European Science Review 2026,
No 1–2
ISSN 2310-5577

PREMIER Publishing
ppublishing.org

Section 2. IT Technology

DOI:10.29013/ESR-26-1.2-41-46

HYBRID DEEP MODEL FOR UZBEK LANGUAGE PUNCTUATION PREDICTION

Maksud S. Sharipov [1]***, Hushnudbek S. Adinaev*** [2]***, Shahzodbek S. Ganijonov*** [3]

1 Department of Computer Sciences, Urgench State University, Urgench, Uzbekistan

[2] Department of CE, Urgench State University, Urgench, Uzbekistan

[3] 10th grade, school No.15, Urgench, Uzbekistan

Cite: *Sharipov, M.S., Adinaev, H.S., Ganijonov, S.S. (2026). Hybrid Deep Model for Uzbek Language Punctuation Prediction. European Science Review 2026, No 1–2.* https://doi.org/10.29013/ESR-26-1.2-41-46

Abstract

Automatic punctuation restoration is one of the important tasks in natural language processing, especially for low-resource languages such as Uzbek, where this problem remains particularly relevant. The lack of sufficiently annotated corpora negatively affects the performance of downstream applications, including speech recognition, machine translation, and semantic text analysis. This study proposes a hybrid deep learning model designed to predict punctuation marks in Uzbek texts. The proposed approach combines BERT-based contextual vector representations, BiLSTM for modeling sequential dependencies, and a rule-based post-processing stage grounded in linguistic knowledge. This architecture effectively leverages the semantic capabilities of transformer models and the temporal dependency modeling strengths of recurrent networks, while the rule-based correction component improves the accuracy of punctuation detection in ambiguous boundary cases. Experimental results obtained on an annotated Uzbek corpus demonstrate that the proposed model outperforms existing neural and statistical approaches in terms of precision, recall, and F1-score. The findings confirm that integrating deep neural architectures with linguistic rules significantly enhances punctuation restoration performance for low-resource languages. This work presents a practical and extensible approach for advancing Uzbek language processing and improving the accuracy of various applied NLP systems.

Keywords: *punctuation marks; NLP; BERT model, BiLSTM model, Rule-based; F1 metrics*

1. Introduction

Natural Language Processing (NLP) has become one of the fastest-growing areas of artificial intelligence in recent years. The effectiveness of many practical applications – such as automatic text analysis,

speech recognition, machine translation, and question-answering systems – directly depends on the correct reconstruction of textual structure. From this perspective, automatic punctuation detection and restoration represents a crucial preprocessing step. This task is particularly important for texts derived from spoken language or written without punctuation marks.

Although punctuation restoration has been extensively studied for high-resource languages, research remains relatively limited for low-resource languages such as Uzbek. The scarcity of annotated corpora, the complexity of linguistic characteristics, and the challenges of directly transferring existing models necessitate the development of effective, language-adaptive approaches. As a result, incorrect or missing punctuation negatively impacts the performance of downstream systems, including speech recognition, machine translation, and semantic analysis.

In recent years, deep learning–based approaches – particularly recurrent neural networks and transformer architectures – have demonstrated strong performance in modeling sequential dependencies. Pretrained language models such as BERT enable the extraction of rich contextual representations, while BiLSTM networks effectively capture temporal dependencies within sequences. However, in certain ambiguous cases, purely neural approaches may fail to achieve sufficient accuracy. Therefore, incorporating linguistic rule-based post-processing mechanisms is considered a practical solution to improve robustness.

This paper proposes a hybrid deep learning model for punctuation prediction in Uzbek texts that integrates BERT, BiLSTM, and rule-based post-processing. The proposed approach unifies the semantic representation capabilities of transformer models, the sequential modeling strengths of recurrent networks, and the disambiguation power of linguistic rules within a single architecture. Experiments are conducted on an annotated Uzbek corpus, and model performance is evaluated using precision, recall, and F1-score metrics.

The main contributions of this study are as follows:

(1) proposing a hybrid neural–linguistic model for punctuation restoration in Uzbek; (2) improving accuracy through the integration of deep learning and rule-based approaches; (3) presenting a practical and extensible solution for punctuation restoration in low-resource languages.

2. Related Work

Punctuation restoration and analysis is considered one of the important tasks in the field of Natural Language Processing (NLP). Proper punctuation improves text readability and has a significant impact on the effectiveness of downstream NLP tasks such as syntactic parsing, machine translation, information extraction, and automatic speech recognition (Guhr O., Schumann A.-K., Bahrmann F., Böhme H.-J., 2021). Although punctuation marks were initially regarded as secondary graphical elements, recent studies have demonstrated that they encode essential syntactic and semantic boundaries within text.

Early research on punctuation detection and correction primarily relied on rule-based approaches. These methods employed manually crafted algorithms grounded in grammatical and stylistic rules of a language. For low-resource languages, such approaches remain relevant. One of the notable works for the Uzbek language is the rule-based algorithm proposed by Sharipov et al., in which the usage of periods and commas was analyzed using linguistic rules (Sharipov M. S., Adinaev H. S., Kuriyozov E. R., 2024). While this study laid the foundational groundwork for automatic punctuation analysis in Uzbek, it was reported to have limitations in handling complex contextual structures.

Subsequently, punctuation restoration began to be formulated as a sequence labeling problem, leading to the widespread adoption of statistical models, particularly Conditional Random Fields (CRF). By modeling dependencies between tokens, CRF-based approaches achieved better performance compared to purely rule-based methods (Attia O. et al., 2014). For the Uzbek language, studies have demonstrated that approaches combining BiLSTM and CRF models can achieve high accuracy and F1-scores (Sharipov M., Adinaev H., Sobirov O., 2025).

In recent years, deep learning–based models, especially Bidirectional Long Short-Term Memory (BiLSTM) networks, have been extensively applied to punctuation restoration tasks. These models jointly consider left and right contextual information, enabling effective learning of long-range dependencies within text (Salimbajevs J., 2018). The integration of a CRF layer with BiLSTM further stabilizes sequence-level predictions of punctuation marks.

With the emergence of transformer architectures, significant advances have been achieved in punctuation restoration. In particular, BERT and its multilingual variant mBERT have demonstrated strong performance in restoring punctuation in Uzbek texts (Adinaev H. S., 2025). These models are pretrained on large-scale multilingual corpora and provide deep contextual semantic representations. Furthermore, transformer models such as XLM-RoBERTa and RoBERTa offer effective transfer learning capabilities for low-resource languages (Shymkovych V. et al., 2025).

Recent studies increasingly adopt hybrid architectures that combine transformer models with recurrent neural networks. Models such as BERT–BiLSTM and XLM-RoBERTa–LSTM jointly capture global contextual semantics and sequential dependencies, resulting in improved accuracy (Zhu X. et al., 2024). In some works, punctuation restoration has been jointly learned with capitalization prediction, further enhancing structural coherence of the reconstructed text.

In addition, hybrid systems integrating acoustic and lexical features have been proposed for automatic speech recognition tasks. For example, an acoustic–lexical approach developed for Spanish demonstrated superior performance in question mark detection compared to text-only models (Qiu J. et al., 2026). This indicates that leveraging multi-source information is a promising direction for punctuation restoration.

Recent research also explores punctuation not only as a final prediction target but as an auxiliary signal. Specifically, punctuation-aware sparse attention mechanisms developed for large language models utilize punctuation marks as indicators of semantic boundaries, significantly improving long-context modeling performance (Qiu J. et al., 2026).

3. Methodology

In this study, the task of automatic punctuation restoration in Uzbek texts is addressed using a **hybrid approach**. The proposed method integrates a contextual neural model **(BERT + BiLSTM)** with a **rule-based post-processing** stage. This approach is specifically designed to reduce errors arising in the prediction of low-frequency punctuation marks.

3.1. Dataset Construction

The corpus used in this research contains more than **122 million tokens** and over **8 million sentences,** and was compiled from publicistic, official, and technical texts. Statistical analysis of punctuation marks within the corpus reveals a highly imbalanced distribution, indicating the necessity of additional mechanisms beyond a purely neural model. **Table 1** presents the statistical overview of the dataset.

Table 1. *Statistical summary of the dataset*

№	Label name	Number
1	Number of words (tokens)	122 161 579
2	Number of sentences	8 196 343
3	Number of commas	7 752 610
4	Number of periods	8 283 883
5	Number of question marks	265 449
6	Number of exclamation marks	249 576
7	Number of colons	1 128 044
8	Number of semicolons	219 920

№	Label name	Number
9	Number of ellipses	144 755
10	Number of parentheses	3 730 689
11	Number of quotation marks	1 799 606
12	Number of dashes	805 807
13	Total number of punctuation marks	24 380 339

3.2. Preprocessing

The texts were subjected to Unicode normalization, Uzbek-specific special characters were standardized, and word-level tokenization was performed. Sentence boundaries were identified, and punctuation-free raw text was prepared as input for the model.

3.3. Annotation and Formatting

For each token, corresponding punctuation labels were assigned. The label set includes periods, commas, question marks, exclamation marks, colons, semicolons, ellipses, dashes, parentheses, and quotation marks. The annotated data were stored in CoNLL format and split into training, validation, and test sets with an 80/10/10 ratio.

3.4. Neural Model Architecture

The punctuation restoration task was formulated as a sequence labeling problem. A BERT model adapted for the Uzbek language was selected as the base encoder. Contextual embeddings generated by BERT were passed to a Bidirectional LSTM (BiLSTM) layer to model long-range dependencies between tokens. In the final stage, a fully connected layer was applied to predict the punctuation label for each token.

Figure 1. *Architecture of the proposed BiLSTM–BERT hybrid model with rule-based post-processing*

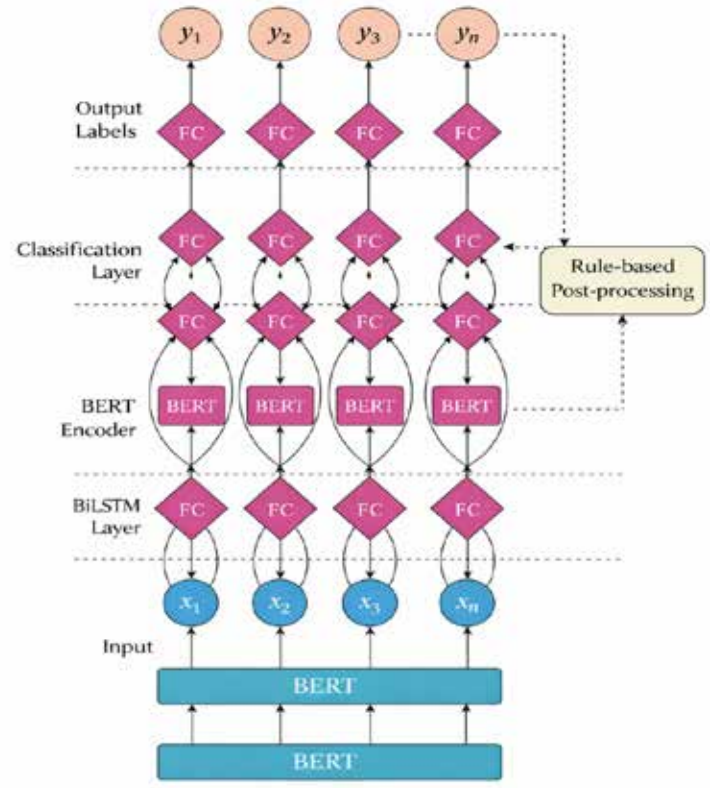

Figure X. Architecture of the proposed BERT–BiLSTM hybrid model for punctuation restoration. Contextual embeddings generated by BERT are processed by a bidirectional LSTM layer and classified at the token level, followed by rule-based post-processsing.

3.5. Rule-Based Post-Processing (Hybrid Stage)

The model outputs were further refined through a rule-based post-processing stage. This stage applies normative rules of the Uzbek language, including the insertion of dashes in reported speech constructions, the use of colons after explanatory and formal expressions, comma placement rules associated with discourse markers, and capitalization at sentence boundaries. This post-processing stage significantly improves the accuracy of low-frequency punctuation marks, which are often challenging for purely neural models.

Figure 2. *Confusion matrix (%) of the proposed BiLSTM–BERT hybrid model with rule-based post-processing*

4. Conclusion

In this study, a hybrid approach was proposed for the task of automatic punctuation restoration in Uzbek texts. The proposed model integrates a contextual language model **(**BERT**)**, a sequence modeling component **(**BiLSTM**),** and a linguistically motivated rule-based post-processing mechanism. This combination effectively unifies the statistical learning capabilities of neural models with the precision of rule-based approaches.

Experimental results obtained on a large-scale, CoNLL-formatted dataset constructed from real Uzbek texts demonstrate that the proposed hybrid model achieves approximately 87% Macro-F1. Confusion matrix analysis shows that the model performs with very high accuracy on frequent punctuation marks such as periods, commas, and question marks. At the same time, the rule-based component plays a crucial role in reducing errors observed in purely neural approaches for punctuation marks such as colons, quotation marks, and parentheses.

Although punctuation marks such as ellipses and dashes remain challenging due to their strong dependence on context, the observed errors can be attributed to inherent semantic and pragmatic ambiguities of the language. Overall, the results confirm that the hybrid approach provides a robust, interpretable, and practically effective solution for punctuation restoration in Uzbek.

5. Future work

Based on the findings of this study, the following directions are considered promising for future research:

Incorporating discourse and pragmatic features to improve the detection of punc-

tuation marks such as ellipses and dashes, it is recommended to introduce mechanisms that model inter-sentential relations and document-level context, such as document-level transformer architectures.

Expanding and adapting the rule system the current set of linguistic rules can be automatically expanded based on statistical observations or enhanced with adaptive mechanisms, including learnable or data-driven rule induction approaches.

Utilizing multi-genre corpora further training the model on literary, journalistic, and social media texts is expected to improve its generalization capability across different writing styles.

Practical integration integrating the proposed model into text editors (e.g., Word add-ins), web applications, or mobile platforms would contribute to the development of automatic text editing tools for the Uzbek language.

Releasing resources as open data Making the dataset, model architecture, and rule sets publicly available would positively impact the advancement of Uzbek NLP research and encourage further studies in this domain.

References

Guhr O., Schumann A.-K., Bahrmann F., Böhme H.-J. “FullStop: Multilingual Deep Models for Punctuation Prediction,” *Proceedings of the SEPP-NLG Shared Task*, 2021.

Sharipov M. S., Adinaev H. S., Kuriyozov E. R. “Rule-Based Punctuation Algorithm for the Uzbek Language,” 2024.

Attia O. et al. “Automatic Spelling and Punctuation Correction for Arabic,” *Computational Linguistics*, 2014.

Sharipov M., Adinaev H., Sobirov O. “Bidirectional LSTM–CRF Models for Punctuation Restoration in Uzbek Texts,” *IEEE UBMK*, 2025.

Salimbajevs J. “Automatic Punctuation Restoration Using Bidirectional LSTM Models,” *IOS Press*, 2018.

Adinaev H. S. “The mBERT Model for Restoring Punctuation in Uzbek-Language Texts,” *European Science Review*, 2025.

Shymkovych V. et al. “Joint Punctuation Restoration and Text Capitalisation with a Hybrid XLM-RoBERTa–LSTM Model,” *IEEE IDAACS*, 2025.

Zhu X. et al. “Resolving Transcription Ambiguity in Spanish: A Hybrid Acoustic-Lexical System for Punctuation Restoration,” *ACL Workshop*, 2024.

Qiu J. et al. “Punctuation-aware Hybrid Trainable Sparse Attention for Large Language Models,” *arXiv:2601.02819*, 2026.

submitted 27.01.2026;
accepted for publication 11.02.2026;
published 28.02.2026

Contact: hushnudbek.adinaev@gmail.com