Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Hybrid Model and Rule-Based Algorithm for Punctuation Prediction in Uzbek Texts

Maksud Sharipov[1], Hushnudbek Adinaev[2*], Tokhir Urazmatov[3], Ortik Ruzibaev[4], Omonboy Khalmuratov[5], Azizbek Ruzmetov[6], Vugar Abdullayev[7]

[1]Department of Computer Sciences and Artificial Intelligence Technologies, Urgench State University, Uzbekistan
[2]Department of Computer Engineering, Urgench State University, Uzbekistan
[3]Department of Digital Technology, Urgench Ranch University of Technology, Urgench, Uzbekistan
[4]Faculty of Software Engineering, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
[5]Educational and Methodological Support Department, Urgench State University, Urgench, Uzbekistan
[6]Department of Applied informatics, Kimyo International University in Tashkent, Tashkent, Uzbekistan
[7]Azerbaijan University of Architecture and Construction, Baku, Azerbaijan

*Corresponding email: abdulvugar@mail.ru

Abstract
Punctuation restoration is a routine task in NLP for well-resourced languages, but the picture changes for Uzbek. Annotated data is thin, the morphology is agglutinative, and correct punctuation directly affects how downstream systems read the text — from syntactic parsing to automatic processing pipelines. This paper describes a hybrid system built for the Uzbek case. The architecture has three parts: a BERT-based encoder (BERTbek) that provides contextual token representations, a BiLSTM layer that models the sequence dependencies around each token, and a post-processing rule set that handles constructions where the neural component tends to be uncertain — those tied to Uzbek grammar and orthography rather than to context alone. The model was trained on a 122-million-word Uzbek corpus drawn from textbooks, fiction, official websites and technical texts, and evaluated on a balanced test set of 30,000 words covering 30 subject areas, each represented by roughly a thousand words. Punctuation-restoration accuracy across the 30 domains reached 76.2%, with token-level accuracy of 87.0% on the confusion matrix. Performance was highest on narrative and topic-specific texts — cinema (98.3%), agriculture (96.6%), stories (94.5%) and culture (94.2%) — and lowest on mathematics (74.6%), sport (76.6%) and mass media (76.7%), where sentence structure varies more sharply and rare punctuation marks are more common. The rule layer contributed most of the improvement on grammatically ambiguous and context-dependent cases. We see this design — a language-specific transformer with a light linguistic layer on top — as a practical route for languages that do not yet have the data volume needed to train purely neural systems to comparable levels. In our setting, the components most likely to benefit are automatic text editing, speech-to-text conversion, machine translation, information retrieval and text indexing.

Keywords: punctuation restoration; Uzbek NLP; BERTbek; BiLSTM; rule-based post-processing.

1. INTRODUCTION

Punctuation analysis is one of the problems of natural language processing (NLP), which is to identify if the placement of period and comma in a text is correct or not. In natural language processing, punctuation analysis can be used for solving problems involving machine translation, text analysis, semantic analysis, syntactic analysis, etc. Like in other languages in general, a period (or a full stop) is one of the most used punctuation marks (.) in the Uzbek language and was used as such when it was added as a writing symbol to Arabic texts in the 11th century. Until the 19th century, it served and was used in a completely different way than its current role and form. The use of the dot as a symbolic sign goes back to ancient Arabic sources. Since the second half of 19th century it has been employed as a punctuation mark in the Uzbek language. This was the case in Western Europe since the 15th century, when the comma mark began to be used as a mark of punctuation. It has been used in Uzbek literature since the start of the 20th century. One of the most frequently used punctuation marks is the comma (,). The place of use, the form was different in different eras and writings of different languages. The Uzbek language also uses the term "inverted comma", "pesh", "half stop", "half pause" to refer to the comma. The use of a comma at first was for a short pause and subsequently its use was extended and its function broadened.

NLP has been one of the fastest-growing fields of AI over the last few years.
Many applications, including automatic text analysis, speech recognition, speech-to-text systems, machine translation and question and answer systems, rely heavily on the correct reconstruction of the text structure. One of the important steps in the process of text processing is automatic prediction and restoration of punctuation marks. This task is particularly relevant to texts that are extracted from spoken language or to texts that lack punctuation.
Punctuation restoration is a thoroughly researched subject in resource-rich languages, but scientific research on this subject is relatively scarce in resource-poor languages like Uzbek. Effective approaches to this task are needed, given the lack of annotated corpora, the complex linguistic aspects of the language and the difficulty of adapting existing models to the situation. This can therefore lead to a negative effect on the performance of subsequent systems, like speech recognition, machine translation or semantic analysis.

The identification of relationships between sequences has been successfully achieved in recent years by deep learning methods, especially neural networks and transformer architectures. Pre-trained language models such as BERT can be used to create more sophisticated representations of text context, and bidirectional long short-term memory (BiLSTM) networks can be used to model temporal dependency in sequences. In certain cases however, it is possible that methods based only on neural networks will not be able to give enough accuracy. So, extra mechanisms using linguistic rules are recommended.
In this research paper, a hybrid deep learning model based on BERT, BiLSTM and rule based techniques is proposed to predict the punctuation marks in uzbek text. We propose a method that is based on the semantic power of transformer models, the sequence modeling power of recurrent neural networks, and the power of linguistic rules to remove ambiguities, all in one. The experiments were carried out on an annotated corpus of Uzbek language and the performance of the model was assessed based on the F1 score.

2. Literature Review And Problem Statement

In natural language processing (NLP), punctuation and capitalization can be automatically restored from the text, which is known as "rich transcription" and plays a crucial role to improve the quality of the output text and the accuracy of the subsequent processes like machine translation, text-to-speech synthesis, and sentiment analysis. There is evidence that grammars that include punctuation marks significantly outperform those that do not when used to process complex real-world language; the punctuation marks offer important information into syntactic structure and semantic disambiguation [1]. Despite its importance, the restoration of punctuation is still a complex task especially for informal writing, such as social media, and the output of automatic speech recognition (ASR) systems which often do not include these orthographic markers.
Development of punctuation restoration techniques. The punctuation restoration literature shows a trend of improvement in the method from the traditional rule-based systems, statistical systems to more sophisticated systems based on deep learning and transformer architecture.

At first, the main approaches were Rule Based and Statistical approaches. Rule-based approaches are based on the syntactic and morphological patterns that are pre-established. Sharipov et al. (2024) proposed an algorithm for commas and periods, which is based on more than 30 linguistic rules, and which has been estimated to be more than 80% accurate on a multi-domain corpus in the context of the Uzbek language. These methods are linguistically interpretable, have high precision in controlled languages, but are hard to maintain and have problems with complexity and variation in natural languages. However, for agglutinative languages such as Uzbek, rule-based systems are difficult to use with regard to the separation of clause boundaries because of words with several morphological affixes [1,2].
The field moved towards data-driven modeling through the statistical approaches, especially Conditional Random Fields (CRF). CRFs are useful for sequence labelling because they can remember the interaction between adjacent labels and are able to deal with complicated sets of features [1,3]. Factorial CRF (F-CRF) models have been proven to be more effective than linear-chain CRFs for capturing bidirectional context by using comparative studies. But the statistical models tend to be less accurate in handling long distance dependency or out of vocabulary tokens in complex sentences [4].

The advent of deep learning has overcome the need for manual feature engineering, particularly through Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. Unlike RNNs, LSTMs are able to remember significant temporal information over longer distances, which helps get rid of the vanishing gradient problem [3]. This was further strengthened via bidirectional LSTMs (BiLSTM) which read sequences in both a forward and backward direction, allowing for a more holistic view of each word throughout the sequence [4,5].
The combination of a CRF layer with a BiLSTM-based structure (BiLSTM-CRF) provided a solid structure for sequence labelling. For the global tag consistency problem in Uzbek, Sharipov et al. (2025) showed that the BiLSTM-CRF model is very successful, with an F1 score of 89.5%. This combination makes it better than

the conventional CRF models due to its sequential dependency modeling capacity of the BiLSTM layer [4,6]. Sequential models, however, have a limitation in that they cannot model global dependencies in parallel, which may lead to poor performance on very long sentences [3,7].

Architectures and context embedding based on transformers. The Transformer architecture introduced the multi-head self-attention mechanism, which made it possible to model long-range dependencies without any regard for their distance, and revolutionized punctuation restoration [8].

The latest models for natural language understanding are Bidirectional Encoder Representations from Transformers (BERT) and its variants. BERT produces words' vector representation based on contextualization instead of word bias, by simultaneously watching both the left and right context [9]. Multiple studies have compared different BERT-like models (such as RoBERTa, XLM-R, and mBERT) with multilingual BERT models, and the trend is that models trained on very large, domain-specific language corpora generally outperform the multilingual models. For example, larger BERT models (from Tiny to Base) resulted in sizable boosts in the F1-scores for capitalization and punctuation in Turkish text correction [5,10]. Likewise, Bakare et al. (2023) used the RoBERTa-Large architectures with BiLSTM decoders, which reached the level of accuracy of 97% on social media data from English [25].

Advanced research in Adversarial and Multi-Task Learning has been investigating how to improve BERT based models with adversarial transfer learning. Yi et al. (2020) suggested a model that leverages Part-of-Speech (POS) tagging for auxiliary task during learning. Using a gradient reversal layer (GRL), the model learns task-invariant features that do not need an external POS tagger during inference to enhance the punctuation prediction task [12]. Later, Yi et al. (2023) showed that this adversarial transfer learning method outperforms previous lexical-based methods by 9.2% absolute in terms of Overall F1-score for punctuation prediction, which highlights that incorporating syntactic knowledge through adversarial training is very effective for punctuation prediction [7,8,12].

Linguistic, Stylistic and Hybrid Perspectives. Punctuation is important for literary and stylistic expression beyond functional restoration. In literary texts, Seagal (2026) discussed the phenomenon of "punctuation transfer", in which authors intentionally translate punctuation marks from one language (such as Spanish) to another (such as Russian) in order to achieve a particular stylistic effect. Here are two kinds of transfer: the former is for marking foreign-language copies in prose, and the latter establishes an original punctuation structure in poetry [9]. This emphasis shows that the purpose of the punctuation is not only grammatical but also it is a resource for stylistic nuance and cultural-linguistic expression.

Hybrid and Multimodal Systems: Hybrid architectures are recent trends because none of the models is complete. The PLASA model (2024) considers punctuation location attention and lexicon-query attention to fine-tune word weights towards human perception and empirical lexical knowledge, which greatly enhances sentiment analysis on short texts from social media. Moreover, in low resource languages such as Basque, multimodal models such as AutoPunct combine acoustic prosodic information, like silence duration, with BERT embeddings to enhance F1-scores [11]. Based on global attention, Shymkovych et al. (2025) also suggested the XLM-RoBERTa–LSTM model, which combines global attention with sequential dependencies, demonstrating the state-of-the-art performance on TED Talks datasets.

Research Gaps—Low Resource and Agglutinative Languages. Existing literature is critically analyzed and found that there are several significant gaps, especially for low-resource and morphologically complex languages such as Uzbek.

Lack of Annotated Resources: Although there are large resources available in English and Modern Standard Arabic (such as Arabic Punctuation Dataset with 312 million words [12]), the resources for Uzbek are very limited. There are not many large-scale datasets to train POS-tagging specialized models that tackle punctuation restoration and UzbekPOS dataset is being developed (4.4K sentences), but this is only a starting point.

Morphological Complexity: Agglutinative languages such as the Uzbek and Turkish languages pose special problems because the function of a word is determined by the affixes. These nuances are not well captured by the standard transformer models that were trained mainly on English [5].

Ambiguities in the Alphabet: The Uzbek alphabet has some diacritics and digraphs (e.g., "o‘", "g‘") which are easy to mistake for punctuation marks, resulting in ambiguity for the model [2,13].

In fields such as history, religion etc. rule-based systems for Uzbek have a significant decrease in accuracy below 60% in comparison to literature [2]. Neural models, on the other hand, may "hallucinate" in certain special areas without fine-tuning [14].

Conclusions and motivation of proposed method. The analysis of the current state of the art shows that transformer models such as BERT give the maximum contextual understanding and BiLSTM-CRF models are able to give the maximum sequence modeling, but none of them are fully suitable for the linguistic profile of the Uzbek language. Traditional rule-based systems for Uzbek are not flexible enough to be easily

applied to different domains, and general-purpose multilingual BERT models do not take into consideration the morphological rules and diacritic ambiguities in the language.
Hence, a hybrid of BERTbek and BiLSTM + Rule-based is required to address these gaps. BERTbek is a monolingual pre-trained model specifically for the Uzbek language, which provides deep bidirectional semantic representations needed for context dependent punctuation. The sequential dependencies and temporal features of long Uzbek sentences are well captured by the BiLSTM layer, and the Rule-based post-processing component is essential for handling some morphological edge cases and alphabet-specific ambiguities, such as separating out digraph markers from apostrophes, that are not captured by the neural models. The multi-layered architecture is especially well suited to provide high precision and recall, essential for precise punctuation in reliable Uzbek.

3. THE AIM AND OBJECTIVES OF THE STUDY

The aim of this research is to create an efficient model in order to boost automatic punctuation recovery accuracy and stability in uzbek texts by fusing the contextual representation of language obtained by BERTbek model with the sequence modeling capability of BiLSTM networks and the post processing model based on linguistic rules in a single hybrid architecture. The proposed method is useful to enhance the quality of punctuation prediction in the grammatically ambiguous and context dependent texts as well as to develop a practical solution for intelligent language processing systems in the case of Uzbek language.
To achieve this goal, the following tasks were set:
– modification of the BERTbek model with BiLSTM network to recover punctuation automatically from the Uzbek text;
– development of a rule-based post-process algorithm including the features of grammar and syntaxis of the Uzbek language and its connection with a hybrid model;
– compilation and creation of the initial processing stages of the Uzbek language corpus for training and validation of the model and implementation of the model;
– evaluate the effectiveness of the proposed model using evaluation indicators like F1 score and Confusion Matrix and compare the results with the results of the existing deep learning models;
– to analyze the effect of the use of deep learning methods along with linguistic rules in the accuracy of punctuation recovery and to evaluate the potential of the proposed approach for implementing in an intelligent system for processing of the Uzbek language [2,15].

4. MATERIALS AND METHODS OF RESEARCH

In this research work, the problem of automatic punctuation restoration in Uzbek texts was solved based on a hybrid approach. The proposed method combines a contextual neural model (BERTbek + BiLSTM) and a rule-based stage. This approach aims to reduce prediction errors, especially for rare punctuation marks and linguistically ambiguous cases.

A. Creating a data set.

The corpus used in this research was compiled from texts from school textbooks, fiction books, official websites, and technical books, and contains over 122 million words, over 8 million sentences, and 24 380 339 punctuation marks. Statistical analysis of punctuation marks in the corpus showed a significant imbalance between their classes [16]. This situation indicates the need to use additional mechanisms, not limited to a neural approach alone. Table 1. presents the statistical distribution of punctuation marks in the data set.

The corpus was initially subjected to the stages of text normalization, tokenization, and segmentation into sentences. Then, labels corresponding to punctuation marks were extracted and associated with appropriate tokens, forming a sequence labeling dataset suitable for training and evaluating deep learning models. The resulting dataset comprehensively reflects the patterns of punctuation usage in the Uzbek language and serves as the basis for developing the proposed hybrid punctuation restoration system.

Table 1. Dataset statistics

№	Item Description	Count
1	Number of Words	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
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

B. A regular algorithm for restoring punctuation marks.

A rule-based algorithm has been developed to automatically restore punctuation marks in Uzbek texts. This algorithm consists of several sequential steps, including text reception, tokenization, rule-based segmentation into sentences, sentence-level analysis, and punctuation. Initially, the incoming text is separated into tokens and divided into sentences based on predefined linguistic rules. Each sentence is then processed separately to identify syntactic and semantic structures that require punctuation [3,17]. The algorithm uses a set of linguistic rules specifically designed to place commas, periods, question marks, exclamation marks, and dashes based on the grammatical features of the Uzbek language. The linguistic rules developed are listed below.

The rule for placing a period punctuation mark can be expressed as follows:
P1=(Q; P; A→B; N)
Where:
Q = Analyze the end of the sentence in Uzbek;
P = The sentence is complete;
A = The last word ends with one of the given suffixes;
B = Puts a period (.);
N = Go to analyze the next sentence.
A={dim, ding, dilar, dik, man, san, miz, siz, yapti, moqchi, moqda, adi, ydi, tilar, sin, sinlar}.

The rule for placing comma punctuation marks can be expressed as follows:
1. Placing commas before words
P2=(Q; P; A→B; N);
Q = Determining whether to place commas before words that act as conjunctions and conjunctions;
P = The word occurs within a sentence;
A = {ammo, biroq, lekin, shuningdek, ammo-da, shu bilan birga, chunki, negaki, toki, go‘yo, shuning uchun, shu sababli, shu bois, ya’ni};
B = Puts a comma (,) before the word;
N = Continue parsing the sentence;

2. Put commas after words
P3=(Q; P; A→B; N);
Q = Determine whether to put a comma after the input words;
P = The word occurs in the sentence;
A = {shubhasiz, albatta, balki, ehtimol, aftidan, nazarimda, chamasi, shekilli, go‘yo, go‘yoki, axir, darhaqiqat, rostdan, rostdan ham, to‘g‘risi, aytmoqchi, afsuski, baxtiga, shunday qilib, xullas, tushunarli, qisqasi, xususan, zero, nahotki, iltimos, ehtiyot bo‘lish kerak, fikrimcha, o‘ylashimcha, ishonchim komilki};
B = Puts a comma (,) after the word;
N = Continue parsing the sentence;

3. Put commas on both sides of the word

P4=(Q; P; A→B; N);
Q = Determine whether to put commas on both sides of the input word;
P = The word occurs in the middle of the sentence;
A = {darvoqe, albatta, avvalo};
B = Put a comma (,) on both sides of the word;
N = Continue analyzing the sentence

4. Put a comma after the word when it comes at the beginning of the sentence
P5=(Q; P; A→B; N);
Q = Determine whether to put a comma when exclamatory and address words come at the beginning of the sentence;
P = The word comes at the beginning of the sentence;
A = {qani, nima, xo‘sh};
B = Puts a comma (,) after the word;
N = Continue analyzing the sentence.

5. Put commas on both sides when it comes in the middle of a sentence
P6=(Q; P; A→B; N);
Q = Determines whether to put commas when exclamatory and address words come in the middle of a sentence;
P = The word came in the middle of a sentence;
A = {qani, nima, xo‘sh};
B = Put a comma (,) on both sides of the word;
N = Continue analyzing the sentence.

6. Put a comma before the word when it comes at the end of the sentence
P7=(Q; P; A→B; N);
Q = Determine whether to put a comma when exclamatory and address words come at the end of the sentence;
P = The word comes at the end of the sentence;
A = {qani, nima, xo‘sh};
B = Puts a comma (,) before a word;
N = Go to the next sentence.

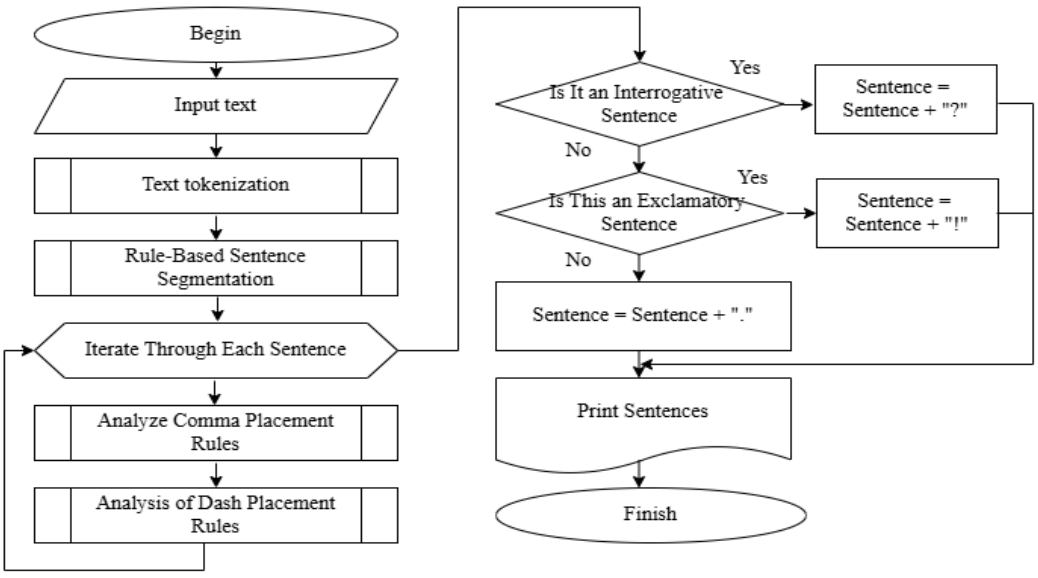
7.Putting a question mark after sentences containing interrogative pronouns
P8=(Q; P; A→B; N);
Q = Determine whether to put a question mark in sentences with interrogative pronouns;
P = Sentence is complete;
A = The sentence contains one of the following interrogative pronouns:
{qaysi biri, qancha vaqt, qancha masofa, qancha pul, qancha, nimalarni, qaysilarini, necha, nimalardan, kim, nimaning, nima qilasiz, kimim, qay holatda, nima qilasan, qayerdan, qanaqasi, qaysisi, nima bo‘ladi, nima qilding, qanaqa, nimada, nega, nima qilasizlar, nima qildingiz, nimalar, qay sababdan, nima bo‘ldi, qandayi, nima qilyapsiz, qay maqsadda, nima qilyapmiz, qayer, nima qilyapti, nima uchun, nima qilyapsan, kimning, qalay, nima qildilar, nima bo‘lyapti, qay vaqt, nima qildik, qay ko‘rinishda, nima qiladi, nechada, nimani, kimni, qay, nechtadan, qaysi, nima qilaman, nima qildim, kimda, qayerga, qay vaqtda, qachon, kimdan, qachonlar, kimlar, nima qil, qani, nima qilyapman, nechinchi, qanday, qay kuni, nimasi, qaysinisi, kimga, nima qilamiz, qayerda, nima, nimadan, qaysilari, qanchadan, nima bo‘l, nima qiladilar, kimi, nima qilyaptilar, qachondan, nimaga};
B = Puts a question mark (?) at the end of the sentence;
N = Go to the next sentence;
If the sentence contains an interrogative pronoun, then puts a question mark (?) at the end of the sentence.
P9=(Q; P; A→B; N);
Q = Determine whether to put a question mark in sentences with interrogative clauses;
P = Sentence is complete;
A = Sentence contains one of the following clauses: {mi, chi, a, ya, mikin, midi};
B = Puts a question mark (?) at the end of the sentence;
N = Proceed to the analysis of the next sentence.

8.The rule for placing an exclamation mark can be expressed as follows:

P10=(Q;P;A→B;N);
Q = Determines whether to place an exclamation mark after words containing exclamations and commands;
P = The word or sentence is complete;
A = The sentence contains one of the following words:
{woy, eh, yashang, barakalla, zo‘r, ajoyib, to‘hta, koch, oh, iye, bas, yetar, voidod, shoshil, to‘tang}
B = Put an exclamation mark (!) after the word or the end of the sentence
N = Go to the analysis of the next sentence
If the sentence contains one of the words with an exclamation or command content such as “woy”, “eh”, “yashang”, “barakalla”, “zo‘r”, “ajoyib”, “to‘hta”, “koch”, “oh”, “iye”, “bas”, “yetar”, “voidod”, “shoshil”, “to‘tang”, then an exclamation mark (!) is put after the word or the end of the sentence.

9.The hyphen punctuation rule can be expressed as follows.
P11=(Q;P;A→B;N)
Q = Determine whether to put a hyphen before a quotation or explanatory units
P = The word occurs in the sentence
A = The sentence contains one of the following words:
{this, that, said}
B = Put a hyphen (–) before the word
N = Continue analyzing the sentence
If the sentence contains one of the words “this”, “that”, “said”, then put a hyphen (–) before this word.
P12=(Q;P;A→B;N)
Q = Determine whether to put a colon after the words that start an explanation or enumeration
P = The word occurs in the sentence
A = The sentence contains one of the following units: {for example, decided made, decided is made, the subject}
B = Put a colon (:) after the word
N = Continue analyzing the sentence
A→B
If the sentence contains one of the units “for example”, “decided”, “is made”, “the subject”, then put a colon (:).
In addition, the type of sentence is determined and whether it is a declarative, interrogative, or exclamatory sentence is determined. The final punctuation mark - a period, question mark, or exclamation mark - is automatically inserted according to the type of sentence detected.
The proposed approach relies on explicit linguistic knowledge, rather than statistical learning. Therefore, it is particularly suitable for resource-constrained languages where large annotated corpora are not available. The algorithm provides a clear, interpretable, and computationally efficient solution for restoring punctuation marks [2,18,19]. It can also be used as a preprocessing component for subsequent natural language processing tasks such as text analysis, machine translation, and automatic speech recognition.

Figure 1. Block diagram of the algorithm for restoring regular punctuation marks for Uzbek texts.



The code for the rule-based algorithm is hosted on GitHub repository[1].

C. Marking and formatting data.

Each token in the corpus is assigned a label corresponding to its corresponding punctuation mark [20,21]. The set of labels included punctuation marks such as periods, commas, question marks, exclamation marks, colons, semicolons, colons, dashes, parentheses, and quotation marks. The annotated data was stored in the CoNLL format, which is widely used for sequence labeling tasks in natural language processing. To ensure reliable training and evaluation of the model, the dataset was split into training, validation, and test parts in a ratio of 80:10:10.

The set of labels is listed in Table 2.

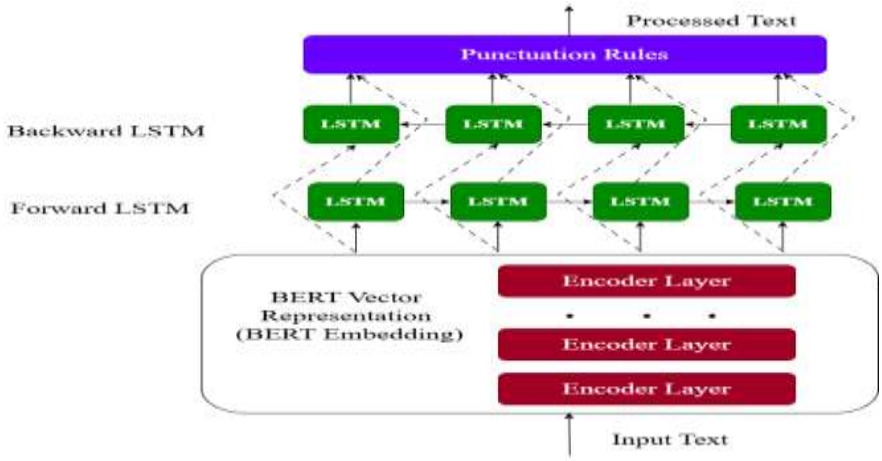
Table 2. Punctuation marks corresponding to a token.

№	Punctuation mark	Label name
1	If there is no punctuation mark	O
2	.	PERIOD
3	,	COMMA
4	:	COLON
5	;	SCOLON
6	?	QMARK
7	!	EMARK
8	…	ELLIPSIS
9	-	DASH
10	“”	QUOTE_L/R
11	()	PAREN_L/R

D. Neural model architecture.

The task of restoring punctuation marks was formulated as a sequence labeling problem.
The BERTbek language model, based on BERT and specifically adapted for the Uzbek language [26], was chosen as the main component that generates contextual images. The contextual vectors generated by BERT were passed to a Bidirectional Long Short-Term Memory (BiLSTM) layer. This layer allows modeling long-range relationships between tokens in both forward and backward directions. The hidden representations from the BiLSTM layer were then fed into a fully connected classification layer to predict the most appropriate punctuation label for each token. This architecture combines the deep context understanding capabilities of transformer models with the sequence modeling advantages of recurrent neural networks. As a result, the efficiency and accuracy of predicting punctuation marks in Uzbek texts is significantly increased.

Figure 2. Architecture of the BERTbek + BiLSTM + Rule-based model.



[1] https://github.com/husha1984/nuqta-va-vergulni-tahlil-qilish-dasturi/

5. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed hybrid model, a balanced Uzbek language corpus consisting of a total of 30,000 words covering 30 different disciplines and fields was used. Each domain contains approximately 1,000 words, which allowed us to objectively assess the quality of the model's performance on texts of various topics. The corpus used for the experiment contained a total of 3,698 punctuation marks, including 1,678 periods and 2,020 commas. The corpus was kept in its original form for evaluation purposes, and in addition, a version was created with all punctuation removed [5,7,22]. The unmarked text was passed as input to the model, while the original text was used as a benchmark to evaluate the prediction results. The performance of the model was evaluated based on the Accuracy indicator. Experimental results showed that the developed BERTbek + BiLSTM + rule-based hybrid model achieved an overall accuracy of 87.0% in restoring punctuation marks in Uzbek texts. This result confirms that combining deep learning methods and linguistic rules significantly improves the efficiency of automatic punctuation recovery. The results of the model's evaluation for individual disciplines and topics are presented in Table 3.

Table 3. Accuracy results of the proposed BERTbek + BiLSTM + rule-based model across different disciplines and domains.

№	Area/Field	Expected	Predicted	Accuracy
1	Literature	142	130	91.5%
2	Anatomy	90	76	84.4%
3	Biology	158	102	84.6%
4	Botany	130	80	89.5%
5	Religion	134	72	83.7%
6	World	96	68	90.8%
7	Physics	150	126	84.0%
8	Geography	110	70	93.6%
9	Stories	110	104	94.5%
10	Law	106	68	84.2%
11	IT	152	132	86.8%
12	Economy	64	60	93.8%
13	Sociology	96	66	88.8%
14	Chemistry	132	100	85.8%
15	Cinema	116	114	98.3%
16	Culture	104	98	94.2%
17	Math	134	100	74.6%
18	Language	150	102	88.0%
19	Agriculture	118	114	96.6%
20	Media	86	66	76.7%
21	Art	118	100	84.7%
22	Politics	100	86	86.0%
23	Sport	154	118	76.6%
24	Treatment	106	90	84.9%
25	Education	114	60	82.6%
26	History	166	96	87.8%
27	Technology	148	118	89.7%
28	Medicine	80	58	82.5%
29	War	148	120	91.1%
30	Zoology	186	122	95.6%
Total/Average:		**3698**	**2922**	**87.0%**

According to the results, the highest accuracy was observed in the fields of Cinema (98.3%), Agriculture (96.6%), Zoology (95.6%), Stories (94.5%), and Culture (94.2%). Relatively low results were recorded in Mathematics (74.6%), Sports (76.6%), and Mass Media (76.7%). Such differences are explained by the syntactic complexity of texts in different fields, the diversity of sentence structure, and the frequency of use of punctuation marks. According to Table 3, out of a total of 3,698 punctuation marks in the corpus, the model correctly predicted the punctuation with an average accuracy rate of 87.0%. These results demonstrate that the proposed hybrid model performs stably on Uzbek texts of various topics and provides

a sufficient level of accuracy for practical application [23]. The punctuation recognition quality of the proposed hybrid model was also evaluated using the Confusion Matrix presented in Figure 2.

Figure 3. Confusion Matrix results of the proposed BERTbek + BiLSTM + rule-based model.

The results show that the model predicts basic punctuation marks such as periods and commas with high accuracy. Also, while some errors were observed mainly in syntactically complex or context-dependent sentences, rule-based post-processing served to eliminate a significant portion of these errors. Overall, the Confusion Matrix results also confirm that the model effectively performed the task of automatically restoring punctuation in Uzbek with an overall accuracy of 87.0%.

6. DISCUSSION OF RESULTS

The aim of this research is to propose a hybrid approach using BERTbek, BiLSTM and rule-based post-processing mechanisms for automatic punctuation recovery in uzbek language texts. The desired model is developed by integrating the power of deep learning techniques for learning the context and sequence dependency with the interpretability and simplicity of linguistic rules. Experiments were carried out on the Uzbek language corpus defined in the CoNLL format to check the effectiveness of the approach. Experiments results indicate that the hybrid model predicted 87.0% accuracy in terms of punctuation marks. The results of the Confusion Matrix analysis showed high accuracy for prediction of high frequency punctuation marks like periods, commas and question marks. The rule-based post-processing mechanism also enhanced the quality of the recovery of some punctuation marks, such as quotation marks, parentheses and colons, which was observed to be less accurate in purely neural network approaches [24]. The study revealed that the punctuation task is still difficult especially for the dashes and full stops which appeared at lower frequencies in the corpus and context-dependent in usage. Most of the observed errors are due to the natural language's semantic and pragmatic ambiguities [25]. In general, the results obtained justifies the proposed hybrid solution as a stable, interpretable and practically viable solution for automatic punctuation recovery for Uzbek texts. This method can be applied to automatic text editing, speech-to-text conversion, machine translation and other natural language processing systems. For future research, the model will be evaluated in larger and more diverse corpora, more punctuation marks will be supported, and the model's efficiency will be further refined using new transformer architectures.

7. CONCLUSION

A hybrid model of BERTbek+ BiLSTM+rule-based post-processing mechanism was created and implemented to automatically recover the punctuation marks in the uzbek text. The model proposed merges the strengths of the language representation learning capabilities of deep learning techniques and the learning of sequence dependencies while retaining the transparency and interpretability of linguistic rules. The principle of this method is the use of the statistical learning properties of neural networks and grammatical properties of rules-based punctuation recovery systems.This approach is trying to combine the advantages of statistical learning by neural networks and grammar properties of rules-based systems to improve the efficiency of the Uzbek punctuation recovery task. The proposed model was tested on the Uzbek language corpus prepared in CoNLL format, and presented various disciplines and topics. The

experimental results revealed that the model obtained an overall accuracy of 87.0%. The accuracy of the model in predicting commonly used punctuation marks like periods, commas, and question marks was confirmed in the confusion matrix analysis.

The rule-based post-processing mechanism further enhanced the quality of character recovery in colons, quotation marks, parentheses, etc., which otherwise had some errors in the neural network based approaches. This is an indication of the effectiveness of the hybrid approach in practice. Based on the obtained results, it is concluded that adding linguistic rules to the deep learning models improves the accuracy and stability of punctuation restoration in the Uzbek language. The prediction of some punctuation symbols, especially hyphens, multiple periods, and rare symbols, however, is still difficult because they are not as frequently used in the corpus and have very tight relationships with semantic and pragmatic context. This situation calls for more improvement of the model. Scientific modeling of the proposed model combines an architecture of contextual language models, sequence modeling, and a rule-based post-processing into a single model, which is scientifically effective for automatic punctuation restoration of the Uzbek language.

In a practical sense, this can enhance the performance of automatic text editing, voice-to-text conversion, machine translation, information retrieval, text indexing and other natural language processing systems. Future research plans involve the expansion of the corpus further, the creation of multi-class models for all major punctuation marks, a comparison of the model with the modern transformer architectures, and the integration of the model into real-time intelligent software systems. Further, it is proposed to apply the developed approach to other Turkic languages and implement multilingual punctuation restoration systems – one of the directions is promising.

Conflicts of interest

The authors declare that there are no financial, personal, authorial, or any other conflicts of interest that could influence the conduct of this study and the results presented in the article.

Funding

This research was conducted without any grant, project, or other financial support.

Data availability

Experimental data, evaluation results, and corpus-related information supporting the results of this study can be provided by the corresponding author upon reasonable request.

Use of artificial intelligence

The authors declare that they used generative artificial intelligence technologies in the preparation of the article only to improve the language style of the scientific text, edit grammatical errors, and improve the style of scientific presentation. The research idea, methodology, experimental work, analysis of results, scientific conclusions, and the final content of the article were fully developed, verified, and approved by the authors. Full responsibility for all scientific results presented in the article lies with the authors.

Authors' contributions

Hushnudbek Adinaev: Conceptualization, methodology, software, data curation, model training and testing, formal analysis, visualization, writing — original draft.

Maksud Sharipov: Methodology, supervision, validation, writing — review and editing, project administration.

Tokhir Urazmatov: Investigation, resources, writing — review and editing.

Ortik Ruzibaev: Software, formal analysis, validation.

Omonboy Khalmuratov: Data curation, resources, writing — review and editing.

Azizbek Ruzmetov: Investigation, validation, writing — review and editing.

Vugar Abdullayev: Conceptualization, methodology, supervision, writing — review and editing.

8. REFERENCES

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

[2] Kim J., Choe Y., Mueller K. Extracting clinical relations in electronic health records using enriched parse trees //Procedia Computer Science. – 2015. – T. 53. – C. 274-283.

[3] Liu G., Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification //Neurocomputing. – 2019. – T. 337. – C. 325-338.

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

[5] O. Attia et al., "Automatic Spelling and Punctuation Correction for Arabic," Computational Linguistics, 2014.
[6] M. S. Sharipov, H. S. Adinaev and E. R. Kuriyozov, 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.
[7] M. Sharipov, H. Adinaev and O. Sobirov, Bidirectional LSTM-CRF Models for Punctuation Restoration in Uzbek Texts, 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025, pp. 354-358, doi: 10.1109/UBMK67458.2025.11206853.
[8] Y. Li, Y. Shao, L. Luo and Z. Han, A pipeline approach for entity and relationship extraction in geological hazard texts using BERT-BiLSTM-CRF with self-attention, Intelligent Geoengineering, vol. 3, no. 2, pp. 169–178, Jun. 2026, doi: 10.1016/j.ige.2026.06.008.
[9] J. Salimbajevs, "Automatic Punctuation Restoration Using Bidirectional LSTM Models," IOS Press, 2018.
[10] V. Shymkovych et al., "Joint Punctuation Restoration and Text Capitalisation with a Hybrid XLM-RoBERTa–LSTM Model," IEEE IDAACS, 2025.
[11] X. Zhu et al., "Resolving Transcription Ambiguity in Spanish: A Hybrid Acoustic-Lexical System for Punctuation Restoration," ACL Workshop, 2024.
[12] Yi, J., et al. (2020). Adversarial Transfer Learning for Punctuation Restoration. arXiv:2004.00248, pp. 1-10, DOI:10.48550/arXiv.2004.00248
[13] Helen A., Pradana A., Afif M. Scientific reference style using rule-based machine learning //International Journal of Advances in Intelligent Informatics. – 2023. – T. 9. – №. 3.
[14] Lei L., Liu D. Conducting sentiment analysis. – Cambridge University Press, 2021.
[15] Duridi, T., et al. (2025). Arabic hate speech detection based on BERT models variants. Egyptian Informatics Journal, pp. 134–143. https://doi.org/10.1016/j.eij.2025.100845
[16] S. Yaghi, A. Elnagar and E. Yaghi, Arabic punctuation dataset, Data in Brief, vol. 53, Art. no. 110118, Apr. 2024, doi: 10.1016/j.dib.2024.110118.

[17] González Docasal, A. García-Pablos, H. Arzelus and A. Álvarez, AutoPunct: A BERT-based Automatic Punctuation and Capitalisation System for Spanish and Basque, Procesamiento del Lenguaje Natural, vol. 67, pp. 59–68, Sep. 2021, doi:10.26342/2021-67-5.
[18] V. Vandeghinste and O. Guhr, FullStop: Punctuation and Segmentation Prediction for Dutch with Transformers, Language Resources and Evaluation, 2023, doi: 10.1007/s10579-023-09676-x.
[19] N. Kemp, R. Kovacic and E. Beyersmann, Is the period really "pissed"? The effect of punctuation and message length on perceptions in digital communication, Telematics and Informatics, vol. 97, Art. no. 102241, Feb. 2025, doi:10.1016/j.tele.2025.102241.
[20] Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, Cunhang Fan, Transfer knowledge for punctuation prediction via adversarial training, Speech Communication, Volume 149, April 2023, Pages 1-10, https://doi.org/10.1016/j.specom.2023.03.003
[21] D. Dolean and N. Prodan, Let's eat grandma: Awareness of punctuation and capitalization rules' violations predicts the development of reading comprehension, Learning and Instruction, vol. 86, Art. no. 101780, Aug. 2023, doi:10.1016/j.learninstruc.2023.101780.
[22] Z. Li and Z. Zou, Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform, Journal of King Saud University – Computer and Information Sciences, vol. 36, Art. no. 102010, 2024, doi:10.1016/j.jksuci.2024.102010.
[23] V. Paiș and D. Tufiș, Capitalization and punctuation restoration: A survey, Artificial Intelligence Review, vol. 55, pp. 1681–1722, 2022, doi:10.1007/s10462-021-10028-8.
[24] G. Tucudean, M. Bucos, B. Dragulescu and C. D. Căleanu, Natural language processing with transformers: a review, PeerJ Computer Science, vol. 10, Art. no. e2222, Aug. 2024, doi:10.7717/peerj-cs.2222.
[25] M. Bakare, K. S. M. Anbananthen, S. Muthaiah, J. Krishnan and S. Kannan, Punctuation Restoration with Transformer Model on Social Media Data, Applied Sciences, vol. 13, no. 3, Art. no. 1685, 2023, doi:10.3390/app13031685.
[26] E. Kuriyozov, D. Vilares and C. Gómez-Rodríguez, BERTbek: A Pretrained Language Model for Uzbek, in Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, Torino, Italy, May 2024, pp. 33–44.