

COMPARATIVE ANALYSIS OF EMBEDDING MODELS FOR MATCHING QUESTIONS AND CONTEXTS IN THE KAZAKH LANGUAGE

¹K.Sh. Mazhitova^{ORCID}, ²A.O. Tleubayeva^{ORCID}✉, ¹S.E. Mukhammediya^{ORCID}, ³A.A. Tanirbergenova^{ORCID}¹A.S. Toktaganova^{ORCID}, ¹Zh.B. Kambarova^{ORCID}¹Higher College "ASTANA POLYTECHNIC", Astana, Kazakhstan,²Astana IT University, Astana, Kazakhstan,³L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

✉Корреспондент-автор: Arailym.tll@gmail.com

This paper presents a comparative analysis of the effectiveness of modern embedding models for matching questions and contexts in the Kazakh language, within the framework of semantic search and question-answering (QA) systems. The study is based on the open dataset Kundyzyka/informatics_kaz, which contains approximately 7,700 question–context pairs covering key concepts in computer science. The analysis focuses on comparing the traditional TF-IDF method with state-of-the-art multilingual models, including mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small, and Snowflake Arctic-Embed v2.0. The evaluation metrics used are Accuracy@1, Mean Reciprocal Rank (MRR), and ROC AUC. The zero-shot experiments demonstrate that embedding-based models significantly outperform TF-IDF, achieving Accuracy@1 in the range of 0.59–0.63, whereas the baseline method yields 0.50. Moreover, visualization results using t-SNE indicate that modern models are able to cluster semantically related questions and contexts more closely, which positively affects the quality of ranking. Among the models considered, Alibaba GTE and Snowflake Arctic-Embed v2.0 achieved the best results, confirming their relevance for developing automatic question-answering systems in educational AI applications. This study highlights that even without fine-tuning on Kazakh data, pretrained multilingual models can substantially enhance semantic matching quality, offering promising prospects for their use in educational systems and other information services.

Keywords: semantic search, sentence embeddings, question-context matching, multilingual transformer models, TF-IDF, Kazakh language, information retrieval.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ ЭМБЕДДИНГОВ ДЛЯ СОПОСТАВЛЕНИЯ ВОПРОСОВ И КОНТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

¹К.Ш. Мажитова✉, ²А.О. Тлеубаева, ¹С.Е. Мухаммедия, ¹А.А. Танирбергенова, ¹А.С.Токтаганова, ¹Ж.Б. Камбарова¹Высший колледж «ASTANA POLYTECHNIC», г. Астана, Казахстан,²Astana IT University, Астана, Казахстан,³Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
e-mail:Arailym.tll@gmail.com

В данной статье представлен сравнительный анализ эффективности современных моделей эмбеддингов для сопоставления вопросов и контекстов на казахском языке в рамках задач семантического поиска и систем вопросов-ответов (QA). Исследование выполнено на открытом датасете Kundyzyka/informatics_kaz, содержащем около 7700 пар «вопрос–контекст», охватывающих основные понятия информатики. Основное внимание уделено сравнению традиционного метода TF-IDF с современными многоязычными моделями, такими как mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small и Snowflake Arctic-Embed v2.0. Для оценки качества сопоставления применялись метрики Accuracy@1, Mean Reciprocal Rank (MRR) и ROC AUC. Эксперименты, проведенные в режиме zero-shot, показали, что модели эмбеддингов значительно превосходят TF-IDF, достигая Accuracy@1 в диапазоне 0.59–0.63, тогда как базовый метод демонстрирует показатель 0.50. Дополнительно, результаты визуализации с использованием t-SNE свидетельствуют о том, что современные модели способны группировать семантически связанные вопросы

и контексты ближе друг к другу, что благоприятно сказывается на качестве ранжирования. Среди рассмотренных моделей наилучшие результаты показали Alibaba GTE и Snowflake Arctic-Embed v2.0, что подтверждает их актуальность для создания систем автоматического поиска ответов на вопросы в образовательных ИИ-приложениях. Работа подчеркивает, что даже без дополнительного обучения на казахских данных предобученные многоязычные модели способны значительно улучшить качество семантического сопоставления, что открывает перспективы для применения данных решений в образовательных системах и других информационных сервисах.

Ключевые слова: семантический поиск, эмбединги предложений, сопоставление вопрос-контекст, мультиязычные трансформерные модели, TF-IDF, казахский язык, информационный поиск.

ҚАЗАҚ ТІЛІНДЕГІ СҰРАҚТАР МЕН КОНТЕКСТТЕРДІ СӘЙКЕСТЕНДІРУГЕ АРНАЛҒАН ЭМБЕДДИНГ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

¹К.Ш. Мажитова , ²А.О. Тлеубаева, ¹С.Е. Мухаммедия, ¹А.А. Танирбергенова, ¹А.С.Токтаганова, ¹Ж.Б. Камбарова

¹«ASTANA POLYTECHNIC» Жоғары колледжі, г. Астана, Қазақстан,

²Astana IT University, Астана, Қазақстан,

³Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
e-mail:Arailym.tll@gmail.com

Бұл мақалада қазақ тіліндегі сұрақтар мен контексттерді сәйкестендіруге арналған қазіргі заманғы эмбединг модельдерінің тиімділігіне салыстырмалы талдау ұсынылған. Зерттеу шамамен 7700 «сұрақ–контекст» жұбынан тұратын ашық Kundyшка/informatics_kaz деректер жиыны негізінде жүргізілді. Зерттеу барысында дәстүрлі TF-IDF әдісі мен заманауи көптілді модельдер - mMini LM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small және Snowflake Arctic-Embed v2.0 - арасындағы айырмашылықтарға ерекше назар аударылды. Сәйкестендіру сапасын бағалау үшін Accuracy@1, Mean Reciprocal Rank (MRR) және ROC AUC метрикалары қолданылды. Zero-shot режимінде жүргізілген эксперименттер эмбединг модельдерінің TF-IDF әдісінен айтарлықтай жақсы нәтиже көрсететінін дәлелдеді: эмбединг модельдері 0.59–0.63 аралығында Accuracy@1 мәнін көрсетті, ал базалық әдіс — 0.50. Сонымен қатар, t-SNE визуализациясының нәтижелері қазіргі модельдердің семантикалық жағынан ұқсас сұрақтар мен контексттерді бір-біріне жақын орналастыра алатынын көрсетті, бұл өз кезегінде рейтинг сапасын арттырады. Зерттелген модельдердің ішінде Alibaba GTE және Snowflake Arctic-Embed v2.0 үздік нәтижелер көрсетті, бұл оларды білім беру бағытындағы жасанды интеллект қосымшалары үшін өзекті құрал етеді. Бұл жұмыс қазақ тіліндегі деректерге арнайы оқытусыз-ақ алдын ала оқытылған көптілді модельдердің семантикалық сәйкестендіру сапасын айтарлықтай арттыра алатынын көрсетіп, оларды білім беру жүйелері мен ақпараттық сервистерде қолдану мүмкіндіктерін ашады.

Түйін сөздер: семантикалық іздеу, сөйлем эмбедингтері, сұрақ–контекст сәйкестігі, көптілді трансформер модельдері, TF-IDF, қазақ тілі, ақпараттық іздеу

Introduction. In the tasks of semantic search and answering questions, bag-of-words and TF-IDF methods are traditionally used, implementing a vector model of the document space [1]. However, such methods often do not take into account the contextual meaning of words, which is especially critical for complex questions in natural language. Modern advances in natural language processing (NLP) suggest using pre-trained language models to obtain dense vector representations (embeddings) of sentences [2,3]. Such models allow you to compare the semantic proximity of phrases and are successfully used to find answers to questions in large collections of texts. The effectiveness of multilingual embeddings has already been shown, for example, the multilingual MiniLM [4] and LaBSE (Language-agnostic Sentence Embedding) models [5], which are able to display sentences of

different languages into a common vector space. This opens up the possibility to use them for the Kazakh language, even if the model was initially trained in other languages.

This research project is aimed at applying such multilingual models in the field of educational AI systems. In particular, we consider the task of comparing educational questions in Kazakh with the corresponding text fragments (contexts) containing the answers. Automating this mapping has practical value for developing intelligent assistants capable of answering students' questions in their native language, or for systems that recommend relevant learning materials. In this paper, we conducted a comparative analysis of several approaches - from classical TF-IDF to modern transformer embeddings – in order to find out which model provides the best quality of searching for "question-context" pairs in the Kazakh language.

The aim of the study is to evaluate the quality of various models of vector representations (embeddings) when comparing questions and contexts in the Kazakh language. This includes comparing the classic TF-IDF method with the latest multilingual embedding models without additional training on this task. We aim to determine which of the considered models most effectively measure the semantic proximity of a question and its corresponding context, and are also suitable for use in a question-and-answer search system in the educational field. The quality metrics of the comparison (Accuracy@1, MRR, ROC AUC) are used for quantitative comparison of models [6].

Materials and methods. For the experiments, an open Kazakh-language dataset of questions and answers Kundyzyka/informatics_kaz [7] was used, containing about 7,700 question–context pairs. Questions are queries in Kazakh on computer science topics, and contexts are excerpts of texts containing an answer or explanation to a given question. The dataset covers the definitions and basic concepts of computer science, which makes it indicative for evaluating models in the academic (educational) field. An example of a pair from a dataset: the question *вопрос “мәліметтер қо-*

ры дегеніміз не?” (“What is a database?”) and the corresponding context with the definition of this term. The presence of such a specialized corpus in Kazakh makes it possible to check how well multilingual models, trained primarily in other languages, transfer their abilities to Kazakh and whether they are able to correctly match semantically related phrases.

Before using the models, the data was divided into question-context pairs, and for each question in the test set there is its correct (appropriate) context and a number of unrelated contexts to assess the quality of the ranking. Thus, the task for the model is to select the text fragment that answers the question among the many candidates. This simulates the real scenario of the question-and-answer search system.: for the user's question, it is necessary to find the most relevant answer in the knowledge base.

Before applying the model, text data is transformed into embeddings, which are dense vector representations. The process of obtaining embeddings for the question and context can be written as follows (1):

$$e^q = \text{Embed}(q), e^c = \text{Embed}(c) \quad (1)$$

where e^q and e^c represent the embeddings of the question and the context, respectively, and $\text{Embed}(\)$ denotes the selected embedding model (e.g., LaBSE, GTE, etc.). A detailed comparison of multilingual question-answering models and their adaptation to the Kazakh language is provided in the study by [8].

To measure semantic similarity, cosine similarity is applied, as shown in Equation (2):

$$\text{sim}(e^q, e^c) = \frac{e^q \cdot e^c}{\|e^q\| \|e^c\|} \quad (2)$$

Candidate contexts are ranked in descending order of similarity value. If the context matches the question with maximum similarity, it is assumed that the model has found the correct answer (Accuracy@1) [6].

Two fundamentally different approaches were used to compare issues and contexts: the TF-

IDF method and transformer models to obtain embeddings. On the TF-IDF side, the question and each candidate context are represented as sparse vectors of term frequencies, and relevance is assessed by the cosine similarity between these vectors [9]. This approach serves as a baseline reflecting the level of classical information search without understanding the context.

As an alternative, five pre-trained multilingual embedding models capable of generating dense vector representations of sentences were selected. Each model independently encodes the question and context into a vector of fixed dimension; then the cosine similarity between these vectors is calculated, based on which the contexts for each question are ranked. The models for the experiment include:

Multilingual MiniLM is a compact multilingual model based on the distillation of the MiniLM transformer [4]. This model is trained to map paraphrases in different languages into close vectors and, as Reimers and Gurevich have shown, can effectively expand monolingual embeddings into new languages through knowledge-distillation [2]. In this experiment, it represents a class of lightweight models with an embedding dimension of ~384.

LaBSE (Language-agnostic BERT Sentence Embedding) is a larger model from Google based on BERT, trained specifically to obtain language-independent embeddings of sentences [5]. LaBSE supports 109 languages, including Kazakh, and is able to form a common vector space where translations in different languages are close to each other. It is assumed that due to large-scale training on parallel data [5], this model will well reflect the semantic similarity of Kazakh questions and answers.

Alibaba GTE (General Text Embedding, multilingual) is a multilingual model from Alibaba DAMO Academy [10]. It is developed as part of the Retrieval-Augmented Generation framework and is able to handle extended contexts. The model is based on a BERT-like architecture with support for 75 languages and has been trained on large bodies of query–document pairs [10]. In our experiment,

the gte-multilingual-base version is used, which provides embeddings of dimension 768. Due to the long context and multilingualism, this model is expected to show high quality matching.

Multilingual E5 (small) is a multilingual model of the E5 family, presented by Microsoft researchers [11]. E5 is a series of embedding models trained on a combination of ranking tasks (for example, MS MARCO) in order to unify the presentation of text for various applications (search, classification, etc.). In particular, multilingual E5-small is a lightweight version (~118M parameters) with support for many languages [11]. Of interest is how this relatively small model will handle Kazakh data compared to larger models.

Snowflake Arctic Embed v2.0 is a modern embedding model from Snowflake [12], aimed at high-quality semantic search. Arctic-embed v2.0 is a second-generation multilingual model that achieves a balance between quality in English and other languages without compromising multilingual capabilities [12]. The model is optimized for retrieval of relevant documents and, according to the authors, is not inferior to larger models in quality, remaining relatively compact [13]. In our experiments, we used its large version, which gives vectors of dimension 1024. Each question and each potential context were transformed into a vector using one of these models. For each question, the cosine similarity with all contexts was calculated, after which the contexts were sorted in descending order of this similarity. If the correct (relevant) context came first, it was assumed that the model had successfully found the answer (case Accuracy@1) [6]. In addition to Accuracy@1, the MRR metric (Mean Reciprocal Rank [14], the average value of $1/\text{rank_correct}$ for all questions) and the area under the ROC curve (ROC AUC) [14] were used to evaluate the ranking, where the correct question-answer pair among negative (inappropriate) pairs is considered a positive label. These metrics are standard for evaluating information retrieval systems and question-and-answer systems: Accuracy@1 shows the proportion of completely correct answers on the first attempt, MRR takes into account the position of

the correct answer in the list, and ROC AUC reflects the overall ability of the model to distinguish the correct context from the wrong one as the similarity threshold changes.

The experiments were conducted without further training the models on a given dataset, i.e. in zero-shot mode: it is assumed that the pre-trained embeddings are versatile enough to capture semantic proximity even for Kazakh questions that the model might not have previously seen. All texts of questions and contexts have been pre-normalized (lowercase, punctuation removal, tokenization if necessary) in order to be correctly represented in TF-IDF and to ensure compatibility with the

requirements of the models (some transformers have entry length limits, but in our dataset the length of questions and contexts does not exceed these limits).

Results and Discussion. The results of the conducted experiment provide insights into the performance of different models in the task of matching questions and contexts in the Kazakh language. Specifically, we analyzed three key evaluation metrics - Accuracy@1, Mean Reciprocal Rank (MRR), and ROC AUC - across six models, including the baseline TF-IDF method and five modern multilingual sentence embedding models. The summarized results are presented in Table 1.

Table 1-Comparison of models based on question-context matching accuracy

Model	Accuracy@1	MRR	ROC AUC
TF-IDF (baseline)	0.4971	0.6407	0.7094
mMiniLM (Multilingual)	0.5419	0.7149	0.6947
LaBSE	0.6192	0.7614	0.7270
Alibaba-nlp/gte-multilingual-base	0.6322	0.7677	0.7611
intfloat/multilingual-e5-small	0.5913	0.7408	0.7188
Snowflake/snowflake-arctic-embed-l-v2.0	0.6309	0.7700	0.7635

Figure 1 presents the values of all three metrics (Accuracy@1, MRR, and ROC AUC) for the six evaluated methods. As shown in the graph, the traditional TF-IDF method (on the left) yields the lowest performance, while more advanced multilingual embedding models - such as mMiniLM, LaBSE, GTE, E5, and Snowflake Arctic - consistently outperform it across all metrics. Among them, GTE-multilingual-base and Snowflake Arctic v2.0 demonstrate the highest scores, indicating their strong effectiveness in the task of ranking relevant question-context pairs, even in a low-resource language like Kazakh.

The results clearly demonstrate the superiority of modern transformer-based embeddings over the classical TF-IDF approach. The baseline TF-IDF method correctly identifies the relevant context in approximately 49.7% of cases (Accuracy@1), while all tested embedding-based models

outperform it on this metric. This is expected, as TF-IDF does not account for synonymy or contextual meaning and may fail when the question and answer use different phrasings for the same concept. Nevertheless, a nearly 50% result for TF-IDF suggests that simple lexical overlap still provides a baseline signal: in about half of the cases, the key terms in the question are directly present in the answer text.

The mMiniLM model achieves an Accuracy@1 of 0.5419, which is roughly 4.5 percentage points higher than the baseline. This improvement stems from its ability to capture semantic relationships: mMiniLM performs better than TF-IDF even without exact word matches between questions and contexts. Furthermore, its MRR (0.7149) is significantly higher than TF-IDF's (0.6407), indicating that even when the top-ranked answer is incorrect, the correct one is often placed near the

top (e.g., 2nd or 3rd). Interestingly, ROC AUC for mMiniLM (0.6947) is slightly lower than that of TF-IDF (0.7094), possibly due to the distribution of similarity scores: TF-IDF may more confidently assign very low scores to clearly irrelevant pairs, while mMiniLM, being semantically “softer,”

sometimes assigns moderate scores even to unrelated pairs. As a result, its ROC curve may slightly underperform. Overall, mMiniLM delivers competitive performance for a lightweight multilingual model, but it understandably lags behind the larger models.

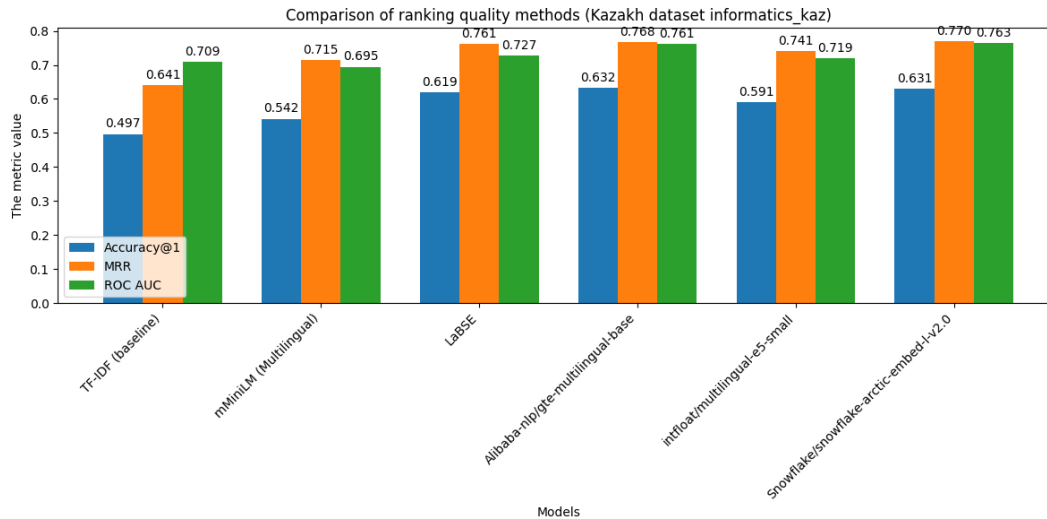


Fig.1 – Comparison of ranking quality methods

LaBSE significantly outperforms both TF-IDF and mMiniLM across all metrics: Accuracy@1 increases to 0.6192 (i.e., about 61.9% of questions are matched with the correct context in the top position), MRR reaches 0.7614, and ROC AUC rises to 0.7270. This improvement - over 12 percentage points higher than mMiniLM in Accuracy@1 - reflects the benefit of LaBSE’s extensive training on parallel multilingual data [3]. The model appears better equipped to interpret Kazakh formulations, likely because the Kazakh language was included in its training data and it has learned to align semantically equivalent phrases across languages. Notably, LaBSE also surpasses previous models in ROC AUC, demonstrating a more consistent ability to distinguish relevant from irrelevant pairs. Thus, LaBSE sets a strong benchmark in our experiment.

Two models - Alibaba GTE and Snowflake Arctic v2.0 - achieved nearly identical and highest overall performance. GTE reached an Accuracy@1 of 0.6322, while Arctic followed

closely with 0.6309, making their performance statistically equivalent and slightly ahead of LaBSE. However, in terms of MRR and ROC AUC, both models slightly outperformed LaBSE: Snowflake Arctic reached an MRR of 0.7700 and ROC AUC of 0.7635, while GTE achieved MRR = 0.7677 and ROC AUC = 0.7611. These are the best results among all tested models. In nearly 63% of cases, the top-ranked context is correct, and on average, the correct answer appears very high in the list (MRR ~0.77 implies it is typically ranked first or second). A high ROC AUC (~0.76) indicates that the similarity scores effectively separate relevant from irrelevant context pairs.

The success of GTE and Arctic can be attributed to their architectures and training data. GTE was specifically designed for document retrieval with long contexts and includes optimizations for multilingual input [10]. Its internal representations are likely more robust to linguistic variation, and it may better leverage token overlap (e.g., through XLM-R-based mixed

tokenization, as discussed in [10]). According to its developers, Snowflake Arctic was built to achieve high-quality multilingual search without compromise - trained on query-document pairs in multiple languages, it captures both language-specific and cross-lingual semantic features [12]. Notably, Arctic v2.0 is promoted as a lightweight model that rivals larger architectures in retrieval quality [12], which is supported by our findings: it matches GTE's performance despite architectural and training differences.

The multilingual E5-small model delivered moderate results: Accuracy@1 = 0.5913, higher than mMiniLM but lower than LaBSE. Its MRR (0.7408) and ROC AUC (0.7188) place it close to mMiniLM, slightly behind LaBSE in overall ranking quality. Considering that E5-small has around 118 million parameters [11] - more than mMiniLM but fewer than LaBSE - and was trained partly on English retrieval tasks, its relatively strong performance compared to mMiniLM is expected. However, it does not reach the levels of LaBSE or GTE/Arctic. This may be due to its smaller capacity and the possibility that its training data contained fewer or simpler Kazakh-language examples. Nonetheless, E5-small clearly surpasses the baseline and shows that even compact multilingual models can be effectively used for Kazakh semantic retrieval.

In summary, the best performance in the Kazakh question-context matching task was achieved by the multilingual embedding models LaBSE, GTE, and Snowflake Arctic, with Accuracy@1 scores in the 0.62–0.63 range (versus ~ 0.50 for TF-IDF) and $MRR \approx 0.76$ –0.77, indicating high-quality ranking. These results support the effectiveness of embeddings for building Kazakh-language QA systems. Importantly, this performance was achieved without fine-tuning— even in a zero-shot setting, pretrained models provided a substantial improvement over classical methods.

In the following section, we present a visual interpretation of how different models structure the semantic space using embedding projections.

Figure 2 illustrates the t-SNE projections of question and context embeddings produced by TF-IDF, mMiniLM, LaBSE, Alibaba GTE-multilingual-base, intfloat/multilingual-e5-small, and Snowflake Arctic Embed v2.0. Blue dots (or a contrasting color) represent question embeddings, while red dots indicate context embeddings. It is evident that the classical TF-IDF approach results in a sparse, disorganized cloud of points, whereas modern models like LaBSE, GTE, and Arctic tend to cluster semantically related questions and contexts closer together, which correlates with their superior ranking performance.

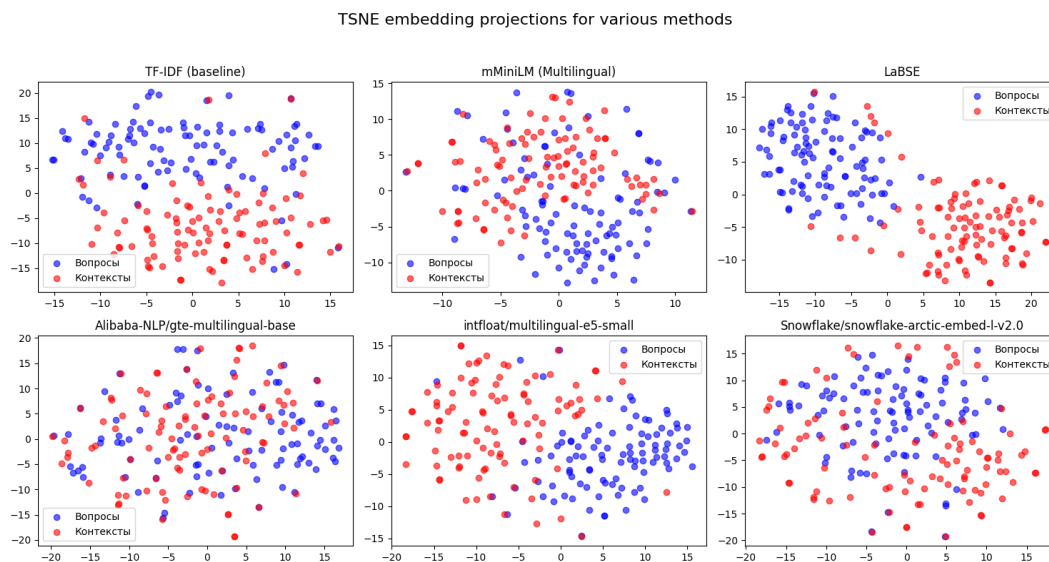


Fig.2 – t-SNE embedding projection for various methods

Conclusion. Within the framework of this study, a comparative analysis of the effectiveness of various methods of comparing questions and contexts in the Kazakh language was carried out – from classical TF-IDF to the latest multilingual embedding models. The experimental results showed that the pre-trained transformer models are significantly superior to the basic TF-IDF approach in terms of the accuracy of extracting answers to questions. The best quality was demonstrated by the LaBSE, Alibaba GTE-multilingual, and Snowflake Arctic v2.0 models, which correctly extracted the appropriate context for ~63% of questions versus ~50% for TF-IDF. The LaBSE model confirmed the high efficiency of language-independent embeddings, previously noted in the works of Google [3], in relation to the Kazakh language. The GTE and Arctic models, relatively new developments in 2024, have shown the ability to achieve even higher metrics without specialized customization to Kazakh, which indicates progress in the field of multilingual semantic representations [4,12]. More compact models (MiniLM, E5-small) also gave a significant advantage over TF-IDF, although they were inferior in quality to the older models, which indicates the still important role of scale and training data for multilingual embeddings [2,9].

Visual analysis using t-SNE projections confirmed that successful models place semantically related questions and answers close to each other in a vector space, whereas less accurate methods show a more chaotic or disjointed distribution. The difference between the discrete nature of TF-IDF and continuous transformer spaces was particularly evident: modern models

create structured embedding clouds that are more suitable for searching query–document pairs.

The practical significance of this work lies in the fact that even without additional training on specialized Kazakh data, pre-trained multilingual models can be used for question-and-answer search tasks. To build an educational AI system that answers students' questions in Kazakh, it is most promising to use models such as LaBSE or GTE/Arctic, which provide the highest quality of comparison. These models can be integrated into e-learning systems and reference systems, where the student asks a question in natural language, and the system finds an appropriate explanation or definition in the knowledge base.

It should be noted that there is room for improvement: although ~63% accuracy is a good result, in the remaining cases the model did not guess the answer on the first attempt. In the future, it is planned to explore the methods of further training (fine-tuning) of the considered models directly on the Kazakh Q&A corpus, as well as try larger versions of the models (for example, multilingual-E5-large or a larger version of GTE) in order to achieve even higher quality. Another direction could be a combination of approaches, for example, using embeddings to initially select candidates and then accurately match them using a cross-encoder. Nevertheless, the results already obtained emphasize the power of modern multilingual models: even for a relatively low-resource language, they can significantly improve the quality of semantic search, which opens up new opportunities for the development of educational and information systems in the Kazakh language.

References

1. Li Z., Zhang X., Zhang Y., Long D., Xie P., Zhang M. Towards General Text Embeddings with Multi-stage Contrastive Learning // arXiv preprint.2023. arXiv:2308.03281.14 p. DOI 10.48550/arXiv.2308.03281.
2. Reimers N., Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020, P.4512-4525. DOI 10.48550/arXiv.2004.09813
- 3.Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers

- for language understanding // Proceedings of NAACL-HLT.- 2019.- Vol.1. P.4171–4186. DOI 10.48550/arXiv.1810.04805.
4. Cross-encoder. mmarco-mMiniLMv2-L12-H384-v1 [Electronic resource] // Hugging Face.2024. URL: <https://huggingface.co> .- Date of address:16.04.2025.
5. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT sentence embedding // arXiv preprint.2020. arXiv:2007.01852. DOI 10.48550/arXiv.2007.01852.
6. Tleubayeva A., Mitroshina A., Arman A., Shokan A., Aigul S. Systemic approach to optimizing natural language processing technologies in Astana IT University' s admissions process // Proceedings of the 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST). Astana, Kazakhstan.2024. -P.520-525. DOI 10.1109/SIST61555.2024.10629481.
7. Maksutova K. Kazakh QA dataset Kundyzka/informatics_kaz [Electronic resource].2023. Available at: https://huggingface.co/datasets/Kundyzka/informatics_kaz.-Date of address:10.04.2025.
8. Tleubayeva A., Shomanov A. Comparative analysis of multilingual QA models and their adaptation to the Kazakh language // Scientific Journal of Astana IT University.-2024.- Vol.19. P.89-97. DOI 10.37943/19WHRK2878.
9. Wang H. Automatic question-answering modeling in English by integrating TF-IDF and segmentation algorithms//Systems and Soft Computing.-2024.-Vol.6:200087. DOI 10.1016/j.sasc.2024.200087.
10. Zhang X., Zhang Y., Long D., et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval//arXiv preprint.- 2024:2407.19669. DOI 10.48550/arXiv.2407.19669
11. Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual E5 text embeddings: a technical report//arXiv preprint.- 2024:2402.05672. DOI 10.48550/arXiv.2402.05672.
12. Yu P., Merrick L., Nuti G., Campos D. Arctic-Embed 2.0:Multilingual retrieval without compromise // arXiv preprint.2024. arXiv:2412.04506. DOI 10.48550/arXiv.2412.04506.
13. Susanti F., Maulidevi N. U., Surendro K. Improving embedding-based link prediction performance using clustering //Journal of King Saud University-Computer and Information Sciences.- 2024.-Vol.36(8): 102181. DOI 10.1016/j.jksuci.2024.102181.
14. Liu Y., Li Y., Xie D. Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks // Journal of Statistical Computation and Simulation.- 2023.- Vol.94. (1). - P. 183-203. DOI 10.1080/00949655.2023.2238235.

Сведения об авторах

Мажитова К. Ш. - магистр технических наук, преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: mazhitovaks@gmail.com;

Тлеубаева А. О.- докторант ТОО «Astana IT University», Астана, Казахстан, e-mail: Arailym.tll@gmail.com;

Мухаммедия С.Е.- магистр технических наук, преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: m.samal219@gmail.com;

Танирбергенова А.А.- доктор PhD, кафедра системного анализа и управления ЕНУ им.Л.Н.Гумилева, Астана, Казахстан, e-mail: alua_15_1982@mail.ru; Токтаганова А.С.- преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: ainur_7t@bk.ru;

Камбарова Ж. Б.- преподаватель специальных дисциплин Высшего колледжа «ASTANA POLYTECHNIC», Астана, Казахстан, e-mail: Kambarova_1976@mail.ru.

Information about the authors

Mazhitova K.Sh. - Master of Technical Sciences, lecturer of special disciplines at the Higher College "ASTANA POLY-

TECHNIC”, Astana, Kazakhstan, e-mail: mazhitovaks@gmail.com;

Tleubayeva A.O. - doctoral student at Astana IT University LLP, Astana, Kazakhstan, e-mail: Arailym.tll@gmail.com;

Mukhammediya S.E.- Master of Technical Sciences, lecturer of special disciplines at the Higher College ”ASTANA POLYTECHNIC”, Astana, Kazakhstan, e-mail: m.samal219@gmail.com;

Tanirbergenova A.A.- PhD, Department of System Analysis and Management, L.N.Gumilyov ENU, Astana, Kazakhstan, Astana, e-mail: alua_15_1982@mail.ru;

Toktaganova A.S.- is a teacher of special disciplines at the Higher College ”ASTANA POLYTECHNIC”, Astana, Kazakhstan, e-mail: ainur_7t@bk.ru;

Kambarova Zh. B.- is a teacher of special disciplines at the Higher College ”ASTANA POLYTECHNIC”, Astana, Kazakhstan, e-mail: Kambarova_1976@mail.ru.