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

İTÜ

1 9 9 6
ATILIM
ÜNİVERSİTESİ

A Hybrid Named-Entity Recognition Algorithm for Ecological Documentation in Uzbekistan

Nilufar Abdurakhmonova
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0001-9195-5723

Davlatyor Mengliev
Urgench State University
Urgench, Uzbekistan
0000-0003-3969-1710

Sarvinoz Mardonova
Bukhara State University
Bukhara, Uzbekistan
0009-0007-1346-0451

Tulkin Asadov
Bukhara State University
Bukhara, Uzbekistan
0009-0007-9975-6236

Farxod Xalilov
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0009-0007-9298-4455

Fayzilat Maxsudova
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0009-0008-9830-9538

***Abstract*—The rapid expansion of environmental documentation in Uzbekistan, from parliamentary minutes to environmental impact assessment reports, requires automated tools to quickly find key actors, territories, and institutions. However, existing named entity extraction systems struggle with the dual script (Cyrillic ↔ Latin) and agglutinative morphology of the Uzbek language. This paper proposes a hybrid NER algorithm that combines two deterministic preprocessing modules with the SpaCy statistical model. The orthographic standardization module converts Cyrillic input to unified Latin, normalizing apostrophes and specific letters. The morphological corrector performs lemmatization and corrects typical typos, eliminating suffix "noise" that masks entity boundaries. A corpus of 5,000 environmental sentences, marked up according to the BIOES scheme for the PER, ORG, LOC classes, was compiled for training. Comparative testing shows that the hybrid approach increases the average F_1-measure by 4–7% relative to the basic SpaCy model. The developed algorithm takes a step towards creating a reliable infrastructure for analyzing environmental data in the Uzbek language and supports decision-making in the field of sustainable development of the country.**

Keywords*—rule-based algorithm, Uzbek language, Turkic languages, named entity recognition, natural language processing.

I. Introduction

Today, the environmental sector of the Republic of Uzbekistan is actively developing, in particular, 2025 has been declared as the "Year of Environmental Protection and the Green Economy"[1]. In 2022, the government adopted a special Green Growth Strategy, which sets a 2030 horizon for increasing energy efficiency, reducing fossil fuel subsidies and scaling up renewable energy, all while tracking progress using an OECD-style indicator framework[2].

It is important to note that these commitments are not just ambitious, they respond to pressing environmental issues such as land degradation around the Aral Sea, recurring dust storms and rising levels of urban air pollution, the health damage from which is already estimated at 6.5% of GDP[3].

Moreover, political and institutional reforms have been carried out in parallel with these policy changes. In particular, the Ecological Movement of Uzbekistan, created in 2008, was transformed into the Ecological Party of Uzbekistan (O'zEP) in 2019[4]. At the same time, 15 reserved seats in the Legislative Chamber gave rise to a permanent eco-faction, which now represents green priorities in parliament[5].

However, effective policy implementation depends on timely access to reliable information. Ministries, non-profit NGOs and private sector entities generate thousands of pages of environmental impact assessments, climate risk assessments and green finance reports every year[6]. Manually reviewing this text stream imposes certain difficulties in the form of time spent, which entails certain costs in this direction[7]. Consequently, key actors - government agencies, industrial polluters, protected areas - remain unindexed, which complicates evidence-based decision-making. This study proposes a solution to this problem, namely, the development of a hybrid solution based on a rule-oriented algorithm and neural networks.

To implement such a solution, the authors examined existing solutions that solve similar problems for the Uzbek language, information on which is contained in the third section of this article. Moreover, the solution itself and the results of its testing are included in sections 4 and 5. In addition, the conclusion and development prospects of the proposed algorithm are described in section 6.

II. Morphology of Uzbek Language

Today, the modern Uzbek language has a Latin alphabet, which was officially established during the years of our country's independence[8]. Despite this, in informal communication, it can be seen how interlocutors actively use the Cyrillic alphabet[9]. In addition, in the appeals of citizens and even legal entities, one can see how the text of the appeal or application is written in the Cyrillic alphabet[10]. However, official documents, including legislative acts and regulatory documents (including those related to ecology), are published in Latin[11]. Speaking about the nature of the language, it can be said that the Uzbek language is an agglutinative language, which implies the formation of word forms by attaching affixes to the root of the word[12]. For example, the word өрмөндө (өрмөндөмөндө) can be divided into such parts as

өрмөн (forest, root of the word) + lar (plural ending) + imiz (first person ending, declension) + dagi (present-modern).

The syntactic pattern of the language, or rather the arrangement of the parts of speech, follows the SOV pattern, where S is the subject, O is the object, V is the verb[13],[14]. However, it should be noted that this pattern is typical, and in practice the structure of the sentence can change. Let's look at two examples:

1) Vazirlik chiqindilarni kamaytirish rejasini tasdiqladi (The Ministry approved the waste plan: *[Ministry-SUBJ] waste-OBJ reduction plan-OBJ approved-VERB* – neutral).

2) Chiqindilarni Vazirlik tasdiqladi (The waste was approved by the Ministry: fronting the object to emphasise waste).

In the first case, we follow the standard SOV pattern, although in the second case, we move the object (the waste) forward to change the focus of attention to it.

Regarding named entities, it is worth noting that there are quite a large number of types (categories) of named entities. In particular, persons, organizations, locations, time, etc. As part of our research work, it is expected that named entities such as persons, organizations, and locations will be identified, which are the most popular and useful in terms of information content today[15].

In addition, it should be noted that the Uzbek language also actively uses dialect lexicons, which are very different from the formal equivalent[16]. In particular, it is possible to separately identify such groups as the Oghuz or Kipchak dialect groups, which are actively used in the northern and western regions of the country (Khorezm region, Karakalpakstan). Moreover, the Oghuz (Khorezm) dialect of the Uzbek language is also actively used in the neighboring country, in particular, in the Tashkhauz region of Turkmenistan.

III. Related Works

In this research paper [17], the authors propose an algorithm for identifying named entities in Uzbek texts. In particular, it is assumed that a dictionary approach (a list of newspapers) will be used, which allows for a word-by-word analysis of the text. At the same time, the authors proposed an additional algorithm that eliminates the drawback of the first. In particular, the first algorithm is based on its base, which means that if there are no words in the base, the algorithm will skip the named entity without recognizing it. Meanwhile, the second algorithm involves the use of syntactic analysis, and thanks to which it is possible to identify location entities due to the identified members of the sentence (adverbial modifier of place). In particular, the keyword (city, village, district, etc.) is identified first, after which the name of the location is identified, which is usually located after the keyword. Thus, the authors developed a universal algorithm for identifying locations in the Uzbek language.

However, it is noted that such a universal solution is difficult to implement for other types of named entities. Although, the possibility of using this approach for other Turkic languages is not excluded. Nevertheless, with such (both algorithms) analyses it is quite difficult to grasp the context of the analyzed text, since in both cases only one word is taken into account. Fig. 1 demonstrates second (upgraded) algorithm.

Meanwhile, in this [18] research paper, the authors examine a similar task, namely, the development of a desktop application for identifying named entities. It should be noted that the authors conducted a fairly detailed study on annotating a text corpus, as well as existing technologies for identifying named entities. Regarding annotating named entities, the authors considered such schemes as BIO, BIOS, BILOU and IOB. Meanwhile, considering the developed algorithm, it should be noted that the authors used a rule-oriented approach to implement this algorithm. In particular, a tokenizer is used that segments the text, and then analyzes each word in the text word by word. Although, in studying existing tasks, the authors provided information not only on traditional algorithms, but also on artificial intelligence technologies (neural networks and machine learning). Moreover, speaking about the chosen approach, it can be noted that this algorithm, like the previous work, does not take into account the context of the entire text, but is limited to one word.

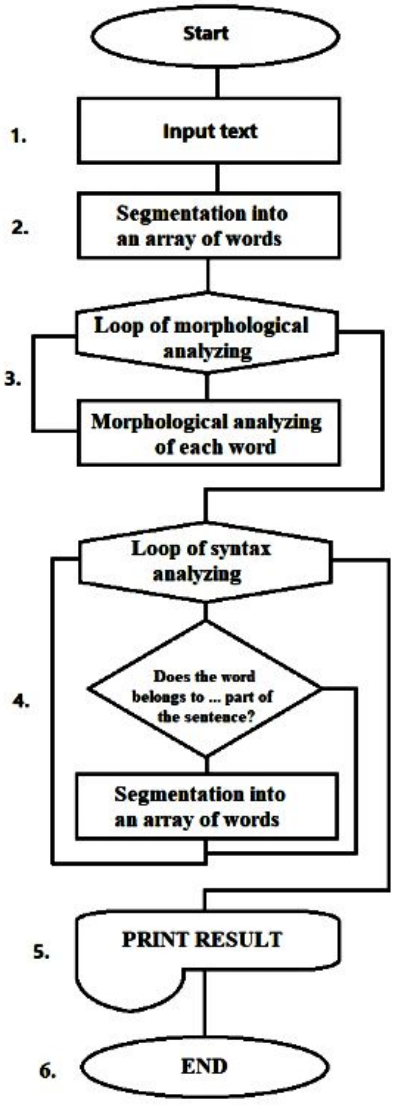

Fig. 1. Scheme of the upgraded algorithm's work.

In addition, this work [19] also solves a similar problem, however, the language for which the algorithm was developed is Tatar. It should be noted that the Tatar language is quite close in nature to the Uzbek language, due to the fact that it, like Uzbek, is agglutinative and belongs to the Turkic languages. Regarding the technical aspect of the algorithm, it is based on a dictionary approach that actively uses n-grams,

which allows taking into account not only single-word entities, but also those that consist of two or more words. The testing results showed how named entities of different categories are recognized by the algorithm differently. In particular, the names of ministries and official services were identified with an accuracy of up to 100%, while the names of restaurants and corporations were within 37.7% - 45.7%.

IV. Proposed Solution

To implement the task, the authors developed a hybrid approach, which consists of rule-oriented algorithms used for text preprocessing, as well as neural networks for identifying named entities.

For the highest quality text analysis, two algorithms based on the rule-oriented approach were developed. In particular, the first algorithm is designed to standardize Cyrillic letters to Latin. To successfully replace these letters, a dictionary is used that contains all the letters of the Uzbek alphabet in two variations (in Cyrillic and Latin). The replacement occurs when Cyrillic is detected with a Latin equivalent.

At the same time, the second algorithm is designed to check the correctness of the spelling of a word in the text. This is checked by stemming, which uses morphological analysis of the word form. To conduct such an analysis, a dictionary of more than 120 affixes is used, as well as a dictionary of 10 thousand word roots. If stemming is successful, the algorithm skips the word, because in this case we do not need to correct the spelling of the word. Otherwise, a counter is triggered that counts incorrectly written words in order to include this information in the algorithm result at the very end.

Meanwhile, a dataset of more than 5,000 sentences was formed to train the model, where each of them was annotated using the BIOES scheme []. This annotation scheme has repeatedly confirmed its advantage over alternative text annotation options for named entity problems []. Online news outlets served as sources for the dataset formation, and for more universal operation of the model, it was decided to extract news articles (ecology related) on various topics.

SpaCy was chosen as the neural network architecture, or rather an empty model (xx), which was trained from scratch on the above-mentioned dataset. This library is quite well-known and popular in light of its ability to quickly scale, which is especially important for low-resource languages such as Uzbek.

The model was trained for 24 epochs, but the weights were rolled back to 21 epochs due to an early stop. It should be noted that the expectation parameter was set to 3, which means that we wait no more than three epochs to improve the model, after which we return to the epoch where the best results were observed. The training results are shown in Fig. 2, which includes such indicators as accuracy, recall, and F1-score in terms of epochs.

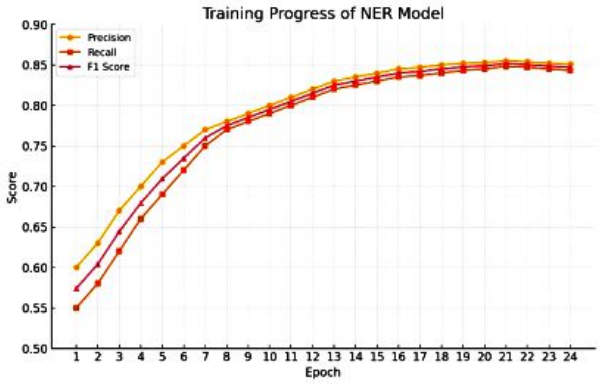

Fig. 2. Process of model traning.

As can be seen from the graph, the best results of precision and recall, as well as f1-measure, reached values of 85.5%, 84.8% and 85.1%, respectively.

V. Testing and Results of the Algorithm

To test the trained model, a dataset of 500 sentences was prepared. Testing was carried out in two formats, where in the first case, the model was tested without preprocessing the text, and in the second case, preprocessing was performed, after which the text was transferred to the model. The results show (Table 1) how the model copes much better with the text after preprocessing. At the same time, Fig. 3 shows a graph of the results before and after using text preprocessing.

TABLE I. Result of Model's Testing

Model	Precision	Recall	F1-Score
SpaCy-only	81.2%	79.0%	80.1%
SpaCy + Preprocessing	85.5%	84.8%	85.2%

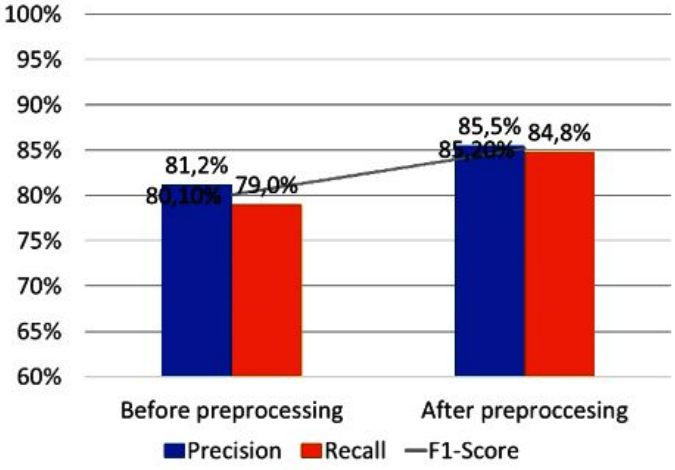

Fig. 3. Process of model training.

In addition to general testing of the model, experimental tests were conducted in the context of named entities. Thus, Table 2 shows the results of testing the model for each type of entity both before and after preprocessing.

TABLE II. Result of Model's Testing Among Named Entities

Model	Precision	Recall	F1-Score
Before preprocessing			
LOC	0.845	0.828	0.836
ORG	0.782	0.765	0.773
PER	0.745	0.721	0.733

After preprocessing			
LOC	0.882	0.874	0.878
ORG	0.824	0.815	0.819
PER	0.792	0.779	0.785

As can be seen from the table above, preprocessing has a positive effect on the final result for all named entity types.

VI. Conclusion

The developed hybrid approach, which involves rule-oriented algorithms in conjunction with a neural network, showed that even lightweight rule-oriented modules can significantly improve entity extraction in Uzbek texts. At the same time, on a corpus of 5,000 environmental sentences, the model increased the F1-measure from 0.80 to 0.85, with the largest increase (+5%) observed for persons. In general, the proposed solution can be considered practical for government and research environmental monitoring systems.

The next steps include expanding the entity sets (species of flora and fauna, pollutants), as well as eliminating the remaining homonymy of toponyms such as Sirdaryo.

References

[1] Official website of the Government of Republic of Uzbekistan, "2025 has been declared the "Year of Environmental Protection and a "Green" Economy"", 11 November 2024, [Online] https://gov.uz/en/eco/news/view/28064

[2] Environment for Development (EfD) research centre, "Uzbekistan measures its green growth strategy", [Online] https://www.efdinitiative.org/news/uzbekistan-measures-its-green-growth-strategy

[3] World bank official report, "Uzbekistan climate adaptation and resilience assessment", [Online] https://documents1.worldbank.org/curated/en/099062724091034716/pdf/P177108 1c619b30dd1a3041e470410c0239.pdf

[4] Wikipedia official website, "Ecological Party of Uzbekistan" [Online] https://en.wikipedia.org/wiki/Ecological_Party_of_Uzbekistan

[5] Officil website of Uzbek laws, "Law of the Republic of Uzbekistan On the regulations of the Legislative Chamber of the Oliy Majlis of the Republic of Uzbekistan", [Online] https://lex.uz/acts/40283

[6] D. B. Mengliev, N. Z. Abdurakhmonova, H. Rahimov, N. Y. Zolotykh, A. A. Ubaydullayev, B. B. Ibragimov, "Automated Recognition of Named Entities and Dialect Standardization in Uzbek Legal Texts", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1050-1053, 2024.

[7] D. B. Mengliev, V. B. Barakhnin, M. O. Eshkulov, O. T. Allamov, B. B. Ibragimov, T. A. Khudaybergenov, "Development of a Legal Document Recognition Algorithm for the Karakalpak Language", 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russian Federation, pp. 323-326, 2024.

[8] B. R. Saidov, J. O. Ruzimov, I. R. Yusupova, A. B. Maksetbaev, A. D. Egamberganova, "Scientometric Analysis of Research Development in Universities From CIS Countries", 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2230-2235, 2024.

[9] U. Salaev, E. Kuriyozov, C. Gomez-Rodrigues, "A Machine Transliteration Tool Between Uzbek Alphabets", The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, Koper, Slovenia. 2022

[10] G. Kurambaeva, "Literary relationships of uzbek and karakalpak in the period of independence", Journal of the Association-Institute for English Language and American Studies, vol. 12, no. 10, pp. 38-46, 2023.

[11] K. Madatov, S. Bekchanov, J. Vičič, "Dataset of stopwords extracted from Uzbek texts", Data in Brief, vol. 43, 108351, 2022.

[12] A. Gatiatullin, D. Suleymanov, N. Prokopyev and K. Karpova, "'Turkic Morpheme" Portal Toolset for Data Augmentation in Low Resource Machine Translation," 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1870-1874, 2024.

[13] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov and J. Cho, "Development of Language Models for Continuous Uzbek Speech Recognition System", Sensors, vol. 23, pp. 1145, 2023

[14] B. B. Ibragimov, A. D. Egamberganova, S. I. Khamraeva, D. A. Fattaxova, Z. Kasimova, D. K. Khudayberganova, "Advancing Oceanology Studies in Karakalpak: A Named Entity Recognition Algorithmic Framework", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1590-1593, 2024.

[15] A. Yusufu, "UZNER: A Benchmark for Named Entity Recognition in Uzbek", Natural Language Processing and Chinese Computing. NLPCC-2023, Springer ed., vol. 14302, 2023.

[16] R. Turaeva, Linguistic Ambiguities of Uzbek and Classification of Uzbek Dialects, Anthropos: International Review of Anthropology and Linguistics, vol. 110, pp. 463-475, 2015.

[17] D. B. Mengliev, V. B. Barakhnin, M. Atakhanov, B. B. Ibragimov, M. Eshkulov, B. Saidov, "Developing Rule-Based and Gazetteer Lists for Named Entity Recognition in Uzbek Language: Geographical Names", 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, pp. 1500-1504, 2023.

[18] B. Elov, M.Samatboyeva, Identifying ner (named entity recognition) objects in Uzbek language texts, Science and innovation international scientific journal, volume 2, issue 4, 2023.

[19] O. Nevzorova, D. Mukhamedshin, A. Galieva, "Named Entity Recognition in Tatar: Corpus-Based Algorithm" CEUR-WS, vol. 2023, 4, 2023.