Extending Monolingual Asymmetric Semantic Search Models for Multilingual Query Processing using Knowledge Distillation

Iryna Yurchuk[1,†] and Danylo Boiko[1,*,†]

[1] *Taras Shevchenko National University of Kyiv, Volodymyrska Street, 60, Kyiv, 01033, Ukraine*

Abstract

Semantic search is a key task in today's world, where the amount of data is growing rapidly. In this work, the main role is devoted to cases when long answers must be found to a short query (known as asymmetric search). The teacher model with 30,522 tokens of vocabulary and the student model with 119,547 tokens of vocabulary are basic for training a multilingual asymmetric semantic search model using multilingual knowledge distillation. The authors used the reciprocal rank (RR), the mean average precision (MAP), and the normalized discounted cumulative gain (NDCG) to evaluate the obtained model.

Keywords

asymmetric semantic search, multilingual embedding, knowledge distillation

1. Introduction

Semantic search is a set of search algorithms that work based on understanding the meaning of text. This approach effectively handles synonyms, abbreviations, and spelling errors, unlike keyword search engines that rely on exact lexical matches to find documents. It is useful for grading and assessing academic work [1], integrating search functionalities on e-commerce [2, 3], information retrieval in the petrochemical sector using a fusion of video transcript data with other data sources [4], helping Human Resources employees to target relevant people for their events and trainings [5], performing social search [6], IoT systems [7], and other different fields [8].

Ontologies and knowledge graphs are classical approaches to the design and implementation of semantic search systems, but recently these approaches have been enriched or replaced by statistical algorithms [9] and AI techniques [10] for query expansion, generating question-precise training statistics, semi-supervised gaining knowledge, etc.

This work aims to study the possibility of using the bi-encoder architecture to implement asymmetric semantic search when storage contains indexed data in English with the necessary to search information in Ukrainian.

2. Background and Algorithms

The intention behind semantic search is to transform passages into a multidimensional vector space. During the search, the query is similarly embedded in this vector space, which allows to identify the

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

i.a.yurchuk@gmail.com (I. Yurchuk); danylo.boiko@knu.ua (D. Boiko)

0000-0001-8206-3395 (I. Yurchuk); 0009-0005-6341-0095 (D. Boiko)

desired quantity of most relevant matches. This approach ensures that even if the wording is different, the meaning remains the same and the system can provide accurate retrievals (Figure 1).

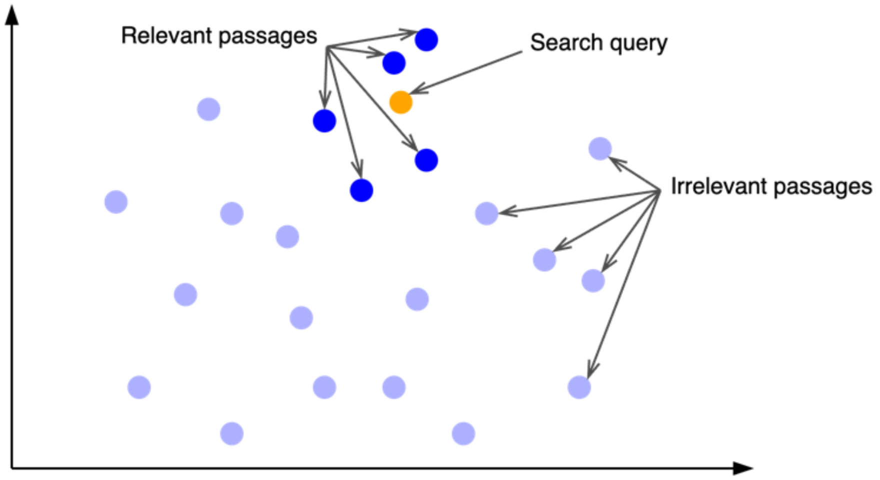

Figure 1: Search in the vector space.

There are two main types of semantic search that work with different types of data. For symmetric search, queries and passages in the corpus have approximately the same length and content. In turn, asymmetric search typically uses short queries (e.g., a question or a few keywords) and longer passages answering those queries.

Nowadays, a lot of trained asymmetric semantic search models based on both bi-encoder and cross-encoder architectures [11] are available for English. Bi-encoder models are efficient for large-scale retrieval due to their ability to use precomputed passage embeddings from storage, making them ideal for speed-critical tasks, while the cross-encoder models are known for higher accuracy by directly capturing similarity between query and passages, suited for accuracy-sensitive scenarios like reranking where computational cost is less of a concern.

To use asymmetric semantic search for less common languages or even a mix of them, we need to train new models, which will require a lot of data and computational power. Fortunately, there is a way to facilitate this using multilingual knowledge distillation [12] (Figure 2).

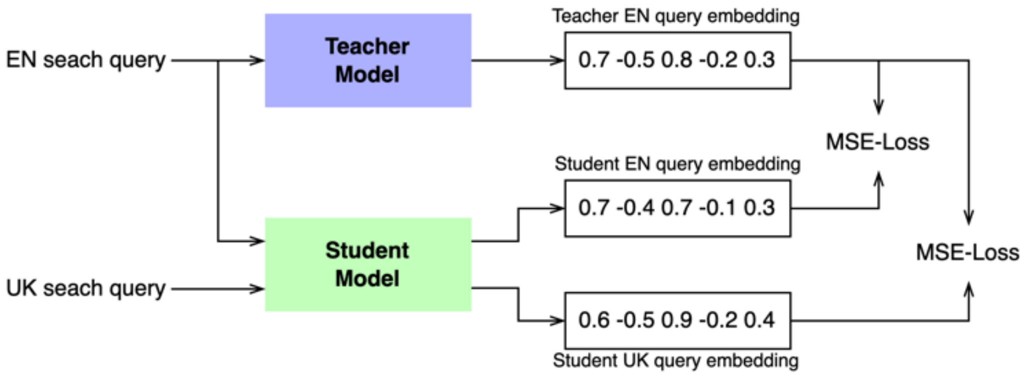

Figure 2: The idea of multilingual knowledge distillation.

This approach requires a teacher model for the source language and a set of pairs (each pair includes a sentence in the source language and its translation). A new student model attempts to approximate the output of a teacher model for both source and target sentences using the mean squared error (MSE) loss. The student model could have the structure and weights of the teacher model, or it could be a different network architecture since the student model learns representation of the teacher model. This allows the student model to achieve robust generalization across languages.

It is critical that identical sentences in different languages have similar vector representations. That is why the vector space properties of the source language, obtained from the teacher model, must be applied to other languages.

From the information above, it is evident that the input parameters of semantic search models are queries and passages. In the field of multilingual search, there are three main paradigms:

- *Multilingual-to-monolingual*: this approach allows to accept queries in multiple languages and compare them with passages in a single language.
- *Monolingual-to-multilingual*: conversely, the idea behind this paradigm is to accept queries in a single language and return passages in multiple languages.
- *Multilingual-to-multilingual*: provides the highest search adaptability and allows users to formulate queries and peruse passages across multiple languages. However, this flexibility may reduce the average accuracy due to the large number of possible language pairs.

This paper explores the scenario when a storage contains indexed data in English and it becomes necessary to search information in other languages, particularly Ukrainian. We will take the original model (data in the index must be produced by this model) trained for asymmetric semantic search, using the bi-encoder architecture, and add capabilities to handle multilingual queries (in our case, bilingual, i.e., English and Ukrainian).

Using knowledge distillation will significantly reduce the training time since we will mimic the teacher model on translated queries. To avoid unnecessary training on passages and to ensure consistency between previously indexed and new documents, we will continue to use the teacher model to create embeddings for new passages. The output of both models can be evaluated using a similarity metric to find the most relevant passages (Figure 3).

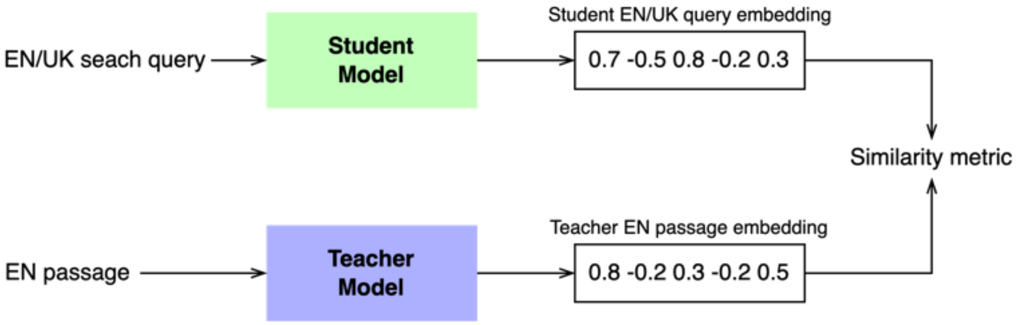

Figure 3: Pipeline of using the trained and original model.

To determine how similar two vectors are, we can use different metrics (cosine similarity, Euclidean distance, and dot product). Measuring Euclidean distance for high dimensional vectors becomes impractical, as they will be very far apart simply because of the vastness of the space they inhabit. Using cosine similarity, which measures the angle between two vectors by paying attention to their direction and ignoring magnitude, or dot product, which measures the overall congruence of two vectors by considering both their direction and magnitude, will help to avoid the "curse of dimensionality".

In practice, we can generate sentence embeddings using specially designed modifications of the Bidirectional Encoder Representations from Transformers (BERT) model, known as Sentence Transformers (SBERT). It provides two types of state-of-the-art asymmetric semantic search models, one tuned for cosine similarity, the other for dot product. Cosine similarity tuned models prefer to retrieve shorter passages, while dot product tuned models prefer to retrieve longer passages.

3. Datasets

3.1. MS MARCO

Microsoft Machine Reading Comprehension (MS MARCO) [13] is a large-scale dataset created by Microsoft for training and evaluating information retrieval systems. It is widely used to benchmark the performance of various models on tasks such as reading comprehension, question answering, and passage ranking.

The dataset comprises over a million anonymous questions extracted from Bing's search logs, offering a collection of short, real-world, natural language queries split into the train (808,731 queries), development (101,093 queries), and evaluation (101,092 queries) subsets. We will use the train and development subsets to train and evaluate the student model respectively, so it makes sense to look at their length and word patterns (Figure 4).

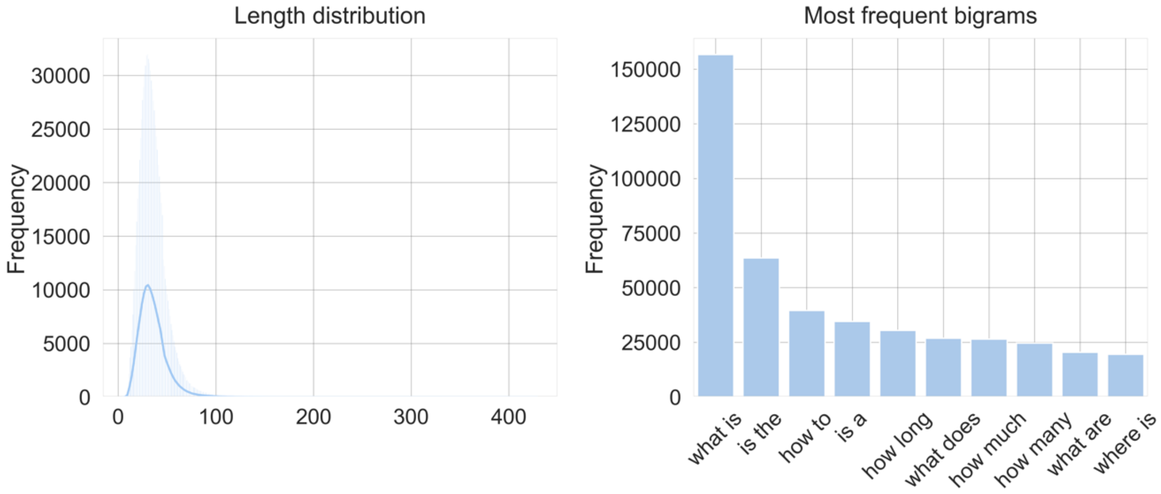

Figure 4: Length and word patterns of the queries from the train and development subsets.

Furthermore, the dataset contains 8,841,823 passages that are required to provide natural language answers (a question may have multiple answers or no answers at all). Although passages are not involved in the knowledge distillation for multilingual-to-monolingual models, it still seems reasonable to analyze them to figure out what the teacher model was trained on (Figure 5).

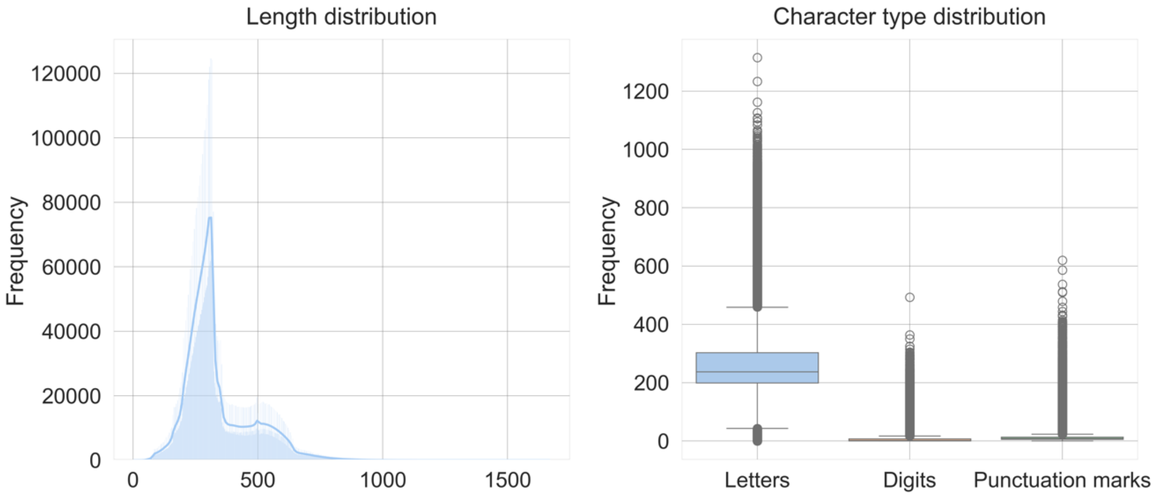

Figure 5: Length and character type distribution of the passages.

The length distribution of the combination of queries from the train and development subsets is markedly skewed, with the majority being quite brief, typically under 100 characters, predominantly concentrated between 15-25 characters. The frequency drops significantly for longer queries, with very few extending beyond 100 characters. The most common bigrams (two-word sequence of words) indicate that the majority of queries are fact-oriented. These patterns reveal that a significant number of queries are short questions asking for factual information, often beginning with “what,” “how,” or “is,” reflecting the nature of queries in the dataset.

The length distribution of the passages has a right-skewed plot with the noticeable peak around 200-300 characters. The character type distribution shows that letters are mostly used, with a median of about 250 characters per passage and a wide interquartile range reflecting considerable variation. Digits and punctuation marks are used less frequently, but there are outliers indicating some passages with a large number of such characters.

The scale and real-world nature of the dataset makes it attractive for training and evaluating machine learning models, but the original MS MARCO contains only English search queries. To train and evaluate models for other languages, the dataset must be translated in one of the available ways. For Ukrainian, the OPUS-MT English to East Slavic neural machine translation model [14] demonstrated high-quality results in a relatively short time.

3.2. TREC 19

In 2019, the National Institute of Standards and Technology (NIST), in collaboration with Microsoft, organized the Text Retrieval Conference (TREC) deep learning track [15] benchmark competition. This event was aimed to foster research in the information retrieval direction using deep learning techniques. As the official evaluation dataset, organizers provided a list of 200 queries and a pool of documents and passages, labeled by NIST assessors using multi-graded judgments.

Although the deliberate selection of the queries at an intuitive level implies high-quality data and a lack of outliers, it still seems reasonable to check directly the length and word patterns (Figure 6).

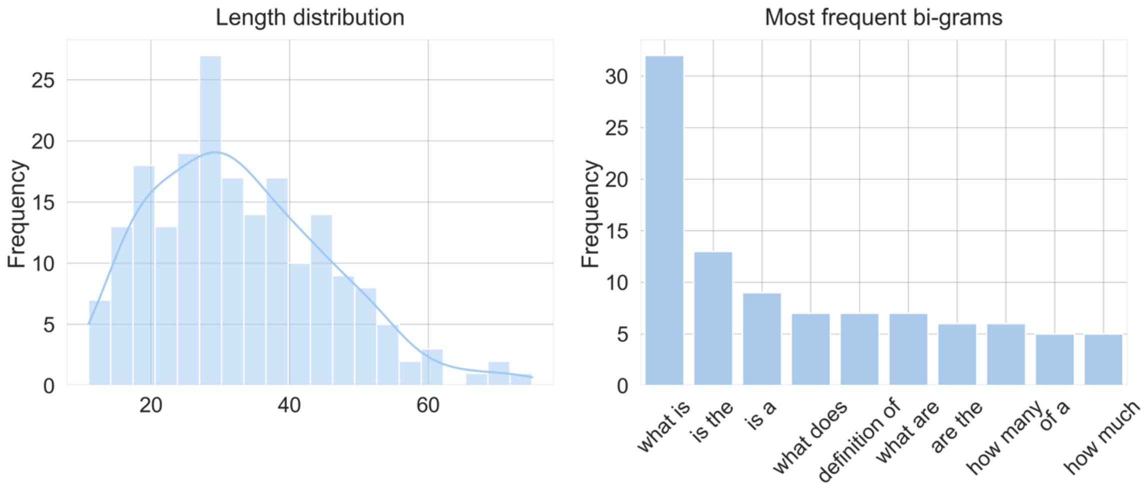

Figure 6: Length and word patterns of the queries from the evaluation dataset.

The length distribution peaks at 20-25 characters, indicating that most queries are concise and right-skewed, with fewer longer queries beyond 60 characters. The most frequent bigrams begin with common phrases such as "what," "definition," and "how," suggesting that queries are structured as questions, often seeking definitions or factual information.

Looking at the queries from both datasets, it is clear that TREC 19 queries are generally shorter and more focused on concrete evidence, while MS MARCO queries might be significantly longer and more complex. It is also important to note that TREC 19 does not have outliers, unlike MS MARCO, where some queries can exceed 400 characters.

These days, the TREC 19 has become a recognized benchmark for various information retrieval and deep learning tasks such as document and passage retrieval. In the task of full search, we can evaluate up to 1,000 passages for each query based on their estimated likelihood of containing the answer.

To evaluate the trained model, all 200 queries were manually translated into Ukrainian by a native speaker, which guarantees the veracity of the results.

4. Training and Evaluation

We will use the DistilBERT base multilingual (cased) [16] as a student model. This model was trained on the concatenation of Wikipedia in 104 different languages (including English and Ukrainian), has 6 layers, 768 dimensions, and 12 heads, totalizing 134 million parameters (compared to 177 million parameters for the BERT base multilingual).

The MS MARCO DistilBERT base v4 [17] was chosen as a teacher model. It embeds text into a 768-dimensional vector space and can be used for clustering and semantic search. This model was fine-tuned on the original MS MARCO passage ranking dataset and optimized to generate embeddings for queries and passages.

We trained the student model for 5 epochs with a batch size of 24, 10,000 warm-up steps, and a learning rate of 2e-5. The entire training process on the train and development (used for intermediate evaluation) subsets took about 8 hours using the Apple M3 Max chip (16-core CPU, 40-core GPU). To measure the difference between computed and target query embeddings we used the MSE loss (Figure 7).

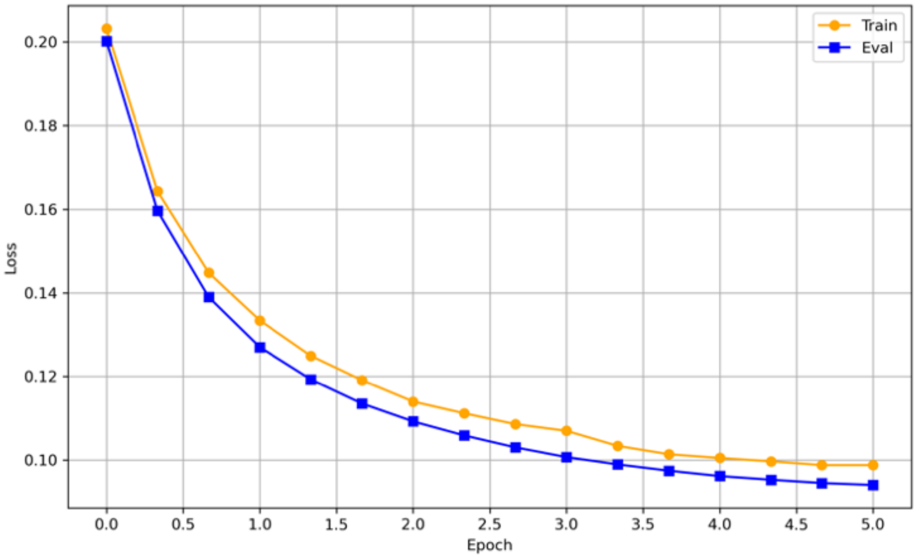

Figure 7: MSE loss on the train and evaluation subsets.

The teacher model has a vocabulary of 30,522 tokens, the same as the original BERT base. This vocabulary includes common words, sub-words, and special tokens to deal with English. The student model, on the other hand, is a distilled version of the BERT multilingual. The extended vocabulary of 119,547 tokens covers a variety of multiple symbols from different languages, allowing to efficiently process and understand text in different linguistic contexts.

Any dataset for semantic search or information retrieval systems has selection bias. It can be related to the data source, the date of publication, and the personal preferences of the publisher. Achieving high performance on cross-lingual tasks depends on the ability to seamlessly map sentences from different languages into a single vector space. The similarity between alphabets also significantly affects accuracy. Languages with similar alphabets, such as English and German, typically produce more accurate results than languages with dissimilar alphabets, such as English and Ukrainian.

Using the original English and machine-translated Ukrainian search queries, we achieved a reasonable level of accuracy (about 93%) for English to Ukrainian and Ukrainian to English translation tasks, indicating a high level of proficiency in both directions (Figure 8).

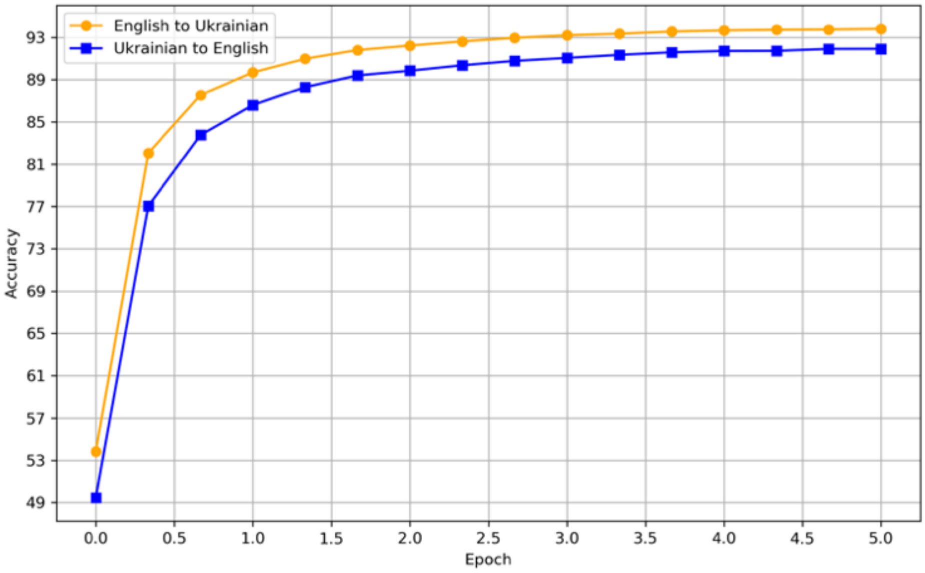

Figure 8: Accuracy of translation between English and Ukrainian.

It is very important to monitor performance on the subset that did not participate in training to avoid overfitting. MSE on the evaluation subset steadily decreases, indicating that the student model learns to more accurately mimic the results of the teacher model (Figure 9).

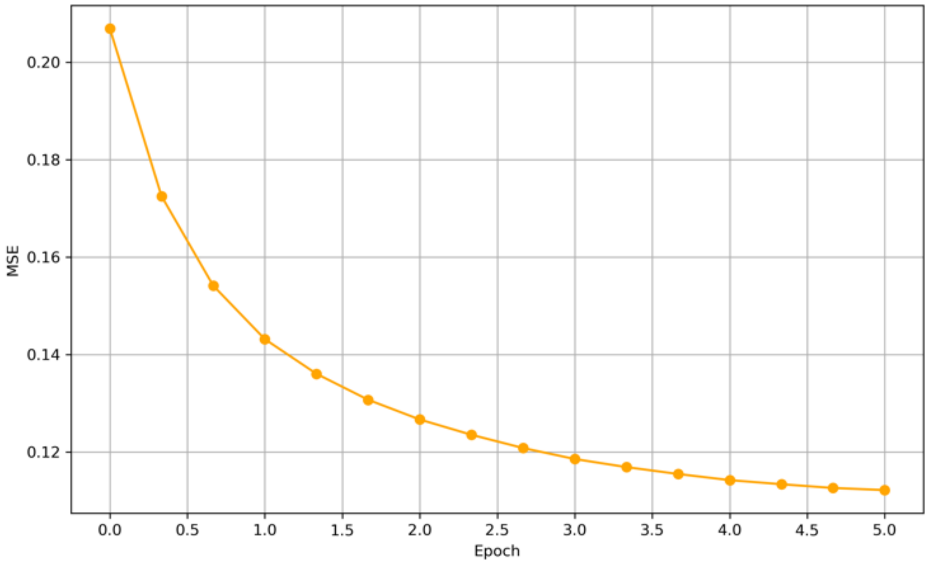

Figure 9: MSE on the evaluation subset.

Evaluating the performance of information retrieval systems is a critical step in improving their efficiency. NIST assessors labeled the TREC 19 using multi-graded judgments, making it easy to measure all the necessary and widely used state-of-the-art metrics.

Reciprocal rank (RR) calculates the score for the first relevant passage in the ranked list. This metric is very important when we need to evaluate the occurrence of the most relevant passage:

$$RR = \frac{1}{rel}, \tag{1}$$

where *rel* – the rank of the first relevant passage in the list.

Mean average precision (MAP) evaluates both the relevance of the suggested passages and the position of the most relevant passages at the top. For each query, the average precision (AP) is determined by calculating the arithmetic mean of the precision scores for every position in which a relevant passage was found. This metric focuses on the ability to distinguish relevant and irrelevant items:

$$MAP@k = \frac{1}{Q}\sum_{i=1}^{Q} AP@k_i, \tag{2}$$

where Q – the total number of queries.

Normalized discounted cumulative gain (NDCG) measures the ability of machine learning algorithms to sort passages by relevance. NDCG is determined by dividing the discounted cumulative gain (DCG) by the ideal DCG, representing the best version of the rating. This is useful in many scenarios where we expect passages to be sorted by relevance:

$$DCG@k = \frac{DCG@k}{IDCG@k} = \frac{\sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}}{\sum_{i=1}^{k} \frac{rel_i^*}{log_2(i+1)}}, \tag{3}$$

where k – the number of items considered in the calculation.

Both MAP and NDCG reflect ranking quality, but account for rank reduction in different ways. MAP gives more weight to relevant passages at the top of the list because this metric is based on precision, making it more sensitive to changes in early positions. DCG assigns decreasing weights to passages proceeding down the hierarchical order, but they are logarithmic and reduce the contribution of passages not very quickly.

In the case of models based on the bi-encoder architecture, it makes sense to evaluate up to 100 passages using the above metrics (Table 1). If the obtained performance is not sufficient, a reranking model based on the cross-encoder architecture can be used to improve results.

Table 1
Performance of the original and trained model on TREC 19

Model	TREC 2019 EN-EN			TREC 2019 UK-EN		
	RR	MAP@100	NDCG@100	RR	MAP@100	NDCG@100
msmarco-distilbert-base-v4	0.96	0.35	0.59	–	–	–
msmarco-distilbert-multilingual-en-uk	0.92	0.29	0.53	0.78	0.25	0.46

The bilingual model trained using knowledge distillation inherited the capabilities of the monolingual teacher model and shows reasonable results for both languages. Obtained 768-dimensional vectors are optimized to work with cosine similarity as expected.

We can notice a slight decrease in performance for English, which is not critical since we got the ability to search passages using multiple languages. A higher score does not necessarily mean higher performance in production. At some point, models can become too specialized on the MS MARCO and its selection bias.

It is impossible to create a perfect dataset for semantic search. Manual creation of a subset even for evaluation is quite expensive and unfortunately always has some selection bias. This is a long-recognized problem, but there is no good solution for it, especially at the scale of millions of queries and passages.

The number of languages may not be limited to two, but it is better to reasonable add only justified languages and keep a balance between performance and multilingual search capabilities. Otherwise, we will face the “curse of multilinguality”, where adding new languages to the model degrades performance because the capacity of the model remains the same.

The performance of the trained model was also affected by the quality of the data, which in our case is a combination of original and machine-translated queries. Cloud solutions such as Google Translate or DeepL could marginally improve the results but would not compare to manual translation by native speakers. The time taken to create pairs using the neural machine translation model was about 50% longer than the time required to train the student model. Intuitively, it makes the knowledge distillation less of a training and more of a translation task.

5. Conclusions

For multilingual processing of asymmetric semantic search queries, the DistilBERT base multilingual (cased) as a student model and the MS MARCO DistilBERT base v4 as a teacher model can be used as components in the scenario when storage contains indexed data in English with the necessity to search for information in Ukrainian up to the reciprocal rank (RR), the mean average precision (MAP), and the normalized discounted cumulative gain (NDCG) as its evaluating.

The authors achieved a reasonable level of accuracy (about 93%) for English to Ukrainian and Ukrainian to English translation tasks, indicating a high level of proficiency in both directions. In the future, this result will be improved by the quality of the data, which can be a combination of original queries and manual translations by native speakers. The resulting model is useful in industries such as finance, healthcare, and e-commerce, where huge data sets are prevalent and asymmetric semantic search plays a key role to quickly retrieve relevant information.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] T. Kgosietsile, E. U. Okike, An Intelligent Semantic Vector Search Model for Grading and Assessing Students, in: 2023 International Conference on Sustainable Technology and Engineering (i-COSTE), Nadi, Fiji, 2023, pp. 1-6. doi: 10.1109/i-COSTE60462.2023.10500811.

[2] S. Shirol, A. Kulkarni, R. Agarwal, Semantic Search for Sustainable Platforms Using Transformers, in: 2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 2023, pp. 112-118. doi: 10.1109/ICETCI58599.2023.10331079.

[3] F. Aamir, R. Sherafgan, T. Arbab, A. Jamil, F. N. Bhatti, A. A. Hameed, Deep Learning-based Semantic Search Techniques for Enhancing Product Matching in E-commerce, in: 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 2024, pp. 1-9. doi: 10.1109/ICMI60790.2024.10586148.

[4] K. P. Saikia, D. Mukherjee, S. Mahapatra, P. Nandy, R. Das, Unveiling Deeper Petrochemical Insights: Navigating Contextual Question Answering with the Power of Semantic Search and LLM Fine-Tuning, in: 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 881-886. doi: 10.1109/ICCCIS60361.2023.10425564.

[5] D. Sheth, A. R. Gupta, L. D’Mello, Using Universal Sentence Encoder for Semantic Search of Employee Data, in: 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-4. doi: 10.1109/ICCICA52458.2021.9697114.

[6] I. Sindhu, F. Shamsi, Semantic Social Searching-An Ontology Based Approach, in: 2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT), Karachi, Pakistan, 2023, pp. 1-4. doi: 10.1109/IMCERT57083.2023.10075145.

[7] P. Singh, K. S. Acharya, M. J. Beliatis, M. Presser, Semantic Search System For Real Time Occupancy, in: 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), Bandung, Indonesia, 2021, pp. 49-55. doi: 10.1109/IoTaIS53735.2021.9628719.

[8] Y. Zheng, An Analysis of the Technical Trend of Semantic Search in Natural Language Processing, in: 2023 9th Annual International Conference on Network and Information Systems for Computers (ICNISC), Wuhan, China, 2023, pp. 51-53. doi: 10.1109/ICNISC60562.2023.00033.

[9] Md, Blending Weighted TF-IDF & BERT for Improving Semantic Search, in: 2022 2nd International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 2022, pp. 154-159. doi: 10.1109/ICARC54489.2022.9753875.

[10] R. V, D. Dhabliya, M. Mathur, S. Das, R. Kumar, S. B. Rao, Ameliorating Semantic Search Through Advanced AI Techniques, in: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-6. doi: 10.1109/SMARTGENCON60755.2023.10442780.

[11] J. Liao, M. Jia, J. Duan, J. Wang, FBC: Fusing Bi-Encoder and Cross-Encoder for Long-Form Text Matching, in: Frontiers in Artificial Intelligence and Applications, Krakow, Poland, 2023, pp. 1473-1480. doi:10.3233/faia230426.

[12] N. Reimers, I. Gurevych, Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4512-4525. doi: 10.18653/v1/2020.emnlp-main.365.

[13] MS MARCO. URL: https://microsoft.github.io/msmarco.

[14] J. Tiedemann, S. Thottingal, OPUS-MT – Building open translation services for the World, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479-480. URL: https://aclanthology.org/2020.eamt-1.61.

[15] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, I. Soboroff, TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2369-2375. doi: 10.1145/3404835.3463249.

[16] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019. doi: 10.48550/arXiv.1910.01108.

[17] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3973-3983. doi: 10.18653/v1/D19-1410.