Automated Recognition of Named Entities and Dialect Standardization in Uzbek Legal Texts

Davlatyor B. Mengliev
Novosibirsk State University
Novosibirsk, Russia;
Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Urgench, Uzbekistan
0000-0003-3969-1710

Nilufar Z. Abdurakhmonova
Department of Computer linguistics of National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0001-9195-5723

Hasanboy Rahimov
Andijan State University
Andijan, Uzbekistan
0009-0009-2414-4635

Nikolai Yu. Zolotykh
Lobachevsky State University
Nizhni Novgorod, Russia
0000-0003-4542-9233

Alisher A. Ubaydullayev
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0009-0003-1476-4386

Bahodir B. Ibragimov
Department of Information Technologies
Urgench State University
Urgench, Uzbekistan
0009-0000-9518-7397

2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE) | 979-8-3315-1632-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/PIERE62470.2024.10804942

***Abstract*—This study presents the development of a tool for identifying named entities in Uzbek legal texts. It should be noted, that besides of detecting named entities, the authors developed an algorithm, which is able to standardize word forms by replacing the detected dialect words (Karluk, Kypchak and Oghuz) with their formal forms. This will help to fix popular grammatical mistakes among native speakers from different regions of the Uzbekistan. The proposed hybrid approach combines the traditional approach, which is used in the preprocessing (standardization of word forms), where a dictionary with more than 10 thousand marked words is actively used. At the same time, a custom language model is used to work with detecting named entities, which was trained on 2000 legal sentences. The testing results showed quite high indicators, in particular, the language model detected named entities with an accuracy of 90%, and the recall reached 94%. Moreover, the algorithm used to standardize dialect word forms showed even higher rates, ranging from 90% to 100% depending on the dialect.**

Keywords—Uzbek language, legal documents, named entity recognition, dialect standardization, Oghuz, Karluk, Kypchak, algorithm development, low-resource languages, text processing, linguistic diversity

I. Introduction

Today, digital analysis of legal documents is one of the popular tasks in the field of natural language processing (NLP)[1]. Moreover, one of the most pressing problems in this task is the admission of grammatical errors in the formation of legal documents[2]. At the same time, the trend is gaining momentum, where the volume of legal documents (contracts, agreements, memorandums) increases, which also imposes an additional need for the introduction of digital tools to solve such problems[3].

In addition, analyzing this problem in the context of the Uzbek language, it is necessary to emphasize another problem such as widespread dialects, where residents of different regions of Uzbekistan can accidentally use dialectal forms of words in official documents, which is also an urgent task of the present time[4]. The researchers opted for three major dialects of the Uzbek language, namely, Karluk (which is used in the official Uzbek language), Oghuz and Kipchak.

Moreover, it should also be noted that in legal documents, one of the most important elements are named entities, in particular, the names of the parties to the agreement, names of organizations, dates and times, locations, etc.[5] The ability of the algorithm to detect such key objects in the document also adds additional relevance in the context of the problem being solved.

The goal of this study is to develop a reliable algorithm that can correctly recognize named entities and standardize dialects in legal documents in the Uzbek language. The authors expect that the solution they propose will help in the tasks of analyzing legal documents, including its correction by standardizing dialect words into formal ones.

The structure of the article consists of 6 sections, where the first two sections present the introduction of the article and brief information on the morphology and dialects of the Uzbek language. The third section includes a comparative analysis of alternative works that solve similar topical problems. The fourth and fifth sections present the algorithm and experiments aimed at assessing the effectiveness of its work. The final section contains the conclusion of the article, as well as reflections on promising directions for the development of the algorithm, including the use of alternative technologies.

II. Morphology of Uzbek Language

The Uzbek language is one of the members of the Turkic language family, which includes more than 10 natural languages with common properties[6]. One of these properties is the agglutinative nature of the language, where words are formed by combining a combination of affixes to the root of the word[7]. In this case, each level of the affix refers to a certain grammatical structure. For example, the most basic level is the plural ending -lar, which usually follows the root of the word. This ending is common to all words, which undoubtedly simplifies the work of linguists [8].

979-8-3315-1632-1/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:16 UTC from IEEE Xplore. Restrictions apply.

Possessive endings

Possessive affixes in the Uzbek language help to determine the belonging of a subject or object to a certain person[9]. These affixes change depending on the person and number of the owner[10]. Table I shows the abbreviations of the categories of affixes that are used in Table II. The second table contains the attractive affixes for each person and number.

TABLE I. AFFIXES' ABBREVATIONS

№	Type of affixes (Full name)	Type of affixes (short)
1	First person, singular	FP-S
2	Second person, singular	SP-S
3	Third person, singular	TP-S
4	First person, plural	FP-P
5	Second person, plural	SP-P
6	Third person, plural	TP-P

TABLE II. POSSESIVE AFFIXES

№	Affixes	Type of affixes	Example
1	-im, -m	FP-S	Uy (home) – uyim (my home)
2	-ing, -ng	SP-S	Kitob (a book) – kitobing (your book)
3	-i, -si	TP-S	Kitob (a book) – kitobi (her/his book)
4	-imiz, -miz	FP-P	Uy (home) – uyimiz (our home)
5	-ingiz, -ngiz	SP-P	Kitob (a book) – kitobingiz (your book)
6	-lari	TP-P	Kitob (a book) – kitoblari (their book)

Dialects

The Uzbek language has many dialects, which can differ significantly from each other in many grammatical parameters[11]. However, the most popular and largest dialects (with over 2 million speakers) are Karluk, Kipchak and Oghuz[12]. The Karluk dialect is used in the modern official Uzbek language, in this regard, the number of speakers of this dialect is the majority in comparison with the other two dialects.The Oghuz dialect, or as people call it, the Khorezm dialect, dominates in the Khorezm region of Uzbekistan. Words and pronunciations are quite close to Turkish and Azerbaijani languages due to historical and cultural ties. The Kipchak dialect is mainly used in the western part of Uzbekistan - in Karakalpakstan, its linguistic properties are very similar to the Kazakh language. In general, you can see the difference in words (spelling) of each dialect in Table III.

TABLE III. POSSESIVE AFFIXES

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

III. USING THE TEMPLATE

The authors [13] of the article conduct research in the field of named entity recognition, emphasizing it as one of the most popular tasks in natural language processing. Moreover, the special role of named entities in extracting the necessary information to identify the semantics of texts for further processing is emphasized. The researchers note that the performance of NER in most cases is limited by the lack of digital resources, in particular, a language corpus with tagged words and sentences. A similar problem can be observed in such low-resource languages as the Uzbek language. Moreover, there are almost no pre-trained language models for the Uzbek language that would be suitable for universal tasks, which also imposes certain difficulties. However, the authors propose their own language corpus, which, in their opinion, can eliminate all the above-mentioned problems, supplementing their argument with a number of experiments in named entity recognition tasks. It should be noted that the authors have done a lot of work, first of all, by collecting their own corpus. However, the proposed solution is based on general concepts, i.e. the model is trained on a corpus that was formed from various sources, and the markup structure is quite complex. Moreover, the proposed solution consists exclusively of a trained model, which in most cases copes with its responsibilities positively. Although, this approach could be improved by implementing a hybrid model, for example, by including pre-processing of word forms in texts, which the analysis was more accurate. However, the authors note that they plan to improve their approach in the future.

The authors [14] of the research study the issues of creating original pre-trained BERT model for the Uzbek language. The article begins with a description of current developments and achievements in the field of text analysis using modern software, including neural network technologies. In particular, the authors especially highlighted the models built on the basis of transformers, used in many tasks. In addition, such popular transformers as BERT and RoBERTa, used for the English language, were highlighted. Continuing to list similar technologies, the authors note the complete absence of such solutions for the Uzbek language, which is undoubtedly a sad fact. To eliminate this problem, the authors pre-trained a model based on the BERT architecture, where they used a large language corpus, which consists of more than 142 million words (625 thousand articles on various topics). The basis for this corpus was news blogs and sites, Wikipedia, as well as an encyclopedia of the Uzbek language. The authors also noted that such a variety of sources caused noise in the data, because in some sources the text was written in Cyrillic, while the official Uzbek in Latin. In addition, the authors also included information on the grammar and morphology of the Uzbek language, demonstrating several examples for readers to understand about this language.

Despite the large amount of research work done, the authors note that there is still a lot of work to do to ensure that the model can work with practical tasks, and its current version is only a working prototype of the proposed product. At the same time, it should be noted that this model is pre-trained on diverse sources, which can undoubtedly make it useful, although in certain tasks (for example, such as the analysis of legal documents) problems may appear. One of the main reasons for this was that the language corpus does not contain legal documents, legislative acts, etc.

In this [15] article, the authors developed 3 pre-trained models for the Uzbek language. The article begins with an overview of alternative works, and during the description of these studies, the authors separately emphasize that one of the

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:16 UTC from IEEE Xplore. Restrictions apply.

most popular approaches for creating such large models are transformer architectures, such as BERT and HLM. Although, the researchers place the main emphasis on BERT, arguing that it appeared much earlier than its analogs. In a further analysis of BERT, the authors note that pre-trained models have already been created for languages such as Russian, Spanish, Portuguese and others, and from the Turkic languages only for Kazakh and Turkish. In this regard, the authors separately note that the models they propose are unique, because the Uzbek language is a low-resource language, despite the fact that it is in second place among the Turkic languages in terms of the number of native speakers. Two sources were selected as sources for forming the dataset: Wikipedia (124 thousand articles) and news blogs (200 thousand). Regarding the topics of these articles, they belong to different categories, in particular - sports, technology, economics, etc. The results of testing the models showed positive ratings, and the percentage of f1-score varies from 70% to 78%. The models are designed for semantic analysis of the text, topic classification and identification of named entities (organization, person and time).

Despite the use of modern technologies for analyzing the text of the Uzbek language, the proposed solution may not show very positive work in the process of analyzing texts that contain dialect words or words from legal terminology. Although, it should be noted that the authors pursued a different goal in their study.

IV. Proposed Solution

The authors implemented the approach in reliable analysis of legal documents by developing two tools, where the first is an algorithm for identifying and standardizing dialect words. This tool is implemented in Python and relies on built-in rules and a lexicon that is actively used to replace dialect words with formal ones. This dictionary consists of more than 10-thousand-word forms. Results of testing the tool is included in fifth section.

The second tool, which completes the work of the proposed approach, is a module for identifying named entities. This module is based on the custom Spacy model (Python library), which was trained on 2000 legal sentences. The main part of these marked sentences are fragments from legislative acts, as well as agreements and contracts. Thus, the possibility of any grammatical errors in the training materials is excluded.

Training corpus was created by BIOES-tagging, where B-beginning of the entity, I-inner of entity, O-outside of entity, E-ending of entity, S-single entity[16]. For example, consider the text: "Raqamli texnologiyalar Vaziri ertaga Xorazmga tashrif qiladi" (The Minister of the digital technologies will visit Khorezm tomorrow). In the table IV is shown detailed tagging the text above.

TABLE IV. Example of BIOES Tagging

Sentence #	A word in Uzbek	BIOES-tag	Type of entity	A word in English
1	**Raqamli**	B	position	Digital
1	**texnologiyalar**	I	position	technologies
1	**Vaziri**	E	position	Minister
1	**ertaga**	O	none	tomorrow
1	**Xorazmga**	S	city	Khorezm
1	**tashrif**	O	none	visit
1	**qiladi**	O	none	will

The model was trained for 48 epochs, where the f1-score value reached 93%. And the percentage of accuracy and recall reached 89% and 96%, respectively. Results of testing the model is included in section V. The algorithm works according to the following scheme:

1) The text is fed to the input

2) The text is divided into an array of sentences, and from there into an array of words.

3) Each word is analyzed by the word form standardization module, with the goal of replacing all dialect words with formal ones.

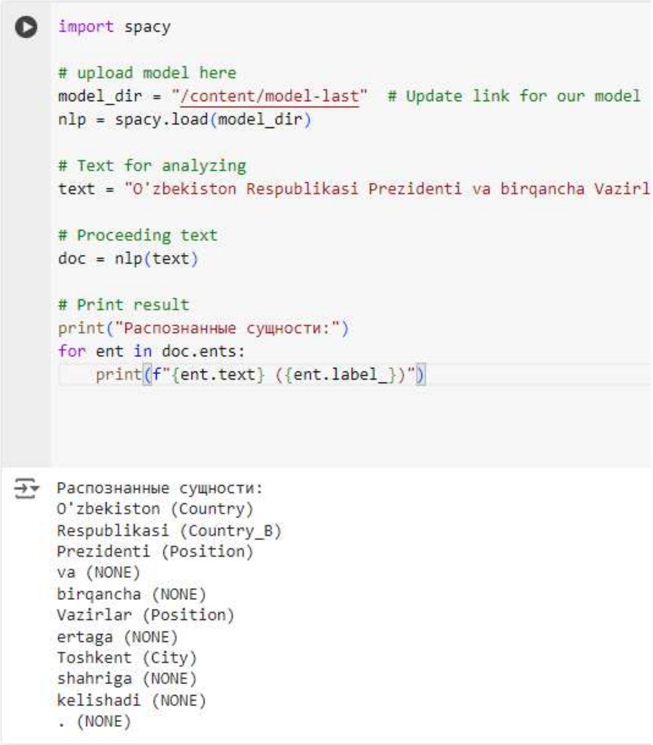
3.1) Each word is searched for in the dictionary of dialect words, if there are matches, the dialect word is replaced with a formal one.

3.2) If the word is not found, morphological analysis is performed, during which the word is stemmed to detect the root of the word, the resulting root is searched for in the dictionary and, if successfully identified, it is replaced with its formal version.

4) The resulting standardized text is analyzed by the named entity detection module, here the Spacey model is triggered.

5) At the output, we obtain the identified named entities

The result of the model's work is shown in Fig. 1.



```
import spacy

# upload model here
model_dir = "/content/model-last"  # Update link for our model
nlp = spacy.load(model_dir)

# Text for analyzing
text = "O'zbekiston Respublikasi Prezidenti va birqancha Vazirl

# Proceeding text
doc = nlp(text)

# Print result
print("Распознанные сущности:")
for ent in doc.ents:
    print(f"{ent.text} ({ent.label_})")
```

```
Распознанные сущности:
O'zbekiston (Country)
Respublikasi (Country_B)
Prezidenti (Position)
va (NONE)
birqancha (NONE)
Vazirlar (Position)
ertaga (NONE)
Toshkent (City)
shahriga (NONE)
kelishadi (NONE)
. (NONE)
```

Fig. 1. Results of model's work.

V. Testing and Results of the Algorithm

The proposed solution was tested in a number of experiments, in particular, for each of the algorithms, appropriate samples were collected for an objective assessment of the effectiveness. The parameters such as accuracy and recall were chosen as evaluation metrics. Accuracy is the ratio of correctly detected dialect words to detected words. While recall implies the ratio of detected dialect words to the total number of dialect words.

Testing began with the algorithm for standardizing word forms from dialect to formal, where the authors prepared four

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:16 UTC from IEEE Xplore. Restrictions apply.

samples. The first three contained dialect words of separate groups, where the first sample contained only Karlu words, the second - Oghuz and the third - Kipchak. And the last one contained wordforms from different dialects. Each sample consist of 100 words. As a result of testing, the algorithm was able to recognize dialect words with an accuracy of 100%, 94%, 93% and 90%, respectively. Moreover, the recall in turn reached the level of 99%, 96%, 97% and 92%.

One of the main reasons for the algorithm's errors is that word forms are missing from the dictionary of tagged words. In addition, certain words have the same form in all dialect variations, i.e. a word has the same root in two or all three dialects, but the meanings of these roots change depending on the combination of affixes. This problem can be solved by increasing the volume of the tagged dictionary to cover as many words as possible. More detailed results of the algorithm testing can be found in Table V.

TABLE V. RESULT OF MODEL'S TESTING

Number of dataset	Correctly detected sentences / Detected sentences	Precision	Correctly identified sentences / total sentences	Recall
1	100 / 100	100%	99 / 100	99%
2	94 / 100	94%	96 / 100	96%
3	93 / 100	93%	97 / 100	97%
4	90 / 100	90%	92 / 100	92%

In addition, the authors conducted a separate experiment to test the trained Spacy model. To achieve this goal, a sample of 200 sentences was prepared, which included 313 named entities. As a result of testing, the model correctly detected 282 entities (accuracy 90%), while covering only 295 entities (recall 94%).

VI. CONCLUSION

The study presented the development of a tool for analyzing Uzbek language texts for detecting legal terms. In addition, an algorithm for standardizing dialects was implemented as an additional module.

The relevance of this work is primarily due to the ability to work with dialect words, where the algorithm not only detects such words, but automatically replaces them with formal versions. In addition, the received (processed) text is analyzed by artificial intelligence to detect named entities in the field of jurisprudence. In particular, the algorithm can classify such named entities as names, names of organizations, dates and places, which significantly facilitates further processing and analysis of legal texts. The developed algorithms have demonstrated high efficiency in solving the tasks.

In the future, further improvement of the algorithm is expected, including expansion of the dictionary base for processing dialects and the introduction of alternative neural network architectures to improve the efficiency of recognizing named entities.

REFERENCES

[1] K. Sugathadasa, B. Ayesha, N. Silva, A. Shehan, V. Jayawardana, “Legal Document Retrieval Using Document Vector Embeddings and Deep Learning”, Advances in Intelligent Systems and Computing, vol 857, 2018.

[2] H. Vardham, N.Surana, B. Tripathy, “Named-Entity Recognition for Legal Documents”, Advances in Intelligent Systems and Computing, vol 1141, 2020.

[3] G. Zhao, Y. Liu, E. Erdun, “Review on Intelligent Processing Technologies of Legal Documents”, Lecture Notes in Computer Science, vol 13338, 2022.

[4] 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.

[5] D. Mengliev, V. Barakhnin, N.Abdurakhmonova, M.Eshkulov, "Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation", Data in Brief, 54, 110413, 2024.

[6] D. Mengliev, V. Barakhnin, B. Ibragimov, "Rule-Based Syntactic Analysis for Uzbek Language: An Alternative Approach to Overcome Data Scarcity and Enhance Interpretability," 2023 IEEE 24th International Conference of Young Professionals in Electron Devices and Materials (EDM), Novosibirsk, Russian Federation, pp. 1910-1915, 2023.

[7] 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.

[8] 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.

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

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

[11] 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.

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

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

[14] B.Mansurov, A.Mansurov, “UzBERT: pretraining a BERT model for Uzbek”, arXiv.org, vol. 2108.09814 22 August 2021.

[15] 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.

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

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:16 UTC from IEEE Xplore. Restrictions apply.