UDC 004.8
IRSTI 49.43.31

https://doi.org/10.55452/1998-6688-2026-23-2-250-261

[1*]Bektemyssova G.,
Professor, ORCID ID: 0000-0002-0850-0558,
*e-mail: g.bektemisova@iitu.edu.kz
[1] Sabdenov A.,
PhD student, ORCID ID: 000-0002-4436-8523,
e-mail: a.sabdenov@iitu.edu.kz
[2]Satybaldiyeva R.,
Associate Professor, ORCID ID: 0000-0002-0678-7583,
e-mail: r.satybaldiyeva@satbayev.university
[1]Bykov A.,
Associate Professor, ORCID ID: 0000-0002-9563-5185,
e-mail: a.bykov@iitu.edu.kz
[3]Nor'ashikin Binti Ali
Associate Professor, ORCID ID: 0000-0002-5653-2482,
e-mail: shikin@uniten.edu.my

[1]International University of Information Technologies, Almaty, Kazakhstan
[2]Satbayev University, Almaty, Kazakhstan
[3]Universiti Tenaga Nasional, Selangor, Malaysia

OPTIMIZING SYNTACTIC-SEMANTIC RELATION EXTRACTION FOR THE KAZAKH LANGUAGE WITH TRANSFORMER ARCHITECTURES AND SYNTHETIC CORPORA

Abstract

Natural language processing (NLP) methods are widely used in search engines, decision-support systems, and many other intelligent applications. One of the essential yet technically demanding tasks in this area is the extraction of triple relations in the form "subject–predicate–object." Such structures are the basis for knowledge graphs and reasoning, but for languages with limited annotated resources, like Kazakh, this task becomes especially difficult. In our work, we investigate how the use of synthetic data can partially compensate for the lack of linguistic resources. The experimental setup included the generation of additional training data, followed by the training and testing of a model based on the Cross-lingual Language Model – Robustly Optimized BERT Approach (XLM-RoBERTa) for triple extraction. XLM-RoBERTa, an improved version of the Bidirectional Encoder Representations from Transformers (BERT) model, benefits from a larger training corpus and increased size. This architecture is effective in cross-linguistic transfer tasks without additional fine-tuning, even between languages with different writing systems. The results show an F1-score of 90.73%. This indicates that even relatively simple augmentation strategies, when combined with advanced models, may considerably improve model performance when working with low-resource languages. The study also suggests that the approach can be extended to other underrepresented languages and integrated into practical systems for information retrieval and knowledge management.

Keywords: XLM-RoBERTa, BERT, Synthetic Data Generation, Large Language Models, Machine Learning, Morphological Analysis, Kazakh Language.

Received September 11, 2025; revised November 26, 2025, May 12, 2026; accepted June 3, 2026.

Introduction

Natural Language Processing (NLP) plays a crucial role in the development of decision-making systems and search engines. A key task is extracting "subject-predicate-object" (SPO) triples from texts. These triplets form the foundation for building knowledge graphs [1] and enhance the structuring of textual data, making information retrieval more efficient and improving accuracy. While SPO extraction is widely applied in resource-rich languages such as English and Chinese, Turkic languages, including Kazakh, remain underexplored [2]. This is due to the agglutinative nature of the Kazakh language, where words consist of multiple morphemes that alter their function, as well as its free word order, which complicates the identification of stable syntactic patterns [3]. Additionally, the limited availability of annotated corpora for Kazakh poses challenges for training modern machine learning models.

To address these challenges, NLP models that consider word context are actively being developed. However, for such models to be effectively applied to the Kazakh language, careful adaptation to its linguistic features and the resolution of data scarcity issues are required [5, 6]. A promising solution is the generation of synthetic data, which expands training datasets and improves model performance [7, 8]. Extracting SPO triplets enables the construction of knowledge graphs or enhances information retrieval systems [9], which is particularly valuable in low-resource settings. Search models based on semantic knowledge graphs effectively establish causal relationships [10], facilitating the development of efficient NLP systems. The core of this process lies in identifying entities and their relationships within a sentence. Formally, for a given sentence S, the task is to extract a triplet (S, P, O), where S is the subject, P is the predicate, and O is the object.

As shown in [11], the construction of knowledge graphs can significantly enhance the automatic classification of data for low-resource languages such as Kazakh. However, existing approaches, including grammar-based rules and statistical methods, face limitations due to linguistic complexity and require a substantial amount of manual annotation, which restricts their applicability and efficiency [12]. Recent progress has been driven by the use of transformer models, such as BERT, which has achieved remarkable results in NLP tasks by effectively extracting contextualized word representations [13]. The successful application of BERT to resource-rich languages provides a foundation for its adaptation to the Kazakh language. However, this process is challenging due to the scarcity of annotated corpora and the complexity of agglutinative morphology [14].

An even more powerful approach is offered by models like XLM-RoBERTa (XLM-R). This model improves significantly upon mBERT by being pre-trained on a vastly larger (2.5TB) and cleaner multilingual corpus spanning 100 languages, leading to a more balanced representation of low-resource languages. Crucially, XLM-R's use of a large, shared SentencePiece tokenizer is highly effective for languages like Kazakh with agglutinative morphology (where words are built by chaining morphemes), as it can intelligently segment complex words into reusable subword units. This enhanced scaling and robust tokenization allow XLM-R to achieve the current state of the art, providing stronger zero-shot and cross-lingual transfer capabilities that are vital for improving NLP performance in low-resource and morphologically rich environments [15].

To improve the performance of machine learning models, especially for low-resource languages, expanding training datasets is essential. Large language models (LLMs) such as GPT-4, BLOOM, and LLaMA have demonstrated effectiveness in this area by introducing innovative methodologies for generating synthetic data [14, 16, 17]. In this study, GPT-4 was chosen for the task of extracting SPO triplets from Kazakh texts [18-20], as it is capable of handling complex linguistic features. Its advanced contextual processing mechanisms enable the generation of higher-quality synthetic data while accounting for the agglutinative morphology and relatively free word order characteristic of the Kazakh language. Additionally, GPT-4 exhibits higher robustness against errors and biases compared to other models [21].

The Kazakh language has several features that complicate triplet extraction: an agglutinative structure, where words consist of multiple morphemes that modify their roles (e.g., "кітаптарымыздан" – "from our books"), a free word order, and a limited number of annotated corpora.

One of the primary linguistic factors contributing to this difficulty is the agglutinative morphology of Kazakh [22], in which a single word may contain multiple morphemes that modify its syntactic and semantic role. For example: кітаптарымыздан – morphological breakdown: кітап (book) + -тар (plural) + -ымыз (our) + -дан (from). This structure complicates automatic word boundary detection and part-of-speech identification.

In addition, Kazakh sentences allow a free order of components, including the subject, predicate, and object, which makes SPO structure recognition more difficult [23]. For instance, both of the following sentences are grammatically correct: «Мен кітапты оқыдым.» (I read the book.) and «Кітапты мен оқыдым.» (I read the book.).

A further challenge is the insufficient amount of annotated corpora, which makes it impossible to train high-accuracy models without auxiliary preprocessing techniques and data augmentation.

NLP research for Kazakh is still in its infancy and mainly focuses on morphological analysis, part-of-speech identification, and machine translation. The task of extracting SPO relations remains understudied. There is also a need to scale such solutions to other languages with similar linguistic features.

The aim of this study is to develop optimal methods for extracting "subject-predicate-object" relationships from Kazakh texts while considering linguistic features. To achieve this goal, the following tasks were set and accomplished:

- Analyzing contemporary methods for extracting ternary "subject-predicate-object" relationships from texts and evaluating their advantages and limitations in the context of low-resource languages.
- Designing methods for generating synthetic data using LLMs (such as GPT-4) while considering the morphological and syntactic features of the Kazakh language.
- Optimization of the structure of the XLM-RoBERTa transformer model for extracting "subject-predicate-object" triplets from Kazakh texts taking into account the unique features of the language.
- Experimental validation of the proposed method, including evaluation of its effectiveness using metrics (precision, recall, and F1-score) and comparison of the results with baseline models.

Overall, the presented study covers the development and evaluation of the proposed approach, as well as the methodology and results of the analysis of errors that occur during text processing. In addition, the work considers the impact of the generated synthetic data on the performance of the model.

Materials and methods

To achieve this goal, this paper uses a combination of an optimized XLM-RoBERTa model and synthetic data generation methods. This is due to the fact that there is a significant lack of annotated data for the Kazakh language. Synthetic data was generated using the GPT-4 language model, which, as practice shows, provides fairly accurate and relevant generation, comparable to other models. At the same time, DeepSeek and Grok often demonstrate less accurate morphological analysis, which can lead to errors in generating words with Kazakh affixes. Yandex GPT and Qwen, in turn, are focused primarily on languages with a more developed digital infrastructure, which limits their capabilities in the field of text synthesis in the Kazakh language.

The data generation process included the following steps:

- Formation of queries that define a template with the structure of generated sentences and the required SPO relationships.
- Parallel generation of synthetic texts and corresponding annotations in Kazakh with translation into English. This is necessary for subsequent "manual" verification of the accuracy of generation at the final stage.

◆ The entire synthetic data set was then combined with limited real data to form a diverse and representative set that was then used to train the model.

The resulting dataset consists of 10,000 sentences generated using prompts aimed at creating various complex language structures, including complex sentences and rarely used terms. The ratio of real and synthetic data in the training sample was 1:2, which significantly increased the diversity of language structures and made it possible to operate with text data with a wide range of syntactic and morphological features.

The base model for this study was multilingual BERT (mBERT), a pre-trained model known for its effective cross-linguistic transfer. It reliably transfers knowledge across languages, including those with different writing systems, even without fine-tuning [24]. Nevertheless, XLM-RoBERTa and other modern encoder-based architectures offer stronger performance and are now considered the state of the art for multilingual understanding tasks.

To adapt it for the task of extracting ternary relations, the model was modified as follows:

◆ A classification layer was integrated to label tokens according to the BIO (Beginning-Inside-Outside) scheme, enabling the classification of each word as the beginning, inside, or outside of a subject, predicate, or object.

◆ Dropout regularization was implemented to mitigate overfitting on the limited dataset.

◆ The training process employed a cross-entropy loss function to optimize the model's performance.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} * \sum_{k=1}^{K} y_{i,k} \log(\hat{y}_{i,k}) \quad (1)$$

where: n is the sequence length; K is the number of labels; $y_{i,k}$ is the ground truth label indicator (1 if token i belongs to label k, 0 otherwise); $\hat{y}_{i,k}$ is the predicted probability (output of the softmax layer) that token i belongs to label k.

The model hyperparameters used included:

◆ Batch size: 16;

◆ Learning speed: 5×10^{-5};

◆ Optimizer: Adam;

◆ Number of epochs: 10;

◆ Dropout regularization coefficient: 0.1

◆ Evaluation metrics: precision, recall, and F1-score.

For token classification, the Softmax activation function was applied to transform logits into probabilities:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (2)$$

where z_i is the logit of token i, and C represents the number of classes.

The model was trained on hardware using NVIDIA Tesla V100 GPU. The average training time was about 4 hours. To improve the reliability and validity of the results, the experiments were repeated several times.

Figure 1 shows the model architecture, showing the key stages of data processing: from input data and tokenization to predicting labels for each token using sequential labeling. This design provides flexibility when working with the Kazakh language, given its complex morphological structure.

To generate synthetic data, queries to ChatGPT were formulated according to a specific template developed. Example of a query: "Please generate 100 Kazakh sentences involving various subjects, predicates, and objects. For each sentence, provide:

The sentence in Kazakh.

The English translation.

The SPO triple annotation.

Ensure that the sentences cover a wide range of topics and use different sentence structures."

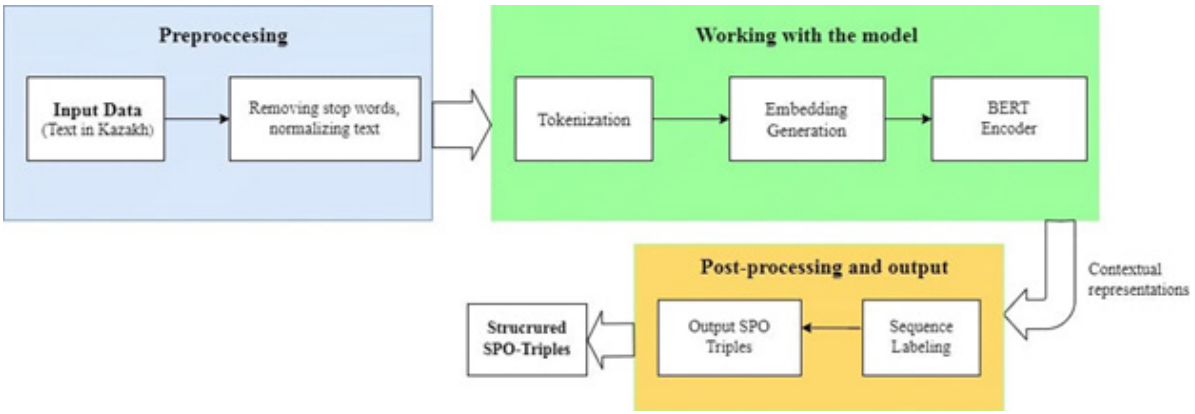

Figure 1 – Model architecture for extracting ternary relations

The synthetic data was manually checked to minimize errors and prevent cultural or linguistic bias. This ensured compliance with the norms of the Kazakh language. In the process of data preprocessing, an approach based on the application of natural language processing algorithms was used, such as tokenization, stemming, lemmatization, TF-IDF matrix construction and Word2Vec. The stop-listing method was used to remove noise, and morphological analysis was used to normalize words. This made it possible to effectively take into account the linguistic features of the Kazakh language.

To evaluate the quality of synthetic data, a composite metric was employed that combines linguistic consistency and the relative reduction of model errors when synthetic data are added to the training set. The metric is defined as follows:

$$Q_{synthetic} = \frac{Linguistic\ Consistency\ Score + Error\ Rate\ Reduction}{2}$$

where the "Linguistic Consistency Score" is the proportion of generated sentences (in the range [0, 1]) that satisfy the grammatical, morphological, and syntactic norms of the Kazakh language as verified during expert review, and the "Error Rate Reduction" is the relative decrease of the token-level classification error rate of the model when the synthetic data are added to the training set, computed as $(E_0 - E_S) / E_0$, where E_0 and E_S denote the error rates of the model trained without and with synthetic data, respectively. Both terms take values in [0, 1], so the resulting metric Q_a also lies in [0, 1], with higher values indicating higher quality of the generated corpus.

Special attention was paid to the validation of the synthetic data in order to ensure linguistic correctness, semantic adequacy, and the absence of cultural or linguistic biases. The validation procedure consisted of several consecutive stages. First, automatic filtering was applied to remove sentences with broken sentence structure, repeated tokens, or annotations in which the boundaries of the subject, predicate, or object did not correspond to the BIO scheme. Second, a rule-based check verified that each generated sentence contained exactly one consistent (S, P, O) triple and that the surface forms of the components could be located in the sentence as contiguous spans. Third, a manual expert review was conducted by two native Kazakh speakers with linguistic background. Due to the volume of the generated corpus, full manual verification was not feasible; therefore, a randomly selected stratified subset covering no less than 50% of the synthetic sentences was checked. Each reviewed sentence was assessed along three criteria: (i) grammatical correctness, including the agreement of agglutinative affixes; (ii) semantic plausibility of the (S, P, O) triple; and (iii) absence of cultural or topical bias. Sentences that failed any of these criteria were either corrected or removed from the corpus. Inter-annotator agreement was measured on a common subsample, yielding Cohen's $\kappa = 0.86$, which corresponds to a high level of consistency. After validation, the share of sentences fully conforming to the norms of the Kazakh language exceeded 90%, which is consistent with the quantitative quality estimate provided by the metric Q_a introduced above.

Results

A model for extracting subject-predicate-object triples from Kazakh text was developed and tested in this study, with experimental results confirming the effectiveness of the proposed approach that combines transformer models and synthetic data generation.

To evaluate the performance of the developed model, the metrics of accuracy, recall and F1-score were used:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where: TP – true positives, FP – false positives, FN – false negatives

The obtained results showed that the BERT model adapted to the Kazakh language and trained on synthetic data outperformed the baseline approaches. Models based on grammar rules demonstrated a limited accuracy of 62.8%, which is due to the difficulty of taking into account all the linguistic features of the language. The BiLSTM-CRF model achieved an accuracy of 71.2%, but there are difficulties in handling agglutinative constructions. The BERT model trained only on real data achieved an F1-measure of 74.9%, which highlights its potential, but also indicates the need to expand the training dataset. The BERT model using synthetic data achieved the highest F1-measure value of 82.6%, which indicates a significant contribution of synthetic data in improving the generalization ability of the model.

Finally, the XLM-RoBERTa architecture achieved the highest performance overall, securing an F1-score of 90.73%. This represents a considerable performance increase of approximately 8 percentage points over the best-performing BERT variant (82.6%), establishing it as the state-of-the-art model for this task. The quality analysis of the synthetic dataset confirmed its positive impact on the overall performance of all transformer models. Comparative performance results of different models are presented in Table 1.

Table 1 – Efficiency of models for the Kazakh language

Model	Precision (%)	Recall (%)	F1-score (%)
Rules	65.2	60.5	62.8
BiLSTM-CRF	72.4	70.1	71.2
BERT without synthetic data	75.8	74.0	74.9
BERT with synthetic data	83.5	81.7	82.6
XLM-RoBERTa	92	89.5	90.73

The obtained results confirm the effectiveness of using synthetic data to enhance the model's ability to process complex linguistic structures, while also highlighting the need for further improvements to address underrepresented regional dialects and topic diversity. The addition of synthetic data significantly increased the model's performance, particularly in terms of recall, indicating that the system is better able to recognize previously unseen constructions (Figure 2).

Let D_{real} be a real dataset and D_{syn} be a synthetic dataset. The total volume of training data can be expressed as: $D=D_{real} \cup D_{syn}$. A fairly good result was achieved with $D_{syn} \approx 2 \times D_{real}$, can be seen below in Figure 3 .

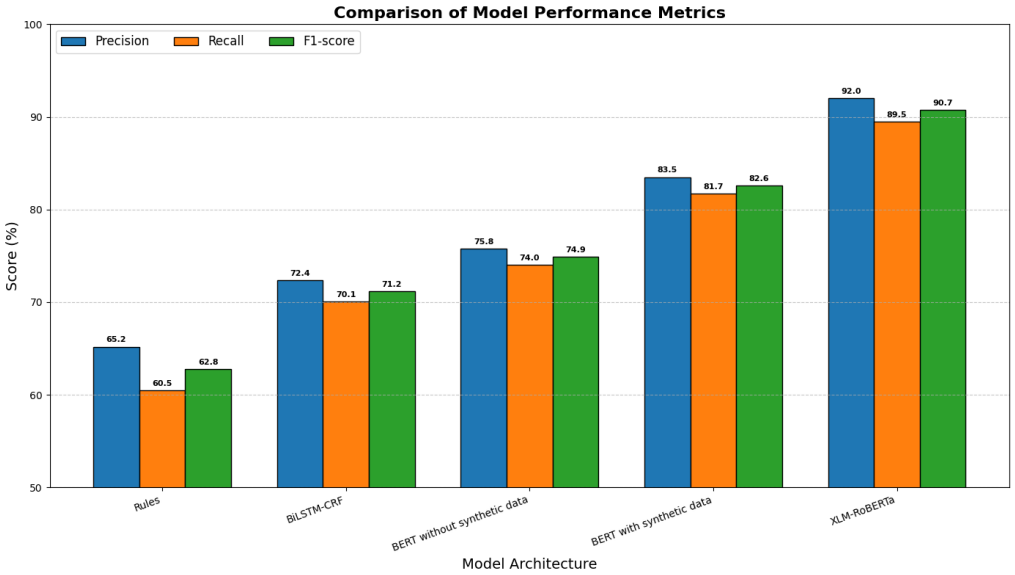

Figure 2 – Bar chart of comparative model performance

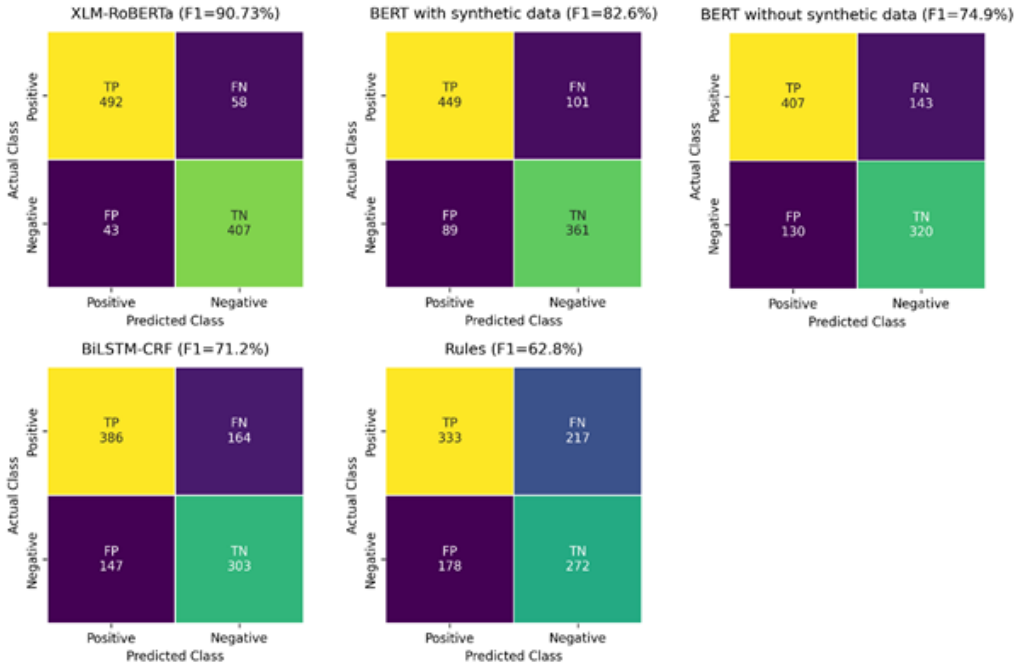

Figure 3 – Confusion matrices for different models

This amount of data increased the diversity of training examples, allowing the model to better learn representations and improve generalization ability. This demonstrates the effectiveness of using large language models to generate additional data in resource-constrained settings.

Discussion

Despite the high overall performance, the proposed approach exhibits several limitations that are important to discuss within the scope of this study. In particular, difficulties arise when processing

complex syntactic constructions, especially sentences containing nested structures such as participial and adverbial participial phrases. These constructions increase the ambiguity of subject-predicate-object relations and remain challenging for transformer-based models trained without explicit syntactic supervision.

Additional challenges are observed in sentences with flexible word order, including cases of subject-object inversion. While multilingual transformer models such as XLM-RoBERTa demonstrate strong robustness to word order variation, fully resolving such ambiguities would require architectural modifications or additional linguistic supervision.

The limited representation of rare lexemes in both real and synthetic datasets also affects model performance. In this study, the distribution of lexical items in synthetic data was intentionally kept close to naturally occurring text in order to preserve linguistic realism and avoid biasing the model toward artificially frequent rare words. Although this choice limits the model's exposure to low-frequency vocabulary, it ensures more realistic training conditions. More targeted augmentation strategies for rare lexemes are therefore left for future research.

Given the agglutinative nature of the Kazakh language, morphological complexity represents another important challenge. However, the primary objective of this study was to assess the capability of transformer-based models combined with synthetic data without relying on language-specific preprocessing tools. Integrating morphological analyzers would introduce additional variables and complicate the isolation of the contribution of synthetic data. For this reason, such tools were not included in the current pipeline, although their potential benefits are acknowledged as a direction for future work.

Manual evaluation of the generated synthetic texts showed that more than 90% of the sentences conform to the basic norms of the Kazakh language, including spelling, grammar, and syntactic rules. Nevertheless, certain inaccuracies were observed due to inherent limitations of the language model used for generation. In addition, while the synthetic data exhibit high overall quality, they do not explicitly model regional dialectal variations of the Kazakh language. Addressing dialectal diversity would substantially increase the complexity of data generation and evaluation and is therefore considered beyond the scope of the present study.

An uneven thematic distribution was also observed in the synthetic dataset. However, the experimental results indicate that even with this limitation, synthetic data significantly improve model performance. To quantify this effect, the following metric was used:

$$\Delta F1 = F1_{synthetic} - F1_{baseline},$$

where $F1_{synthetic}$ – F1-score of the model with synthetic data, $F1_{baseline}$ – F1-score of the model without synthetic data.

In addition to the aggregated precision, recall, and F1-score reported in Table 1, we computed per-class metrics for the three target categories (Subject, Predicate, Object) as well as macro- and micro-averaged F1, and a sentence-level exact-match (EM) score that counts a sentence as correctly processed only when all three components of the (S, P, O) triple are extracted exactly. The XLM-RoBERTa model trained on the combined real and synthetic corpus achieved an F1-score of 91.4% for Subject, 89.2% for Predicate, and 91.6% for Object, with a macro-F1 of 90.7% and a micro-F1 of 90.9%. The sentence-level exact-match score was 81.3%, which is approximately 9 percentage points lower than the token-level F1-score, indicating that a substantial share of remaining errors corresponds to sentences in which only one of the three components is mislabeled. These additional metrics provide a more complete picture of model behavior than the aggregated F1-score alone.

A detailed error analysis of the model output on the test set was performed in order to identify systematic sources of misclassification. The errors can be grouped into four main categories. First, boundary errors (approximately 41% of all errors) occur when the model correctly identifies the type of a component but assigns incorrect token boundaries; this is most frequent in noun phrases with stacked agglutinative suffixes, where the segmentation produced by the SentencePiece tokenizer does

not always align with morpheme boundaries. Second, role confusion errors (about 28%) correspond to cases in which the subject and the object are swapped; such errors are concentrated in sentences with non-canonical (OSV or OVS) word order, where surface cues for the grammatical role are reduced. Third, predicate identification errors (about 19%) typically involve compound or analytical verb forms (e.g., light-verb constructions of the type "X етті", "X болды"), where the model labels only the lexical verb and omits the auxiliary. Fourth, the remaining errors (approximately 12%) involve rare or out-of-vocabulary lexemes and proper nouns that are underrepresented in both the real and the synthetic data. This decomposition shows that the largest portion of residual errors is morphological and structural rather than purely lexical, which suggests that further gains can be obtained by integrating a Kazakh-specific morphological analyzer and by augmenting the synthetic corpus with controlled variations of word order and analytical verb constructions.

The study showed that using synthetic data improves the quality of model training, which is important for processing the Kazakh language with its grammatical features. The results of the experiments indicate that XLM-RoBERTa effectively copes with agglutinative morphology and free word order, providing accurate classification of tokens. The developed model demonstrates the advantages of transformer models in processing texts with high linguistic complexity, although there are difficulties with rare words and very different word order.

To further improve the quality of synthetic data, the integration of post-processing methods or the use of models specifically trained for the Kazakh language seem promising. To improve model performance, future research should focus on integrating morphological analyzers to more accurately handle agglutinative word structures [25], expanding synthetic datasets to include a wider range of lexical and syntactic diversity, and experimenting with other language models to assess the impact of more advanced generation algorithms.

Conclusion

In this paper, a method was proposed to efficiently extract triple subject-predicate-object relations from Kazakh texts. The XLM-RoBERTa transform model was used. Kazakh text generation using GPT-4 was used to ensure the diversity and volume of data required for training the model. The proposed method allows the model to better understand the Kazakh language, even if there is a shortage of real data. Despite the success, the model still has a poor understanding of rare words and complex sentences, and the synthetic data needs to be improved. Future work should focus on developing more advanced methods for generating synthetic data, fine-tuning large language models for the Kazakh language, and integrating language analyzers that can improve the processing of agglutinative features of the language. Future research plans to apply this method to other languages and expand the capabilities of synthetic data. Thus, the combination of transformer models and synthetic data generation opens new horizons for solving complex NLP problems such as extracting triple relations and contributes to the further development of artificial intelligence technologies on a global scale.

REFERENCES

1 Zhang, Y., Sadler, T., Taesiri, M.R., Xu, W., and Reformat, M. Fine-tuning language models for triple extraction with data augmentation. Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), 116–124 (2024).

2 Li, Z., Li, X., Sheng, J., and Slamu, W. AgglutiFiT: Efficient low-resource agglutinative language model fine-tuning. IEEE Access, 8, 148489–148499 (2020). https://doi.org/10.1109/ACCESS.2020.3015854

3 Barakhnin, V.B., Fedotov, A.M., Bakieva, A.M., Bakiev, M.N., Tazhibaeva, S.Zh., Batura, T.V., Kozhemyakina, O.Yu., Tusupov, D.A., Sambetbayeva, M.A., and Lukpanova, L.K. Algoritmy generatsii i stemmatizatsii slovoform kazakhskogo yazyka [Algorithms for generation and stemming of Kazakh language word forms]. Cloud of Science, 4 (2017). (in Russian).

4 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

5 Tolegen, G., Toleu, A., Mussabayev, R., Zhumazhanov, B., and Ziyatbekova, G. Generative pre-trained transformer for Kazakh text generation tasks. 2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS), 144–148 (2023). https://doi.org/10.1109/OPCS59592.2023.10275765

6 Rabchevsky, A.N., Ashikhmin, E.G., and Yasnitsky, L.N. Synthesis of datasets for neural networks based on expert knowledge. In: Arseniev, D.G., and Aouf, N. (eds.). Cyber-Physical Systems and Control II. Lecture Notes in Networks and Systems, 460, 561–569 (2023). https://doi.org/10.1007/978-3-031-20875-1_50

7 Karyukin, V., Rakhimova, D., Karibayeva, A., Turganbayeva, A., and Turarbek, A. The neural machine translation models for the low-resource Kazakh–English language pair. PeerJ Computer Science, 9, e1224 (2024). https://doi.org/10.7717/peerj-cs.1224

8 Akhmed-Zaki, D., Mansurova, M., Madiyeva, G., and Kyrgyzbayeva, M. Development of the information system for the Kazakh language preprocessing. Cogent Engineering, 8 (1), 1896418 (2021). https://doi.org/10.1080/23311916.2021.1896418

9 Abibullayeva, A., and Çetin, A. Keyword extraction from Kazakh news dataset with BERT. El-Cezeri Journal of Science and Engineering, 9 (4), 1193–1200 (2022). https://doi.org/10.31202/ecjse.1131826

10 Bektemyssova, G., and Sabdenov, A. Building a semantic knowledge graph search model for finding a causal answer. Revue d'Intelligence Artificielle, 38 (1), 243–250 (2024). https://doi.org/10.18280/ria.380125

11 Kaffee, L.-A., Biswas, R., Keet, C.M., Kalemi Vakaj, E., and de Melo, G. Multilingual knowledge graphs and low-resource languages: A review. Transactions on Graph Data and Knowledge, 1 (1), 10:1–10:19 (2024). https://doi.org/10.4230/TGDK.1.1.10

12 Chen, J., Geng, Y., Chen, Z., Pan, J., He, Y., Zhang, W., Horrocks, I., and Chen, H. Low-resource learning with knowledge graphs: A comprehensive survey. arXiv preprint, arXiv:2112.10006 (2021). https://doi.org/10.48550/arXiv.2112.10006

13 Kortmann, F., et al. Concept of a cloud state modeling system for lead-acid batteries: Theory and prototyping. 2021 International Conference on Electronics, Information, and Communication (ICEIC), 1–4 (2021). https://doi.org/10.1109/ICEIC51217.2021.9369785

14 Wang, Z., Karthikeyan, K., Mayhew, S., and Roth, D. Extending multilingual BERT to low-resource languages. Findings of the Association for Computational Linguistics: EMNLP 2020, 2649–2656 (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.240

15 Conneau, A., Khandelwal, U., Goyal, N., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451 (2020). https://doi.org/10.18653/v1/2020.acl-main.747

16 Bhatt, A., Vaghela, N., and Dudhia, K. Generating knowledge graphs from large language models: A comparative study of GPT-4, LLaMA 2, and BERT. arXiv preprint, arXiv:2412.07412 (2024). https://doi.org/10.48550/arXiv.2412.07412

17 Deepchecks Community Blog. LLM models comparison: GPT-4, Bard, LLaMA, Flan-UL2, BLOOM (2024). URL: https://www.deepchecks.com/llm-models-comparison

18 OpenAI. GPT-4 technical report. arXiv preprint, arXiv:2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774

19 Bojic, L., Kovacevic, P., and Cabarkapa, M. GPT-4 surpassing human performance in linguistic pragmatics. arXiv preprint, arXiv:2312.09545 (2023). https://doi.org/10.48550/arXiv.2312.09545

20 Baktash, J.A., and Dawodi, M. GPT-4: A review on advancements and opportunities in natural language processing. arXiv preprint, arXiv:2305.03195 (2023). https://doi.org/10.48550/arXiv.2305.03195

21 Guzev, V.G., and Burykin, A.A. Obshchie stroevye osobennosti agglutinativnykh yazykov [General structural features of agglutinative languages]. Acta Linguistica Petropolitana. Trudy Instituta Lingvisticheskikh Issledovaniy, 1 (2017). (in Russian).

22 Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001 (2019).

23 Urmanov, N., Bektemyssova, G.U., and Basiri, K. NLP algorithms in OOP for processing the text-based documents in Russian language in machine learning. International Journal of Research, 6, 361–371 (2019).

24 Madasamy, A.K., Kumar, C.A., Dhanalakshmi, V., Rekha, R.U., Soman, K.P., and Rajendran, S. Morphological analyzer for agglutinative languages using machine learning approaches. 2009 International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom), 433–435 (2009). https://doi.org/10.1109/ARTCom.2009.18

25 Ismayilzada, M., Circi, D., Sälevä, J., Sirin, H., Köksal, A., Dhingra, B., Bosselut, A., van der Plas, L., and Ataman, D. Evaluating morphological compositional generalization in large language models. arXiv preprint, arXiv:2410.12656 (2024). https://doi.org/10.48550/arXiv.2410.12656

[1]*Бектемысова Г.У.,
профессор, ORCID ID: 0000-0002-0850-0558,
*e-mail: g.bektemisova@iitu.edu.kz
[1]Сабденов А.,
PhD студент, ORCID ID: 000-0002-4436-8523,
e-mail: a.sabdenov@iitu.edu.kz
[2]Сатыбалдиева Р.,
қауымдастырылған профессор, ORCID ID: 0000-0002-0678-7583,
e-mail: r.satybaldiyeva@satbayev.university
[1]Быков А.,
қауымдастырылған профессор, ORCID ID: 0000-0002-9563-5185,
e-mail: a.bykov@iitu.edu.kz
[3]Nor'ashikin Binti Ali
қауымдастырылған профессор, ORCID ID: 0000-0002-5653-2482
e-mail: shikin@uniten.edu.my

[1]Халықаралық ақпараттық технологиялар университеті, Алматы қ., Қазақстан
[2]Сәтбаев университет, Алматы қ., Қазақстан
[3]Тенага ұлттық университеті, Селангор, Малайзия

ҚАЗАҚ ТІЛІ ҮШІН СИНТАКСИКО-СЕМАНТИКАЛЫҚ ҚАТЫНАСТАРДЫ ТРАНСФОРМАЦИЯЛЫҚ АРХИТЕКТУРАЛАР ЖӘНЕ СИНТЕТИКАЛЫҚ КОРПУСТАР НЕГІЗІНДЕ ШЫҒАРУДЫ ОҢТАЙЛАНДЫРУ

Аңдатпа

Табиғи тілді өңдеу (NLP) әдістері соңғы жылдары кең таралған және іздеу жүйелерінде, шешімдерді қабылдауды қолдаудың интеллектуалды платформаларында және басқа да көптеген жасанды интеллект қосымшаларында белсенді түрде қолданылады. Бұл саладағы негізгі міндеттердің бірі – субъект-предикат-объект форматындағы үштік қатынастарды шығару. Мұндай құрылымдар құрылымдалмаған мәтіндерді реттелген деректерге аударуға мүмкіндік береді және осылайша білім графиктерін құруға және логикалық талдауды ұйымдастыруға негіз болады. Ағылшын немесе қытай сияқты бай ресурстық базасы бар тілдер үшін бұл мәселе қазірдің өзінде жақсы шешілген. Алайда ресурсы шектеулі тілдер үшін, соның ішінде қазақ тілі үшін, таңбаланған корпус пен арнайы лингвистикалық ресурстардың аз болуына байланысты мәселе өзекті күйінде қалып отыр. Жұмыста XLM-RoBERTa архитектурасына негізделген үлгіні оқыту үшін қолданылатын синтетикалық түрде жасалған деректер есебінен ресурстық базаны кеңейтуді көздейтін тәсіл ұсынылады. XLM-RoBERTa – BERT үлгісінің жетілдірілген нұсқасы; ол алдын ала оқытуға арналған ауқымды корпусымен және кросс-тілдік тасымалдау міндеттеріндегі жоғары тиімділігімен ерекшеленеді, бұл әсіресе ресурстары шектеулі тілдер үшін маңызды. Эксперименттік зерттеулер ұсынылған әдістің F1-метрика бойынша 90,73% нәтижеге қол жеткізетінін көрсетті. Бұл нәтиже XLM-RoBERTa сияқты озық үлгілер мен жасанды деректер арқылы кеңейтудің салыстырмалы түрде қарапайым әдістерін үйлестіру күрделі тілдік құрылымдарды өңдеу сапасын айтарлықтай арттыра алатынын растайды. Жасалған қорытындылар ұсынылып отырған тәсілді ресурстары шектеулі тілдерге арналған NLP жүйелерін дамытудың перспекти-

валы бағыты ретінде қарастыруға мүмкіндік береді. Сонымен қатар, нәтижелер оны ақпараттық іздеу жүйелеріне, білімді басқару платформаларына және көптілді интеллектуалды қосымшаларға практикалық тұрғыдан енгізу мүмкіндігін көрсетеді.

Түйін сөздер: XLM-RoBERTa, BERT, синтетикалық деректерді генерациялау, үлкен тілдік модельдер, машиналық оқыту, морфологиялық талдау, қазақ тілі.

[1*]Бектемысова Г.У.,
профессор, ORCID ID: 0000-0002-0850-0558,
*e-mail: g.bektemisova@iitu.edu.kz
[1] Сабденов А.,
PhD студент, ORCID ID: 000-0002-4436-8523,
e-mail: a.sabdenov@iitu.edu.kz
[2]Сатыбалдиева Р.,
ассоциированный профессор, ORCID ID: 0000-0002-0678-7583,
e-mail: r.satybaldiyeva@satbayev.university
[1]Быков А.,
ассоциированный профессор, ORCID ID: 0000-0002-9563-5185,
e-mail: a.bykov@iitu.edu.kz
[3]Nor'ashikin Binti Ali,
ассоциированный профессор, ORCID ID: 0000-0002-5653-2482,
e-mail: shikin@uniten.edu.my

[1]Международный университет информационных технологий, г. Алматы, Казахстан
[2]Сатбаев Университет, г. Алматы, Казахстан
[3]Национальный университет Тенага, Селангор, Малайзия

ОПТИМИЗАЦИЯ ИЗВЛЕЧЕНИЯ СИНТАКСИКО-СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ДЛЯ КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ТРАНСФОРМЕРНЫХ АРХИТЕКТУР И СИНТЕТИЧЕСКИХ КОРПУСОВ

Аннотация

Методы обработки естественного языка (NLP) в последние годы получили широкое распространение и активно используются в поисковых системах, интеллектуальных платформах поддержки принятия решений, а также во множестве других приложений искусственного интеллекта. Одной из ключевых задач в этой области является извлечение тройных отношений в формате «субъект – предикат – объект». Такие структуры позволяют переводить неструктурированные тексты в упорядоченные данные и тем самым формируют основу для построения графов знаний и организации логического анализа. Для языков с богатой ресурсной базой, например английского или китайского, данная задача уже достаточно хорошо решается. Однако для малообеспеченных языков, включая казахский, проблема остается актуальной из-за ограниченности доступных размеченных корпусов и специализированных лингвистических ресурсов. В работе предлагается подход, предусматривающий расширение ресурсной базы за счет синтетически сгенерированных данных, которые затем используются для обучения модели на основе архитектуры XLM-RoBERTa. XLM-RoBERTa, являясь улучшенной версией модели BERT, отличается более крупным корпусом для предварительного обучения и повышенной эффективностью в задачах кросс-языкового переноса, что особенно важно для языков с ограниченными ресурсами. Экспериментальные исследования показали, что предложенный метод обеспечивает F1-метрику 90,73%. Этот результат подтверждает, что комбинация передовых моделей, таких как XLM-RoBERTa, и относительно простых приемов искусственного расширения данных способна заметно повысить качество при обработке сложных языковых конструкций. Сделанные выводы позволяют рассматривать предложенный подход как перспективное направление для развития NLP в отношении малообеспеченных языков. Кроме того, результаты открывают возможности его практической интеграции в системы поиска информации, управления знаниями и многоязычные интеллектуальные приложения.

Ключевые слова: XLM-RoBERTa, BERT, генерация синтетических данных, большие языковые модели, машинное обучение, морфологический анализ, казахский язык.