



EVALUATION METHODOLOGY FOR SEMANTIC SEARCH MODELS FOR UZBEK-LANGUAGE TEXTS

Adxamov Doniyorbek Abdulxodi o'g'li

"Datatruck" LLC, Foreign Enterprise

doniyorbek.adhamov@gmail.com

Abstract

This thesis proposes a methodology for evaluating semantic search (embedding) models for Uzbek-language texts. In addition to measuring search accuracy, the methodology examines factors specific to the Uzbek language, including script type, apostrophe characters, word forms, and model robustness to paraphrasing.

Keywords

semantic search, embedding models, Uzbek language, evaluation metrics, RAG, robustness.

O'ZBEK TILIDAGI MATNLAR UCHUN SEMANTIK QIDIRUV MODELLARINI BAHOLASH METODIKASI

Adxamov Doniyorbek Abdulxodi o'g'li

"Datatruck" MChJ Xorijiy Korxona

doniyorbek.adhamov@gmail.com

Annotatsiya. Tezisda o'zbek tilidagi matnlar uchun semantik qidiruv (embedding) modellarini baholash metodikasi taklif etilgan. Metodika qidiruv aniqligini o'lchash bilan birga o'zbek tiliga xos omillar – yozuv turi, apostrof belgilari, so'z shakllari va qayta ifodalashga modelning barqarorligini tekshirishni ham o'z ichiga oladi.

Kalit so'zlar: semantik qidiruv, embedding modellari, o'zbek tili, baholash metrikalari, RAG, barqarorlik.

Semantik qidiruv modellari matnni ko'p o'lchovli vektorga aylantiradi va ma'nosi yaqin matnlarni topish imkonini beradi. Kalit so'zlarga asoslangan an'anaviy qidiruvdan





"KELAJAK TEKNOLOGIYALARI VA SUN'IY INTELLEKT"
nomli respublika ilmiy-amaliy masofaviy konferensiyasi
VOLUME-1, ISSUE-7, 2026

farqli ravishda, ular savoldagi so'zlar hujjatda aynan uchramasa ham mazmunan mos matnni topa oladi. Bunday modellar qidiruv bilan boyitilgan generatsiya (RAG) tizimlarining asosiy tarkibiy qismi hisoblanadi: qidiruv bosqichida topilmagan ma'lumot keyingi bosqichda ham to'g'ri javobga aylanmaydi [1]. Shu sababli o'zbek tilida ishonchli RAG tizimini yaratish uchun avvalo qaysi embedding modeli o'zbek matnlari bilan yaxshiroq ishlashini aniqlash kerak.

Hozirda o'zbek tilini qo'llab-quvvatlaydigan bir nechta ko'p tilli modellar mavjud. LaBSE modeli 109 tilda parallel gaplar asosida o'qitilgan va tillararo moslikni topishda yaxshi natija beradi [2]. Multilingual-E5 modellari 100 dan ortiq tilni qo'llab-quvvatlaydi va ikki bosqichda: avval katta hajmdagi matn juftliklarida, so'ng belgilangan ma'lumotlarda o'qitilgan [3]. BGE-M3 modeli esa bitta modelda zich, siyrak va ko'p vektorli qidiruvni birlashtiradi hamda uzun hujjatlar bilan ishlay oladi [4]. Biroq bu modellar asosan ingliz tilidagi ma'lumotlar ustunligi bilan o'qitilgan va xalqaro benchmarklarda o'zbek tili bo'yicha vazifalar kam. Natijada bu modellarning o'zbek tilidagi haqiqiy samaradorligi yetarlicha o'rganilmagan.

Taklif etilayotgan metodika

Metodika uch bosqichdan iborat: test to'plamini shakllantirish, aniqlikni o'lchash va o'zbek tiliga xos barqarorlikni tekshirish.

1-bosqich: test to'plamini shakllantirish. To'plam "savol – tegishli matn bo'lagi" juftliklaridan iborat bo'ladi. Buning uchun ikki manbadan foydalanish taklif etiladi. Birinchisi – mavjud ko'p tilli resurslar, masalan, o'zbek tilini ham o'z ichiga olgan Belebele o'qib tushunish to'plami [5]. Undagi savollar va matn parchalari qidiruv vazifasiga moslashtiriladi: har bir savol uchun to'g'ri parcha barcha parchalar orasidan topilishi kerak. Ikkinchisi – aniq soha (masalan, me'yoriy hujjatlar) bo'yicha mutaxassislar tomonidan tuzilgan savollar. Ikkinchi manba amaliy tizimga yaqinroq natija beradi.



"KELAJAK TEKNOLOGIYALARI VA SUN'IY INTELLEKT"
nomli respublika ilmiy-amaliy masofaviy konferensiyasi
VOLUME-1, ISSUE-7, 2026

2-bosqich: aniqlikni o'lchash. Har bir model uchun standart qidiruv metrikalari hisoblanadi [6]: Recall@K (to'g'ri matn dastlabki K ta natija ichida topilgan savollar ulushi), MRR (birinchi to'g'ri natija o'rnining teskari qiymati o'rtachasi) va nDCG@10 (natijalar tartibining sifati). Taqqoslash uchun asos sifatida leksik qidiruv usuli BM25 [7] olinadi. Agar embedding modeli BM25 dan yaxshi natija bermasa, uni o'zbek tili uchun qo'llash asossiz hisoblanadi. Bundan tashqari, BM25 va embedding modeli natijalarini birlashtiruvchi gibril variant ham sinovdan o'tkaziladi.

3-bosqich: o'zbek tiliga xos barqarorlikni tekshirish. Bu bosqich metodikaning asosiy yangiligi hisoblanadi. Har bir savol bir necha ko'rinishda qayta yoziladi va natijalar asl savol natijasi bilan solishtiriladi (1-jadval).

1-jadval

O'zbek tiliga xos barqarorlik testlari

Test turi	Misol	Nimani tekshiradi
Yozuv turi	"Mehnat ta'tili" va "Мехнат таътили"	Lotin va kirill yozuvidagi savollarga bir xil javob berilishi
Apostrof belgisi	"o'quv", "o‘quv", "o’quv", "o`quv"	Turli apostrof belgilariga sezgirlik
So'z shakli	"shartnoma", "shartnomaning", "shartnomalarga"	Qo'shimchalar qo'shilganda ma'no saqlanishi
Qayta ifodalash	"Ta'til necha kun?" va "Yillik dam olish muddati qancha?"	Bir xil ma'noni turli so'zlar bilan ifodalashni tushunish

Barqarorlik ko'rsatkichi sifatida asl savol va uning variantlari uchun Recall@K qiymatlari o'rtasidagi farq olinadi:

$$\Delta_t = \text{Recall}@K(\text{asl}) - \text{Recall}@K(t), \quad (1)$$



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

bu yerda t – test turi (yozuv, apostrof, so'z shakli yoki qayta ifodalash). Δ_t qancha kichik bo'lsa, model shu omilga shuncha barqaror hisoblanadi. Agar Δ_t katta bo'lsa, bu omil qidiruvdan oldin dasturiy ravishda bartaraf etilishi kerak.

Tajriba o'tkazish tartibi

Metodikani amalda qo'llash quyidagi tartibda amalga oshiriladi:

1. test to'plami shakllantiriladi va har bir savol uchun 1-jadvaldagi variantlar tayyorlanadi;
2. barcha matn parchalari har bir model yordamida vektorlarga aylantiriladi va indekslanadi;
3. har bir savol va uning variantlari bo'yicha qidiruv bajariladi, natijalar saqlanadi;
4. Recall@K, MRR, nDCG@10 va Δ_t ko'rsatkichlari hisoblanadi;
5. natijalar asosida o'zbek tili uchun eng mos model va zarur oldindan ishlov berish qadamlari tanlanadi.

Tajribani bepul hisoblash resurslari (masalan, Google Colab) yordamida ham o'tkazish mumkin, chunki embedding modellari faqat matnni vektorga aylantirish uchun ishlatiladi va qayta o'qitishni talab qilmaydi.

Kutilayotgan amaliy ahamiyati

Metodika o'zbek tilidagi RAG tizimlarini ishlab chiquvchilarga modelni asosli tanlash imkonini beradi. Shuningdek, u qaysi oldindan ishlov berish qadamlari (masalan, apostroflarni normallashtirish yoki transliteratsiya) zarur ekanini ko'rsatadi. Natijalar o'zbek tili uchun ochiq benchmark yaratishga ham asos bo'lishi mumkin.

Xulosa

O'zbek tili uchun embedding modellarini baholashda faqat umumiy aniqlik ko'rsatkichlari yetarli emas. Yozuv turi, apostrof belgilari, so'z shakllari va qayta





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

ifodalashga barqarorlikni tekshirish ham zarur. Taklif etilgan metodika o'zbek tilidagi semantik qidiruv va RAG tizimlari sifatini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar

1. Lewis P., Perez E., Piktus A. va boshq. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – P. 9459–9474.
2. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT Sentence Embedding // Proceedings of ACL. – 2022. – P. 878–891.
3. Wang L., Yang N., Huang X. va boshq. Multilingual E5 Text Embeddings: A Technical Report // arXiv preprint arXiv:2402.05672. – 2024.
4. Chen J., Xiao S., Zhang P. va boshq. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation // arXiv preprint arXiv:2402.03216. – 2024.
5. Bandarkar L., Liang D., Muller B. va boshq. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants // Proceedings of ACL. – 2024. – P. 749–775.
6. Thakur N., Reimers N., Rüklé A., Srivastava A., Gurevych I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models // NeurIPS Datasets and Benchmarks Track. – 2021.
7. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. – 2009. – Vol. 3, No. 4. – P. 333–389.