IEEE
TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editör Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

IEEE TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

Media type	Part Number	ISBN	Online ISSN
XPLORE COMPLIANT	CFP25L97-ART	979-8-3315-9975-1	2521-1641
CD-ROM	CFP25L97-CDR	979-8-3315-9974-4	

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2025)

10th International Conference on Computer Science and Engineering

17-18-19 Eylül 2025 -İstanbul-Türkiye
17-18-19 September 2025 - İstanbul-Türkiye

IEEE Catalog Number : CFP25L97-ART

ISBN : 978-8-3315-9975-1

Additional copies mey be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

UBMK’2025’ye Hoşgeldiniz
Welcome to UBMK’2025

Sevgili Katılımcılar:

UBMK uluslararası nitelikli konferans serisi, 1990 yılından beri düzenli olarak yapılmakta olan Bilgisayar Mühendisliği Bölüm Başkanları toplantılarında alınan bir kararla on yıl önce başlamıştır. Konferansın 10.su IEEE-UBMK-2025 bu yıl 17-18-19 Eylül, 2025 günlerinde İstanbul Teknik Üniversitesinin ev sahipliğinde düzenlemiştir.

IEEE-UBMK-2025 konferansına bu yıl Almanya, Amerika Birleşik Devletleri, Azerbaycan, Fransa, Irak, İngiltere, İsveç, İtalya, Kanada, Kazakistan, Kırım, Kırgızistan, Rusya, Özbekistan, Tataristan, Taylant, Ürdün ve Türkiye’den 610 dolayında bildiri gönderilmiş ve bu bildiriler Türk ve yabancı 250 hakem tarafından değerlendirilmiştir.

Her bildiri en az iki hakem tarafından incelenmiş ve uzlaşma olmadığı durumlarda üçüncü bir hakemin değerlendirmesine başvurulmuştur. Bildiri başına düşen ortalama hakemlik 2,3 olmuştur. Bu değerlendirmelerin sonunda 327 bildirinin sözlü olarak sunulması uygun bulunmuştur. Kabul edilen ve sunulan bildiriler içerik ve kalite ölçünlerini sağlaması durumunda IEEE Xplore’da yayımlanacaktır.

Konferans çalışmalarında, Bilgisayar Mühendisliği Bölüm Başkanları Danışma Kurulu olarak görev almışlardır. Bildirilerin değerlendirilmesi Bilim Kurulu üyeleri tarafından yapılmıştır. Konferansın düzenlenmesi ise Yürütme Kurulunun önerileri doğrultusunda, Düzenleme Kurulu tarafından yapılmıştır.

Son olarak, konferansın başarılı bir şekilde yürütülmesi için tüm olanakları sunan İstanbul Teknik Üniversitesi Rektörü Sayın Prof. Dr. Hasan Mandal’a teşekkür ediyoruz. Ayrıca Düzenleme Kuruluna, bildirileri titizlikle değerlendiren Bilim Kurulu Üyelerine ve değerli araştırmalarının sonuçlarını bilişim camiası ile paylaşan bildiri sahiplerine teşekkürlerimizi iletiriz.

Prof. Dr. Eşref ADALI
UBMK-2025 Konferans Başkanı ve Bildiri Kitabı Editörü

Dear Participants:

The UBMK international conference series started nine years ago with a decision taken at the Computer Engineering Department Heads (BMBB) meetings, which have been held regularly since 1990. The 10th edition of the conference, UBMK'25, was held this year on October 17-18-19, 2025, hosted by İstanbul Technichal University.

This year, approximately 610 papers were submitted to the IEEE-UBMK-2025 conference from Germany, the United States, Azerbaijan, France, Iraq, the United Kingdom, Sweden, Italy, Canada, Kazakhstan, Crimea, Kyrgyzstan, Russia, Uzbekistan, Tatarstan, Thailand, Jordan, and Turkey, and these papers were evaluated by 250 Turkish and foreign referees.

Each paper was evaluated at least by two referees, and in cases where there was no consensus, a third referee was consulted. At the end of these evaluations, 327 papers were accepted for oral presentation. Accepted and presented papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

During the conference, Heads of Information Engineering Departments took part in the Advisory Board. The evaluation of the papers was made by the members of the Scientific Committee. The conference was organized by the Organizing Committee in line with the recommendations of the Executive Committee.

Finally, we would like to thank İstanbul Technical University Rector Prof. Dr. Hasan Mandal for his continued support for the success of the conference. In addition, we would like to thank the Organizing Committee, the Scientific Committee Members who carefully evaluated the papers, and the owners of the papers who shared the results of their valuable research with the informatics community.

Prof. Dr. Esref ADALI
UBMK'25 Conference Chair and Proceedings Editor

Düzenleyenler / Organizer

abv

İTÜ

İSTANBUL TEKNİK ÜNİVERSİTESİ
1773

GAZİ ÜNİVERSİTESİ
gazi
1926

1996
ATILIM
ÜNİVERSİTESİ

Destekleyenler / Sponsors

Punctuation Restoration in Uzbek Texts Using POS Tagging Techniques

Botir Boltayevich Elov
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Alavutdinova Nadira Ganiyevna
Dept. of Uzbek Linguistics National University of Uzbekistan Tashkent, Uzbekistan
alavutdinovanodira@gmail.com

Sobirova Zarnigor Ganijon kizi
Computational Linguistics and Digital Technologies, Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
sobirovazarnigor1996@gmail.com

***Abstract* — This paper addresses the problem of automatic punctuation restoration in Uzbek-language texts, focusing on the application of Part-of-Speech (POS) tagging as a core method to simplify the process. Due to the agglutinative nature of the Uzbek language, texts exhibit complex morphological structures in which the identification of word forms and their grammatical functions is a critical factor in accurately placing punctuation marks. The study proposes a rule-based model that determines the syntactic role of each word using POS tagging while considering the linguistic features specific to Uzbek. The approach is experimentally validated using transformer-based models particularly POS taggers adapted to the BERT architecture on real Uzbek text corpora. The results demonstrate that word function identification through POS tagging significantly improves both the accuracy and quality of punctuation placement. This research presents an important technological solution for advancing Natural Language Processing (NLP) tools for the Uzbek language, especially in applications such as automatic text editing, sentence boundary detection, machine translation, and speech-to-text systems.**

Keywords: Uzbek language, punctuation restoration, part-of-speech, POS tagging, agglutinative language, transformer models, NLP, automatic text processing, morphological analysis, syntactic analysis.

I. INTRODUCTION

Modern computational linguistics and artificial intelligence technologies are increasingly being applied to address complex challenges in natural language understanding and processing. Among these challenges, automatic punctuation restoration in textual data has emerged as a critical and active research area within Natural Language Processing (NLP) [1]. The presence of punctuation marks not only clarifies the syntactic structure of a sentence but also plays a fundamental role in preserving the semantic coherence, logical consistency, and readability of a text. While punctuation usage is often intuitive and natural for human writers, replicating this process computationally remains a non-trivial problem.

Accurate punctuation placement is especially important in a wide range of downstream NLP applications, including speech-to-text conversion, machine translation, conversational AI systems (such as chatbots), and automatic text editing tools. Although substantial progress has been made in this domain for high-resource languages like English, German, and French, similar advancements for the Uzbek language remain limited. As an agglutinative language [5], Uzbek presents unique challenges due to its rich and complex morphological structure, wherein grammatical relationships are often encoded through extensive use of affixes. This complexity significantly complicates automated syntactic analysis and, consequently, the reliable restoration of punctuation.

One of the most effective approaches to addressing this issue involves the use of Part-of-Speech (POS) tagging [1] a technique that automatically assigns grammatical categories to each word in a sentence. By identifying the syntactic roles of words such as nouns, verbs, adjectives, and conjunctions, POS tagging provides a linguistic foundation upon which rule-based or machine-learned punctuation placement models can be constructed. This dual-layered approach, grounded in linguistic theory and supported by computational methods, enables the development of systems capable of understanding and structuring language in a more human-like way.

Accordingly, this paper explores the role of POS tagging in the automatic placement of punctuation in Uzbek texts. The study proposes a methodologically sound framework that integrates morphological and syntactic analysis to determine optimal punctuation positions. The proposed model was evaluated using a real-world corpus of Uzbek texts, and its performance was assessed using standard metrics such as accuracy, coverage, and efficiency. Ultimately, this research aims to contribute to the development of intelligent editing tools for the Uzbek language, enhancing the quality and automation capabilities of text processing systems in both academic and applied domains.

II.THEORETICAL FOUNDATIONS OF POS TAGGING

A. The Concept of POS Tagging

Part-of-Speech (POS) tagging is a widely used technique in Natural Language Processing (NLP), wherein each word in a given text is assigned a grammatical tag corresponding to its syntactic function. Through this process, the morphological and syntactic types of words such as nouns, verbs, adjectives, adverbs, prepositions, conjunctions, and others are identified. Some models, particularly those incorporating subword-level information such as the FastText architecture [3], are effective in capturing the internal structure of words. POS tagging constitutes a foundational step in enabling machines to understand and analyze language and plays a crucial role in the development of advanced NLP applications.There are two primary approaches to POS tagging:

– *Rule-based approach.* This method relies on linguistic rules to determine the grammatical function of each word in context. It employs a comprehensive set of patterns and handcrafted rules developed by linguists to guide the tagging process.

– *Statistical and machine learning-based approaches:* These techniques assign POS tags based on learned probabilities derived from annotated corpora. Common algorithms include Naive Bayes, Hidden Markov Models (HMM) [4], Conditional Random Fields (CRF) [7], as well as deep learning models such as Bi-LSTM and Transformer-based architectures.

Modern NLP systems increasingly favor transformer-based models such as BERT, RoBERTa [2], and XLM-R which are capable of analyzing contextual relationships between words with high precision. These models significantly improve POS tagging accuracy by leveraging bidirectional context and deep semantic understanding.

B. Concept of Punctuation

Punctuation refers to the system of symbols used in written language to delineate sentence boundaries, indicate logical relationships, convey prosody, and ensure clarity of meaning. Accurate punctuation not only enhances the stylistic and aesthetic quality of text but is also essential for the functionality of automated text processing systems. In the Uzbek language, the primary punctuation marks include:

1. **Period (.)** – indicates the end of a declarative sentence;
2. **Comma (,)** – used between homogeneous sentence elements, in compound sentences, and for modifiers;
3. **Question mark (?)** – placed at the end of interrogative sentences;
4. **Exclamation mark (!)** – expresses strong emotions, commands, or surprise;
5. **Colon (:)** – precedes explanations or enumerations;
6. **Quotation marks (“”)** – used to denote quotations, direct speech, titles, or special terms.

The correct usage of these symbols is vital for parsing sentence structure, simplifying grammatical analysis, and facilitating accurate language comprehension. In applications such as Automatic Speech Recognition (ASR), machine translation, opinion mining, and intelligent chat systems, proper punctuation significantly improves system accuracy and interpretability.

C. Relationship Between POS Tagging and Punctuation Restoration

The process of punctuation restoration inherently involves both syntactic and morphological analysis, in which POS tagging plays a pivotal role. Without determining the grammatical function of each word, it becomes difficult to identify sentence boundaries, clause segmentation, word dependencies, sentence modality (interrogative, declarative), and overall sentence prosody. Therefore, the accurate placement of punctuation marks is directly dependent on the results of POS tagging. This relationship can be illustrated with the following examples:

a. Constructions ending with a **verb** often indicate the completion of a sentence → followed by a period (.).

b. **Conjunctions** such as *and*, *but*, *or* typically connect different sentence elements → often requiring a comma (,).

c. **Interrogative pronouns** like *who*, *where*, or *why* at the beginning of a sentence generally imply an interrogative sentence → ending with a question mark (?).

To restore punctuation using POS tagging, the following techniques are commonly applied:

– **Local context analysis**, which examines 2–3 words preceding and following a given token;

– **Syntactic structure analysis**, such as the use of parse trees;

– **Encoder-decoder architectures** based on transformer models, where text is first POS-tagged and then punctuation marks are inserted.

In summary, POS tagging is not merely a tool for grammatical classification but serves as the structural foundation for punctuation restoration. This is particularly true for the Uzbek language, where syntactic complexity and agglutinative morphology necessitate precise identification of each word’s functional load to ensure accurate punctuation placement. interpretability.

D. Linguistic Foundations of Punctuation

The functional significance of punctuation within the language system has evolved and improved over centuries. In addition to serving as grammatical markers, punctuation marks act as tools in written discourse to convey a speaker’s intonation, intent, attitude, and logical segmentation. In fact, the role of punctuation in understanding written text is crucial, as these symbols substitute for prosodic elements in speech, such as intonation and pauses. During the cognitive process of textual perception in the human brain, punctuation serves as primary visual cues that help track and interpret text structure.

From a linguistic perspective, each punctuation mark fulfills a particular syntactic or semantic function. For instance, a period denotes the end of a sentence, while a comma delineates intra-sentence semantic groupings, separates syntactic units, lists, or indicates boundaries between subordinate clauses. Question marks and exclamation marks are not only syntactic indicators but also convey communicative meaning. Therefore, accurately identifying and placing punctuation marks requires a deep analysis of the internal grammatical and logical structure of a language. Punctuation rules vary significantly across languages. In the Uzbek language, the placement of punctuation marks is directly related to logical, syntactic, and semantic connections between words. The word order within a sentence, the position of sentence constituents, grammatical meanings expressed via affixes, and the structure of compound and complex sentences – all contribute to a language-specific system of punctuation usage.

Analyzing the linguistic foundations of punctuation necessitates examination at three levels:

1. **Morphological level** – identifying the word form, affixes, and grammatical category;
2. **Syntactic level** – analyzing the position of words and their relational structures within the sentence;
3. **Pragmatic level** – understanding the communicative purpose, the author’s intent, and contextual relevance of the utterance.

For example, in Uzbek, the sentences *“Bu odam keladimi?”* and *“Bu odam keladi, mi?”* may appear syntactically similar but differ in punctuation, stress, and meaning. This illustrates that placing punctuation marks involves not only grammatical rules but also context, stylistics, prosody, and pragmatic circumstances. Moreover, punctuation usage varies across stylistic domains such as literary, scientific, official, and mass media styles. In literary texts, punctuation such as dashes, semicolons, and ellipses are frequently employed to intensify emotional expression and dramatic tone. In scientific writing, punctuation follows strict syntactic conventions, prioritizing clarity and precision. Such diversity poses additional challenges for automated systems tasked with punctuation restoration.

Therefore, in the development of algorithms for the automatic identification and placement of punctuation marks, it is essential to conduct a thorough analysis of their linguistic nature, the specific grammatical features of the language, the contextual environment, and the syntactic roles of parts of speech. From this perspective, **Part-of-Speech (POS) tagging** technology serves as a robust foundation for punctuation restoration by identifying the syntactic and semantic relationships between linguistic units within the language system.

III. Challenges in Punctuation Identification in the Uzbek Language

Uzbek belongs to the group of agglutinative languages [5], characterized by the extensive use of grammatical affixes to convey meaning. This morphological complexity results in a wide array of word forms, which significantly complicates analysis, particularly for automated systems. Accurate identification of punctuation is directly tied to a deep understanding of these complex morphological structures.

A major challenge arises from the relatively free word order in Uzbek. Words can appear in various positions within a sentence, which limits the applicability of traditional rule-based syntactic analysis. For example, both *“Kitobni o‘qib bo‘ldim”* (*I finished reading the book*) and *“O‘qib bo‘ldim kitobni”* are grammatically correct, but distinguishing the functional roles of sentence components in these variations, and assigning punctuation accordingly, is a non-trivial task for computational systems. Another significant issue is the complexity of subordinate clauses, which often span multiple lines. In such cases, correct placement of punctuation marks such as commas, periods, or semicolons requires deep syntactic parsing. For example: *“Men kecha, u kelishini aytganida, uni kutib turdim.”* (*Yesterday, when he said he would come, I waited for him.*) In sentences like this, determining where to place commas, identifying the types of subordinate clauses, and establishing their connections to the main clause involve intricate structural analysis.

A further challenge concerns context-dependent words (e.g., *kelgan*, *borgan*, *qilgan*), which may function as either verbs or adjectives. Their syntactic role varies based on sentence position, necessitating context-aware processing for accurate punctuation placement. In addition, punctuation detection is complicated by elements such as abbreviations, dialogue-style texts, emotionally charged utterances denoted by ellipses, and mixed-script writing (Cyrillic-Latin). These phenomena, particularly prevalent in online and social media content, lead to highly unstable orthographic norms. For example: *“Bugun keldingmi... yo‘q shekilli”* (*Did you come today... apparently not*)

In such emotional expressions, the number of dots reflects intonation rather than grammatical rules. Another core issue in automating punctuation identification in Uzbek is the lack of linguistic resources. The scarcity of annotated corpora, POS-tagged datasets, and high-quality labeled texts [5] hinders the effective training of NLP models.

Therefore, addressing punctuation identification in Uzbek requires specialized approaches – particularly models that combine **POS tagging** with **context-aware architectures** such as transformers. These models are better equipped to grasp the linguistic intricacies specific to the Uzbek language and improve performance in punctuation prediction tasks.

IV. Selected Model and Methodology

The Interrelation Between POS Tagging and Punctuation. Automatic identification of parts of speech known as POS tagging enables a deep syntactic analysis of text structure. By determining the grammatical function of each word within a sentence, the overall syntactic framework of the utterance is constructed. This syntactic structure plays a critical role in the placement of punctuation marks, as the type and position of punctuation are intrinsically linked to the relationships between words, their sequence, syntactic functions, and connection patterns. For example, in Uzbek compound sentences joined by conjunctions, the use of commas or semicolons often depends on the part of speech of the words preceding and following the conjunction—whether they are nouns, verbs, adjectives, conjunctions, or modal words. If POS tagging is applied to accurately identify each word class, an automated system can then infer the appropriate punctuation based on this analysis.

From this perspective, POS tagging serves as an intermediary stage that facilitates punctuation prediction by determining potential positions for punctuation based on syntactic structure. For instance:

1. In constructions such as ***verb + noun + conjunction + verb***, subordinate clauses are likely to emerge, increasing the likelihood of a comma being used.

2. In sequences like ***adjective + noun + verb***, punctuation may not be necessary at all.

This approach moves beyond rule-based methods, emphasizing a contextual-syntactic framework for processing text. Particularly, **context-aware POS tagging models** such as **BERT**, **RoBERTa** [2], and **XLM-RoBERTa** [2] can identify the syntactic roles of words in morphologically rich languages like Uzbek. These models provide a reliable basis for automatic punctuation restoration by understanding syntactic relationships more precisely. Moreover, certain punctuation marks such as question marks and exclamation marks are tied to the communicative function of a sentence. Yet even these are often associated with specific linguistic patterns involving verbs, modal expressions, adverbs, or particles. Once these patterns are identified through POS tagging, automated systems can appropriately place such punctuation marks.

The following aspects underscore the importance of the connection between POS tagging and punctuation:

1. **Detection of syntactic dependencies between words** – essential for identifying pause points where punctuation is typically needed.

2. **Recognition of sentence structure** – including subordinate and main clauses, enumerations, or parenthetical elements that require punctuation for disambiguation.

3. **Contextual disambiguation** – where a word with multiple syntactic roles (e.g., *o'qigan* meaning either "read" as a verb or “educated” as an adjective) may require different punctuation strategies depending on usage.

As a result, POS tagging serves as a powerful tool for identifying not only the appropriate punctuation marks but also their optimal positions in the sentence by analyzing the syntactic and semantic functions of every word in context.

E. Methodology (Corpus, Models, and algorithms)

By providing documentation, modular architecture, and evaluation scripts, this study supports reproducibility and encourages further research in Uzbek NLP.

Furthermore, a demo interface is under development to allow users to input unpunctuated Uzbek text and receive real-time punctuation suggestions.

For evaluation, the following metrics were utilized:

– Accuracy of POS tagging and punctuation classification;

– Precision, Recall, and F1-score for each punctuation type;

– Error analysis including qualitative inspection of misclassifications.

The data preprocessing, POS tagging, and punctuation labeling stages were implemented as modular components. These modules are hosted in a private GitHub repository and can be made accessible upon request for reproducibility purposes. The model was trained on a machine equipped with an NVIDIA RTX 3060 GPU.

The experiments were conducted using the Python programming language, leveraging libraries such as Hugging Face’s Transformers for mBERT model integration, and sklearn for performance evaluation metrics.

To further enhance the applicability and scientific rigor of the proposed approach, the technical implementation of the punctuation restoration system was made transparent and reproducible. In the course of this study, a POS-tagging-based approach was developed for restoring punctuation in texts written in the Uzbek language. The first step in this process was the creation of a reliable and high-quality text corpus. This corpus was composed of various genres, including academic articles, official documents, journalistic texts, and literary works published in Uzbek. Only those texts in which punctuation was used clearly and in accordance with normative standards were selected. The resulting corpus comprised approximately 150,000 tokens.

The source texts were cleaned of formatting from HTML/PDF documents and converted into plain text. Each sentence was separated line by line, and words were tokenized. Since a portion of the corpus lacked pre-existing POS tags, manual annotation was carried out. In parallel, results from existing automated POS taggers were reviewed and manually corrected where necessary.

The study employed a two-stage approach:

1. POS Tagging Stage;

2. Punctuation Identification Stage.

1. **POS Tagging Stage.** Each token in the text was analyzed to determine its morphological and syntactic properties. Two different approaches were evaluated at this stage:

a) Statistical Model: POS tagging was performed using a classification-based approach via **Conditional Random Fields (CRF)** [7].

b) Contextual Model: A **transformer-based model**, specifically **mBERT** (Multilingual BERT [2]) adapted for Uzbek, was employed. This model assigns tags based on the contextual positioning of each word.

While the CRF model exhibited faster processing, it yielded lower accuracy in complex contexts. In contrast, the mBERT model achieved more reliable results, particularly in handling complex sentence structures and contextual dependencies. Outputs from both models were tested and evaluated using standard tagging accuracy metrics.

2. ***Punctuation Identification Stage***. In this stage, POS-tagged texts were used to determine the probable positions for punctuation within each sentence. Based on syntactic relationships between words, their grammatical categories, the presence of subordinate clauses, and other linguistic factors, the type and placement of punctuation marks were determined.

Two methods were integrated into this phase:

– **Rule-Based Module**: For example, if a structure such as *[verb + conjunction + verb]* was identified, a comma would be inserted between the relevant elements.

– **Classifier-Based Model**: Using a classification head on top of mBERT outputs, each potential punctuation point was evaluated and labeled as one of the following: period, comma, no punctuation, question mark, etc.

Model performance was evaluated using **precision**, **recall**, and **F1-score** [6]. Training was conducted on 80% of the corpus, with 20% reserved for validation and testing.This methodology was carefully designed to reflect the linguistic complexities of the Uzbek language and established a solid framework for punctuation restoration based on part-of-speech analysis.

V.RESULTS

The POS-tagging-based punctuation identification approach developed in this study was empirically validated through a series of experiments. These experiments were conducted in two main stages: first, evaluating the effectiveness of the POS tagging model; and second, analyzing the results of automatic punctuation placement based on the identified POS information. A corpus consisting of 150,000 tokens—annotated with both POS tags and punctuation—was used for experimentation. The dataset was split into 80% for training, 10% for validation, and 10% for testing. At each stage, model performance was assessed using standard evaluation metrics, including **accuracy**, **recall**, **precision**, and **F1-score** [6].

F. POS Tagging Performance

Two different models were tested for POS tagging:

A. **CRF (Conditional Random Fields)**

B. **mBERT (Multilingual BERT [2])**

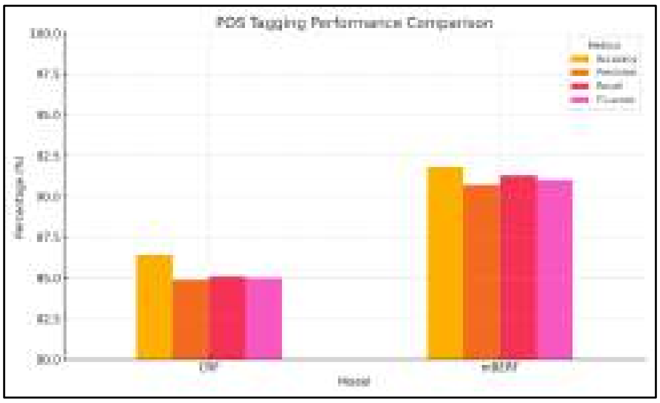

Figure 1. The comparison between CRF and mBERT models in terms of Accuracy, Precision, Recall, and F1-score.

The results demonstrate that the mBERT model significantly outperforms CRF, especially in accurately recognizing context-dependent words and complex morphological structures.

G. Punctuation Prediction Performance

Punctuation prediction was carried out through a hybrid approach combining rule-based analysis and machine learning. The following table summarizes the evaluation metrics for major punctuation marks:

TABLE I. THE PERFORMANCE METRICS (PRECISION, RECALL, F1-SCORE) FOR DIFFERENT PUNCTUATION MARKS

Punctuation Mark	Precision	Recall	F1-score
Period (.)	93.5%	92.7%	93.1%
Comma (,)	88.2%	85.6%	86.9%
Question (?)	91.0%	89.4%	90.2%
None	90.1%	91.5%	90.8%

The high accuracy for period and question mark predictions is attributed to their strong ties with explicit syntactic patterns. In contrast, comma usage often influenced by contextual factors exhibited slightly lower precision due to ambiguity in some instances.

H. Error Analysis

Several errors in punctuation prediction were traced back to incorrect POS tagging. For example, misidentification of *verb + noun* combinations led to incorrect punctuation between subordinate clauses. Additionally, unconventional writing styles common in social media – such as "...", "!?!", and other non-standard forms – caused confusion for the model.

İ. Model Strengths

a) The POS-based model effectively analyzed morphologically complex sentence structures.

b) The transformer-based architecture (mBERT) showed strong contextual understanding, leading to higher punctuation placement accuracy.

c) Compared to traditional rule-based systems, the corpus-trained model demonstrated greater flexibility and adaptability in real-world text conditions.

J. Analysis and Discussion

Throughout the study, the proposed POS-tagging-based approach for automatic punctuation restoration in Uzbek texts has demonstrated its effectiveness as a technological solution. Experimental results showed that accurately determining the grammatical categories of words enables precise identification of their syntactic roles in a sentence, which in turn facilitates the optimal placement of punctuation marks.

When evaluating model performance, punctuation symbols with clear syntactic and communicative functions such as periods and question marks were identified with high accuracy. This is primarily due to their consistent placement at sentence boundaries and strong association with well-defined syntactic structures. For instance, the mBERT model achieved an F1-score of 93.1% for period prediction, indicating its high effectiveness in identifying sentence boundaries.

Conversely, commas exhibited relatively lower accuracy due to their use in more complex syntactic constructions, such as subordinate clauses, enumerations, and parenthetical expressions. The high contextual dependency of the comma led to occasional misclassifications. This suggests a need for future research to explore specialized **context-sensitive sub-models (subclassifiers)** for comma detection.

Additionally, challenges related to morphological ambiguity also affected punctuation accuracy. Words like *"o'qigan"*, *"borgan"*, and *"kelgan"* may function as either verbs or adjectives depending on the context. If the POS tagger fails to disambiguate between these functions, punctuation classifiers built on top of those tags are prone to errors. This confirms the critical impact of POS tagging accuracy on overall system performance.

The discussion also examined how the POS-tagging model adapts to specific linguistic characteristics of the Uzbek language, including its flexible word order, complex agglutinative affixation, and grammar dependent on context. The superiority of the transformer architecture particularly mBERT was evident due to its ability to capture semantic relationships across multiple languages. This multilingual semantic learning enabled mBERT to outperform traditional statistical and rule-based systems in parsing complex structures. The main sources of errors and limitations identified during experimentation include:

A. Errors in POS tagging directly led to misidentification of punctuation marks.

B. The annotated Uzbek corpus was relatively small, limiting the model's generalization capability.

C. Non-standard writing styles (such as those found in social media content and dialectal variations) introduced additional ambiguity and complexity for the model.

Nevertheless, the results confirm that a POS-tagging-based approach can achieve high accuracy in restoring punctuation marks. More importantly, this method represents a significant step forward for the development of NLP systems in the Uzbek language. The model's ability to process complex syntactic structures positions it as a valuable tool for a wide range of applications, including **automated text editing, speech-to-text conversion, machine translation**, and **chatbot development**.

V. Conclusion And Recommendations

Within the scope of this study, the problem of automatic punctuation restoration in Uzbek texts was comprehensively addressed using **Part-of-Speech (POS) tagging** technology [1]. The research findings indicate that identifying the grammatical function of each word via POS tagging enables accurate prediction of punctuation placement. This is particularly crucial for Uzbek, a morphologically rich and agglutinative language.

The analysis yielded the following key conclusions:

1. **There exists a strong syntactic correlation between POS tagging and punctuation restoration**. Once the grammatical role of each word is determined, punctuation marks associated with it can be assigned based on contextual cues.

2. **The agglutinative nature of the Uzbek language** – with its extensive and variable use of suffixes – reduces the effectiveness of simple rule-based or statistical approaches. Therefore, **semantic-rich methods based on POS tagging** prove to be more efficient.

3. **Utilizing transformer-based architectures** significantly improves the accuracy of POS tagging, which, in turn, contributes to more reliable punctuation placement.

4. **Experimental results** confirm that POS-tagging-based punctuation restoration yields substantially higher precision, particularly for **periods, commas, and question marks**, compared to traditional methods – based on F1-score metrics [6]

The following indicated recommendations:

1. **POS tagging should be implemented as a mandatory preprocessing step** in the development of punctuation restoration systems for the Uzbek language.

2. It is essential to **develop large-scale POS-tagged corpora and annotated datasets** for Uzbek, as existing open-source resources are limited and not fully compatible with other Turkic languages.

3. It is recommended to explore **multimodal transformer-based architectures** for punctuation restoration. These architectures allow **joint integration of POS tagging and punctuation prediction** within a single system.

4. For practical deployment, **this technology should be integrated into automatic speech recognition (ASR) systems**, machine translation tools, and intelligent text editors to enhance functionality and linguistic quality.

5. Further research should focus on **adapting POS tagging and punctuation models** for dialectal variations, irregular syntax, and non-standard texts common in **social media and informal communication**.

In conclusion, the use of POS tagging for punctuation restoration in the Uzbek language represents a significant advancement in the field of natural language processing. These research efforts are not only scientifically valuable but also contribute meaningfully to the development of linguistic technologies. This approach plays a critical role in increasing the digital presence of the Uzbek language, developing language-learning tools, and enabling intelligent text analysis systems.

REFERENCES:

[1] D. Jurafsky, J. H. Martin. *Speech and Language Processing* (3rd ed.). Stanford University, 2021. [Online draft].

[2] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[3] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.*, 2017, vol. 5, pp. 135–146.

[4] S. Xolmurodov. Automatic Morphological Analysis Systems in the Uzbek Language: Challenges and Solutions. *Issues of Philology*, 2022, 1(3), pp. 52–61.

[5] G. Nasrullayeva, B. To'rayev. The Relevance of Developing POS Tagging Systems for Agglutinative Languages. *Uzbek Language and Literature*, 2021, 5(2), pp. 33–38.

[6] Tashkent NLP Group. POS Tagging and Text Analysis Based on Transformer Models for the Uzbek Language. *TATU Scientific Bulletin*, 2023, 2(1), pp. 89–97.

[7] J. D. Lafferty, A. McCallum, F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. ICML*, 2001, pp. 282–289.