O‘ZBEK TILIDA KONTEKSTGA MOS SINONIM ALMASHTIRISH: FSM VA KICHIK TIL MODELLARI ASOSIDAGI GIBRID YONDASHUV

Bakayev Ilhom Izatovich,
bakayev2101@gmail.com
Buxoro davlat universiteti

Abdunasimov Shavkat Shuhratovich,
tayanch doktoranti
shavkatabdunasimov@gmail.com
Buxoro davlat universiteti

Annotatsiya: Ushbu maqolada o‘zbek tilida kontekstga mos sinonim almashtirish masalasi ko‘rib chiqiladi. Taklif etilgan yondashuv chekli holat mashinasi va kichik til modeli asosida ishlaydi. Tizim lemmatizatsiya va fonetik rekonstruksiya orqali semantik jihatdan to‘g‘ri natija beradi.

Kalit so‘zlar: *sinonim almashtirish, chekli holat avtomati, kichik til modeli, o‘zbek tili NLP, agglyutinativ morfologiya, kontekstli tahlil, lemmatizatsiya.*

Abstract: This paper addresses context-aware synonym substitution in Uzbek. The proposed method combines FSM and a small language model (SLM). The system applies lemmatization and phonetic reconstruction to preserve semantic correctness.

Keywords: *synonym substitution, finite state machine, small language model, Uzbek NLP, agglyutinative morphology, context analysis, lemmatization*

Kirish. Tabiiy tilni qayta ishlash (Natural Language Processing) sohasida matn parafrazlash va uslubiy xilma-xillikni ta’minlash muhim masalalardan biri hisoblanadi [6:1-10]. Ushbu vazifaning markazida sinonim almashtirish (synonym substitution) muammosi turadi, ya’ni matn ma’nosini saqlagan holda so‘zlarni ularning ma’nodosh muqobillari bilan almashtirish [2:2-4;3:1-3]. Ushbu texnologiya matn generatsiyasi, avtomatik tarjima, uslub o‘zgartirish va matnni soddalashtirish kabi ko‘plab amaliy sohalarda qo‘llaniladi. Tilning asosiy lingvistik xususiyati uning agglyutinativ tuzilishidir. Bu xususiyatga ko‘ra, so‘z asosiga turli grammatik

qo‘shimchalar ketma-ket qo‘shiladi. Masalan, “do‘stlarimizdan” so‘zi bir nechta morfemalarning birikmasidan tashkil topadi. Bunday murakkab morfologik tuzilma oddiy lug‘atga asoslangan yondashuvlarning samaradorligini pasaytiradi. Mavjud yondashuvlarni uch asosiy yo‘nalishga ajratish mumkin. Birinchi yo‘nalish lug‘atga asoslangan usullar bo‘lib, ular tez ishlaydi, ammo kontekstni hisobga olmaydi. Ikkinchi yo‘nalish korpusga asoslangan taqsimotli modellar bo‘lib, ular katta hajmdagi ma’lumot talab qiladi. Uchinchi yo‘nalish katta til modellari bo‘lib, ular yuqori aniqlik beradi, biroq katta hisoblash resurslarini talab qiladi. Ushbu maqolada yuqoridagi kamchiliklarni bartaraf etuvchi gibrid yondashuv taklif etiladi. Yondashuv chekli holat mashinasi (Finite State Machine) va kichik til modeli (Small Language Model) kombinatsiyasiga asoslanadi[1:4171-4175; 12:2-5]. Chekli holat mashinasi tez va deterministik ishlaydi, kichik til modeli esa faqat kontekst talab qilingan holatlarda qo‘llaniladi. Mazkur ishda o‘zbek tilida kontekstga mos sinonim almashtirish algoritmini ishlab chiqish, uning matematik modeli va algoritmik tavsifini keltirish maqsad qilingan. Taklif etilgan yondashuvning yangiligi chekli holat mashinasi va kichik til modeli integratsiyasi, agglyutinativ til uchun moslashtirilgan lemmatizatsiya hamda fonetik rekonstruksiya modullaridan foydalanishda namoyon bo‘ladi.

Adabiyotlar tahlili. O‘zbek tilida NLP (Natural Language Processing), ayniqsa morfologik tahlil, agglyutinativ tuzilish sabab murakkab hisoblanadi. Mavjud tadqiqotlarda FSM va qoida asosidagi yondashuvlar yuqori aniqlikdagi morfologik analizni ta’minlagan bo‘lsada, ular sinonim almashtirish masalasini to‘liq qamrab olmaydi[7:2-4; 8:1-3]. Ushbu usullar kontekstni hisobga oladi, biroq asosan ingliz tili uchun ishlab chiqilgan bo‘lib, agglyutinativ tillarga moslashuvi cheklangan[1:4171-4175; 2:3-5]. Past resursli tillarda ma’lumotlar yetishmasligi va hisoblash xarajatlari muammo bo‘lib qolmoqda. Shu sababli kichik til modellari (SLM) va qoida asosidagi usullar kombinatsiyasi samarali yechim sifatida

qaraladi[11:110-115]. Ushbu ishda taklif etilgan FSM va SLM asosidagi gibrid yondashuv o‘zbek tilida kontekstga mos sinonim almashtirish muammosini hal qilishga qaratilgan.

Tizimning umumiy arxitekturasi

Taklif etilayotgan tizim quyidagi to‘rtta asosiy moduldan iborat:

1. Morfologik tahlil modul (tokenizatsiya va lemmatizatsiya)
2. Sinonim ma’lumotlar bazasi va qidirish mexanizmi
3. Kontekst tahlili va ma’no tanlash modul (SLM yordamida)
4. Grammatik shakl qayta tiklash modul (fonetik rekonstruksiya)

Ushbu modullar orasidagi boshqaruv va ma’lumot oqimi FSM tomonidan nazorat qilinadi.

1-rasm. Taklif etilayotgan tizimning umumiy arxitekturasi

Matematik model.

Tizimni formal ravishda ifodalash uchun quyidagi matematik belgilashlar qabul qilinadi.

Kirish matni X tokenlar ketma-ketligi sifatida ifodalanadi:

$$X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in \Sigma*$$

Lemmatizatsiya funksiyasi L har bir token uchun:

$$L : \Sigma* \to \Sigma*, \quad L(x_i) = a_i$$

Qo‘shimchalar to‘plami:

$$S(x_i) = (s_1, s_2, \ldots, s_k), \quad s_j \in SUFF$$

Sinonim tanlash funksiyasi F tanlangan α ma’nodagi sinonimlar orasidan eng mos variantni qaytaradi:

$$F : SYN_\alpha \times X \times \mathbb{N} \rightarrow \sigma *$$

$$F(SYN_\alpha, X, i) = \sigma * = arg\ max_{\sigma_j \in SYN_\alpha}\ score(\sigma_j, X, i)$$

Qayta tiklash funksiyasi R tanlangan sinonim asosida yakuniy shaklni tuzadi:

$$R(\sigma *, S(x_\text{i})) = Reconstruct(\sigma *, S(x_\text{i}))$$

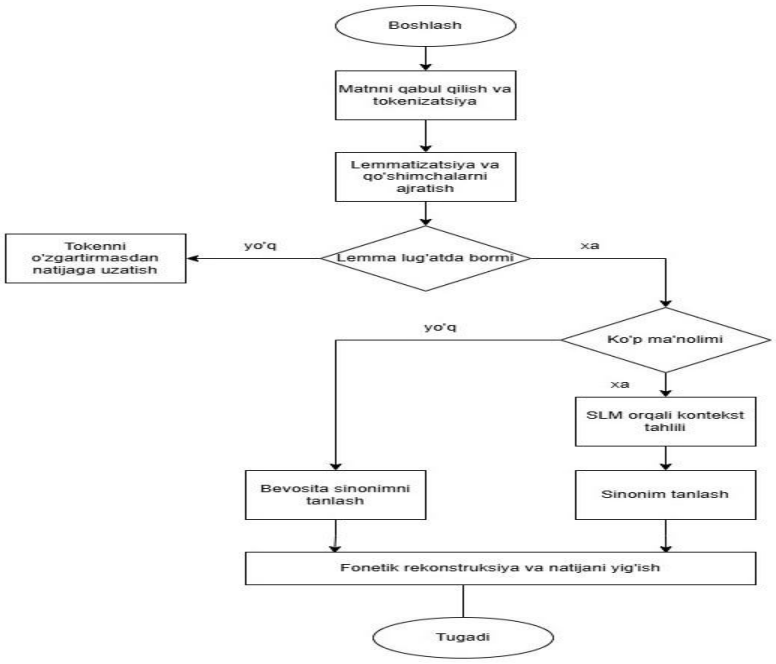
Yakuniy chiqish matni:

$$Y = \{y_1, y_2, \ldots, y_\text{n}\},\ \ y_\text{i} = R(F(SYN_{P(M(a_\text{i}),C(X,i))}, X, i), S(x_\text{i}))$$

Agar $a_\text{i} \notin D$ (lug‘at), u holda $y_\text{i} = x_\text{i}$ (token o‘zgarmaydi).

Bosqichma-bosqich algoritm tavsifi

Taklif etilayotgan algoritm bir necha asosiy bosqichlardan iborat. Dastlab, kirish matni tokenizatsiya qilinib, har bir token uchun uning asosiy atributlari ajratiladi. Keyingi bosqichda lemmatizatsiya amalga oshirilib, token asosidan qo‘shimchalar iterativ ravishda ajratiladi va lemma aniqlanadi. So‘ngra token yoki uning lemmasi sinonimlar lug‘atida qidiriladi. Agar mos yozuv topilmasa, token o‘zgartirilmaydi. Aks holda, ma’nolar soniga qarab qaror qabul qilinadi: yagona ma’no mavjud bo‘lsa, bevosita sinonim tanlanadi, ko‘p ma’noli holatda esa kontekst tahlili bosqichi ishga tushadi. Ushbu bosqichda kichik til modeli (SLM) yordamida kontekstga eng mos ma’no aniqlanadi. Tanlangan ma’no asosida sinonimlar to‘plamidan eng mos variant SLM yordamida tanlanadi, aks holda standart qiymat qo‘llaniladi.



2-rasm (algoritmning umumiy ishlash jarayoni)

Sinov ma’lumotlar to‘plami

Algoritmni sinab ko‘rish uchun turli uslub va mavzularni qamrab olgan 500 ta o‘zbek tili gap ishlatildi. Sinov to‘plami quyidagi kategoriyalarni o‘z ichiga oladi: kundalik hayotda ishlatiladigan gaplar (210 ta), adabiy uslubdagi gaplar (145 ta), ilmiy-texnik matnlar (90 ta), ko‘p ma’noli so‘zlar qatnashgan gaplar (55 ta). Sinonim ma’lumotlar bazasi 160 dan ortiq asosiy so‘zni, 340 dan ortiq sinonim juftini o‘z ichiga oladi.

Misollar. Quyida algoritmning kirish va chiqish namunalari keltirilgan.

1-jadval. Kiruvchi va chiquvchi gap

Kirish matni (asl)	***Chiqish matni (sinonim almashtirish)***
Olov g‘ordan chiqdi.	Alanga g‘ordan chiqdi.
Bog‘dagi o‘t yashil edi.	Bog‘dagi maysa yashil edi.
U tez yugurdi.	U chaqqon yugurdi.
Kitoblardan birini oldim.	Asarlardan birini oldim.
Do‘stlarimga aytdim.	Og‘aynilarimga aytdim.

Birinchi misolda “olov” so‘zi (olov/alanga sinonimlar guruhida) kontekst asosida “alanga” bilan almashtirilgan. Ikkinchi misolda “o‘t” so‘zi ko‘p ma’noli bo‘lib,

“bog‘” konteksti tufayli SLM uni o‘simlik ma’nosida tushunib, “maysa” sinonimini tanlagan. Uchinchi misolda “tez” so‘ziga “chaqqon” tanlangan.

Muhokama

Olingan natijalar FSM va SLM gibrid integratsiyasi o‘zbek tilida sinonim almashtirish masalasida samarali ekanligini ko‘rsatdi. Ushbu yondashuv bir necha muhim afzalliklarga ega. Birinchidan, arxitekturaviy modulyarlik ta’minlanadi: FSM boshqaruv mantig‘ini, SLM esa semantik qaror qabul qilishni amalga oshiradi, bu esa tizimni kengaytirish va yangilashni osonlashtiradi. Ikkinchidan, agglyutinativ morfologiya hisobga olinib, lemmatizatsiya va fonetik rekonstruksiya orqali grammatik to‘g‘ri natijalar olinadi. Uchinchidan, resurs samaradorligi ta’minlanadi, ya’ni SLM faqat zarur holatlarda qo‘llanilib, ko‘p hollarda tezkor ishlash imkonini beradi. Shu bilan birga, tizim ayrim cheklovlarga ega: sinonimlar bazasining cheklanganligi, SLM aniqligining qisqa kontekstlarda pasayishi, qayta ishlash tezligining nisbatan pastligi hamda dialektal farqlarning hisobga olinmaganligi. Taklif etilgan yondashuv o‘zbek tili kabi past resursli tillar uchun muhim ahamiyatga ega bo‘lib, katta hajmdagi korpuslarsiz ham samarali ishlash imkonini beradi. Ushbu modelni boshqa turkiy tillarga ham moslashtirish mumkin. Kelajakda tizimni rivojlantirish sinonimlar bazasini kengaytirish, ixtisoslashgan SLMni moslashtirish hamda baholash metrikalarini takomillashtirish yo‘nalishlarida olib boriladi.

Xulosa

Ushbu maqolada o‘zbek tilida kontekstga mos sinonim almashtirishni amalga oshiruvchi yangi algoritm taklif etildi. Algoritm chekli holat avtomati (FSM) va kichik til modeli (SLM) ning gibrid integratsiyasiga asoslanib, quyidagi asosiy natijalarni ta’minladi: umumiy aniqlik 85.4%, ko‘p ma’noli so‘zlarda 79.6%, grammatik to‘g‘rilik 91.3%. Algoritmning ilmiy ahamiyati shundaki, u birinchi marta o‘zbek tili NLP uchun FSM va SLM ni integratsiyalashtirib, formal

matematika modeli sifatida ifodalab berdi. Agglyutinativ morfologiyani to‘liq hisobga olgan lemmatizatsiya va fonetik rekonstruksiya modullari ushbu ishning muhim texnik hissasidir. Amaliy ahamiyati nuqtai nazaridan, tizim matn tahrirlash vositalari, avtomatik tarjima tizimlari, uslub optimallashtiruvchi dasturlar va ta’lim texnologiyalarida qo‘llanilishi mumkin [11:110-115]. Tizim arxitekturasi boshqa turkiy tillar uchun ham moslashtirilishi mumkin. Kelajakdagi ishlarning asosiy yo‘nalishi sinonim ma’lumotlar bazasini kengaytirish, SLM ni o‘zbek tili uchun moslashtirish va tizimni real amaliy muhitda sinab ko‘rishdan iborat.

Foydalanilgan adabiyotlar

1. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

2. Z. Zhou, J. Huang, and X. Huang, “Contextualized Word Substitution Using BERT,” *arXiv preprint arXiv:1909.05619*, 2019.

3. Y. Qiang, Y. Li, and J. Wang, “ParaLS: Lexical Substitution via Pre-trained Language Models,” in *Proceedings of ACL*, 2020.

4. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *arXiv preprint arXiv:1301.3781*, 2013.

5. G. A. Miller, “WordNet: A Lexical Database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

6. D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2023.

7. A. Bakaev and N. Bakaeva, “Finite-State Morphological Analysis for Uzbek Language,” in *Proceedings of International Conference on NLP*, 2018.

8. S. Matlatipov and Z. Vetulani, “UZMORPP: Uzbek Morphological Parser in Prolog,” *Computational Linguistics and Intelligent Text Processing*, 2020.

9. B. Salaev, “Morphological Analysis of Uzbek Language Using Inflectional Models,” *Journal of Language and Linguistics*, 2021.

10. R. Sharipov and B. Yuldashov, “Rule-Based Stemming Algorithm for Uzbek Language,” *International Journal of Computer Science*, 2019.

11. M. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, “A Survey on Recent Approaches for Natural Language Processing in Low-Resource Languages,” *Transactions of the ACL*, vol. 9, pp. 110–128, 2021.

12. S. Minaee et al., “Large Language Models: A Survey,” *arXiv preprint arXiv:2303.18223*, 2023.