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

Process of Automatic NER Identification in the Uzbek Language Corpus

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

Samatboyeva Madina To'lqinjon qizi
Department of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
samatboyevamadina@navoiy-uni.uz

Kilichev Bayramali Ergashovich
Department of Uzbek Linguistics and Journalism
Bukhara State University
Bukhara, Uzbekistan
b.e.kilichev@buxdu.uz

Ishonkhanova Iroda Abduvakhitovna
Dept of Uzbek language and literature education
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
irodaishonxanova@gmail.com

Kuldasheva Sanobar Khayotovna
Dep of Uzbek language and literature education
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan
kuldasheva1777@gmail.com

***Abstract* — This article discusses the issues of Named Entity Recognition (NER) and their interpretation in the Uzbek language corpus. Based on the morphological and syntactic rules of Uzbek language grammar, a lexical database (in Uzbek) has been developed, and a rule-based system has been formulated for the automatic identification of Named Entities (NERs) within texts. The article classifies named entities (such as personal names, place names, organization names, etc.) for NER models based on the Uzbek language corpus and describes the methodology used for their interpretation. Additionally, the process of identifying NER entities is presented through specialized diagrams, and the entities themselves are illustrated in tables.**

Keywords: NER, corpus linguistics, natural language processing, NLP, dictionary, anthroponym, toponym.

I. Introduction

In NLP, the concept of NER has been thoroughly studied, and proper nouns can be recognized as NER entities. Existing models, programs, and ready-made libraries available online, developed abroad, were examined. Specifically, the concept of NER was adapted for the Uzbek language and its grammar.

In this context, proper nouns in the Uzbek language can be considered as NER objects. This is because Named Entity Recognition (NER) refers to the re-identification of names. For example, the lexical item "odam" (person) semantically refers to an animate being and is understood as a common noun.

However, when the unit "odam" is renamed as "Alisher", a specific personal name is assigned to it, thus transforming it into a named entity. In this case, the "renamed word" represents a proper noun. Proper nouns, by their nature, consist of lexical items that are not typically found in standard dictionaries. One such dictionary is the Explanatory Dictionary of the Uzbek Language, which contains more than 80,000 entries but generally does not include personal names or place names. Therefore, for Uzbek-language texts, proper nouns can appropriately be treated as NER objects.

In foreign programs based on English, not all units recognized as entities are specific to Uzbek. For example, seasons and colors. However, not all proper nouns in Uzbek can fully form a complete list of NER entities for a program. Based on studying existing and sufficiently functional programs, we have compiled a list of NER entities specific not only to Uzbek but also to English. This list consists of 14 major categories, each combining smaller entities (e.g., the Location category alone includes nearly 30 entities). Thus, NER entities form a very large and broad category. Our research focuses on the three largest categories: anthroponyms (Person), toponyms (Location), and organization names (Organization).

II. Literature Review

NER is one of the most important tasks in NLP. Research on NER can be divided into periods: the dominance of statistical models (CRF, HMM) from 2000–2010, the combination of LSTM, BiLSTM, and CRF from 2015–2020, and the dominance of Transformer-based models (BERT, RoBERTa, XLM-R) post-2020.

Significant research on NER was conducted by Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer in their paper "Neural Architectures for Named Entity Recognition," published in 2016 in the proceedings of the NAACL conference in San Diego. This paper was one of the first foundational studies demonstrating the practical superiority of deep learning-based NER models. The LSTM+CRF combination proved highly effective for NER, and Lample's team scientifically demonstrated the differences between traditional and neural approaches. All research results were presented with high accuracy using diagrams and mathematical models [1].

Another prominent figure in NER and NLP is Dan Jurafsky, who contributed to the development of the Stanford NER system. Jurafsky is a pioneer in NLP, and his work

"Speech and Language Processing" (co-authored with James H. Martin) is notable for its broad coverage. His research is deeply theoretical yet practical. Classical NER methods, such as Conditional Random Fields (CRF), became widely adopted due to Jurafsky's work [2].

Christopher Manning, the creator of the Stanford Named Entity Recognizer, is another key figure. His work primarily relies on statistical and deep learning models. The Stanford NER system is widely used in both academic and industrial settings, with his CRF-based model becoming a standard NER tool in the late 2000s and early 2010s. His most famous works include "Foundations of Statistical Natural Language Processing" and "Introduction to Information Retrieval" [3].

III. Main Section

NER is an NLP method that identifies key entities in text and categorizes them into predefined classes. The process of identifying named entities such as personal names, locations, company names, and other out-of-vocabulary items is a crucial step in solving many NLP tasks. In NLP, named entity recognition is also referred to as entity identification, entity extraction, or chunking.

For the Uzbek language, the concept of NER has been thoroughly studied, and proper nouns are recognized as NER entities. Foreign-developed models, programs, and ready-made libraries available online were examined, and the NER concept was adapted for Uzbek language and grammar. Proper nouns in Uzbek can be accepted as NER entities because NER is the "process of re-identifying names," where a "renamed word" represents a proper noun. Proper nouns are lexical items outside explanatory dictionaries. For Uzbek texts, proper nouns can be recognized as NER entities.

The process of identifying NER objects in the Uzbek language can be illustrated through Figure 1:

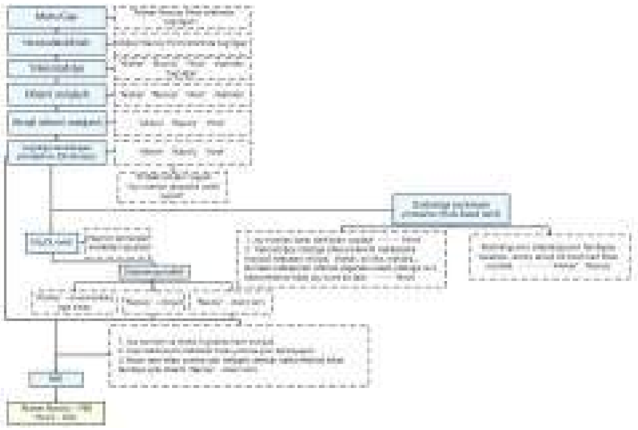

Fig. 1 Automatic Identification of NER Entities

The process of identifying NER objects in the Uzbek language is as follows:

– **Text Selection:** Identifying Named Entity Recognition (NER) objects within a text is a complex task (see Figure 1). The identification process includes the following stages: (1) selection of a text or sentence. The selected texts are automatically collected for the corpus. The texts within the corpus may consist of newspaper articles, literary works, online publications, and similar sources. Before identifying NER objects, a second step – text cleaning – is carried out.

– **Normalization Process:** This is the stage in the NLP process where the text is brought into a single, consistent form, aiming to reduce variability in the text and make it convenient and standardized for analysis. This stage eliminates unnecessary variations in the text and also simplifies the analysis of words and sentences. This stage is particularly important in cases where there are various written forms, misspelled words, abbreviations, and synonyms. Additionally, during this process, the text and sentences are cleaned of unnecessary elements (Removing extra spaces and empty lines – extra spaces and blank lines, Cleaning special characters (excluding certain necessary ones) – special characters; Removing URLs, emails, and HTML tags – internet addresses and tags) [4].

The benefits of normalization: the accuracy of models, reduces unnecessary variability in the data, and prepares a solid foundation for stages such as tokenization, vectorization, NER, and sentiment analysis.

In the provided example, punctuation marks and unnecessary units in the sentence were removed during the normalization process. At the same time, normalization is also important for tokenization and lemmatization.

– **Tokenization**: One of the key roles in understanding human language is performed by tokenizers. Each type of tokenizer helps machines process text in a different way [5]. Tokenizers can be developed based on various approaches depending on the given problem. Tokenization refers to transforming unstructured natural language text into a more structured form (from a computer's perspective) during the initial processing stage [6]. In the tokenization process, the text is divided into parts called tokens, which contain useful information [7]. In this sentence, the boundaries of the tokens are determined, and the text is divided into segments (tokens).

– **Part-of-Speech (POS) Tagging Process**: POS tagging is the task of assigning each word form in a given sentence to its grammatical category (noun, verb, adjective, numeral, adverb, or pronoun). POS tagging is one of the fundamental tasks in Natural Language Processing (NLP) and is an important stage in the pipeline conveyor. It is a crucial step for NLP applications such as machine translation, text summarization, question-answering, and sentiment analysis [8]. POS tagging can be performed with or without a dictionary. Most scientific research on POS tagging [9], [10], [11] is word-based and does not implement morphological segmentation of words. Based on POS tagging, the nouns in the sentence are separated (Alisher, Navoiy, Hirot, shahrida). The word "tug'ilgan" belongs to the verb category, so it is not accepted in the next stage.

– **Proper Noun Identification Process**: Here, the basic rules specific to proper nouns ("Proper nouns are written with a capital letter," "A person's first name, father's name, surname, nickname, or symbolic proper noun begins with a capital letter," "Place names begin with a capital letter") [12] are checked. As a result, three lexemes written with capital letters in this sentence (Alisher, Navoiy, Hirotda) are extracted.

A. *NER Identification Methods – Primary Process*

This is the first stage based on the four main methods (Dictionary-based, Rule-based, Machine learning, Deep learning). The dictionary-based approach is a method based on a pre-compiled dictionary or word list in language processing or text analysis processes. The dictionary-based approach is also known as the “Lesk method” (introduced by Michael Lesk in 1986). The Lesk algorithm-based Lesk method is defined as “measuring the similarity between the definitions of all words in the text.” However, in 2000, Kilgarriff and Rosensweig simplified the Lesk definition as “a measure of similarity between a word's definitions and the current text,” which simultaneously means determining the correct meaning for a single word [13]. In this approach, each word or phrase is linked to specific information in the dictionary, and based on this, analysis or identification tasks are performed. This approach is applied in many fields.

TABLE I. FIELDS WHERE THE DICTIONARY-BASED APPROACH IS APPLIED

Dictionary-Based Approach	Text Analysis
	Language Identification
	Part-of-Speech Tagging
	Spam Detection
	Emotion Detection (Sentiment Analysis)

The dictionary-based approach is an understandable and simple working method. Additionally, it allows obtaining precise results based on specific words or phrases and enables manual or automatic expansion of the dictionary. However, this approach may not be a fully perfect solution for certain programs. For example, the dictionary must be constantly updated. Ambiguous or new words may not be identified. Ambiguities may also arise with homonymous words.

In the analyzed sentence, the words “Alisher,” “Navoiy,” and “Hirot” are searched in dictionaries using this approach. In reality, NERs represent “out-of-vocabulary” lexemes – this refers to the “Explanatory Dictionary of the Uzbek Language.” That is, the absence of these words in the dictionary is only relevant for the explanatory dictionary. However, in Uzbek, dictionaries of anthroponyms (personal names) and toponyms (place names) have been compiled within the field of onomastics. The proper nouns extracted from the sentence exist in the “Dictionary of Uzbek Names” [14] and the “Brief Explanatory Dictionary of Place Names” [15].

– **Rule-based Approach – Analysis Process:** This process is based on a set of predefined rules. Rule-based text data analysis and processing is one of the earliest NLP methods that applies predefined linguistic rules. The rule-based approach involves using specific rules or models to achieve certain structures, extract information, or classify text. Some common rule-based methods include regular expressions that conform to philological rules.

In NLP, the rule-based approach is applied through several steps:

- **Rule Creation**: Here, specific linguistic rules related to grammar, syntax, and semantics required for the models are created.
- **Rule Application**: Predefined rules are applied to the input data to obtain matching results.
- **Rule Processing**: Text data is processed according to the results of the rules for information extraction, decision-making, or other tasks.
- **Rule Improvement**: Created rules are refined through iterative processing to improve accuracy and performance. Rules are modified and updated as needed based on previous considerations [16]. This method is based on philological rules (lexicology, morphology, syntax), and models are also formed based on these rules.

In this approach, there is a special language template created by philologists. That is, it uses a template like “if this... is so, then... will be so.” Additionally, rules can be predefined, such as “if words like 'viloyat' (region), 'shahar' (city), 'qishloq' (village), 'tuman' (district) are present in the sentence, it is highly likely that the words around them may be toponyms.”

This approach is also not fully effective because a rule must be written for each case, which takes a lot of time. It is not suitable for large or variable datasets and may not work in unexpected or new scenarios.

For the analyzed sentence, based on the philological language rule that “place names are written with a capital letter,” the word “Hirot” is identified as a place name. Additionally, its (Hirot) classification, i.e., that it is a place name, is confirmed by the indicator (“shahrida” – “in the city of”) appearing next to it. This also confirms that it is a toponym – a place name. Based on the rule that “a person's name is written with a capital letter,” the words “Alisher” and “Navoiy” are accepted as anthroponyms – personal names.

– **Machine Learning (ML) and Deep Learning (DL) Approaches**: This analysis process is based on Machine Learning and Deep Learning approaches, where the sentence under analysis is identified using neural network models, specialized algorithms, and pre-constructed language models designed specifically for the system. In this case, special language libraries on the internet (SpaCY, GATE, OpenNLP, CoreNLP, NLTK, CogcompNLP) can be used.

TABLE II. ML METHODS

№	ML Methods	Definition
1	Conditional Random Fields (CRF)	One of the most popular ML methods. Predicts the sequence of labels based on words and their features in the text. Special features are manually selected: word type, starts with a capital letter, previous/next words, etc.
2	Hidden Markov Models (HMM)	A good statistical model for analysis. Used less now but historically played an important role.
3	Support Vector Machines (SVM)	Classifies each token into classes based on dependent features. Does not fully consider the sequence compared to CRF and HMM.
4	Decision Trees, Naive Bayes, k-NN	Simpler models. Weaker in analyzing large contexts but convenient for learning.

TABLE III. DL METHODS

№	DL Methods	Definition
1	BiLSTM (Bidirectional LSTM)	Learns the preceding and following context of words in the text. Often used with CRF: BiLSTM + CRF

2	CNN (Convolutional Neural Networks)	Used to identify morphological features of words (e.g., vowels, suffixes). Often used with other models.
3	Transformers (BERT, RoBERTa, XLM-R, GPT, etc.)	The most advanced and efficient approaches. Deeply understands context, works on pre-trained models. Examples: BERT for NER, XLM-R for multilingual NER, BioBERT – for medical texts, FinBERT – for financial texts.
4	Pretrained Language Models + Fine-tuning	Retraining a pre-trained model (e.g., BERT) on a specific NER dataset. Yields very good results, especially with large corpora.

The goal of NER is to identify named entities in the text and assign them to appropriate categories. Three main approaches are important for NER: dictionary-based, rule-based, and machine learning-based. However, an NER system can combine several of these categories [17]. The ML method is almost the same as the DL method, but there are significant differences in performance efficiency. ML works better with small datasets, while DL requires large datasets. However, in terms of result accuracy, DL provides efficient and precise results.

– **Homonymy Analysis**: After completing the checks according to NLP methods, the analysis process for the biggest problem in NER identification – homonymy (homography) – is performed. Correctly analyzing homonymy is crucial for assigning the right category to the identified NER entities. The word “Alisher” does not have homonyms. However, the word “Navoiy” is homonymous: in one meaning, it refers to a person, while in another, it refers to a place (Navoiy region). The next step is to clarify in which sense this word is used in the text – as a personal name or a place name. For this, a pre-prepared database, including dictionary databases, philological rules, and models based on them, is required.

The dictionary database is the first to operate and shows that the word exists in both meanings (in the dictionaries of personal names and place names). The model that checks with indicators then checks whether there are any indicator words around it. In this case, there are no indicators around the word “Navoiy” that would refer to a place name. Additionally, the word “Navoiy” appears alongside an anthroponym. Therefore, it is highly likely that the model will recognize this lexeme as a personal name.

The final stages (10-11) involve demonstrating that the identified units are indeed NER and correctly expressing their categories. The result shows that "Alisher Navoiy" is PER (person), and “Hirot” is LOC (location).

In the Uzbek language corpus, the sentence “Biz bugun Abdulla akan bilan birga Navoiy viloyati Bahor qishlog‘iga jo‘nadik” contains several proper nouns – NER entities. The system automatically identifies them. First, the sentence is tokenized, and then all words are POS-tagged. During the morphological process, the identified nouns are separated, and NER is extracted. As a result, the words “Abdulla,” “Navoiy viloyati,” and “Bahor qishlog‘I” in the sentence are shown to be nouns and NER entities. This process is performed in the “Morphological Analysis" section of the corpus. (See: Figure 2)

Fig. 2 Identification of NER in the Uzbek Language Corpus

If we pay attention to all the NER entities identified in the statistical section of the analyzed text in the corpus, we can see that only their word categories are observed. The table (See: Figure 3) does not show the lemma or root part. Because the lemma is free from all formative suffixes, i.e., it is the dictionary form of the word. Lemma is the name of the lexeme in corpus linguistics. Returning to the above thought, the absence of the lemma and root form in the table indicates that they are out-of-vocabulary lexemes. This is a characteristic feature of NER entities.

Fig. 3 Statistics of Analyzed NER in the Uzbek Language Corpus

Let us test the operation of this analysis process (in Run state) using the Python programming language (See: Figure 3). The proposed method currently only checks based on capitalization rules. More complete solutions are implemented based on texts of various forms and contents in the corpus, and the database is also expanded.

Fig. 4 Testing NER in Python Programming Language

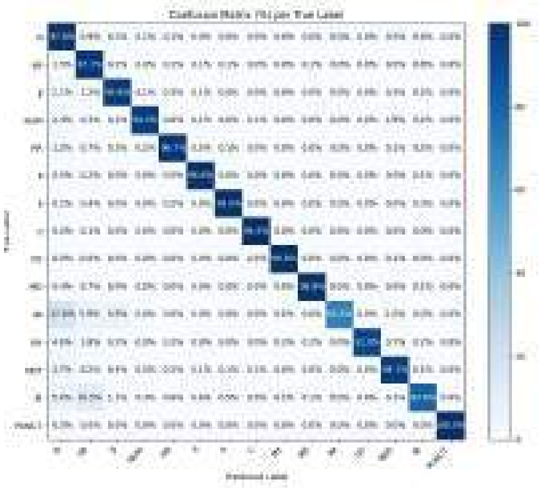

Fig. 5 Automatic NER Analysis in Uzbek Language Corpus

– **Correct Identification (Recall / Sensitivity):** NER → NER: The model correctly identifies named entities in 94.7% of cases. This indicates that the NER module can be effectively used in practical NLP applications (e.g., document analysis, chatbots, information retrieval systems).

– **Confusion Analysis:**

NER → N (noun): 3.7%

(NER → N = 3.7% → This means that 3.7% of NER entities were incorrectly labeled as “nouns.”)

NER → JJ (adjective): 0.4%

NER → RR (pronoun): 0.5%

In these cases, the NER class is sometimes confused with common nouns (N), adjectives (JJ), or pronouns (RR) — showing that distinguishing NER from general nouns or grammatical units can be challenging. This issue mainly arises due to the presence of homonymous units in the Uzbek language.

IV. Conclusion

This article analyzes the process of automatically identifying named entities (NER — Named Entity Recognition) in the Uzbek language corpus. Examples were analyzed based on the Uzbek language corpus, and the results were verified using programming languages. Dictionary-based and rule-based approaches were studied with examples. Neural network models were classified and analyzed. These works can serve as a basis for creating new models. Additionally, considering language-specific lexical and grammatical peculiarities significantly impacts the quality of the model. In the future, expanding the corpus, multi-level tagging, and working with multimodal data can improve the system's efficiency. This research is an important step in digitizing the Uzbek language and integrating it with artificial intelligence technologies.

References

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. NAACL, 2016. [Online]. Available: https://doi.org/10.18653/v1/N16-1030

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, draft of January 12, 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/cd3book.pdf

[3] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in Proc. 43rd Annual Meeting of the ACL, 2005, pp. 363–370.

[4] B. Elov and M. Samatboyeva, "General rules for NER tagging in the Uzbek language corpus," News of the National University of Uzbekistan, Tashkent, 2025, [1/3/1]. ISSN: 2181-7324.

[5] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in Proc. ACL, 1992. [Online]. Available: https://doi.org/10.3115/992424.992434

[6] A. Rai and S. Borah, "Study of various methods for tokenization," in Lecture Notes in Networks and Systems, vol. 137, 2021. [Online]. Available: https://doi.org/10.1007/978-981-15-6198-6_18

[7] Z. Khusainova, "Implementing tokenization based on the BPE algorithm," News of the National University of Uzbekistan, Tashkent, 2023.

[8] B. Elov, Z. Khusainova, O. Abdullayeva, and N. Khudayberganov, "POS tagging and stemming for agglutinative languages (examples of Turkish, Uyghur, and Uzbek)," Uzbekistan. Language and Culture, vol. 2, Tashkent, 2023. ISSN: 2181-922X.

[9] J. Gao and M. Johnson, "A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers," in Proc. EMNLP, 2008. [Online]. Available: https://doi.org/10.3115/1613715.1613761

[10] S. Goldwater and T. L. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in Proc. 45th Annual Meeting of the ACL, 2007.

[11] J. van Gael, A. Vlachos, and Z. Ghahramani, "The infinite HMM for unsupervised POS tagging," in Proc. EMNLP, 2009. [Online]. Available: https://doi.org/10.3115/1699571.1699601

[12] Explanatory Dictionary of the Uzbek Language, Tashkent: Kafolat Print Company, 2023. ISBN: 978-9943-389-19-9.

[13] "Word sense disambiguation in NLP," [Online]. Available: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_word_sense_disambiguation.htm

[14] E. Begmatov, Uzbek Names, Tashkent: National Encyclopedia of Uzbekistan, 1998.

[15] Z. Dosimov and Kh. Egamov, Brief Explanatory Dictionary of Place Names, Tashkent, 1977.

[16] "Rule-based approach in NLP," [Online]. Available: https://www.geeksforgeeks.org/rule-based-approach-in-nlp/

[17] N. Keretna, "NER systems. Open source libraries," 2014.