), Wroclaw, Poland, 2019, pp. 8–15.

6. Kuriyozov, Y. Doval, and C. Gómez-Rodríguez, “Cross-Lingual Word Embeddings for Turkic Languages,” in Proc. 12th Language Resources and Evaluation Conf. (LREC), Marseille, France, May 2020, pp. 4054–4062.