Enhancing Kazakh Text Embeddings through Sentence Pair Classification on Translated Datasets

Alim Tleuliyev
Nazarbayev University / Astana, Kazakhstan
alim.tleuliyev@nu.edu.kz

Abstract

This study focuses on enhancing Natural Language Processing (NLP) capabilities for the Kazakh language, a relatively low-resource language, by employing transformer-based models fine-tuned on datasets translated from high-resource languages. Leveraging the Quora Question Pair and Microsoft Research Paraphrase Corpus we aimed to bridge the technological gap by adapting these resources to Kazakh. Our approach involved fine-tuning several transformer models, including BERT, RoBERTa, and LaBSE, on these translated datasets to assess their efficacy in sentence pair classification and semantic similarity tasks specific to Kazakh. Although the performance improvements were modest, our results demonstrate the potential of using translated datasets to enhance NLP capabilities in low-resource languages. This research not only sheds light on the challenges faced but also highlights future directions for improving semantic understanding and language technology accessibility for the Kazakh language and beyond.

1 Introduction

In recent years, the field of Natural Language Processing (NLP) has seen remarkable advancements, largely due to the development and application of transformer models. These models have set new benchmarks across a variety of tasks, including but not limited to, sentiment analysis, text summarization, and sentence pair classification. However, the preponderance of research and model development has been centered around high-resource languages, such as English, leaving low-resource languages, including Kazakh, with limited tools for advanced NLP tasks.

This project aims to bridge this gap by translating high-quality datasets known in English NLP research — specifically, the Quora Question Pair (QQP) dataset (Quora, 2017) and the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) from the GLUE benchmark (Wang et al., 2019) — into Kazakh language. We prepare these datasets for training and evaluating transformer-based models (Vaswani et al., 2023) on sentence pair classification tasks and semantic similarity assessments in Kazakh. The goal is to understand how well transformer models, fine-tuned on these translated datasets, can capture semantic relations between sentence pairs in Kazakh, a language with significantly fewer resources compared to English.

2 Data

This project employs three significant datasets to train and evaluate the performance of transformer models on the Kazakh language. Each dataset has been translated from English to Kazakh. Below, we detail the origin, composition, and purpose of these datasets.

2.1 Microsoft Research Paraphrase Corpus

The Microsoft Research Paraphrase Corpus (MRPC) is a benchmark dataset for the task of identifying whether two sentences are paraphrases of each other (Dolan and Brockett, 2005). It consists of sentence pairs automatically extracted from online news sources, with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship. We are going to use it for fine-tuning. Table 1 shows the distribution of the dataset.

Split	Length	0	1
Training	3,668	1,194	2,474
Validation	408	129	279
Test	1,725	578	1,147

Table 1: MRPC dataset splits and label distributions. Labels indicate whether the questions are paraphrases (1) or not (0).

2.2 Quora Question Pair Dataset

Quora Question Pairs (QQP) dataset comprises over 400,000 question pairs (Quora, 2017). Each pair is annotated with a binary label indicating whether the two questions are considered paraphrases of each other. This dataset serves as a substantial resource for training and evaluating NLP models on the task of semantic equivalence detection between sentences. For the purposes of our project, we utilized a subset of 10,000 question pairs randomly selected from the original dataset to fine-tune transformer models. Table 2 shows the distribution of the dataset.

Split	Length	0	1
Training	7,000	4,391	2,609
Validation	989	603	386
Test	2,011	1,285	726

Table 2: QQP dataset splits and label distributions. Labels indicate whether the questions are paraphrases (1) or not (0).

3 Methods

3.1 Datasets

The MRPC and QQP datasets were merged to form a comprehensive Sentence Similarity Dataset (SSD), which served as the foundation for model fine-tuning. Table 3 shows the distribution of the dataset.

Split	Length	0	1
Training	10,668	5,585	5,083
Validation	1,397	732	665
Test	3,736	1,863	1,873

Table 3: Sentence Similarity Dataset splits and label distributions. Labels indicate whether the questions are paraphrases (1) or not (0).

The translation of the datasets into Kazakh was facilitated through the use of the `yandexfreetranslate` library on GitHub, which is a free API that uses Yandex Translate. All datasets were initially accessed via Hugging Face's `datasets` library.

3.2 Models

Our study utilizes a selection of transformer-based models, each accessed through the Hugging Face's `transformers` library. These models were chosen for their unique features and adaptability to multilingual understanding and Kazakh-specific contexts.

3.2.1 BERT Base Multilingual Cased

A variant of BERT (Devlin et al., 2019) designed for multilingual applications, pre-trained on the Fill Mask task. This model covers over a hundred languages, including Kazakh, making it a foundational tool for multilingual NLP tasks (Devlin et al., 2018). This model is accessible at Hugging Face under the identifier `google-bert/bert-base-multilingual-cased`.

3.2.2 Kazakh RoBERTa Conversational

Offered by Beeline Kazakhstan, this RoBERTa-based model (Liu et al., 2019) has been specifically fine-tuned on a large Kazakh corpus for conversational understanding. Its training on the Fill Mask task enhances its capabilities in semantic understanding within the Kazakh context. This model is accessible at Hugging Face under the identifier `kz-transformers/kaz-roberta-conversational`.

3.2.3 KazakhBERTmulti

This model, based on BERT (Devlin et al., 2019), underwent pre-training on the Fill Mask task and was further refined on a comprehensive Kazakh corpus. The fine-tuning process aimed to adapt it more closely to the nuances of the Kazakh language (Bekbulatov and Kartbayev, 2014). This model is accessible at Hugging Face under the identifier `amandyk/KazakhBERTmulti`.

3.2.4 Multilingual E5 Large

Leveraging the XLM-RoBERTa architecture (Conneau et al., 2020), this model was trained on sentence similarity tasks across multiple languages. Its design is suitable for applications requiring understanding and comparison of sentence-level semantic similarities in a multilingual setting (Wang et al., 2024). This model is accessible at Hugging Face under the identifier `intfloat/multilingual-e5-base`.

3.2.5 Language-agnostic BERT Sentence Embedding

Employing BERT's robust architecture (Devlin et al., 2019), LaBSE has been specifically trained for sentence similarity, creating embeddings that perform well across various languages. This model is pivotal for tasks that require language-agnostic sentence embeddings for accurate semantic similarity assessment (Feng et al., 2022). This model is accessible at Hugging Face under the identifier `sentence-transformers/LaBSE`.

Each model's selection was informed by its pretraining background and its anticipated compatibility with or adaptability to the Kazakh language for our NLP tasks.

3.3 Fine-Tuning

The models were fine-tuned using Hugging Face's `Trainer` API for sentence pair classification, focusing on semantic similarity. To accommodate our task, we replaced the models' original heads with a classification head that outputs two classes: semantically similar (1) and not similar (0). This adjustment allowed the models to assess sentence pairs for semantic similarity, a feature not originally included in their pre-training.

Fine-tuning was comprehensive, updating all model weights to ensure they adapted well to the nuances of our task. This approach leveraged the models' pre-trained knowledge while optimizing them for sentence pair classification on our custom Sentence Similarity Dataset.

3.4 Evaluation

For the evaluation process, we utilized the fine-tuned models to extract sentence embeddings from the last hidden layer using mean pooling, specifically employing the last element for this purpose. These embeddings were then leveraged to compute cosine similarities between sentence pairs. To align the similarity scores with our binary classification task, we converted these cosine similarities to probabilities ranging from 0 to 1.

The model's performance was assessed using the ROC AUC metric, comparing the probability scores against the binary labels of the SSD test set. This metric was selected for its effectiveness in evaluating the models' ability to distinguish between semantically similar and dissimilar sentence pairs, providing a nuanced understanding of model accuracy in semantic similarity detection.

4 Experiments

4.1 Experimental Setup

A generic training setup function, `setup_training`, was defined to configure model training parameters, including epochs, batch size, warmup steps, and evaluation strategies. This function dynamically creates a directory for each model to store training results, logs, and the best-performing model based on loss. It also initializes a data collator for padding and a metric for evaluation from the GLUE MRPC benchmark.

4.2 Training Configurations

Experiments were conducted with five different models, each selected for its potential in multilingual understanding or specific tuning for the Kazakh language. The configurations varied in terms of epochs and batch sizes, as summarized in Table 4.

Model	Batch Size	Epochs
BERT Multilingual	64	4
Kazakh RoBERTa	96	4
KazakhBERTmulti	64	4
Multilingual E5	64	5
LaBSE	32	5

Table 4: Training configurations for each model.

The default learning rate for all experiments was set to 5×10^{-5}.

4.3 Evaluation Methodology

Our evaluation methodology centers on the extraction of mean-pooled embeddings from the last hidden layer of each model, followed by the computation of cosine similarities between sentence pairs. This process was implemented for both base models and their fine-tuned counterparts, facilitating a direct comparison of their performance on the sentence similarity task.

The `get_mean_pooled_embeddings` function is instrumental in this process, leveraging the Hugging Face's `AutoTokenizer` and `AutoModel` to process texts in batches. This function calculates the mean-pooled embeddings by accounting for the attention mask, ensuring that only the meaningful parts of each sentence contribute to the final embedding.

5 Results and Analysis

5.1 Fine-Tuning Results

We present a comparative analysis of Fine-Tuning Results in Figure 1. The figure illustrates each model's accuracy and F1 score, providing a visual representation of their effectiveness on the sentence similarity classification task.

The LaBSE model outperformed others, achieving the highest accuracy and F1 score. This demonstrates the efficacy of language-agnostic embed-

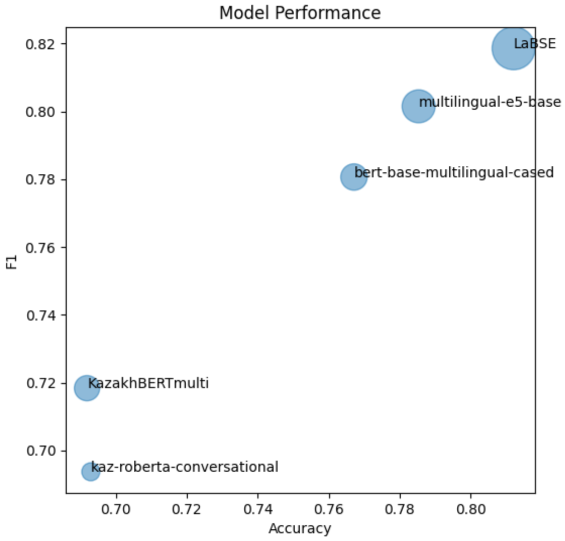

Figure 1: Fine-tuning results, showcasing the trade-off between accuracy and F1 score for each model. The size of the point indicates the number of parameters.

dings in handling sentence similarity tasks across diverse languages, likely benefiting from its extensive training on a large-scale multilingual corpus.

Following LaBSE, the Multilingual XLM-RoBERTa model (multilingual-e5-base) showed commendable performance, suggesting that its architecture is well-suited to the nuances of cross-linguistic semantic understanding. The BERT Base Multilingual Cased model also delivered solid results, reinforcing the versatility of the BERT architecture in multilingual settings.

Models explicitly fine-tuned for Kazakh, such as the Kazakh RoBERTa Conversational and KazakhBERTmulti, exhibited lower accuracy and F1 scores in comparison. While these models are tailored specifically for the Kazakh language, their specialized training does not necessarily translate to superior performance in this semantic similarity task. This could point to the need for more diverse training data or more refined tuning to fully capture the semantic nuances required for the task.

5.2 Evaluation Results

In the evaluation phase, models were assessed based on the Area Under the ROC Curve (AUC) metric. Figure 2 displays the AUC scores for both the base and fine-tuned models. The scatter plot differentiates the scores of the base models from those of the fine-tuned models, allowing for a visual comparison of their performance in terms of AUC score.

In assessing the impact of fine-tuning on the performance of various models, our results, as detailed in Table 5. The data reveals that while fine-tuning has generally led to improvements in AUC scores, the degree and direction of change are model-specific.

The **Kazakh RoBERTa Conversational** model's slight decrease in AUC score post fine-tuning suggests potential overfitting or a misalignment between the fine-tuning procedure and the evaluation dataset. Given the model's specialization in the Kazakh language, the discrepancy may also be attributed to the nature of the training data, which could be less diverse or comprehensive in capturing the broader semantic patterns necessary for the task.

For the **KazakhBERTmulti**, the small increase in performance indicates that fine-tuning had a positive effect, though it was not very significant. This could be due to the model's initial pre-training already positioning it close to an optimal state for the task, leaving limited room for improvement.

The **bert-base-multilingual-cased** model demonstrated a clear benefit from fine-tuning, improving its AUC score significantly. This improvement likely stems from the BERT architecture's robust foundation, which, when coupled with task-specific tuning, can more effectively model the intricacies of sentence similarity.

Contrastingly, the **multilingual-e5-base** model experienced a decline in AUC score following fine-

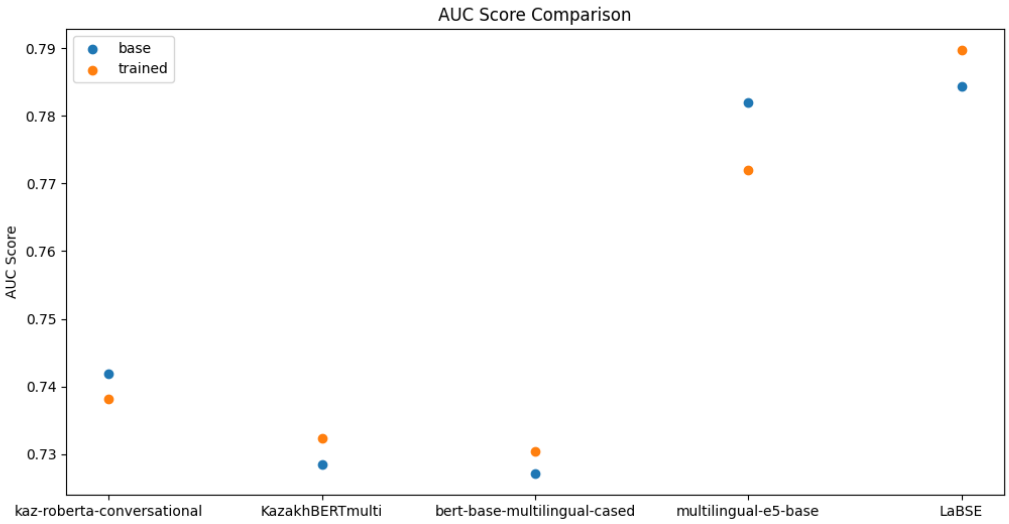

Figure 2: Comparison of AUC scores between base models and fine-tuned models. Each point represents the AUC score for a model, with the blue points indicating base models and the orange points indicating fine-tuned models.

tuning. This could imply that the model's base configuration was better suited to the evaluation task than its fine-tuned counterpart, or that the fine-tuning diverged from the optimal trajectory for the given task—potentially due to an overly narrow focus during the fine-tuning phase.

Most notably, the **LaBSE** model showcased an uptick in AUC score after fine-tuning, underlining the capability of advanced, language-agnostic embeddings to adapt and excel when further refined with task-specific data. The high base AUC score of LaBSE indicates a strong pre-existing capacity for semantic understanding, which fine-tuning has successfully built upon.

Model	**Base**	**Trained**
Kaz RoBERTa	0.74189	0.73820
KazakhBERTmulti	0.72848	0.73240
bert-base-multilingual	0.72714	0.73042
multilingual-e5-base	0.78200	0.77201
LaBSE	0.78431	0.78972

Table 5: AUC scores for base and fine-tuned models.

Despite the variations in model performance post fine-tuning, the observed increases in AUC scores were modest. This can likely be attributed to the limited size of the Sentence Similarity Dataset, which, at 10,000 pairs for training, pales in comparison to the extensive corpora on which these models were originally pre-trained. The models require a larger volume of task-specific data to achieve significant improvements. This premise, coupled with the small yet positive gains witnessed, suggests a promising potential for enhancing the quality of Kazakh embeddings. Future research should focus on compiling a more substantial dataset, tailored to the Kazakh language, to further refine the fine-tuning process. Such efforts are expected to yield more pronounced benefits and could set a new benchmark for Kazakh language processing.

6 Conclusion

This project has made significant strides in advancing Natural Language Processing capabilities for the Kazakh language by fine-tuning transformer-based models on translated datasets. Through meticulous experimentation with a variety of transformer architectures, including BERT, RoBERTa, and language-agnostic models such as LaBSE, we have explored the potential for enhancing Kazakh text embeddings and semantic understanding in a low-resource language context.

Our findings reveal that, although the improvements in model performance post-fine-tuning were not as pronounced as hoped, there was a demonstrable enhancement across several metrics. The LaBSE model, in particular, exhibited notable gains, underscoring the effectiveness of language-agnostic embeddings in capturing semantic nuances across languages. This success points towards the promising applicability of advanced embedding techniques in improving NLP tasks for languages with limited digital resources.

However, the limited size of the translated Sentence Similarity Dataset, and the slight performance variations among the models, underscore the challenges inherent in NLP research for low-resource languages. These findings highlight the critical need for larger, more diverse datasets tailored to the specific linguistic features and nuances of the Kazakh language.

Moreover, our research demonstrates the feasibility and potential benefits of leveraging translated datasets to enhance NLP capabilities for low-resource languages. By fine-tuning advanced transformer models on such datasets, we can begin to close the gap in language technology between high-resource and low-resource languages, fostering greater inclusivity and accessibility in the digital realm.

In conclusion, while the performance gains observed in this study were modest, they represent a critical step forward in the development of NLP capabilities for the Kazakh language. This research not only contributes to the body of knowledge in the field but also highlights the potential for further advancements and the importance of continued investment in NLP research for low-resource languages. As we move forward, it is imperative to build upon these foundations, leveraging the power of machine learning and the richness of linguistic diversity to create more inclusive and effective technology solutions for all languages.

References

Eldar Bekbulatov and Amandyk Kartbayev. 2014. A study of certain morphological structures of kazakh and their impact on the machine translation quality. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Quora. 2017. First Quora Dataset Release: Question Pairs. `https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.