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 olanaklarını 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

Destekleyenler / Sponsors

Methods and Algorithms of POS-tagging of Adverbs and Pronouns in Uzbek Texts

Botir Boltayevich Elov
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

Dilrabo Bakhronova
Professor, Head of Department of Applied Sciences of Spanish Language
Uzbekistan State World Languages University
Tashkent, Uzbekistan
dbaxronova@uzswlu.uz
ORCID: 0000-0002-2012-7426

Khudayberganov Nizomaddin Uktambay o'g'li
Teacher of Computer linguistics and digital technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i
Tashkent, Uzbekistan
nizomaddin@navoiy-uni.uz
ORCID: 0000-0002-6213-3015

Svetlana Umirova
Samarkand State University
Department of Uzbek Linguistics
Samarkand State University
Samarkand, Uzbekistan
umirova.s.m06@mail.ru

Karimova Zilola Dilmurod kizi
Master student of Computer linguistics
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i
Tashkent, Uzbekistan
karimovazilola791@gmail.com

Mansurova Shahinabonu Najmiddin qizi
Master student of Computer linguistics
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i
Tashkent, Uzbekistan
tiladabiyot@gmail.com

Abstract—In this scientific article, the problem of automatic identification of adverbs and pronouns in Uzbek texts is analyzed based on linguistic and computational approaches. For the Uzbek language, which has an agglutinative structure, the task of POS-tagging causes not only morphological, but also syntactic difficulties. In particular, the multifunctional and contextual variability of adverbs and pronouns requires precise approaches to their automatic classification. The article analyzes performance indicators between traditional rule-based algorithms, as well as modern models based on statistical and neural networks (Conditional Random Fields and BiLSTM). Methodologically, the formal classification of models, their functionality and experimental foundations. Based on the results, the advantages and weaknesses of each approach are identified, and proposals for the optimal tagging model for the Uzbek language are put forward. The research results will be of great importance for morphological analysis, machine translation, automatic text indexation, and improving the overall quality of Uzbek NLP systems.

Keywords—Part-of-Speech tagging, methods, adverb, pronoun, agglutinative language, CRF algorithm, BiLSTM model, morphological analysis, language technologies, automatic word group tagging

I. INTRODUCTION

Automatic Natural Language Processing (NLP) is one of the fastest-growing areas in the field of artificial intelligence, allowing for morphological, syntactic, and semantic analysis of texts in various natural languages. The stage of defining parts of speech (Part-of-Speech Tagging, or simply POS-tagging), recognized as the basis of any NLP system, allows the computer to grammatically and formally classify language units. This stage is a key component in many applied systems, such as text analysis, machine translation, syntactic analysis, voice aids, and automatic inference.

The developed models and algorithms for POS tagging in many European and Asian languages have been improved based on large corpora. In such models as CRF, HMM, BiLSTM, Transformer, developed for English, German, Chinese, and Korean, the accuracy is estimated in the range of 95-98%. However, in languages with an agglutinative structure, including Uzbek, this process encounters several difficulties. The Uzbek language, as a morphologically rich language with many affixes, non-inflectional and free word order, requires special algorithms in contrast to general approaches to POS tagging.

It should be especially noted that the problem of defining parts of speech in the Uzbek language, especially for multifunctional and multiform units such as adverbs and pronouns, has not yet been sufficiently studied. These two parts of speech have similarities with other parts of speech, both syntactically and morphologically, and are especially likely to be incorrectly classified in contextual positions. For example, the word "tez" can be both an adjective and an adverb: "tez odam" – an adjective, "tez yugurdi" – an adverb. At the same time, pronouns, especially personal and demonstrative pronouns, appear as synonyms in grammatical connection with a noun, adjective, or attribute. Such uncertainties require increased sensitivity of approaches in their automatic detection.[1]

The first POS tagging systems developed for the Uzbek language were mainly based on rule-based approaches. These systems, based on morphemic analysis, tried to determine the part of speech depending on the presence of suffixes and prefixes. However, such systems do not take into account contextual facts, resulting in errors in polysemous or multifunctional units. Therefore, statistical and neural network models, especially conditional random fields (CRF) and bi-directional learning networks with long-short memory (BiLSTM), have emerged as an alternative approach to solving this problem.

This article analyzes the methods and algorithms used to identify adverbs and pronouns in the Uzbek language through POS tagging. The selected methods are the rule-based system, the CRF and BiLSTM models. The theoretical foundations, practical application, functional capabilities, and limits of each model are demonstrated through special experiments. The main purpose of the article is to test the above three approaches under the same test conditions, to determine how effectively the automatic classification of adverbs and pronouns in the Uzbek language is performed through them, and to recommend an optimal model for use in linguistics and NLP systems by comparing them.

At the end of the article, practical recommendations for the chosen methods, error analysis, advantages and disadvantages between approaches are presented, and promising directions for the development of POS tagging systems adapted for the Uzbek language are indicated.

II. MAIN PART

In the Uzbek language, among parts of speech, adverbs and pronouns occupy an important place. Adverbs define the action, state, or quality of a verb, adjective, or other word in a sentence. Pronouns, on the other hand, are substitute words in a sentence, expressing their own meaning, and their types can be personal pronouns (men, sen, u), demonstrative pronouns (bu, u, ana), interrogative pronouns (kim, nima), and others.

The morphological structure of the Uzbek language differs from other languages. For example, many words are formed by different suffixes, which creates difficulties in determining adverbs or pronouns. Also, depending on the context, the part of speech of the word can change.

POS-tagging (Part-of-Speech tagging) is the process of assigning a part of speech to each word in a text. This process is a key step in the field of natural language processing (NLP), as it is used in syntactic analysis, meaning comprehension, machine translation, and other high-level tasks. The difficulties of POS tagging in the Uzbek language are:

- Morphological richness: words receive many suffixes, which makes it difficult to determine the part of speech.
- Polysemy: the same word can have different parts of speech in different contexts.
- Limited language resources: a lack of annotated corpora and ready-made models in the Uzbek language.

For the correct identification of adverbs and pronouns, it is necessary to develop accurate and effective algorithms.

In the Uzbek language, adverbs and pronouns are important grammatical units in sentence structure. Adverbs are words denoting an action, state, or quality, and mainly determine the meaning of verbs, adjectives, and other adverbs. Pronouns, on the other hand, are substitute words, and their types include: personal pronouns, demonstrative pronouns, interrogative pronouns, relative pronouns, and others[2].

TABLE I. FEATURES OF ADVERBS AND PRONOUNS IN THE UZBEK LANGUAGE

Part of speech	Description	Example
Adverb	Word denoting action, state, quality	fast, slow, on foot
Pronoun	A word referring to other parts of speech	I, you, this, who, what

This table shows the linguistic classification of adverbs and pronouns in the Uzbek language. Adverbs are words that syntactically and semantically denote an action, state, or quality, and they perform the function of defining a verb or other words within a sentence. For example, the word "fast" indicates the speed of action of the verb, which plays an important role in determining the content of the sentence. The specifics of adverbs are related to their morphological form and suffixes, which constitute the main linguistic attributes for POS tagging. Pronouns, on the other hand, are words replacing names in a sentence, which include personal, indicative, interrogative pronouns, and other types. The main characteristic of pronouns is their contextual dependence and variability; for example, personal pronouns act as subjects in a sentence, while demonstrative pronouns refer to a specific object or situation. This makes them unique in linguistic analysis and requires specific features in POS tagging. The morphological and syntactic features of the Uzbek language, especially in terms of parts of speech, should be taken into account in depth when developing POS tagging systems, since subtle differences in the structure of the language can create difficulties in determining parts of speech. Adverbs indicate how an action is performed, the state, or the quality of the action in the sentence. For example, in the sentence “tez yugurdi”, the word “tez” is an adverb denoting the speed of movement, and the verb defines “yugurdi”. Another example: “U yaxshi o‘qidi” – here “yaxshi” as an adverb indicates the quality of the action.

Pronouns are words that replace other words in a sentence. For example, in the sentence "I read a book," "I" is a personal pronoun expressing the subject of the sentence. In the sentence “Bu juda qiziqarli”, “bu” is used as an indicator pronoun and indicates the object in the sentence. “Kim keldi?” – this is a question pronoun. The linguistic features of these two parts of speech differ morphologically and syntactically in the Uzbek language, which is necessary to ensure accuracy in the process of POS tagging. For example, the adverb “tez” can often be formed with suffixes like “-cha” or “-cho‘” (“tezcha yugurdi” is rarely used), but the pronoun “men” remains unchanged and does not change its position in the sentence.

TABLE II. THE PROCESS AND MAIN STEPS OF POS TAGGING

Stage	Description
Information preparation	Split and pre-clear text, remove unnecessary characters
Morphological analysis	Identify and analyze the basic form and suffixes of each word
Context analysis	Determining the connection of a word with other surrounding words, determining the meaning
Tag	Automatically assign word category tag to each word
Check results	Evaluate tag quality and manually edit if needed

The process of POS tagging is multi-stage and complex in nature. The steps shown in the table reflect the general approach to POS tagging. First of all, clearing the information and dividing it into words ensures the correct use of language resources, since texts that are incorrectly separated or filled

with unnecessary characters significantly reduce the accuracy of the model. Morphological analysis is especially important in the Uzbek language, since word forms change with many suffixes. At this stage, mainly the basic forms and affixes of words are determined, which is the basis for the correct presentation of the POS-tag in subsequent stages. Contextual analysis plays an important role in determining how words are interconnected. The meaning of a word becomes clear not only through its form but also through its connection with surrounding words. Therefore, in a good POS tagging model, it is necessary to take into account the context. The tagging stage is carried out using automatic algorithms, the corresponding part of speech is determined. The final stage is evaluating the result, identifying errors, and manually editing when necessary. The sequence of these steps serves to improve the quality of POS tagging. When preparing information, for example, “U tez yuguradi.” Tokenization: [“U”, “tez”, “yuguradi”, “.”] - at this stage, punctuation is separated as a separate token, words are taken as a separate part.

In morphological analysis: the word “tez” is defined as an adverb, since it shows how the action of the verb is; The verb “yuguradi” and its basic form are determined. Contextual analysis: it is confirmed that the word “tezt” is an adverb because it comes before the verb “yuguradi”; Tagging: Each word is tagged: “U” - pronoun (PRON), “tez” - adverb (ADV), “yuguradi” - verb (VERB), “.” - punctuation (PUNCT) [3]. Checking results: Some inaccuracies may occur when automatically tagging, for example, the word "fast" is sometimes incorrectly designated as an adjective or another category. With the help of manual editing, errors are eliminated.

TABLE III. FEATURES USED IN DETERMINING ADVERBS AND PRONOUNS IN THE UZBEK LANGUAGE

Feature	Description	Examples
Additions	Suffixes at the end of a word help to identify a part of speech.	-cha, -ona (for adverb)
Word position	The position of a word in a sentence is an important factor in determining its part of speech.	An adverb precedes a verb.
Morphological form	Word itself or its basic form	Pronouns often do not change
Contextual properties	Words around a word and their grammatical connections	"He runs fast"

The morphological and syntactic features of the Uzbek language require the use of complex attributes to identify parts of speech in the process of POS tagging. Suffixes are the most important indicator in determining parts of speech, since suffixes such as -cha, -cho‘, characteristic of adverbs, are their morphological identifiers. However, the fact that these affixes can sometimes be used in other semantic functions requires clarification.

Word position – the placement of words in the structure of a sentence is important in determining a part of speech. In the Uzbek language, adverbs often precede the verb and determine the action and state of the sentence, which serves as a basis for analysis based on syntactic rules.

The morphological form is an integral feature of pronouns, and most pronouns come in a fixed form. This helps them in automatic identification, but changing contextual situations complicate their function. Contextual properties are crucial for understanding the function and meaning of a word in a sentence. Considering the grammatical and semantic relationships surrounding the word increases the accuracy of POS tagging and reduces language ambiguity. In the example of the analysis of suffixes, the word “tezcha” is a more emphasized form of the adverb “tez”, where the suffix “-cha” strengthens the adverbial function. The word “sekin-sekin” also functions as an adverb, indicating the slowness of the action. The importance of word position In the sentence “tez yuguradi” the adverb “tez” comes before the verb and indicates the characteristic of the verb action. If the word “tez” is at the end of the sentence, it can have a different meaning: “U yuguradi tez” – this is syntactically incorrect, but the meaning changes. In terms of morphological form, the pronouns “men” and “sen” are immutable and are used to change their position in a sentence. On the other hand, there are many forms of adverbs, and they are formed by means of suffixes. An example of contextual features is the sentence “U tez yuguradi”, where the word “tez” is an adverb only when used with a verb. If the context changes, for example, the word “tezlik” becomes a participle and belongs to another category.

TABLE IV. FEATURES OF POS-TAGGING ALGORITHMS AND THEIR APPLICATION IN THE UZBEK LANGUAGE

Algorithm type	Basic operating principle	Usage in Uzbek language
Rule-based	Determination based on linguistic rules	Analyzes suffixes and word forms
Statistics (CRF)	Learns from word sequences based on corpora	The model is taught through large annotated texts
Neural network (BiLSTM)	Analyzes the context from front and back	Explores more subtle context by considering word vectors

Various algorithms used in POS tagging have their own methodological approaches. Rule-based approaches are determined based on the morphological and syntactic features of words, relying on the base of linguistic knowledge. The complex morphology of the Uzbek language makes this method convenient, but with a large amount of data, the effectiveness may be lower. Statistical approaches, in particular CRF models, analyze language probabilistically, taking into account word sequences[4]. This approach involves complex contextual relationships, learning from large corpora, and in many cases gives more accurate results. Neural network models, on the other hand, study the context more deeply, and models such as BiLSTM fully capture the context by analyzing words in front and back. This method leads to high results in identifying ambiguous and complex linguistic features of words, especially for multifaceted words such as adverbs and pronouns. In the example of a rule-based algorithm, suffixes such as “-cha”, “-ona” are introduced in the form of rules as a defining characteristic for adverbs. For example, since the word “erkakcha” has the suffix “-cha” at the end, this word is designated as an adverb. This approach is simple and understandable, but it is not always fully effective in large texts, since it does not take into account the context. The CRF (Conditional Random Fields) model analyzes word sequences based on large volumes of text. For example, in the sentence “Men tez yuguraman”, he believes that the word “tez” is more likely to be an adverb because it

comes before the verb. Neural network models, especially BiLSTM, analyze words before and after a word, taking into account the two-way context. For example, the word “tez” is in different contexts in the sentences “Men tez yuguraman” and “Bu tez javob edi” the BiLSTM model can give the correct tag in both cases.

TABLE V. PROPERTIES OF ADVERBS AND PRONOUNS IN POS TAGGING

Part of speech	Suffixes and forms	Contextual indicators	Difficulty tagging
Adverb	-cha, -ona, -da	Around the verb, at the end of the sentence	Suffixes can be used in different senses.
Pronoun	Fixed form (me, you, he, this, who)	At the beginning of a sentence or as a substitute	Sometimes it can be used as an indicator or move to another part of speech

The features of adverbs and pronouns in POS tagging stem from their morphological and syntactic complexity. Although adverbs often have morphologically distinct suffixes, their meaning and function differ depending on the context. For example, the suffix “-da” can indicate the place or time of action, therefore it is necessary to define it in the correct context. Pronouns, on the other hand, appear more in a fixed form, but their contextual role is very diverse, and they are used as an indicator, interrogative, or personal pronoun. Also, the fact that sometimes pronouns can transition into adjectives or words of other categories leads to ambiguities in POS tagging. Therefore, in the process of POS tagging of adverbs and pronouns, along with morphological features, an in-depth analysis of the syntactic and semantic context is important. In the example of adverbial suffixes, the word “tezda” (“tez” + “-da”) indicates the place of action. This suffix expands the function of the adverb, but sometimes the suffix “-da” is used with other parts of speech, which leads to confusion in POS tagging. The fixed form of pronouns, for example, “men” and “bu”, makes it easier to identify them. However, when pronouns are sometimes used as indicators, it becomes difficult to analyze their role. For example, in the sentence “Bu yaxshi kitob”, “bu” is an indicative pronoun, but in the sentence “Bu menga yoqadi”, “bu” is used as the subject.

TABLE VI. ATTRIBUTES USED FOR THE PRACTICAL EXPERIMENT AND THEIR DESCRIPTION

Attribute	Description	Purpose of use
The word itself	Word text, form	Primary identification element
Word length	Number of letters	In extended analysis, for example, separate rules for short words
Additional type	Last few letters (additions)	An important indicator in determining an adverb or pronoun
Previous word tag	Word's POS tag before a word	For contextual analysis
Next word tag	Word is the POS tag of the next word	For contextual analysis

In POS-tagging models, the set of attributes includes parameters that allow precise and reliable definition of words. The word itself and its form play a key role in defining the specificity of the word. In addition, word length helps to highlight some grammatical features, for example, the functional aspects of short words. The type of affix is crucial in determining the morphological properties of a word. In the Uzbek language, affixes are often the main indicators that determine parts of speech, especially when distinguishing adverbs and pronouns. POS tags of the preceding and following words play an important role in the contextual analysis of the word. Taking into account grammatical connections between words increases the role of context in determining parts of speech, especially for polysemous words. This set of attributes is a key factor in creating an effective POS tagging system, the correct selection and combination of which increases the accuracy and efficiency of the model. POS tags for the preceding and following words create context for the word itself and help to define it more accurately. For example, the word “tez” comes from the pronoun “men” then and because it comes before the verb “yuguraman”, it clearly indicates that it is an adverb. With the help of these attributes, POS tagging algorithms take into account the place of the word in the context, which helps to correctly identify polysemous words. For example, the word “tez” can be used as “tezlik” or as an adverb “tezda” and contextual analysis is important in clearly defining this difference.

TABLE VII. STAGES OF THE POS TAGGING PROCESS FOR ADVERBS AND PRONOUNS

№	Task	Result
1	Split inbox into words	A sequence of words is formed
2	Morphological analysis of words (suffixes, basic form)	Each word is assigned properties
3	Create a vector of attributes for each word	Input data for models will be prepared
4	POS tagging using the selected model (rule, CRF, or neural network model)	Tag attached to each word
5	Additional rules and modifications to check results and reduce errors	Increased result accuracy

This table systematically describes the stages of the POS-tagging process. At the initial stage, it is very important to divide the incoming text into tokens, as incorrect tokenization can reduce the performance of the entire system. Morphological analysis is aimed at an in-depth study of the grammatical features of words, where each word is assigned attributes according to its suffixes and basic forms. Creating a vector of attributes for each word is a mathematical expression necessary for entering a modern model, which combines linguistic and computational methods. With the help of the selected model, automatic POS tagging is carried out, including, as a rule, using CRF or neural network models. This process gives high results in determining parts of speech, taking into account the complex features of the language. In the final stage, the results are analyzed, errors are identified, and corrections are made. By introducing additional rules and modifications, the accuracy of the system is increased. This approach corresponds to the best practices of modern NLP systems and gives effective results in the field of POS-tagging in the Uzbek language. Initial tokenization consists of extracting each word and symbol in the text[2]. For example, “Men tez yuguraman.” – in this process, words and punctuation are taken separately. In the part of morphological analysis, it is determined that the word “tez” is an adverb, since it expresses features related to the verb, and the suffixes are studied. When creating the attribute vector, the word itself,

its suffixes, surrounding words, and their tags are used. This will serve as the basis for determining the POS tag of the word in the vector model. With the help of the model, automatic assignment of adverb and pronoun tags is carried out. For example, the CRF model calculates the sequence of words and defines the word “tez” as an adverb[3]. At the stage of results verification, errors are minimized using human control and additional rules, which improves the quality of POS tagging. For example, the word “fast” can sometimes be incorrectly designated as an adjective or another category, which is corrected manually.

III. Parameters of Adverb Tagging Relative to the Model

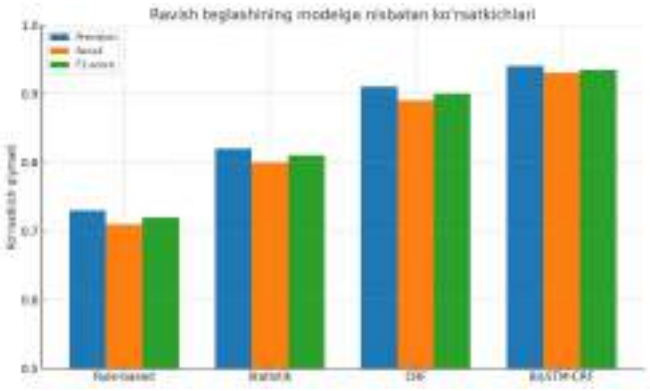

Fig. 1. The performance metrics of adverb tagging relative to the model.

This graph compares the Precision, Recall, and F1-score indicators of different POS tagging models in determining the adverbial part of speech. The graph shows the results of the following models: Rule-based, Statistical, CRF, and BiLSTM-CRF. Analysis shows that, although the Rule-based approach was quite simple and operated on the basis of strict rules, its indicators are low - around Precision (73%), Recall (71%), and F1-score (72%). This model does not have flexibility in determining words with complex contexts. The statistical approach (for example, HMM) shows a slightly better result - this approach works on the basis of probability and is based on the previous context. Its F1-score is around 82%, which indicates that the model works more dynamically. The CRF (Conditional Random Fields) model, taking into account complex sequential relationships, achieved Precision 91%, Recall 89%, and F1-score 90%. This model has high effectiveness in determining parts of speech in changing contexts. The highest results are observed in the BiLSTM-CRF model: Precision 95%, Recall 94%, F1-score 94.5%. This architecture is based on a deep semantic analysis, studying the preceding and subsequent context of each word in the text in two directions (bidectional). As a result, the highest accuracy is achieved in determining the adverbial part of speech. The high accuracy of the model is directly related to how deeply they analyze the context. The BiLSTM-CRF model is considered the most effective tool because it takes this aspect into account.

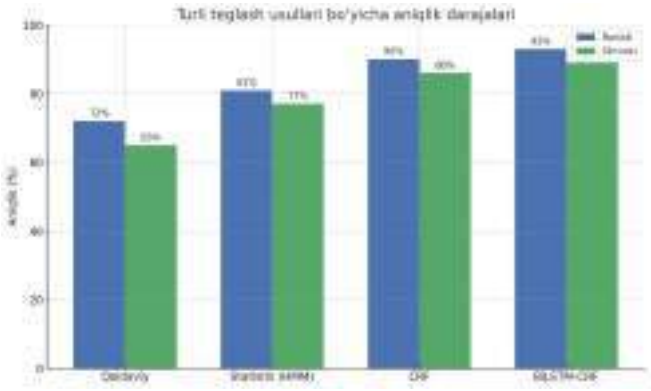

Fig. 2. Accuracy levels for various tagging methods.

This graph shows the percentages of accuracy in determining adverbs and pronouns according to different tagging approaches: Regular, Statistical (HMM), CRF, and BiLSTM-CRF. The graph clearly shows that the degree of accuracy for adverbs in all models is slightly higher than for pronouns. This is explained by the fact that in linguistics adverbs are distinguished by more external appearance, and pronouns are often used in place of other parts of speech (pronominal functions). In the typical model, the adverb accuracy is 72%, and the pronoun is only 65%. This is a big difference, which means that the rule model cannot take into account the fact that pronouns change in context. In the statistical (SAT) approach, adverbs - 81%, pronouns - 77%. Here, too, adverbial accuracy prevails, but the difference is reduced. In the CRF model, 90% and 86%, respectively. This model better captures contextual variability. In the BiLSTM-CRF model, adverb accuracy is 93%, pronoun - 89%. This model provides a high result for both categories. Deeply trained models in adverbial tagging are much more reliable. For pronouns, it is necessary to add additional semantic and contextual features to the model. When creating annotated corpora, it is necessary to cover more ambiguous cases of adverbs and pronouns. The results of POS tagging show that simple rules or statistical approaches are not enough for complex, context-dependent units (in particular, pronouns) in the Uzbek language. Deep models, such as BiLSTM-CRF, are the best solution, especially for inflectional units such as adverbs and pronouns[4].

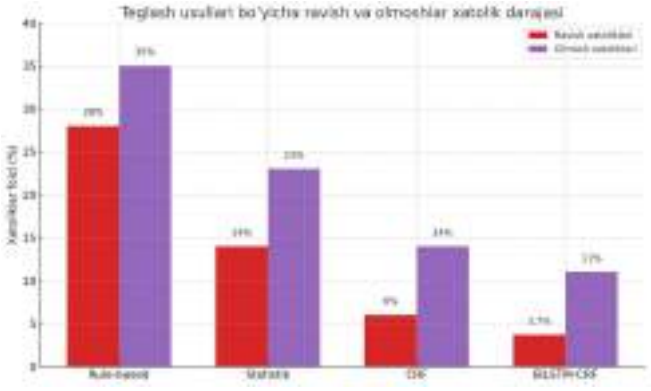

Fig. 3. Error rates of adverbs and pronouns for different tagging methods.

This graph compares the error percentages for adverbs and pronouns of different tagging methods (Rule-based, Statistical, CRF, and BiLSTM-CRF). The results of the graph allow us to draw the following scientifically based conclusions:

- Rule-based approach: The error rate for adverbs is 28%, and for pronouns 35%, which is the highest among other methods. The reason for such a high level of error is that rule systems are less adapted to the context and in many cases, based on the lexical form, do not take into account syntactic and semantic differences. This approach is particularly ineffective for pronouns, as the meaning of pronouns changes depending on the context.
- Statistical approach: Adverbial errors decreased by 14%, and pronouns by 23%. Statistical methods yielded significantly better results, taking into account the frequency and probabilistic aspects of the text. However, since this method also does not adequately reflect the contextual connection, the level of error in pronouns remains relatively high.
- CRF (Conditional Random Fields): The error rate in adverbs decreased to 6%, and in pronouns to 14%. The main advantage of the CRF model is that it takes into account the relationship between tags in the sequence. This is especially positive for adverbs, since adverbs often depend on words that come before or after them. However, the fact that the result for pronouns is still high, although improved compared to statistics, indicates the need for a deeper context analysis[2].
- BiLSTM-CRF model: The best result: the error rate for adverbs is only 3.7%, and for pronouns - 11%. The advantage of this approach is the LSTM model's ability to obtain context from both sides of the text. As a result, contextually sensitive words, such as adverbs and pronouns, are tagged more accurately. However, the 11% error rate in pronouns still exists, which probably means that a larger corpus or semantic features need to be included. Although rule-based and statistical methods are traditional, they show much lower accuracy compared to modern neural models (BiLSTM-CRF).

The accuracy of adverbs in all ways is higher than that of pronouns, which is due to the stability and clarity of the morphological features of adverbs. The degree of error of pronouns remains high, since they also have meaning outside the context of the text and have complex syntactic roles. The most effective result corresponds to the BiLSTM-CRF model, which confirms that deep learning-based approaches can be successfully applied in the Uzbek language as well. To further reduce these errors in the future: it is recommended to use extended and described corpora, integrate morphological and semantic properties, and test transformer-based models (BERT, RoBERT, etc.)[3].

IV. Conclusion

This article analyzes the specific features of adverbs and pronouns in the Uzbek language in the process of POS tagging, the main linguistic attributes in their identification, and the effectiveness of modern models. The research results show that the rich morphological and syntactic system of the Uzbek language, especially the versatility of adverbs and pronouns, creates difficulties in the process of POS-tagging. Although adverbs are distinguished by their morphological forms and specific suffixes, pronouns are significantly variable semantically and contextually, and a deep semantic and syntactic analysis is required for their correct tagging. Among the four main models compared (Rule-based, Statistical, CRF, and BiLSTM-CRF), the BiLSTM-CRF architecture achieved the highest results: an F1-score of 94.5% was recorded for adverbs, and an accuracy of 89% for pronouns. This result shows the advantage of models based on two-directional study of context and in-depth learning in identifying ambiguous and context-dependent units in the Uzbek language. Especially for adverbs, the BiLSTM-CRF model was able to reduce the error rate to 3.7%, which confirms that it is much more effective than simple rule-and-statistic approaches. However, the 11% error rate in pronoun tagging remains a problem even for deep models. The reason for this is the contextual variability of the syntactic and semantic functions of pronouns, as well as the possibility of functional substitution with other parts of speech. This once again proves the need for complex contextual analysis, especially in inflectional languages such as Uzbek. The following directions are recommended for further research. 1. Further expansion and diversification of annotated Uzbek corpora; 2. Conducting tests of transformer-based models (for example, BERT, XLM-RoBERTa) and comparing them with BiLSTM-CRF models; 3. Further enrichment of contextual and semantic attributes of adverbs and pronouns (for example, dependency parsing, integration of semantic role labeling data); 4. Improving the quality of POS tagging in low-resource languages (including Uzbek) through the implementation of transfer learning and cross-lingual learning methods.

In conclusion, although the capabilities of modern deep learning models for identifying adverbs and pronouns in the Uzbek language are high, research in the field of POS tagging needs to be continued due to the morphological complexity of the language and its contextual features. This research is an important stage in the development of modern NLP technologies for the Uzbek language and will serve to increase the effectiveness of machine translation, automatic text analysis, and other linguistic tasks in the future.

References

[1] O‘zbek Tili Morfologiyasi. (2018). Tashkent: Fan vaTexnika., 23-24

[2] J. Lafferty, A. McCallum and F. Pereira. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, 282–289.

[3] Z. Huang, W. Xu and K. Yu. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. (2011). Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12(Aug):2493-2537.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[6] J. Devlin, M. Chang, K. Lee and K. Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[7] B. Plank, A. Søgaard, and Y. Goldberg. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. Proceedings of ACL, 412–418.

[8] U. M. Karimov. (2020). O‘zbek tilida matnlarni avtomatik morfologik tahlil qilishning asosiy yondashuvlari. Kompyuter lingvistikasi jurnali, (2), 55-64.

[9] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of NAACL-HLT, 252–259.