



IMPROVING SEARCH FOR SPELLING VARIANTS AND INFLECTED FORMS IN AN UZBEK ELECTRONIC TERMINOLOGICAL DICTIONARY

Janpulatova Dilorom

English Teacher, Philology School, Mirzo Ulugbek District

Abstract

This paper proposes a search model for an Uzbek electronic terminological dictionary that accounts for spelling variants and affixed forms. Based on research in information retrieval and Uzbek morphological analysis, the framework grounds the stages of spelling normalization, morphological candidate identification, and their verification against the terminological dictionary. The model is designed to preserve the dictionary form of multi-word terms and prioritize exact matches. The work is theoretical and methodological in nature; the software implementation of the model and its experimental evaluation pertain to the next stage.

Keywords

terminological dictionary, information retrieval, morphological analysis, lemmatization, finite-state automaton, Uzbek language.

O'ZBEKCHA ELEKTRON TERMINOLOGIK LUG'ATDA IMLO VARIANTLARI VA QO'SHIMCHALI SHAKLLAR BO'YICHA QIDIRUVNI TAKOMILLASHTIRISH

Annotatsiya. Tezisdagi o'zbekcha elektron terminologik lug'atda imlo variantlari va qo'shimchali shakllarni hisobga oluvchi qidiruv modeli taklif etiladi. Axborot qidiruvi hamda o'zbek tili morfologik tahliliga oid tadqiqotlar asosida yozuvni me'yorlashtirish, morfologik nomzodlarni aniqlash va ularni terminologik lug'at bilan tekshirish bosqichlari asoslanadi. Modelda ko'p so'zli terminning lug'aviy shaklini saqlash va aniq



“KELAJAK TEKNOLOGIYALARI VA SUN’IY INTELLEKT”
nomli respublika ilmiy-amaliy masofaviy konferensiyasi
VOLUME-1, ISSUE-7, 2026

moslikka ustuvorlik berish ko‘zda tutiladi. Ish nazariy-metodik xususiyatga ega; modelni dasturiy amalga oshirish va tajribada baholash keyingi bosqichga tegishli.

Kalit so‘zlar: terminologik lug‘at, axborot qidiruvi, morfologik tahlil, lemmatizatsiya, chekli holatli avtomat, o‘zbek tili.

Kirish

Inglizcha-o‘zbekcha informatika terminlari lug‘atida foydalanuvchi so‘rovini tegishli lug‘at yozuviga bog‘lash muhim amaliy vazifadir. “Algoritm”, “algoritmlar” va “algoritmning” shakllari bir lug‘aviy birlikka murojaat qilishi mumkin. Tutuq belgisining turli kodlar bilan kiritilishi ham belgilar ketma-ketligini o‘zgartiradi. Tezisning maqsadi ushbu farqlarni hisobga oluvchi, terminning asl yozuvi va birikma tarkibini saqlaydigan qidiruv tartibini ishlab chiqishdan iborat.

Mavzuga oid tadqiqotlar

C. D. Manning, P. Raghavan va H. Schütze stemming bilan lemmatizatsiya o‘rtasidagi farqni tushuntiradi: birinchisi ko‘pincha evristik qisqartirishga, ikkinchisi esa lug‘at va morfologik tahlilga tayanadi. Mualliflar so‘z shakllarini umumlashtirish ayrim so‘rovlarda noo‘rin mosliklarni ko‘paytirishi mumkinligini ham ko‘rsatadi [1]. Shu sabab bu tezisda morfologik ishlov aniq qidiruvdan keyingi bosqich sifatida taklif etiladi.

K. R. Beesley va L. Karttunen morfotaktik hamda morfofonologik qoidalarni chekli holatli vositalar yordamida modellashtirish asoslarini yoritgan [2]. M. Sharipov va U. Salaev o‘zbekcha affikslar ketma-ketligini chekli holatli avtomat bilan ifodalab, lug‘atsiz stemming yondashuvini taqdim etgan [3]. U. Salaevning UzMorphAnalyser ishida esa o‘zbekcha so‘zlarning stemmingi, lemmatizatsiyasi va morfologik axborotini aniqlash, jumladan morfofonetik istisnolar ko‘rib chiqilgan [4].

Ushbu ishlarga tayangan holda, mazkur tezis umumiy morfologik analizator yaratishdan ko‘ra, tahlil natijalarini terminologik qidiruvga ulash tartibini taklif qiladi.



“KELAJAK TEKNOLOGIYALARI VA SUN’IY INTELLEKT”
nomli respublika ilmiy-amaliy masofaviy konferensiyasi
VOLUME-1, ISSUE-7, 2026

Qo‘shimchalarni ajratish mavjud yondashuvlarga asoslanadi; lug‘at yozuvini tanlash va birikmani saqlash esa qidiruv modulining vazifasi sifatida belgilanadi.

Taklif etilayotgan qidiruv modeli

Model uch asosiy qatlamdan iborat: yozuvni me’yorlashtirish, morfologik nomzodlarni hosil qilish va lug‘atga moslashtirish. Terminologik bazada termin identifikatori, o‘zbekcha bosh shakl, inglizcha muqobil, ta’rif, tasdiqlangan variantlar va manba saqlanadi. Asl yozuv bilan qidiruv kaliti alohida maydonlarda yuritiladi. Bu lug‘at matnini o‘zgartirmasdan turib variantlar bo‘yicha qidirish imkonini beradi.

Birinchi bosqichda so‘rovning asl nusxasi saqlanadi, ishchi nusxasi esa Unicode NFC shakliga keltiriladi. NFC kanonik ekvivalent belgilar ketma-ketligi uchun yagona ifoda olishga xizmat qiladi [5]. Biroq apostrofga o‘xshash barcha belgilar NFC orqali tenglashmaydi. Shu bois so‘z ichidagi tutuq va o‘, g‘ harflari tarkibidagi belgi variantlari uchun alohida almashtirish qoidalari belgilanadi. Bixillashtirish faqat qidiruv kalitida bajariladi; belgi butunlay o‘chirilmaydi.

Ikkinchi bosqichda asl so‘rov lug‘atdagi bosh shakl bilan aynan solishtiriladi. Moslik topilmasa, so‘rov va lug‘at yozuvining bir xil qoidalar bilan olingan kalitlari solishtiriladi. Ortiqcha bo‘shliqlar tartibga keltiriladi, registr farqlari esa belgilangan siyosat asosida hisobga olinadi. Registrga sezgir qisqartmalar va dastur kodlari umumiy almashtirishdan mustasno qilinadi.

Uchinchi bosqichga dastlabki qidiruvlar natija bermaganda o‘tiladi. Morfologik modul so‘rov uchun ehtimoliy asoslar va grammatik belgilarni qaytarishi ko‘zda tutiladi. Masalan, “algoritmlarning” shakli “algoritm + lar + ning” tarzida tahlil qilinadi. Bu yerda qo‘shimchalar ketma-ketligini tekshirish uchun chekli holatli qoidalar qo‘llanishi mumkin [2; 3]. Olingan “algoritm” nomzodi bazada mavjud bo‘lsagina tegishli termin yozuvi chiqariladi. Bir nechta mos nomzod aniqlansa, ular muqobil natijalar sifatida ko‘rsatiladi.



Ko‘p so‘zli terminlarni qayta ishlash

Termin birikmasining har bir qismini alohida qisqartirish birikmaning lug‘aviy tuzilishini buzishi mumkin. Masalan, “ma’lumotlar bazasida” so‘rovini “ma’lumotlar bazasi” shakliga bog‘lashda “bazasi + da” tahlili yetarli bo‘lishi mumkin. Bunda “ma’lumotlar” qismi hamda “bazasi” shaklidagi egalik ko‘rsatkichi bosh shakl tarkibida saqlanadi. Shuning uchun avval butun birikma tekshiriladi, so‘ng lug‘atdagi birikma andozasiga mos grammatik o‘zgarishlar ko‘rib chiqiladi.

Natijalarni chiqarishda aniq moslik birinchi, me‘yorlashtirilgan moslik ikkinchi, morfologik moslik uchinchi o‘rinda turadi. Bu tartib mavjud aniq natijani kengaytirilgan qidiruv bilan almashtirmaslikka xizmat qiladi. Modeldagi ushbu ustuvorlik samaradorlik haqidagi tayyor xulosa emas, balki tajribada tekshiriladigan loyihaviy qarordir.

Baholash mezonlari va muhokama

Tekshirish uchun bosh shakllar, belgi variantlari, qo‘shimchali so‘zlar, termin birikmalari va lug‘atda mavjud bo‘lmagan so‘rovlardan iborat to‘plam tuzish taklif etiladi. Har bir so‘rovga mutaxassis tomonidan to‘g‘ri termin identifikatori yoki “mos termin yo‘q” belgisi biriktiriladi. Qoidalarni sozlash va yakuniy baholash to‘plamlari terminlar bo‘yicha ajratiladi: bitta terminning variantlari ikkala to‘plamga ham kiritilmaydi.

Uch konfiguratsiya taqqoslanadi: faqat aniq qidiruv; aniq qidiruv va yozuvni me‘yorlashtirish; ularga morfologik tahlilni qo‘shish. Birinchi natijaning to‘g‘riligi, to‘g‘ri terminning dastlabki besh natijaga kirishi, lug‘atda yo‘q so‘rovlarga noto‘g‘ri moslik berish ulushi va javob vaqti qayd etiladi. Har bir guruhning hajmi hamda baholash uchun ishlatilgan dastur versiyasi ko‘rsatilishi lozim.

Manning va hammualliflar muhokama qilgan ortiqcha umumlashtirish muammosi [1] terminologik lug‘atda ayniqsa ehtiyotkor tekshiruvni talab qiladi. Sharipov va



“KELAJAK TEXNOLOGIYALARI VA SUN’IY INTELLEKT”
nomli respublika ilmiy-amaliy masofaviy konferensiyasi
VOLUME-1, ISSUE-7, 2026

Salaevning lug‘atsiz stemming modelidan [3] farqli ravishda, bu taklifda morfologik nomzod termin bazasi bilan tasdiqlanadi. UzMorphAnalyser kabi vosita [4] morfologik modul sifatida ko‘rib chiqilishi mumkin, biroq uning mavjudligi terminologik qidiruv sifati oldindan kafolatlanishini anglatmaydi. Omonimiya, qisqartmalar va lug‘at qamrovi alohida xato manbalari bo‘lib qoladi.

Xulosa

O‘zbekcha elektron terminologik lug‘at uchun yozuvni me’yorlashtirish va lug‘at bilan tekshiriladigan morfologik tahlilni birlashtiruvchi qidiruv modeli taklif etildi. Uning metodik jihati aniq moslikka ustuvorlik berish, ko‘p so‘zli termin tarkibini saqlash va morfologik tahlilni nomzod hosil qilish vositasi sifatida qo‘llashdan iborat. Bu umumiy morfologik usullarning terminologik qidiruvga moslashtirilishidir. Modelning amaliy foydasi taklif etilgan taqqoslash tartibi asosida keyingi tajribada aniqlanishi kerak.

Foydalanilgan adabiyotlar

1. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. § 2.2.4. <https://nlp.stanford.edu/IR-book/>
2. Beesley K. R., Karttunen L. Finite State Morphology. Stanford: CSLI Publications, 2003. <https://web.stanford.edu/group/cslipublications/cslipublications/site/1575864347.shtml>
3. Sharipov M., Salaev U. Uzbek affix finite state machine for stemming. 2022. arXiv preprint. DOI: 10.48550/arXiv.2205.10078.
4. Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings. 2024. arXiv preprint, v2. DOI: 10.48550/arXiv.2405.14179.
5. Unicode Consortium. Unicode Standard Annex #15: Unicode Normalization Forms. <https://www.unicode.org/reports/tr15/> (murojaat sanasi: 17.10.2026).