A HYBRID TRANSFORMER-BILSTM MODEL FOR PUNCTUATION RESTORATION IN LOW-RESOURCE LANGUAGES: UZBEK AS A CASE STUDY

*Maksud Sharipov, Hushnudbek Adinaev**

Urgench State University named after Abu Rayhan Biruni, Urgench
Uzbekistan

*Corresponding Author, e-mail: hushnudbek.adinaev@gmail.com

Abstract: Automatic punctuation restoration is an important natural language processing task that improves the readability of unpunctuated text and supports automatic speech recognition, machine translation and text editing systems. This study proposes a language-adaptable hybrid Transformer+BiLSTM framework for token-level punctuation restoration in low-resource languages and evaluates it using Uzbek as a case study. The task was formulated as a sequence-labeling problem with eleven classes. RoBERTa, XLM-RoBERTa, BERTbek and TahrirchiBERT were fine-tuned also BiLSTM layers were added to selected transformer encoders. The models were evaluated using accuracy, precision, recall and F1-score. XLM-RoBERTa+BiLSTM achieved the best overall results, with 0.8588 accuracy, 0.8493 precision, 0.8588 recall and 0.8393 F1-score. The findings indicate that bidirectional sequential modeling can complement transformer representations in low-resource punctuation restoration.

Key words: NLP, transformer models, Punctuation restoration, BiLSTM, low-resource languages.

1. INTRODUCTION

Punctuation marks are an integral part of written speech and play an important role in expressing syntactic boundaries, sentence completion, the speaker's attitude, pauses, and tone in the text. Marks such as periods, commas, question marks, exclamation marks, colons, semicolons, dashes, quotation marks, brackets, and colons directly affect the correct understanding of the content of the text. Incorrect punctuation or complete omission of punctuation marks reduces the meaning of the text and leads to errors in subsequent automatic processing stages. In the field of natural language processing, the issue of automatic punctuation recovery and prediction is particularly important in the processes of automatic speech recognition, machine translation, text editing, sentence boundary detection, and syntactic analysis. Because many automatic speech recognition systems generate text without punctuation marks.

To make such texts user-friendly and semantically understandable, it is necessary to automatically restore punctuation marks. Although this issue has been studied for the

Uzbek language, it has not been studied enough. The agglutinative structure of the Uzbek language, the relative freedom of word order, the morphological richness, and the limited digital resources complicate the task of automatic punctuation prediction. Therefore, predicting punctuation marks in Uzbek texts based on modern deep learning and transformer models is a relevant scientific and practical issue.

This study addresses token-level punctuation restoration in low-resource languages, using Uzbek as a case study. The task is formulated as a sequence-labeling problem in which each input word is assigned either a punctuation class or the class O, indicating that no punctuation mark follows the word. To address this task, a generalizable hybrid Transformer+BiLSTM framework is proposed. The transformer encoder generates contextual representations for the input tokens, while the BiLSTM layer further models their bidirectional sequential dependencies before punctuation classification. The framework is evaluated using both multilingual and language-specific transformer encoders to assess its effectiveness and adaptability in a low-resource setting.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on punctuation restoration, transformer-based approaches and hybrid neural architectures. Section 3 presents the generalized hybrid Transformer+BiLSTM framework, including its mathematical formulation, model architecture and adaptation procedure for low-resource languages. Section 4 describes the Uzbek case study, including corpus preparation, the punctuation label set, and the experimental configuration. Section 5 presents and discusses the experimental results, including overall performance, class-level evaluation, and error analysis. Section 6 discusses the applicability, limitations, and transferability of the proposed framework to other low-resource languages. Finally, Section 7 concludes the paper and outlines directions for future research.

2. RELATED WORK

Automatic punctuation recovery and prediction is one of the important issues in natural language processing. This task is especially important in automatic speech-to-text conversion, text preprocessing, grammar checking, machine translation, text segmentation, and information retrieval systems. The study presented in [1] extensively reviewed approaches to capitalisation and punctuation restoration, including rule-based methods, n-gram models, latent-event models, boosting techniques, conditional random fields and neural-network-based approaches. In recent years, transformer architectures, in particular BERT and its multilingual variants, have been effective in punctuation restoration tasks due to their dual awareness of context [2]. In early studies, punctuation recovery was performed using more statistical and sequence-based methods. For example, in CRF models, for each word or token, a possible punctuation mark is predicted that will follow it. While these approaches are useful for describing sequential data, they may not adequately capture long-range dependencies, complex syntactic relationships, and context-dependent meanings [3]. In the next stage, recurrent neural networks such as LSTM and BiLSTM began to be used. Tilk and Alumäe studied the problem of restoring punctuation marks in speech transcripts using a model based on BiRNN and the attention mechanism [4]. These types of models perform better than

traditional statistical methods because they are able to process context from left to right and right to left.

The emergence of transformer models marked a new stage in the task of restoring punctuation. The BERT model proposed by Devlin and co-authors allows encoding text in a two-way context [2]. Alam and co-authors investigated the punctuation recovery task using transformer models for high- and low-resource languages [5]. Their work shows that transformer models can be applied not only to resource-rich languages like English, but also to resource-poor languages. The FullStop project predicts punctuation and sentence segmentation using multilingual BERT-type models [6]. A RobBERT-based model for predicting the punctuation mark following each word in Dutch texts and for segmenting automatic speech recognition output is proposed in [7].

Omission or incorrect use of punctuation marks is common in social media texts, user comments, and informal correspondence. Article [8] studied punctuation recovery in social media texts using transformer models and showed that the combination of RoBERTa-large and LSTM yielded high accuracy on Amazon and Telecom data. In [9] is proposed an investigation of punctuation and capitalisation restoration for Spanish and Basque using the AutoPunct system, which combines a BERT-based model with acoustic features, such as pause duration, and lexical information. These studies show that not only text context, but also additional acoustic and structural cues can improve performance in the punctuation retrieval task.

Recent work has also paid great attention to hybrid architectures. For example.in [10] is proposed a model that combines XLM-RoBERTa and LSTM layers, taking advantage of the ability of transformers to capture global context and the ability of LSTM to model sequential dependencies. For Turkish, Saoud et al. compared different sizes of BERT models and showed that the quality of punctuation and capitalization tasks improved as the model size increased [11]. For Arabic, Yagi and co-authors created the large Arabic Punctuation Dataset, justifying the need for a specially annotated corpus for identifying punctuation and sentence boundaries [12]. These works demonstrate the importance of specialized corpora, appropriate tokenization, and language-specific models for low-resource or morphologically complex languages.

Punctuation processing in Uzbek remains an insufficiently studied research area. A rule-based algorithm for analysing the use of periods and commas in Uzbek texts is proposed in [13]. The algorithm incorporates more than 30 rules and was evaluated on a 30,000-word corpus covering 30 different domains. The authors also emphasised the need to extend this research by applying machine-learning and deep-learning approaches. In a subsequent study [14] is applied the BiLSTM-CRF model to the task of restoring punctuation in Uzbek texts. In addition, the creation of resources enriched with morphological and POS tags, such as UzbekPOS, serves as an important basis for punctuation, syntactic analysis and other NLP tasks for the Uzbek language [15].

The present study is also related to previous research on Uzbek punctuation prediction, where BERT-based contextual representations, BiLSTM and rule-based post-processing were combined in a hybrid model [16]. In contrast, the current paper evaluates several pretrained transformer encoders and their BiLSTM-based hybrid variants, and presents Uzbek as a case study for punctuation restoration in low-resource languages. The differences in dataset size and reported results are mainly caused by

differences in corpus preparation, punctuation label inventory, preprocessing, tokenisation and evaluation protocol. In addition, hybrid recurrent-transformer architectures have also been used in other sequence modelling tasks. For example, parallel BiLSTM-Transformer networks were applied to chaotic dynamics forecasting in [17], showing that recurrent and attention-based components can complement each other in modelling sequential dependencies.

3. GENERALIZED HYBRID TRANSFORMER-BILSTM FRAMEWORK

3.1. Mathematical formulation of the problem

The input text is represented as a sequence of words as follows:

$$X=\{x_1, x_2, ..., x_n\} \quad (1)$$

where x_i denotes the i-th word in the input sequence.

The punctuation restoration task is formulated as a token-level sequence-labelling problem. For each input word x_i, the model predicts a label y_i representing the punctuation mark that should follow that word:

$$Y=\{y_1, y_2, ... ,y_n\} \quad (2)$$

where $y_i \in Z$, and the punctuation label set is defined as:

$$Z=\{O, period, comma, question, exclamation, colon, semicolon, dash, ellipsis, quote, parent\} \quad (3)$$

The label O indicates that no punctuation mark follows the corresponding word. The remaining labels represent a period, comma, question mark, exclamation mark, colon, semicolon, dash, ellipsis, quotation mark, and parenthesis, respectively.

Given an input sequence X, the objective of the model is to estimate the most probable punctuation label for every word:

$$\overline{y_i} = argmax\ P(y_i = c|X)\ c \in Z \quad (4)$$

Although the framework is evaluated using Uzbek data, the formulation itself is not restricted to a specific language. The same sequence-labelling formulation can be applied to another low-resource language by adapting the tokenizer, transformer encoder, annotated corpus, and punctuation label inventory.

3.2. Hybrid Transformer+BiLSTM architecture

The proposed framework combines the contextual representation capability of a pretrained transformer encoder with the sequential modelling capability of a bidirectional long short-term memory network. The transformer component captures long-range semantic and syntactic dependencies, whereas the BiLSTM component further processes the contextualized token representations in both forward and backward directions.

First, the input word sequence is processed by the tokenizer associated with the selected transformer encoder. Because transformer tokenizers may divide a word into multiple subword units, the tokenized sequence may contain more elements than the original word sequence. Let the resulting subword sequence be:

$$\overline{X} = (t_1, t_2, \ldots, t_m) \qquad where\ m \geq n \quad (5)$$

The transformer encoder converts the subword tokens into contextual representations:

$$H^T = TransformerEncoder(\overline{X}) \quad (6)$$

where $H^T = (h_1^T, h_2^T, \ldots, h_{m1}^T)$. For a word divided into several subwords, the representation of its first subword is used for punctuation classification. The remaining subwords and special tokens are excluded from the loss calculation.

The word-level contextual representations obtained from the transformer are then passed to the BiLSTM layer:

$$\vec{h}_i = \mathrm{LSTM_f}(h^T_i, \vec{h}_{i-1}) \quad (7)$$

$$\overleftarrow{h}_i = \mathrm{LSTM_b}(h^T_i, \overleftarrow{h}_{i-1}) \quad (8)$$

where $\vec{h}_I$ and $\overleftarrow{h}_I$ represent the forward and backward hidden states, respectively. They are concatenated to form the bidirectional representation:

$$H^B_i = [\vec{h}_I; \overleftarrow{h}_i] \quad (9)$$

The BiLSTM output is passed through a dropout layer and a linear classification layer. The probability distribution over the punctuation classes is calculated using the softmax function:

$$P(y_i|X) = \mathrm{softmax}(W\, h^B_i + b) \quad (10)$$

where W and b are trainable parameters. The class with the highest probability is selected as the punctuation prediction for the corresponding word.

The model is trained by minimizing the token-level cross-entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^{N} log\ P(y_i|X) \quad (11)$$

where N is the number of valid word-level tokens included in the loss calculation.

The architecture of the proposed hybrid Transformer-BiLSTM framework for punctuation restoration is illustrated in Figure 1.

Input text → Tokenization → Transformer encoder → Word-level representation selection → BiLSTM layer → Dropout → Linear classifier → Softmax → Punctuation labels

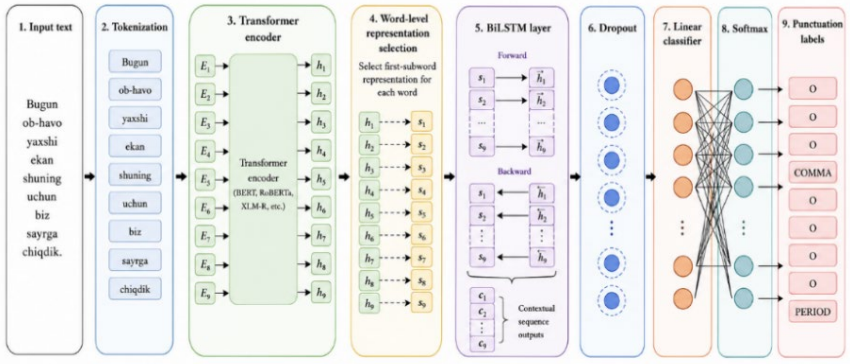


Figure 1. General architecture of the proposed hybrid Transformer+BiLSTM framework for punctuation restoration.

3.3. Adaptation procedure for low-resource languages

The proposed framework is language-adaptable rather than being restricted to a single language. Its application to a new low-resource language consists of several main stages.

First, the punctuation inventory is defined according to the orthographic conventions of the target language. A corpus containing naturally punctuated texts is then collected, cleaned and converted into a word-level sequence-labelling format. Each word is assigned either a punctuation label or the label O.

Second, an appropriate pretrained encoder is selected. A language-specific transformer may be used when a sufficiently reliable monolingual model is available. Otherwise, a multilingual encoder, such as XLM-RoBERTa, can be employed to transfer knowledge acquired from multiple languages.

Third, the word-level punctuation labels are aligned with the subword tokens generated by the transformer tokenizer. The label is assigned to the first subword of each word, whereas subsequent subwords and special tokens are masked during loss computation.

Fourth, the transformer encoder and the BiLSTM classification architecture are fine-tuned on the annotated corpus. The transformer captures contextual relationships, while the BiLSTM explicitly models sequential dependencies between neighbouring contextual representations.

Finally, the model is evaluated using both overall and class-level metrics. Because punctuation datasets are generally dominated by the O class and frequently occurring punctuation marks, macro-averaged scores and class-level results should be reported together with weighted metrics.

In this procedure, the overall architecture and training pipeline remain unchanged across languages. The language-dependent components are the annotated corpus, tokenizer, pretrained encoder, punctuation inventory and text-normalization rules. This separation enables the framework to be adapted to other morphologically rich and low-resource languages.

4. UZBEK CASE STUDY AND EXPERIMENTAL SETUP

4.1. Corpus preparation

To evaluate the framework in a low-resource setting, Uzbek texts written in the Latin script were used as a case study. Texts from 30 thematic domains were collected, cleaned and segmented into sentences. Punctuation marks were converted into word-level labels, with O assigned to words not followed by punctuation. The corpus was stored in CoNLL format and split at sentence level into training, validation and test subsets. Corpus statistics are given in Table 1.

Table 1. Sentence-level distribution of the Uzbek punctuation corpus

Dataset subset	***Number of sentences***	***Number of tokens***	***Share of sentences***
Train	*719 893*	*11 888 213*	*80%*
Validation	*89 986*	*2 720 684*	*10%*
Test	*89 986*	*2 720 186*	*10%*
Total	***899 865***	***17 329 083***	***100%***

4.2. Punctuation label distribution

Eleven labels were used in the experiments: one non-punctuation class and ten punctuation classes. The distribution of the labels in the training subset is presented in Table 2. The O class constitutes 75.73% of the training tokens, indicating a substantial class imbalance. Frequently occurring punctuation classes, including PERIOD and COMMA, contain considerably more examples than EXCLAMATION, ELLIPSIS, and other less frequent classes. This imbalance is considered in the class-level evaluation and error analysis.

Table 2. Distribution of punctuation labels in the training subset

Label	***Punctuation mark***	***Number***	***Share***
O	*No punctuation marks*	*9 003 523*	*75.73%*
PERIOD	*PERIOD*	*845 944*	*7.12%*
COMMA	*COMMA*	*1 068 023*	*8.98%*
QUESTION	*QUESTION*	*98 928*	*0.83%*
EXCLAMATION	*EXCLAMATION*	*87 695*	*0.74%*
COLON	*COLON*	*132 328*	*1.11%*
SEMICOLON	*SEMICOLON*	*102 292*	*0.86%*
DASH	*DASH*	*179 845*	*1.51%*
ELLIPSIS	*ELLIPSIS*	*95 246*	*0.80%*
QUOTE	*QUOTE*	*93 984*	*0.79%*
PAREN	*PAREN*	*180 405*	*1.52%*
Total		***11 888 213***	***100%***

4.3. Experimental configuration

The experiments compared a BiLSTM baseline, four pretrained transformers and three hybrid Transformer+BiLSTM models. RoBERTa, XLM-RoBERTa, BERTbek and TahrirchiBERT were evaluated, while hybrid variants added a BiLSTM layer to XLM-RoBERTa, BERTbek and TahrirchiBERT. The models were implemented in Python using PyTorch and Hugging Face Transformers. Models were trained with AdamW and cross-entropy loss, selected on the validation set and finally assessed on the test set. The evaluation script computes accuracy, precision, recall, F1-score, confusion matrix for all punctuation labels. The main settings are summarised in Table 3.

Table 3. Experimental and training configuration

Parameter	***Base transformer models***	***Transformer-BiLSTM models***
Transformer encoders	*RoBERTa, XLM-RoBERTa, TahrirchiBERT*	*XLM-RoBERTa, BERTbek, TahrirchiBERT*
Maximum sequence length	*256*	*256*
Training batch size	*4*	*Actual value from training logs*
Evaluation batch size	*8*	*Actual value from training logs*
Gradient accumulation steps	*Actual value*	*Actual value*
Learning rate	*2e -5*	*2e -5*
Number of epochs	*3*	*3*
BiLSTM hidden size	*–*	*Actual value*

Number of BiLSTM layers	–	*Actual value*
Dropout	–	*Actual value*
Optimizer	*AdamW*	*AdamW*
Loss function	*Cross-entropy*	*Cross-entropy*
Evaluation metrics	*Accuracy, Precision, recall, F1-score*	*Same metrics*

5. EXPERIMENTAL RESULTS AND DISCUSSION

The models were evaluated using accuracy, weighted precision, weighted recall and weighted F1-score. All eleven classes, including the non-punctuation class O, were included in the weighted metrics. Eight models were tested under the same corpus partitioning and evaluation protocol. The overall results are presented in Table 4.

Table 4. Overall comparison of the evaluated models

Model	***Accuracy***	***Precision***	***Recall***	***F1-score***
BiLSTM	*0.8229*	*0.8068*	*0.8229*	*0.7910*
RoBERTa	*0.8434*	*0.8303*	*0.8434*	*0.8204*
XLM-RoBERTa	*0.8580*	*0.8473*	*0.8580*	*0.8379*
XLM-RoBERTa+BiLSTM	***0.8588***	***0.8493***	***0.8588***	***0.8393***
BERTbek	*0.8370*	*0.8370*	*0.8370*	*0.8370*
BERTbek+BiLSTM	*0.8464*	*0.8373*	*0.8463*	*0.8379*
TahrirchiBERT	*0.8491*	*0.8366*	*0.8491*	*0.8276*
TahrirchiBERT+BiLSTM	*0.8494*	*0.8369*	*0.8494*	*0.8282*

As shown in Table 4, all transformer-based models outperformed the BiLSTM baseline in weighted F1-score. The best overall result was achieved by XLM-RoBERTa+BiLSTM, with an accuracy of 0.8588 and a weighted F1-score of 0.8393. Adding BiLSTM produced modest improvements for XLM-RoBERTa, BERTbek and TahrirchiBERT, indicating that sequential modelling can complement transformer-based contextual representations. For example, the unpunctuated input "***bugun talabalar darsga keldi lekin o‘qituvchi hali kelmagan edi***" was restored by the final hybrid model as "***Bugun talabalar darsga keldi, lekin o‘qituvchi hali kelmagan edi.***"

5.1. Class-level performance analysis

Table 5 presents the class-level precision, recall, and F1-score values of the best-performing XLM-RoBERTa + BiLSTM model.

Table 5. Class-level results of the XLM-RoBERTa + BiLSTM model

Label	***Precision***	***Recall***	***F1-score***
O	*0.8924*	*0.9803*	*0.9343*
PERIOD	*0.8076*	*0.7933*	*0.8004*
COMMA	*0.6665*	*0.7249*	*0.6945*
QUESTION	*0.9248*	*0.5523*	*0.6916*
EXCLAMATION	*0.3555*	*0.1011*	*0.1574*
COLON	*0.7893*	*0.4797*	*0.5967*
SEMICOLON	*0.7707*	*0.2787*	*0.4094*

ELLIPSIS	*0.7066*	*0.2135*	*0.3279*
DASH	*0.6695*	*0.2815*	*0.3963*
QUOTE	*0.8357*	*0.2601*	*0.3967*
PAREN	*0.7228*	*0.2604*	*0.3829*

The model achieved its highest F1-score for O (0.9343), followed by PERIOD (0.8004), COMMA (0.6945) and QUESTION (0.6916). PERIOD showed balanced precision and recall, whereas COMMA had higher recall than precision. QUESTION achieved very high precision (0.9248) but lower recall (0.5523), indicating that many question marks were missed. Rare and context-dependent classes performed less effectively, with EXCLAMATION obtaining the lowest F1-score (0.1574), mainly because of class imbalance and variation across syntactic, stylistic and genre-specific contexts.

5.2. Confusion matrix and error analysis

Figure 2 shows the row-normalised confusion matrix of the XLM-RoBERTa + BiLSTM model. Most errors involve O, COMMA and PERIOD, with rare punctuation marks often misclassified as O. Confusion between O and COMMA reflects difficulties in detecting clause boundaries, while PERIOD–COMMA confusion indicates uncertainty about boundary strength. Overall, class imbalance and the context-dependent nature of rare punctuation remain the main challenges.

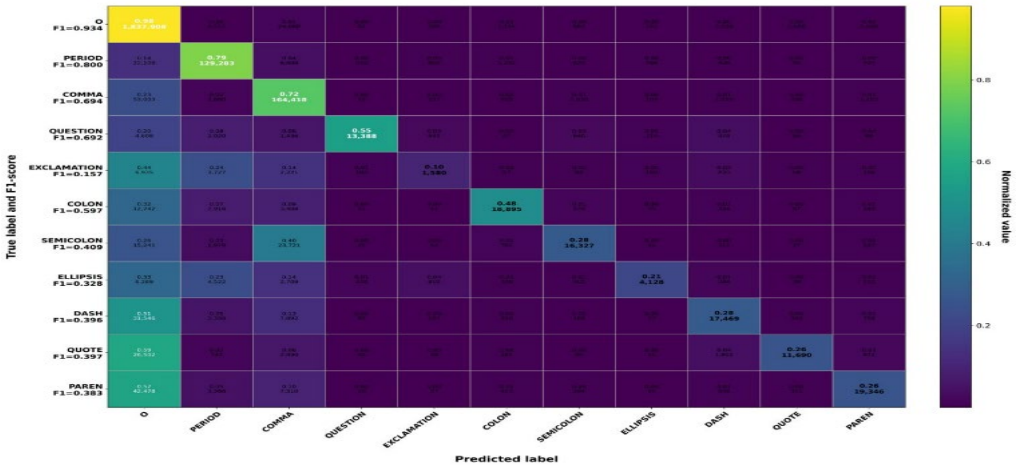


Figure 2. Row-normalized confusion matrix of the XLM-RoBERTa + BiLSTM model on the test subset.

5.3. Visualization of class-level metrics

Figure 3 illustrates the precision, recall, and F1-score values for each class. The visualization confirms that the model performs reliably for O and PERIOD and obtains comparatively satisfactory results for COMMA, QUESTION and COLON. For QUESTION, QUOTE, SEMICOLON, and ELLIPSIS, precision is considerably higher than recall. This pattern indicates that predictions for these classes are often correct when they are produced, but the model fails to identify many of their actual occurrences.

The largest imbalance between precision and recall is observed for QUESTION and QUOTE. The model achieved precision values of 0.9248 and 0.8357 for these classes, whereas their recall values were only 0.5523 and 0.2601, respectively. These results demonstrate that evaluation based only on accuracy or weighted F1-score would not adequately reflect the model's performance on less frequent punctuation classes.

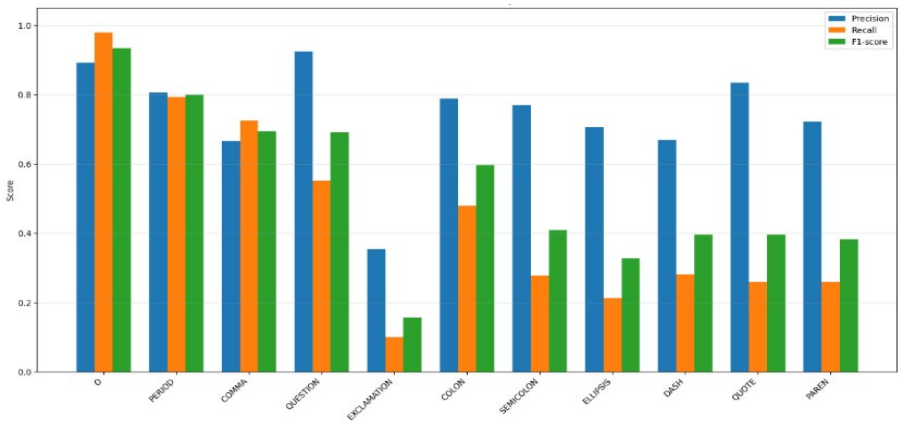


Figure 3. Class-level precision, recall, and F1-score of the XLM-RoBERTa + BiLSTM model on the test subset.

6. CONCLUSION

This study presented a language-adaptable hybrid Transformer+BiLSTM framework for token-level punctuation restoration in low-resource languages and evaluated it using Uzbek as a case study. The task was formulated as a sequence-labeling problem in which each input word was assigned either one of ten punctuation classes or the non-punctuation class O. The proposed architecture combines the contextual representation capability of pretrained transformer encoders with the bidirectional sequential modelling capability of BiLSTM.

The experimental evaluation compared a standalone BiLSTM model, multilingual and language-specific transformer encoders, and their hybrid Transformer+BiLSTM variants. The best overall performance was obtained by XLM-RoBERTa + BiLSTM, which achieved an accuracy of 0.8588, weighted precision of 0.8493, weighted recall of 0.8588, and weighted F1-score of 0.8393. Adding the BiLSTM layer produced numerical improvements for all three evaluated transformer families, although the magnitude of the improvement varied depending on the underlying encoder. These findings indicate that explicit bidirectional sequential modelling can complement the contextual representations generated by transformer models.

The class-level analysis showed that the model performed best for the O, PERIOD, COMMA, and QUESTION classes. However, substantially lower recall was observed for EXCLAMATION, ELLIPSIS, DASH, QUOTE, PAREN, and SEMICOLON. These results demonstrate that overall and weighted metrics alone are insufficient for evaluating punctuation restoration systems when the corpus is dominated by the O class. Rare and context-dependent punctuation marks remain difficult to identify because their use depends on syntactic structure, discourse context, genre, and authorial style.

Although the framework was experimentally validated only on Uzbek, its architecture and training procedure are not restricted to a particular language. The same pipeline can be adapted to another low-resource language by replacing or fine-tuning the tokenizer and transformer encoder, preparing a language-specific annotated corpus, defining the relevant punctuation inventory, and aligning word-level labels with subword tokens. The strong performance of XLM-RoBERTa also suggests that multilingual encoders can provide an effective basis for languages with limited pretrained resources.

The present study has several limitations. The experiments were conducted on a single language and with one random seed, while some improvements between the base and hybrid models were relatively small. In addition, no acoustic features, explicit syntactic information, or statistical significance tests were included. Future work will therefore focus on evaluating the framework on additional low-resource languages, conducting repeated experiments with multiple random seeds, improving class balance through weighted loss functions and targeted data augmentation, and integrating CRF layers and language-specific rule-based post-processing. The resulting models may be incorporated into automatic speech recognition, text editing, sentence segmentation, machine translation, and academic writing support systems.

REFERENCES

[1] Paiș, V., Tufiș, D. Capitalization and punctuation restoration: A survey. *Artificial Intelligence Review*, vol. 55, 2022, pp. 1681-1722. DOI: 10.1007/s10462-021-10028-8.

[2] Devlin, J., Chang, M.-W., Lee, K.,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*, vol. 1 (Long and Short Papers), Minneapolis, Minnesota, USA, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[3] Lafferty, J. D., McCallum, A., Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. *18th International Conference on Machine Learning, Massachusetts*, USA, June 2001, pp. 282-289.

[4] Tilk, O., Alumäe, T. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. *Proceedings of Interspeech 2016*, San Francisco, USA, 2016, pp. 3047-3051. DOI: 10.21437/Interspeech.2016-1517.

[5] Alam, T., Khan, A., Alam, F. Punctuation restoration using transformer models for high- and low-resource languages. *6th Workshop on Noisy User-generated Text (W-NUT 2020)*, Online, November 2020, pp. 132-142. DOI: 10.18653/v1/2020.wnut-1.18.

[6] Guhr, O., Schumann, A.-K., Bahrmann, F., Böhme, H.-J. FullStop: Multilingual deep models for punctuation prediction. *Proceedings of the Swiss Text Analytics Conference 2021 (SwissText 2021)*, CEUR Workshop Proceedings, vol. 2957, Winterthur, Switzerland, 14–16 June 2021, pp. 1–9.

[7] Vandeghinste, V., Guhr, O. FullStop: Punctuation and segmentation prediction for Dutch with transformers. *Language Resources and Evaluation*, vol. 58, 2024. DOI: 10.1007/s10579-023-09676-x.

[8] Bakare, M., Anbananthen, K. S. M., Muthaiah, S., Krishnan, J., Kannan, S. Punctuation restoration with transformer model on social media data. *Applied Sciences*, vol. 13, no. 3, 2023, Art. no. 1685. DOI: 10.3390/app13031685.

[9] González-Docasal, A., García-Pablos, A., Arzelus, H., Álvarez, A. AutoPunct: A BERT-based automatic punctuation and capitalisation system for Spanish and Basque. *Procesamiento del Lenguaje Natural*, no. 67, 2021, pp. 59-68. DOI: 10.26342/2021-67-5.

[10] Shymkovych, V., Nowakowski, G., Telenyk, S. Joint punctuation restoration and text capitalisation with a hybrid XLM-RoBERTa-LSTM model. *13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Gliwice, Poland, September 2025, pp. 990-997. DOI: 10.1109/IDAACS68557.2025.11322226.

[11] Saoud, A., Alomeyr, M., Kesgin, H. T. and Amasyali, M. F. Scaling BERT models for Turkish automatic punctuation and capitalization correction. *arXiv preprint*, arXiv:2412.02698, 2024. DOI: 10.48550/arXiv.2412.02698.

[12] Yagi, S., Elnagar, A., Yaghi, E. Arabic punctuation dataset. *Data in Brief*, vol. 53, 2024, Art. no. 110118. DOI: 10.1016/j.dib.2024.110118.

[13] Sharipov, M., Adinaev, H., Kuriyozov, E. Rule-based punctuation algorithm for the Uzbek language. *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, Altai, Russian Federation, 2024, pp. 2410-2414. DOI: 10.1109/EDM61683.2024.10615061.

[14] Sharipov, M., Adinaev, H., Sobirov, O. Bidirectional LSTM-CRF models for punctuation restoration in Uzbek texts. *2025 10th International Conference on Computer Science and Engineering (UBMK)*, Istanbul, Türkiye, 2025, pp. 354-358. DOI: 10.1109/UBMK67458.2025.11206853.

[15] Sharipov, M., Kuriyozov, E., Vičič, J. UzbekPOS: A multi-domain dataset for Uzbek part-of-speech tagging. *Data in Brief*, vol. 66, 2026, Art. no. 112640. DOI: 10.1016/j.dib.2026.112640.

[16] Sharipov, M., Adinaev, H., Ganijonov, S. Hybrid deep model for Uzbek language punctuation prediction. *European Science Review*, No. 1–2, 2026, pp. 41–46. DOI: 10.29013/ESR-26-1.2-41-46.

[17] Ma, J., et al. Parallel BiLSTM-Transformer networks for forecasting chaotic dynamics. *arXiv preprint*, arXiv:2510.23685, 2025.

Information about the authors:

Hushnudbek Adinaev is a lecturer at the Department of Computer Engineering, Urgench State University named after Abu Rayhan Beruni. His research interests include natural language processing, automatic analysis of Uzbek texts, punctuation restoration, and machine learning methods.

Maksud Sharipov is an associate professor at Urgench State University named after Abu Rayhan Beruni. His research interests include computational linguistics, natural language processing, machine learning, and the development of intelligent systems for the Uzbek language.

Manuscript received on 18 July 2026