2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS) | 979-8-3503-8749-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICTACS62700.2024.10840741

Development of Named Entity Recognition Model for Analysis of Oceanographic Texts in Uzbek Language

Davlatyor B. Mengliev[1], Nilufar Z. Abdurakhmonova[2], Vladimir B. Barakhnin[3], Gavhar I. Kuvondikova[4], Zebo G. Kadirova[5], Bahodir B. Ibragimov[6]

[1]*Novosibirsk State University, Novosibirsk, Russia*
[1]*Urgench Branch of Tashkent University, of Information Technologies Named after Muhammad al-Khwarizmi Urgench, Uzbekistan*
[2]*Department of Computer Linguistics, National University of Uzbekistan Named after Mirzo Ulugbek Tashkent, Uzbekistan 0000-0001-9195-5723*
[3]*Federal Research Center for Information and Computational Technologies Novosibirsk, Russia 0000-0003-3299-0507*
[4]*National University of Uzbekistan Named after Mirzo Ulugbek Tashkent, Uzbekistan*
[5]*National University of Uzbekistan Named after Mirzo Ulugbek Tashkent, Uzbekistan, 0009-0006-6509-7255*
[6]*Urgench State University Urgench, Uzbekistan 0009-0000-9518-7397*

***Abstract*—This paper presents the development of a language model for recognizing named entities in Uzbek-language texts on oceanology and navigation. The study included a corpus of 5,000 sentences related to oceanology. These sentences contained more than 33,000 manually annotated words. The BIOES scheme was used to label the data, which allowed labeling both single-word entities and entire phrases. The trained model demonstrated effectiveness in recognizing entities such as geographic features, natural phenomena, vehicles, etc. The accuracy of the model when analyzing test texts was 88%, and the recall was 94%. Despite these results, the model showed a decrease in accuracy when analyzing texts from other areas, indicating the need for further improvement. In addition, the authors also conduct a comparative analysis with existing scientific research in this area to create a more relevant solution to the problem. The article discusses the prospects for improving the model and expanding the scope of its application.**

Keywords: Uzbek Language, Named Entity Recognition, Oceanography, Text Processing, Low-Resource Languages, Machine Learning Model, Natural Language Processing

I. Introduction

Today, there is an active development of the field of oceanology, which is one of the most dynamically developing scientific disciplines[1]. It is associated with solving such global problems as climate change, marine environment protection and rational use of ocean resources[2]. In light of the growing need for accurate and timely information that contributes to making informed decisions, the role of reliable data processing algorithms is increasing[3]. In this context, one of the key tasks is the recognition and further processing of information related to navigation systems and movement control. It should be noted that the sources of such information can be scientific articles, reports, technical documentation, ship logs and much more. The objects that need to be identified within the framework of text analysis are named entities, in particular, these are geographic objects, vehicles, natural phenomena, date and time.

Moreover, it should be noted that research conducted in the field of natural language processing (NLP) is often carried out in such common world languages as English, Chinese or Spanish[4]. However, such scientific solutions for such low-resource languages as Uzbek are either absent or insufficient due to the poor study of these languages[5]. In particular, despite the existence of a certain number of research works on text processing in the Uzbek language, they are insufficient to solve specific problems related to oceanography and navigation systems.

At the same time, it should be noted that in text analysis tasks, recognition of named entities is one of the most popular and significant, since they represent key elements from the content of the text[6]. In particular, thanks to named entities, it is possible to understand the approximate context, because within the framework of such an analysis we can identify (in the context of oceanography) such key objects as geographical objects (seas, oceans, ports), transport vehicles (ships, submarines), meteorological phenomena (cyclones, storms) and technological systems (GPS, sonars). In addition, by identifying these entities, it is possible not only to structure information, but also to automate many tasks that are somehow related to navigation, ship traffic

management, or oceanography in general. Moreover, such a solution can be used in the tasks of forecasting and rapid response to natural disasters, which can contribute to safety at sea.

Regarding the main goal of the current research work, the authors are faced with the task of developing a model for recognizing named entities in Uzbek-language texts related to oceanography and navigation. To achieve this goal, the authors have formed a number of subtasks:

1. Analysis of similar scientific works
2. Collection and preparation of a language corpus in the Uzbek language
3. Development, in particular training of a language model on the collected corpus
4. Evaluation of the efficiency of the trained model.

II. Morphology of Uzbek Language

The Uzbek language is a member of the Turkic family of languages and is also the official language of the Republic of Uzbekistan[7]. More than 40 million people speak this language, and in addition, due to the language belonging to the Turkic family, it has the property of agglutinativity[8]. In addition, the modern Uzbek language, according to the adopted Law on Language (in 1993), uses the Latin alphabet, although the Cyrillic alphabet, introduced during the Soviet period, also remains widespread[9]. At the same time, Uzbek has a fairly large number of dialects, some of which differ not only in pronunciation, but also in spelling[10].

The Uzbek language, as mentioned above, has an agglutinative nature, which means that words are formed by concatenating affixes to the root of the word. Due to this property, it is possible to generate words of different lengths, while the meaning of the word will change each time, regardless of the root of the word. This phenomenon (change of meaning) occurs due to the fact that each affix is a grammatical autonomous unit capable of changing a part of speech from one word to another[11]. Example: the noun “qor” (translated as snow) changes to the verb “qorish” (translated as mix) thanks to the affix “ish”.

Sentences in the Uzbek language are formed according to the fairly popular SOV (subject - object - predicate) scheme[12]. Although, there are exceptions, but this rather depends on either the genre of the text (for example, poetry), or special cases (for example, informal dialogue). Example:

- Men kitobni o‘qidim - standard word order (I read the book).
- Kitobni men o‘qidim - change of structure to highlight “I” (I read the book).

In addition, as mentioned earlier, the Uzbek language has several dialects, in particular, they can be divided into three main groups: Karluk, Oghuz and Kipchak dialects[13]. Examples of the difference between words can be seen in Table 1.

Table 1: Differences of Wordforms in Uzbek Dialects

English	Karluk	Oghuz	Kypchak
a carrot	sabzi	gashir	geshir
a mother	ona	opa	one
like that	unaqa	bundin	unday
how?	qanday	nichik	qalay

III. Related Works

The authors [14] of the article propose relevant solutions for identifying named entities for the Uzbek language, which are based on traditional methods. In particular, two algorithms were proposed that operate on the basis of rules, where the linguistic rules of the Uzbek language, a dictionary of marked words and a dictionary of affixes, necessary for more correct identification of the necessary words, are built in. The scheme of the first algorithm itself is quite simple and is presented below:

1. Text is fed to the input
2. The text is segmented into an array of sentences, and then - an array of words.
3. Morphological analysis of word forms occurs
4. Search for a word from a dictionary of geographical locations
5. Output of the result

In addition, the authors conducted a number of experimental tests on the algorithm, where the accuracy of identifying locations reached 100%. Theoretically, this is an expected result due to the fact that typical search algorithm is used for looking for a word from the database. So, the algorithm will find the word in case the word is present in the database. On the other hand, if the word is not in the database, it will not be found.

The second algorithm has almost the same scheme of work, except for step 4, where instead of searching for a word from the dictionary, the algorithm performs syntactic analysis and, based on grammatical rules, outputs a word that can be a geographic location. Meanwhile, during the testing of the second algorithm, the efficiency was significantly lower - 68%. This is a more interesting result, since in this case the process is based not on a simple search for a word from the dictionary, but on grammatical rules, which can make mistakes in complex sentences that do not meet the requirements of typical sentence formation in the Uzbek language. In addition, the authors conducted a comparative analysis of existing works to show the relevance of the study.

This research paper [15] proposes software for identifying named entities in the Uzbek language. The authors of the article conducted a study of existing technologies that can be used to implement the NER algorithm. In particular, well-known technologies such as BERT, RoBERTa, and several neural network architectures actively used for NLP tasks were studied. At the same time, the researchers attached screenshots of the developed software for NER, demonstrating its performance on several sentences.

However, it should be noted that the authors did not include information on the basis of which this software was developed. Based on the information provided, it is not entirely clear how the NER algorithm was implemented (a machine learning model or an algorithm that works on the basis of rules). Although, the authors mention the BIO tagging scheme, which is actively used in online guides and Internet blogs of developers[16].

In this research paper[17], the authors propose three pre-trained language models for Turkish in medical topics. Each of these models has its own specific features, in particular, BioBERTurkcon uses only Turkish biomedical texts for continuous training. At the same time, weights from the publicly available BERTurk dictionary are introduced. The second model differs in that it was pre-trained on various scientific articles on radiology. The last, third model was trained on the basis of the classical BERT, and data from various sources were used for training in order to create a model from scratch.

To test these models, a corpus of 45,304 radiology reports provided by Ege University Hospital was formed. The reporting period was from 2016 to 2022. According to the testing result, it was found that the f1-scores of these models ranged from 89% to 93% inclusive, which indicates good efficiency. However, despite the proposed approach, such pre-trained models are difficult to apply to our problem. In particular, each of these models is aimed at working with the Turkish language, which, despite its kinship with the Uzbek language (Turkic language family), has a significant number of different grammatical rules. Moreover, the subject matter of the data (biomedicine) on which the models are pre-trained differs from the area we are considering (oceanology).

The article [18] is devoted to the development of an algorithm for recognizing named entities in the Tatar language. The authors proposed an iterative method based on n-gram comparison, which is implemented in the Tugan Tel corpus management system. The algorithm showed different results depending on the type of entities: from 37.7% to 100% accuracy. It is most successful in recognizing the names of ministries, but requires additional filters and extended dictionaries to improve work with more complex categories. In conclusion, the authors note that the algorithm is promising for working with low-resource languages, but needs further tuning to improve accuracy. In addition, it should be noted that this dictionary approach does not take into account the context of sentences, which will definitely lower the quality of the accuracy of the work. For example, consider the text in the Uzbek language: Kecha **vazirlik tizimidagi rahbarlar** *Khorazmlik talaba* ni tabriklashdi (Yesterday management of Ministry system congratulated a student from Khorezm). There are two entities, where first is marked in bold (vazirlik tizimidagi rahbarlar) and the second is marked in italic (Khorazmlik talaba). However, proposed n-Gram based algorithm will detect these entities wrong and total entities' quantity will be 3 (vazirlik, rahbarlar, talaba).

The paper [19] presents a hybrid approach that combines a dictionary-based approach with a machine learning-based AI model for named entity extraction from semi-structured data. The authors focus on text preprocessing using morphological analysis, using a random walk method to extract semantically similar word pairs, and applying neural networks to recognize linguistic constructions.The key components include text lemmatization, a random walk method, and a neural network model for named entity extraction. The experiments showed effective results with an accuracy of 85% for detecting geographical names, demonstrating that machine learning can significantly improve named entity recognition in semi-structured data. However, such an approach is difficult to implement for the Uzbek language, since, unlike the Kazakh language (which can handle inflectional classes), it is more susceptible to agglutination. Due to this property, in the Uzbek language there are often special cases when one affix can completely change the meaning of a word, for example: ok (white color) and ok+lash (to prove someone's innocence).

IV. Proposed Solution

To implement the algorithm, a multilingual Spacy model from the Python library was chosen. This model can be retrained for any other language, in particular, for such a low-resource language as Uzbek.

In addition, data from various sources, in particular, Internet news blogs and reports, were collected as training data. Moreover, to form a wider corpus on oceanography, the authors manually translated articles and some reports from other foreign languages (Russian, English) into Uzbek. This made it possible to build a fairly large (for educational purposes) corpus, which has various terms on oceanography. In particular, the corpus contains a total of 5,000 sentences or 33,667 words, which were manually annotated by the authors. Words were marked up using the BIOES[20] method, which assumes various variants of named entities. In particular, the letter B means the beginning, I is the internal part and E is the end of the named entity. Besides, the letter O means that this element is outside the entity, and C is a named entity consisting of one word. It should be noted that there are several categories of entities in the dictionary: person, position, location, navigation, natural phenomena, vehicles, date and time.

Example: *Kecha Vazir bilan uchrashuvda yomg'ir yog'di, texnologik universiteti talabalari, kollej talabalariga nisbatan o'z navbatini kutishdi*. Translation: Yesterday it rained at the meeting with the Minister, but the students of the technological university waited for their turn, unlike the college students. Type of entities for each word are demonstrated in the Table II.

Table II: Example of BiOES Tagging

Sentence #	A word in Uzbek	BIOES-tag	Type of entity	A word in English
1	**Kecha**	O	none	Yesterday
1	**Vazir**	S	position	Minister
1	**bilan**	O	none	with
1	**uchrashuvda**	O	none	meeting
1	**yomg'ir**	O	none	rain
1	**yog'di**	O	none	was
1	**texnologik**	B	organization	technology
1	**universiteti**	I	organization	universitety
1	**talabalari,**	E	position	students
1	**kollej**	B	organization	college
1	**talabalariga**	E	position	students
1	**nisbatan**	O	none	unlike
1	**o'z**	O	none	their
1	**navbatini**	O	none	turn
1	**kutishdi**	O	none	waited

Results of testing the model is included in section V. In terms of the algorithm scheme, it is described in Fig 1.

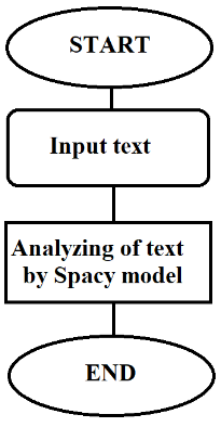

Fig. 1: Algorithm's scheme.

As it can be seen, there are not many steps in algorithm's work, due to trained model.

It should be noted that the model was trained for 50 epochs, where the f1-score value reached 91%. And the percentage of accuracy and recall reached 88% and 94%, respectively.

V. Testing and Results of the Algorithm

The authors of the article conducted a number of experiments, in particular, for the most objective evaluation of the algorithm, two samples were created. The first sample contained 250 sentences related to oceanology. The sample included 421 named entities. The second sample included sentences from different sources, including different topics, which also consisted of 250 sentences. The second dataset contained 388 named entities, but not all of them were in the training sample, and therefore, it is assumed that the model simply does not recognize them.

The testing results showed that the model was able to identify named entities from the first sample with an accuracy of 88% (345 out of 392 named entities), and the completeness reached 93% (392 out of 421 named entities). Moreover, in the case of the second sample, the accuracy was noticeably lower - 84% (216 of 257 named entities), and the recall dropped significantly, reaching 66% (257 of 388 named entities). The results are presented in more detail in Table 3 and Fig 2.

Table III: Result of Model's Testing

Number of dataset	Correctly detected entities / Detected entities	Precision	Identified entities / total entities	Recall
1	345 /392	88%	392 / 421	99%
2	216 / 257	84%	257 / 388	66%

It should be noted that the results do show positive ratings, although in the case of the second dataset the indicators dropped slightly. However, as was mentioned earlier, this sample also contains named entities from other topics (for example, medicine or literature). As you can see, when using the model for its intended purpose (for the field of oceanography), it will show positive success in identifying the objects we need from the text.

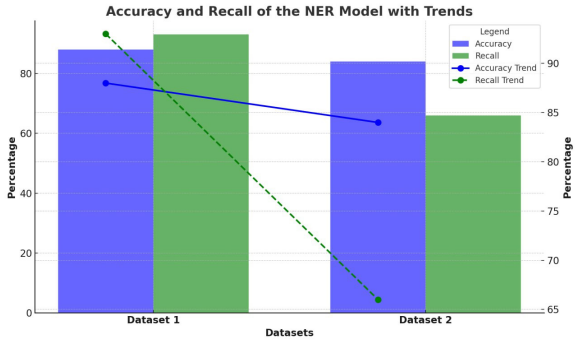

Fig. 2: Results of model's testing.

VI. Conclusion

This study presented the development of a model for recognizing named entities in Uzbek-language texts related to oceanography and navigation. The main objective was to create a tool that could effectively identify such named entities as geographic objects, date, time, persons, positions (job title), navigation, natural phenomena, and vehicles in Uzbek-language texts.

A corpus of 5,000 sentences, consisting of 33,667 words, was collected during the work. Each of these words was manually annotated to a specific category of named entity, and if it does not belong to any of the categories, it is marked as a word outside the entity. The BIOES scheme was used to mark the entities, which allowed marking both individual words and entire phrases. The model trained on this data showed high results, especially when analyzing texts directly related to oceanography, where the accuracy reached 88%, and the recall - 94%.

Despite the results achieved, the study also revealed some limitations. In particular, when analyzing texts containing entities from other fields (e.g. medicine or literature), the accuracy of the model significantly decreased. This indicates the need for further improvement of the model and expansion of the training corpus to increase its universality. In the future, it is planned to improve the algorithm by adding more data and using alternative neural network architectures. This will expand the scope of the model and improve its effectiveness in a broader context. In addition, it should be noted that the proposed algorithm might be used for other Turkic languages, which are also low-resourced (Karakalpak, Kyrgyz, Kazakh and e.g.).

References

[1] M. Visbeck, "Oceanography in the decade of digital science and sustainable development: New opportunities for TOS?", Oceanography, vol. 33, issue 1, 2020.

[2] B. Cushman-Roisin, J. Beckers, "Introduction to Physical Oceanography", Wiley-Blackwell Publishing house, 343 pp., 2010.

[3] E.Kuriyozov, D. Vilares, C. Gomez-Rodriguez, "BERTbek: A Pretrained Language Model for Uzbek", Special Interest Group on Under-resourced Languages workshop (SIGUL-2024), Torino, Italy, May 2024.

[4] D. Mengliev, M. Eshkulov, V. Barakhnin, R. Abdullayev, N. Boltayev, B. Ibragimov, "Linguistic Nuances in Text Analysis: TF-IDF Metric's Algorithm Implementation for the Karakalpak Language Recognition", 2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Yekaterinburg, Russian Federation, pp. 019-022, 2024.

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

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

[7] G. Dushaeva, "Phonological System of Modern Uzbek Language", Pindus Journal of Culture, Literature, and ELT, vol. 2, no. 5, 2022.

[8] M. Sharipov, J. Mattiev, J. Sobirov, R. Baltayev, "Creating a morphological and syntactic tagged corpus for the Uzbek language", The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing(ALTNLP), June 7-8, 2022.

[9] I. Bakaev, T. Shafiyev, "Morphemic analysis of Uzbek nouns with Finite State Techniques", Journal of Physics: Conference Series, 1546, 2020.

[10] S. Raxmatova, M. Kuzibayeva, "Generality and specificity of dialectics and its reflection in the morphology of the Uzbek language", Economy and society, vol. 9, issue 88, 2021.

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

[12] Kh. Madatov, S.Sattarova, "Creation of a Corpus for Determining the Intellectual Potential of Primary School Students", 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2420-2423, 2024.

[13] D. Mengliev, N. Abdurakhmonova, D. Hayitbayeva, V. Barakhnin, "Automating the Transition from Dialectal to Literary Forms in Uzbek Language Texts: An Algorithmic Perspective", 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, pp. 1440-1443, 2023.

[14] D.Mengliev, V.Barakhnin, M.Atakhanov, 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), pp. 1500-1504, 2023.

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

[16] Charudatta Manwatkar, "A Beginner's Guide to Named Entity Recognition (NER)," online guide, 2020 December, [Online]. Available: https://medium.com/swlh/a-beginners-introduction-tonamed-entity-recognition-ner-2002b1a010c1.

[17] H. Türkmen, O. Dikenelli, C. Eraslan, M. C. Çalli and S. S. Ozbek, "Developing Pretrained Language Models for Turkish Biomedical Domain," 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, pp. 597-598, 2022.

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

[19] M, Mansurova, V.Barakhnin, Y.Khibatkhanuly, I. Pastushkin, "Named Entity Extraction from Semistructured Data Using Machine Learning Algorithms", Computational Collective Intelligence (ICCCI 2019), 2019.

[20] E. Sang, J. Veenstra, "Representing text chunks", Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 173-179, 1999.