European Science Review 2025,
No 7–8
ISSN 2310-5577

PREMIER Publishing
ppublishing.org

Section 4. Informatics

DOI:10.29013/ESR-25-7.8-28-33

THE MBERT MODEL FOR RESTORING PUNCTUATION IN UZBEK-LANGUAGE TEXTS

Hushnudbek S. Adinaev [1]

[1] Department of IT, Urgench State University, Urgench, Uzbekistan

Cite: *Adinaev H.S. (2025). The mBERT model for restoring punctuation in Uzbek-language texts. European Science Review 2025, No 7–8.* https://doi.org/10.29013/ESR-25-7.8-28-33

Abstract

This study proposes an mBERT-based approach for restoring punctuation in Uzbek-language texts. The main objective is to ensure the structural coherence of Uzbek texts by accurately reinserting punctuation marks. Using the mBERT model, we first predict punctuation for each token and then compare the prediction with any existing punctuation in the text to determine whether each mark is correctly or incorrectly placed.

Within the project, we construct a dedicated Uzbek corpus in which the relationship between every word and its surrounding punctuation is explicitly annotated. Each text is labelled according to its morphological and syntactic features. A dataset derived from this corpus is then prepared for training the model.

Keywords: *punctuation marks; NLP; mBERT model; F1 metrics*

Introduction

Punctuation marks are essential graphic tools that ensure written speech in a given language is presented correctly, expressively, and logically; they help condense text and indicate the logical-grammatical relationships between its parts (clauses and phrases). Punctuation belongs to the core system of symbols—together with letters—and differs in several respects from auxiliary symbol systems such as numerals, scientific notations, and printing marks.

The use of punctuation follows a specific system: the total number of marks, the order in which they are placed, and the principles governing their use together constitute punctuation rules. These marks convey various conceptual relationships and psychological nuances that cannot be expressed by other written means (letters, numerals, diacritics) or by linguistic units alone (words, morphemes), thereby making written text easier to understand.

The primary functions of punctuation are to indicate the semantic segmentation of speech and to clarify its syntactic structure and intonational pattern. Modern Uzbek orthography employs ten punctuation marks: the period, question mark, exclamation mark, comma, parentheses, dash, ellipsis, semicolon,

quotation marks, and colon. Most of these marks entered Uzbek writing in the second half of the 19th century, coinciding with the appearance of certain newspapers and lithographed books.

Predicting punctuation marks is a major challenge in natural-language processing (NLP). Correct and consistent punctuation greatly simplifies reading and comprehension and also enhances downstream tasks such as part-of-speech tagging, syntactic parsing, information extraction, and machine translation. Consequently, adding punctuation to transcribed speech or to texts that lack it remains a pressing problem (Pham, Q. H., Nguyen, B. T., Cuong, N. V. 2014).

Literature Review

Automatic speech-recognition (ASR) systems typically produce transcripts without punctuation and sentence boundaries. The absence of punctuation has been shown to reduce readability and to degrade the accuracy of downstream tasks such as neural machine translation, sentiment analysis, and information extraction. The authors recast punctuation restoration as a sequence-labeling problem and propose BERT-based models for English and Hungarian. Their approach extends an existing contextual model and boosts accuracy by averaging overlapping (sliding-window) predictions for each token. On the TED Talks corpus (English) and the Szeged Treebank (Hungarian), the proposed model achieved macro-F1 scores of 79.8 and 82.2, respectively, matching or surpassing previous methods (Nagy, A., Bial, B., va Ács, J., 2021).

Punctuation analysis plays a crucial role in natural language processing, requiring a model that can accurately predict proper punctuation for text preprocessing, spelling correction, grammar checking, information retrieval, and other applications. The task of accurately placing punctuation is context-dependent, which makes universal, language-independent tools less effective for this purpose. While many languages already have implemented such tools, Uzbek is a low-resource language, and, to our knowledge, there are no algorithms yet developed for analyzing or predicting punctuation in Uzbek texts. This paper proposes a rule-based algorithm and a model for analyzing periods and commas in Uzbek texts (Sharipov, M. S., Adinaev, H. S., and Kuriyozov, E. R. 2024).

The article develops an algorithm based on an N-gram statistical language model for the automatic analysis of punctuation in Uzbek texts. The author builds a corpus of 137 848 sentences (1 642 860 words) covering 30 different topics and, with a bigram model, estimates the probability of each punctuation mark, predicting the mark in place of the next word. Test results show an overall accuracy of roughly 81 percent; the F1 score for the period, comma, and question mark hovers around 0.87, while rarer marks (e.g., parentheses and quotation marks) perform slightly worse. The author suggests future work that combines the N-gram model with rule-based methods and AI approaches, and explores applying the method to other languages (Adinaev, H. S. 2025).

The article proposes a rule-based algorithm for automatically extracting interrogative sentences (questions) from Uzbek-language texts. The algorithm relies on predefined lists of interrogative pronouns (e.g., *kim* "who," *nima* "what," *qayerda* "where," *qachon* "when," *nima uchun* "why," *qanday* "how") and interrogative clitics (*-mi*, *-chi*, *-a*, *-ya*). Each word in a sentence is compared against these lists; if a match is found, the sentence is labelled "question" and added to the output list.

For evaluation, a corpus of 137 848 sentences (1 642 860 words) was compiled from 30 different text genres. Out of 7 186 genuine questions, the algorithm correctly identified 7 180, while it mistakenly labelled 554 out of 130 368 declarative sentences as questions. This yields an overall accuracy of 87.9 %, with F1 scores ranging from 76 % to 98 % across subject areas. Detailed results show that the highest accuracy was achieved on literature and mathematics texts, whereas the lowest was observed on anatomy and economics texts. The authors plan to improve accuracy by 2 hybridizing the system with N-gram language models and AI-based approaches (BiLSTM, CNN) and by expanding the corpus (Sharipov, M. S. and Adinaev, H. S. 2025).

The article describes the creation of a dedicated punctuation corpus for studying

and automatically restoring punctuation in Uzbek texts. The authors collected 30 types of material—literary works, textbooks, popular-science articles, and web content—building a corpus of 137 848 sentences (1 642 860 words). After cleaning the texts, they annotated them with ten different punctuation labels, and the frequency of each mark is presented in Table 1. Using Python scripts, they implemented functions to count the marks and save the results in CSV format. Statistical analysis shows that the period and comma are the most frequent marks, while less conventional marks—such as the dash, quotation marks, and ellipsis—occur relatively rarely. The authors plan to use the corpus in future NLP tasks such as adding punctuation to ASR outputs, developing grammar checkers, and training transformer-based models, as well as to improve accuracy through artificial-intelligence techniques (Sharipov, M. S. and Adinaev, H. S. va Yusupova, M. M. 2025).

The article offers an in-depth analysis of punctuation marks in written discourse from three perspectives: syntactic, logical-semantic, and emotional-intonational. For each of the ten primary marks-period, comma, semicolon, colon, question mark, exclamation mark, quotation marks, parentheses, dash, and ellipsis—the authors supply examples and emphasize their indispensable role in ensuring clear and precise reading. They show how incorrect punctuation can distort meaning, cause reader confusion, and disrupt written communication. Problems observed in education-such as an inability to apply rules in practice and difficulties with complex sentences—are discussed along with remedies, including interactive platforms and regular text-analysis exercises.

The final section focuses on applying modern NLP and AI technologies (ASR systems, BERT, RoBERTa, transformer models) to punctuation restoration. It cites literature confirming that hybrid approaches yield high accuracy and that combining lexical and prosodic features is effective. Thus, the study frames punctuation not only as a theoretical issue but also as a practical and technological challenge (Elov, B. B. va Sobirova, Z., 2025)

Materials and Methods

In this study, we developed a prepared dataset required for adapting the multilingual BERT (mBERT) model to the task of restoring punctuation in Uzbek texts. The process was carried out in three stages:

1. **Corpus collection —** A raw corpus of 1.5 million words was assembled from textbooks, articles, and online publications in the fields of programming, information technology, and engineering.

2. **Pre-processing —** spelling errors and non-standard punctuation were automatically corrected, Unicode normalization was applied, and all text was converted to UTF-8 encoding; sentence boundaries were detected and the text was tokenized.

3. **Annotation and formatting —** Each token was assigned a label (PERIOD, COMMA, QUESTION, EXCLAMATION, etc.); the data were stored in an mBERT-friendly layout using SentencePiece tokens and the BIO-style CoNLL format. The corpus was then split into 80 % training, 10 % validation, and 10 % test sets, with class balance preserved.

As a result, a clean and well-structured Uzbek corpus—labelled for ten core punctuation marks—has been created and is fully ready for training and evaluating mBERT or other transformer-based models. This dataset is expected to markedly improve the quality of automatic punctuation restoration in Uzbek.

3.1. The program code for creating the dataset is as follows

```
import re
from pathlib import Path
# Kirish va chiqish fayllari
RAW_FILE = “korpus.txt”
OUT_FILE = “dataset.conll”
# 10 ta tinish belgisi va ularning teg nomlari
PUNCTUATION_TAGS = {
    “.”: “PERIOD”,
    “,”: “COMMA”,
    “?”: “QMARK”,
    “!”: “EMARK”,
```

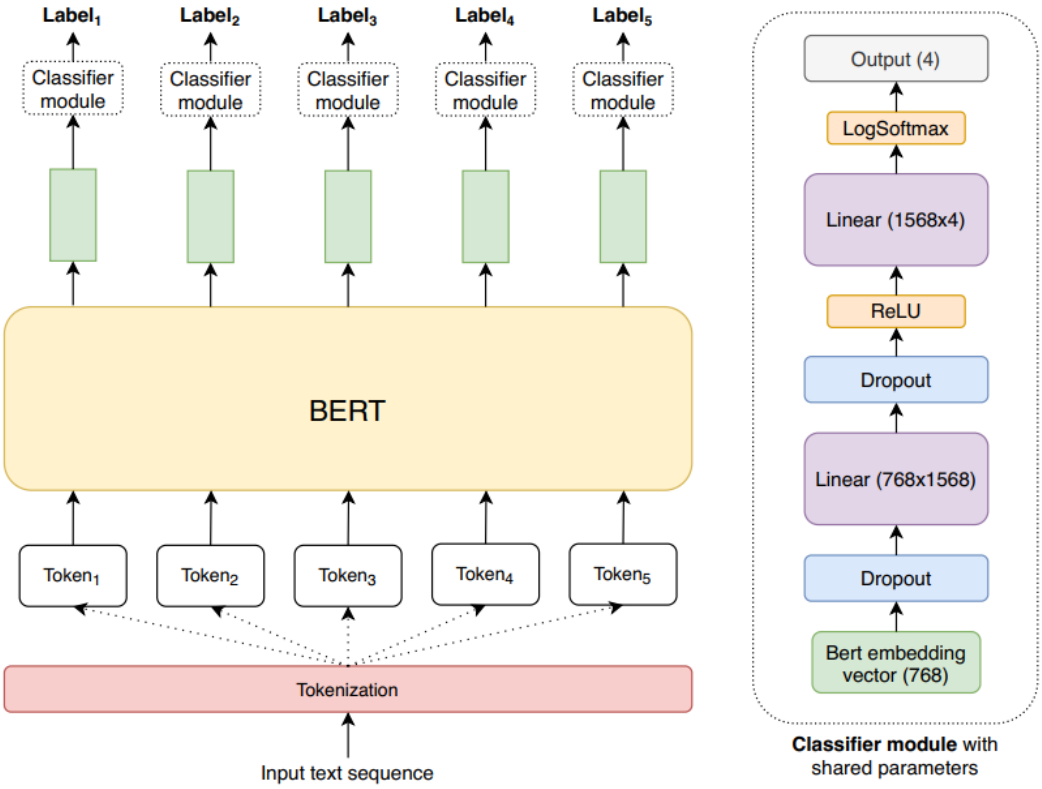
```
    “:”: “COLON”,
    “;”: “SEMICOLON”,
    “—”: “DASH”,
    “...”: “ELLIPSIS”,
    “\””: “QUOTE”,
    “(“: “PAREN”,
    “)”: “PAREN”
}
# Belgilarni uzunligi bo‘yicha tartiblab olish (masalan, “...” avvalroq)
SORTED_PUNCTS = sorted(PUNCTUATION_TAGS.keys(), key=lambda x: -len(x))
# Belgilarga mos regex (harflar orasidagi apostroflarni inobatga olamiz)
TOKEN_RE = re.compile(r”\w+[‘’‘‘‘’’]*\w*|[“+ re.escape (“”.join (SORTED_PUNCTS)) +
r”]|…”)
def tokenize_and_label(text):
    tokens = TOKEN_RE.findall(text)
    labeled_tokens = []
    for token in tokens:
        if token in PUNCTUATION_TAGS:
            if labeled_tokens:
                prev_token, _ = labeled_tokens[-1]
                labeled_tokens[-1] = (prev_token, PUNCTUATION_TAGS[token])
        else:
            labeled_tokens.append((token, “EMPTY”))
    return labeled_tokens
def process_file(in_path, out_path):
    with open(in_path, “r”, encoding=”utf-8”) as infile, open(out_path, “w”, encod-
ing=”utf-8”) as outfile:
        for line in infile:
            line = line.strip()
            if not line:
                continue
            labeled = tokenize_and_label(line)
            for word, label in labeled:
                outfile.write(f”{word}\t{label}\n”)
            outfile.write(“\n”)
if __name__ == “__main__”:
    process_file(RAW_FILE, OUT_FILE)
    print(f”[OK] Dataset yaratildi: {OUT_FILE}”)
```

3.2. The created dataset is tokenized as follows

Table 1. *Sample dataset created for mBERT*

Sequence of words	Sequence of tokens
Aksariyat	EMPTY
dasturlash	EMPTY
tillari	COMMA
xususan	COMMA
C++	COMMA
Python	COMMA
va	EMPTY
boshqalar	EMPTY
integrallashgan	EMPTY
dasturlash	EMPTY
muhiti	PAREN
IDE	PAREN
ga	EMPTY
ega	PERIOD

Figure 3.1. *The complete architecture used for punctuation restoration*



Findings and Analysis

For the experiment, a balanced Uzbek corpus of over 1.5 million words in the IT domain was selected. The resulting corpus contained 295 012 punctuation marks in total. The mBERT-based punctuation-prediction model achieved an overall accuracy of 85 percent—a noteworthy result.

Table 2. *Evaluation of models based on the F1-score*

№	Model	F1-Score
1	Rule based	0.752
2	N-gram	0.794
3	CRF	0.823
4	CRF+Rule	0.844
5	**mBERT**	**0.853**

Conclusion

Developing an Uzbek punctuation corpus is a crucial step for advancing natural-language processing (NLP) and automated punctuation systems. A properly collected and annotated corpus enables the creation of high-quality NLP models. Our punctuation corpus is a curated collection of texts designed to study and automate the use of punctuation marks. Provided in a machine-readable format, it plays a key role in tasks such as automatic punctuation restoration, speech-to-text conversion, grammar checking, and machine translation. Consequently, the corpus can bring significant benefits to AI and NLP research.

The corpus we created contains more than 1 500 000 words from IT-domain texts in Uzbek and includes 295 012 punctuation marks. As a continuation of this work, we aim to tackle Uzbek punctuation analysis with artificial-intelligence models and achieve even higher accuracy in future experiments.

References

Pham, Q.H., Nguyen, B.T., Cuong, N.V. (2014). Punctuation Prediction for Vietnamese Texts Using Conditional Random Fields // *ACML Workshop: Machine Learning and Its Applications in Vietnam (MLAVN 2014)*. – P. 1–9.

Nagy, A., Bial, B., va Ács, J., (2021). Automatic punctuation restoration with BERT models, *arXiv* preprint arXiv: 2101.07343, Jan.

Sharipov, M.S., Adinaev, H.S., and Kuriyozov, E.R. (2024). Rule-Based Punctuation Algorithm for the Uzbek Language, in International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM, – P. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.

Adinaev, H.S. (2025). "Punctuation Analysis of Uzbek Texts Based on the N-gram Model," Electronic Journal of Actual Problems of Modern Science, Education and Training, – vol. 2025. – No. 2. – P. 2025. ISSN 2181-9750.

Sharipov, M.S.va Adinaev, H.S. (2025). Development of models for punctuation analysis of Uzbek language texts, Bulletin of TUIT: Management and Communication Technologies, – vol. 1. – no. 4. 2025. ISSN 2181-1083.

Sharipov, M.S. and Adinaev, H.S. (2025). O‘zbek tili matnlarida so‘roq gaplarni aniqlashning qoidaga asoslangan algoritmini ishlab chiqish, Management and Future Technologies, – vol. 2. – no. 1. – P. 182 – 189. Mar. 2025.

Sharipov, M.S. and Adinaev, H.S. (2025). Shartli tasodifiy maydonlar modeli asosida o‘zbek tili matnlarini punktuatsion tahlil qilish, Al-Farg‘oniy avlodlari elektron ilmiy jurnali, – vol. 1. – no. 2. – P. 66–70, 2025.

Sharipov, M.S. and Adinaev, H.S. va Yusupova, M.M. (2025). O‘zbek tili matnlarida punktuatsion tahlil qilish uchun korpus yaratish, Development of Science, – vol. 2. – no. 3. – P. 128–133, Mar. 2025. ISSN 3030-3907.

Elov, B.B. va Sobirova, Z. (2025). Tinish belgilarining matndagi ahamiyati va vazifalari, *Proc. V Xalqaro ilmiy-amaliy konferensiya "Kompyuter lingvistikasi: muammolar, yechim, istiqbollar"*, – vol. 1. – no. 1. – P. 540–547, May 2025.

Sharipov, M. and Vičič, J. "Dataset of Uzbek verbs with formation and suffixes," *Data in Brief*, – vol. 61. 2025. doi: 10.1016/j.dib.2025.111731.

Sharipov, M., Kuriyozov, E., (2024). Yuldashov, O., and Sobirov, O. "UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts," in *Proc. of the 2024 Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, – P. 17343–17347.

submitted 12.07.2025;
accepted for publication 26.07.2025;
published 30.09.2025

Contact: hushnudbek.adinaev@gmail.com