QAYTA IFODALANGAN O'ZBEKCHA SAVOLLARDA DALIL QIDIRISH XATOLARI TAHLILI

Erkayeva Laylo Odilbek qizi
Urganch davlat universiteti bitiruvchisi
"Kompyuter injiniringi" bo'yicha magistr
ORCID: 0009-0001-5380-5721
layloerkayeva0@gmail.com

Annotatsiya

RAG tizimida savol manbadagi ma'lumotni boshqacha so'zlar bilan ifodalasa, javobga asos bo'ladigan parchani topish qiyinlashishi mumkin. Ushbu tezis o'zbekcha ta'lim hujjatlarida qayta ifodalangan savollar bo'yicha dalil qidirish xatolarini o'rganadi. Sakkizta rasmiy sahifadan olingan 111 parchada 25 ta qayta ifodalangan savol tekshirildi; yana 25 ta manba matniga yaqin savol tavsifiy qiyoslash uchun ishlatildi. Qayta ifodalangan savollarda so'zli TF-IDF to'g'ri dalilni dastlabki uch natijadan 15/25 holatda, belgili TF-IDF esa 21/25 holatda topdi. Ikkala usul ham topa olmagan to'rtta savolning tahlili qisqa bandlar, turlicha atamalar va mavzusi o'xshash sahifalar xatoga hissa qo'shishi mumkinligini ko'rsatdi. Xulosalar ushbu kichik to'plam bilan cheklanadi.

Kalit so'zlar: o'zbek tili, qayta ifodalangan savol, dalil qidirish, TF-IDF, xato tahlili, RAG.

Abstract

When a question uses different wording from its source, a RAG system may struggle to find the passage needed to support an answer. This thesis examines evidence retrieval errors for rephrased questions about Uzbek education documents. It evaluates 25 rephrased questions against 111 passages from eight official pages; another 25 questions closer to the source wording provide a descriptive comparison. For the rephrased questions, word-level TF-IDF retrieved the designated passage among the first three results in 15/25 cases, compared with 21/25 for character-level TF-IDF. An examination of the four questions missed by both methods suggests that short list items, different terms, and competing pages on similar topics may contribute to retrieval errors. The findings are limited to this small collection.

Keywords: Uzbek, rephrased questions, evidence retrieval, TF-IDF, error analysis, RAG.

Аннотация

Если вопрос сформулирован иначе, чем сведения в источнике, системе RAG может быть трудно найти фрагмент для обоснования ответа. В тезисе изучаются ошибки поиска таких фрагментов по перефразированным вопросам об образовании на узбекском языке. На 111 фрагментах из восьми официальных страниц проверены 25 перефразированных вопросов; ещё 25 вопросов, близких к тексту источника, использованы для описательного сравнения. Для перефразированных вопросов словный TF-IDF нашёл заданный фрагмент среди первых трёх результатов в 15/25 случаях, символьный TF-IDF — в 21/25. Анализ четырёх вопросов, по

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

которым оба метода не нашли фрагмент, показал возможное влияние коротких пунктов, различий в терминах и страниц со схожей тематикой. Выводы ограничены этим небольшим корпусом.

Ключевые слова: узбекский язык, перефразирование, поиск фрагментов, TF-IDF, анализ ошибок, RAG.

KIRISH

Hujjatga tayangan savol-javob tizimida avval savolga tegishli matn parchasi topiladi, so‘ng shu dalil asosida javob tuzilishi mumkin [1]. «Yotoqxona» va «talabalar turar joyi», «birinchi kurs» va «birinchi bosqich» kabi ifodalar yaqin mazmunni anglatishi mumkin, biroq satrda bir xil so‘z sifatida qatnashmaydi. O‘zbek tilining qo‘shimchalarga boy tuzilishi ham so‘z shakllarining turlicha bo‘lishiga olib keladi [2]. Boshqa sohalardagi qidiruv ishlari savol bilan kerakli parcha o‘rtasidagi so‘z tafovuti dolzarb masala ekanini qayd etgan [3]; bu tezisda o‘sha tadqiqotlarning natijasi o‘zbekcha ta’lim hujjatlariga avtomatik ko‘chirilmaydi.

Maqsad — mavjud kichik o‘zbekcha ta’lim matnlari to‘plamida qayta ifodalangan savollar qaysi holatlarda to‘g‘ri dalilni dastlabki uch natijaga olib chiqmaganini hisoblash va ayrim xatolarni matn darajasida izohlash. Tadqiqot savoli: so‘zli va belgili TF-IDF xatolari qayta ifodalangan 25 savolda qanday kesishadi va ikkala usul qaysi savollarda birgalikda adashadi? Bu ish tayyor javobning to‘g‘riligini baholamaydi.

MATERIALLAR VA METODLAR

2026-yil 24-sentabr holatida gov.uz portalidagi sakkiz ochiq ta’lim sahifasidan [5–12] ajratilgan 111 parcha ishlatildi. Muallif uchun tayyorlangan baholash to‘plamida 50 savol va har savolga bittadan javobni o‘z ichiga oluvchi belgilangan parcha mavjud: 25 tasi manba ifodasiga yaqin («bevosita»), 25 tasi boshqa ifodadan foydalanadi («qayta ifodalangan»). Savollar manbalar va yordamchi raqamli vosita ko‘magida tuzilgan, real foydalanuvchilardan yig‘ilmagan. Ikki guruhda bir xil ma’nodagi savollarning juftlari tuzilmagan; guruhlar o‘rtasidagi farq qayta ifodalashning sababiy ta’siri sifatida talqin qilinmaydi.

Har parchaga tegishli sahifa sarlavhasi va nuqta oldindan qo‘shilgan; 1-tezisdan farqli ravishda sarlavha sharti bu yerda o‘zgartirilmaydi. scikit-learn 1.8.0 kutubxonasida [4] quyidagi to‘rt variant hisoblandi: so‘zli TF-IDF (1–2 so‘zli n-gramm), butun satrli belgili TF-IDF (3–5 belgi), so‘z chegarali belgili TF-IDF (3–5 belgi) va so‘zli TF-IDF vektorlarini 20 o‘lchamga qisqartiruvchi LSA. TF-IDF faqat 111 parcha bo‘yicha moslandi; savollar tayyor fazoga o‘tkazildi. O‘zaklash, sinonimlar lug‘ati va javob yozuvchi til modeli qo‘llanmadi. Normallashgan vektorlar yaqinligi bo‘yicha parchalar tartiblandi; teng ballar dastlabki tartib bilan yechildi.

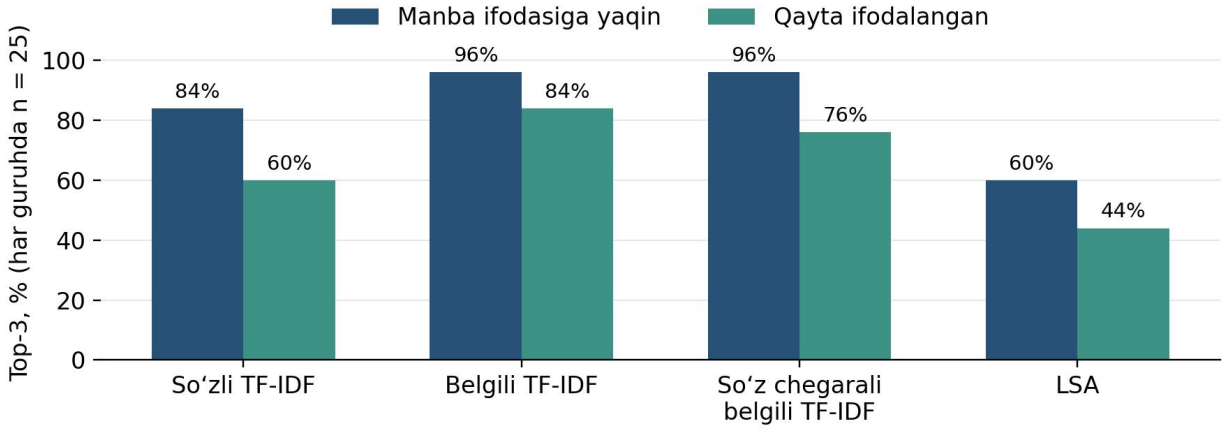
Top-1 birinchi o‘rinda, Top-3 dastlabki uch o‘rinda belgilangan yagona dalil chiqishini bildiradi. Har bir tur uchun maxraj 25; bir savolning boshqa mos parchasi dastlabki uchlikda chiqsa ham, belgilangan parcha yo‘q bo‘lsa, u «topilmadi» deb hisoblanadi. So‘zli va butun satrli belgili usullar uchun 25 qayta ifodalangan savolda Top-3 muvaffaqiyatlari o‘zaro kesishma sifatida sanaldi. Ikkalasi ham topmagan to‘rtta savolning manba matnlari va dastlabki uch natijasi qo‘lda ko‘rib chiqildi. Xato izohlari kuzatuv va ehtimoliy mexanizmlardir; alohida nazorat tajribasida sabab sifatida tasdiqlanmagan.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

NATIJALAR

1-jadval. Savol turiga ko'ra to'g'ri dalil topilgan savollar soni (har guruhda n = 25).

Usul	Bevosita Top-1	Qayta ifodalangan Top-1	Bevosita Top-3	Qayta ifodalangan Top-3
So'zli TF-IDF (1–2)	20/25	10/25	21/25	15/25
Belgili TF-IDF (3–5)	20/25	14/25	24/25	21/25
So'z chegarali belgili TF-IDF (3–5)	22/25	14/25	24/25	19/25
So'zli LSA (20 o'lcham)	9/25	8/25	15/25	11/25



1-rasm. Savol turiga ko'ra Top-3 ko'rsatkichlari. Guruhlar mazmunan juftlashtirilmagan.

So'zli TF-IDF qayta ifodalangan savollarda Top-3 bo'yicha 15/25 (60%), butun satrli belgili variant 21/25 (84%) natija berdi (1-jadval, 1-rasm). So'zli usulning barcha 15 muvaffaqiyati belgili usulda ham qayd etildi; faqat belgili usul topgan olti savol Q001, Q003, Q019, Q031, Q035 va Q039. Faqat so'zli usul topgan savol yo'q; ikkala usul ham topmaganlari Q005, Q009, Q017 va Q025. Bu kesishma 15 + 6 + 0 + 4 = 25 savolni to'liq qamrab oladi.

2-jadval. Ikkala asosiy usul Top-3 da topmagan qayta ifodalangan savollarni ko'rib chiqish.

Savol	Kerakli dalil va kuzatilgan raqobat
Q005	Harbiy xizmat sabab ta'til arizasiga ilova qilinadigan chaqiruv qog'ozi D01-L29 da. Qidiruv yuqoriga shu sahifadagi «arizaga hujjatlar ilova qilinadi» kirish bandini D01-L28 va «harbiy xizmat» haqidagi boshqa bandni D01-L11 olib chiqdi.
Q009	Ijara shartnomasini davlat soliq xizmati hisobga qo'yishi D02-L11 da. «Qaysi idora» va «ro'yxatga olish» savoliga ro'yxatdan o'tishga doir boshqa bandlar, shu jumladan D08 sahifasi raqobat qildi.
Q017	1-avgustdan ariza berish D03-L8 da. «Universitet yotoqxonasi» savoliga boshqa yotoqxona xizmati haqidagi D08-L27 hamda D03 dagi qisqa xizmat bandlari yuqori chiqdi.
Q025	Birinchi bosqich talabalari uchun to'lov-kontrakt summasi miqdoridagi kredit

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Savol	Kerakli dalil va kuzatilgan raqobat
	D04-L13 da. Savoldagi «birinchi kurs» va «shartnoma qiymatining qancha qismi» ifodasi manbadagi «birinchi bosqich» va «summasi miqdorida» bilan aynan bir xil emas.

2-jadvaldagi izohlar dastlabki uchlik va belgilangan parcha matnini tavsiflaydi. Masalan, Q005 bo‘yicha yuqori chiqqan umumiy «hujjatlar ilova qilinadi» bandi savolga kerakli aniq hujjat nomini bermaydi. Q017 bo‘yicha boshqa sahifa yotoqxona xizmatini tasvirlasa-da, belgilangan D03-L8 dagi boshlanish sanasini bildirmaydi. Demak, mavzusi mos parchaning topilishi o‘zi kifoya emas: unda savolga kerakli fakt bo‘lishi lozim.

MUHOKAMA

Belgili n-grammlar ayrim so‘z shakllaridagi umumiy qismlarni hisobga olishi mumkin, lekin sinonimning mazmunini mustaqil anglamaydi. Olti savolning faqat belgili usulda topilishi shu kichik to‘plamdagi kuzatuv; qaysi belgi bo‘laklari har bir yaxshilanishni keltirganini sababiy tajriba aniqlamagan. To‘rt qolgan xatoda qisqa ro‘yxat bandlari, savol va manba orasidagi atama tafovuti yoki o‘xshash xizmatlar haqidagi sahifalar raqobati ko‘rinadi. Shuning uchun keyingi tajribada ro‘yxatning kirish bandi bilan tegishli bandini birlashtirish va sinonim ifodalarni alohida tekshirish asosli taxmin sifatida qo‘yilishi mumkin; mazkur ish ularning samarasini o‘lchamagan.

Cheklovlar: sakkiz sahifa, sun’iy tuzilgan 50 savol va savolga bitta belgilangan parcha; muqobil to‘g‘ri parchalarni mustaqil baholovchilar tekshirmagan. Manba ifodasiga yaqin va qayta ifodalangan guruhlar bir xil axborot ehtiyojining juft versiyalari emas. Oldingi qidiruv usullari ayni kichik to‘plamda ko‘rib chiqilgan, mustaqil yakuniy test yo‘q; to‘rt savol bo‘yicha qo‘lda tahlil ham izohlovchi xarakterda. Foizlar statistik ahamiyat yoki boshqa hujjatlar uchun ustunlik da’vosini bermaydi. Olingan parchadan avtomatik javob tuzish va javobning faktlarga mosligi baholanmadi.

XULOSA

Qayta ifodalangan 25 savolda belgilangan dalilni dastlabki uch natijaga so‘zli TF-IDF 15 marta, butun satrli belgili TF-IDF 21 marta chiqardi. Belgili variant so‘zli usul topmagan olti dalilni topdi; to‘rtta savol ikkala usulda ham yechilmadi. Xatolarni ko‘rib chiqish qidiruv uchun mos mavzuni aniqlash bilan aynan so‘ralgan faktni topish alohida vazifa ekanini ko‘rsatadi. Natijalarni mustahkamlash uchun bir xil savolning bevosita va qayta ifodalangan juftlarini yig‘ish, bir nechta mos dalilni belgilash va mustaqil baholash to‘plamida yangi usullarni sinash kerak.

FOYDALANILGAN ADABIYOTLAR

1. Lewis P. va boshq. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
2. Sharipov M., Kuriyozov E., Yuldashev O., Sobirov O. UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts. LREC-COLING, 2024, 17343–17347. https://aclanthology.org/2024.lrec-main.1506/
3. Kim D. va boshq. GuRE: Generative Query REwriter for Legal Passage Retrieval. Natural Legal Language Processing Workshop, 2025, 424–438. https://doi.org/10.18653/v1/2025.nllp-1.31
4. scikit-learn. TfidfVectorizer texnik hujjatlari. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

5. O'zbekiston Respublikasi Hukumat portali. Akademik ta'til. https://gov.uz/oz/advice/65/document/1755
6. O'zbekiston Respublikasi Hukumat portali. Talabalarga ijara to'lovining bir qismini qoplab berish tartibi. https://gov.uz/oz/advice/64/document/2562
7. O'zbekiston Respublikasi Hukumat portali. Davlat OTMlarining talabalar turar joyiga joylashish uchun ariza yuborish. https://gov.uz/oz/advice/65/document/3106
8. O'zbekiston Respublikasi Hukumat portali. Ta'lim krediti. https://gov.uz/oz/advice/65/document/3248
9. O'zbekiston Respublikasi Hukumat portali. Talabalar bilimini baholash. https://gov.uz/oz/advice/65/document/1758
10. O'zbekiston Respublikasi Hukumat portali. Ta'lim grantida tahsil olgan talabalarning 3 yil ishlab berish majburiyati bekor qilindi. https://gov.uz/oz/advice/65/document/3250
11. O'zbekiston Respublikasi Hukumat portali. O'qishini ko'chirishga oid: muhim 7 savolga javob. https://gov.uz/oz/advice/65/document/4173
12. O'zbekiston Respublikasi Raqamli texnologiyalar vazirligi. Endi my.gov.uz orqali talabalar turar joyiga joylashish va vaqtincha ro'yxatdan o'tish mumkin. https://gov.uz/oz/digital/news/view/200526

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.