Automated Detection of Allusions in Uzbek Language: A Computational Approach

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

Raima Kh. Shirinova
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0003-3860-6684

Mohira F. Saparova
Mamun University of Khorazm
Khiva, Uzbekistan
0009-0004-4833-6774

Inomjon M. Azimov
Tashkent State University of Uzbek language and literature
Tashkent, Uzbekistan
0009-0001-4803-7166

Bahodir B. Ibragimov
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/PIERE62470.2024.10804911

***Abstract*— Allusions play a significant role in literary and cultural works, serving as a tool for conveying deep meanings embedded by authors. In the Uzbek language, as in many other languages, allusions are commonly used to create more complex narratives, which allows us to connect the work with broader cultural or historical traditions. However, identifying these allusions, and even more so interpreting them, presents significant challenges for readers.**

In recent years, it might be seen the active development of algorithms and different kind of solutions for various text analysis tasks. Among these tasks, identifying metaphors, idioms, and allusions can be singled out as one of the most difficult tasks, due to the need for a deeper understanding of the text. Moreover, interpreting allusions requires interdisciplinary knowledge in the field of literature, cultural studies, linguistics, and more.

This article discusses the development of an algorithm for the automatic detection of allusions in the Uzbek language, aimed at helping researchers and readers fully understand the meaning conveyed by authors. In addition, the proposed solution can help linguists and those who study the Uzbek language. The algorithm uses two dictionaries, where first consist of 10,000 tagged sentences, while second contains over 80 thousand word forms. At the same time, the algorithm was tested and as a result of a number of experiments, the efficiency of the work was revealed, where the accuracy of the algorithm reached range from 82,6% to 100%.

Keywords*— *Allusion detection, Uzbek language, natural language processing, dictionary based approach, cultural references, corpus annotation, BIO-tagging, historical allusions, literary analysis, custom algorithm.

I. INTRODUCTION

Allusions play an important role in various literary and cultural works, being a means of conveying the deep meaning conveyed by the authors of the works[1]. In the Uzbek language, as in many other languages, allusions are often used to create a more complex and interesting narrative that can connect the work with a broad cultural or historical tradition[2]. However, identifying allusions, and even more so, interpreting them, is quite a difficult task for readers[3]. In the context of automatic text processing systems, this can impose additional complexity (model development, training, data collection, etc.).

In addition, in recent years there has been an increase in interest in the development of algorithms and solutions based on artificial intelligence for solving various text analysis problems[4]. In particular, this is the detection of named entities in texts, syntactic and semantic analysis, etc[5]. However, along with such tasks, there are also those that require much more effort to solve, for example, identifying metaphors, idioms and allusions. Such tasks require a deeper understanding of the text, as they are based on more interdisciplinary knowledge, including literature, cultural studies, linguistics, etc. This article discusses the development of an algorithm for the automatic detection of allusions in the Uzbek language, which will help researchers and readers fully understand the text presented by the author of the work. In addition, the proposed solution can help linguists and those who study the Uzbek language.

Moreover, in literature, an allusion can refer to famous works or authors, for example, Shakespeare, Navoi or Pushkin[6]. In a cultural context, an allusion can refer to traditional customs and folk tales, as well as various mythological plots. As a rule, historical allusions often refer to important events or personalities that in one way or another had a certain impact on society. One example of an allusion in the Uzbek language is a reference to Layli and Majnun. For example, "Bu ham bir Layli-Mazhnun qissasi", this phrase implies a situation with tragic love, which speaks of impossible or doomed relationships.

The aim of this work is to develop an algorithm for automatic detection of allusions in the Uzbek language based on a custom machine learning model. To achieve this goal, the following tasks are planned:

1. Collecting and annotating a corpus of texts in the Uzbek language, which contain various allusions.

2. Training a custom machine learning model using the annotated corpus.

3. Testing the language model to evaluate its performance.

The authors of the article divided the content into 6 sections, where the first two sections consist of introductory materials and about the morphology of the Uzbek language, for the easiest understanding of the problem posed, which is solved within the framework of the study. The third section suggests getting acquainted with similar works or studies that are either related or directly similar to the problem being solved. The fourth and fifth sections contain information on the language model and its testing. The sixth section is the conclusion of the article, where the authors also discuss the further development of the proposed solution.

II. Morphology of Uzbek Language

The Uzbek is the official language of Uzbekistan, spoken by over 40 million people[7]. Moreover, this language belongs to the Turkic language group and is the second most spoken language[8]. Throughout its history, the Uzbek language has used several alphabets: Latin and Cyrillic[9]. After gaining independence, the government of the country decided to return to the Latin alphabet.

Morphology

The Uzbek language, like other Turkic languages, has a property called agglutinativity, where words are formed by concatenating an affix(es) to the root of the word to express different grammatical meanings[10]. For example:

- uy (house) + lar (plural ending) = uylar (houses).
- ket (to leave) + di (past tense ending) = ketdi (left).

In Uzbek morphology there are three levels of affixes that are successively connected to the root, namely - the plural ending (-lar, it is universal), adjective endings and case ending[11]. Regarding cases, there are six of them in the Uzbek language: nominative, genitive, accusative, dative, locative and original. For example:

- kitob (book) + ni (accusative case) = kitobni (book).
- uy (house) + ga (dative case) = uyga (to the house).

Dialects

In the Uzbek language, dialect forms are actively used, where residents of different regions have their own dialects[12]. The largest dialects are Kipchak, Karluk and Oguz[13]. These dialects differ in many properties, including phonetics, vocabulary and some grammatical features. The Karluk dialect (formal Uzbek) dominates in literature and official speech.

Sentence members and syntax

In the Uzbek language, sentences are constructed in a certain order, so the most common scheme is "subject - object - predicate"[14]. For example: Men maktabga boraman (I go to school), where Men (I) is the subject, maktabga (to school) is the object, boraman (go) is the predicate.

Moreover, adjectives and attributes come before the word they define, which is a characteristic feature of many Turkic languages. For example: Katta uy (Big house), where Katta (Big) is an adjective, uy (house) is a noun.

III. Related Works

The authors of the article [15] consider two main groups of approaches to text representation: discrete and distributed text representations. The first group discusses the following types of representation:

- **One-Hot Encoding:** The essence of this method is to encode words as vectors, where each word is represented as a unique vector with one unit (1) and the rest of the zeros (0). The advantage of this method is its simplicity, however, it does not preserve information about the order of words or semantic relationships between them.
- **Bag-of-Words:** In this approach, words from the corpus are put into the so-called "bag of words", and then the frequency of occurrence of these words in the text is calculated. This method, like the previous one, ignores the word order and lexical information. Although, this approach can be useful for tasks where it is necessary to calculate the frequency of words or classify documents.
- **CountVectorizer:** This approach works on a very similar principle to BoW. It is useful for tasks that require calculating the importance of words in a document.
- **TF-IDF:** This method is used to calculate the importance of words in a document. TF-IDF is often used in document retrieval and classification tasks.

The second group contains distributed text representations:

- **Co-Occurrence Matrix:** This method takes into account the co-occurrence of words in the text and helps to identify the relationships between words. It preserves the word order and can be used to analyze the relationship between different words in a corpus.
- **Word2Vec:** One of the most popular methods for creating vector representations of words, which is based on context-based word prediction (CBOW) or context-based word prediction (Skip-Gram). Like other complex approaches, Word2Vec preserves semantic and syntactic relationships between words and is widely used in various NLP tasks.

In general, the authors do not offer their own solution, but only provide the results of experiments on adapting the above-mentioned approaches to the Uzbek language. Judging by the results, these methods have shown their effectiveness in processing texts in the Uzbek language, despite some limitations.

Moreover, it was found that the use of these methods in a joint combination will provide greater efficiency than if they were used individually. In addition, the authors noted that the adapted methods work well for basic tasks (text classification and word frequency analysis), although they require additional research and improvements in more complex tasks. In particular, the use of these approaches for such tasks as context analysis, recognition of named entities, semantic analysis in Uzbek texts seems difficult due to the morphology and grammar of the Uzbek language. An additional complication is that the Uzbek language is low-resource, which indicates an insufficient amount of the necessary language corpus for training more advanced language models.

The article [16] focuses on creating an educational corpus that matches the intellectual potential of primary school students in Uzbekistan. The authors examine the impact of their corpus on the quality of education, including its improvement. The researchers begin by describing the

relevance of the work, including emphasizing the importance of primary education in the formation of basic literacy skills in adolescents. At the same time, the authors emphasize that students who are provided with educational materials that do not match their intellectual capabilities experience difficulties in understanding and assimilating the material, which reduces their interest in learning. The main goal of the article is to create an educational corpus based on primary school textbooks (grades 1-4). The authors describe in detail the process of creating a corpus, where school textbooks approved by the Ministry of Preschool and School Education of Uzbekistan were used. Such text processing methods as tokenization, punctuation removal, and term frequency counting (TF-IDF) were used as methods for processing primary (raw) information. An algorithm was also developed to create a dictionary of unique words, which is used to assess the vocabulary of students at each stage of learning.

Moreover, the study found that a first-grade student must master about 7,980 unique words, while by the end of the fourth grade this figure increases to 24,730 words. In addition, the authors also conducted a comparative analysis of existing works, where one of the main tasks was the creation of (educational) corpora. It should be noted that of the existing works on corpus development, not all are focused on the school audience, which undoubtedly adds to the relevance of the work. Despite the positive aspects of the work, it has certain shortcomings, in particular, the proposed algorithms are implemented on the basis of traditional methods (based on rules). This approach undoubtedly works well in basic text processing tasks (stemming, TF-IDF), but in more complex ones (identification of named entities, semantic analysis, allusions, etc.) its efficiency will be quite low. The main reason for this is that the proposed algorithm works individually with each word in a sentence, while it is necessary to analyze the entire sentence.

The authors of the article[6] examine allusions in Uzbek and English languages, and also conduct a brief comparative analysis. The purpose of the article is to compare the use of allusions in poetry of British and Uzbek literatures. The article emphasizes the importance of understanding the cultural context to fully perceive the meaning of allusions in literature. Moreover, the article emphasizes that language not only serves as a means of communication, but also embodies the history and spiritual view of a particular people.

In addition, the article provides examples of allusions in English literature, demonstrating how these references enhance the depth and understanding of the poem. However, the researchers also emphasize the risk that readers may not understand the meaning of certain fragments of works where allusions are present, and this, in turn, can alienate readers. In Uzbek literature, there is a similar concept known as "Talmeh", which functions as an allusion, referring to famous historical events, myths or literary works. One such example is the famous love stories "Farhad and Shirin" and "Leyli and Majnun". These allusions refer to vivid images and deep understanding in readers familiar with these stories.

The article concludes that allusions can create difficulties in ensuring universal understanding of literary works. Despite the detailed analysis of allusions in the Uzbek and English languages, the authors conduct a theoretical study without addressing the issues of their automatic detection by means of information technology.

IV. Proposed Solution

To implement the algorithm for identifying allusions in the Uzbek language, an approach based on the use of a dictionary of allusions and sequential search in the text is proposed. This method does not involve the use of machine learning and relies on deterministic steps and morphological analysis.

The proposed algorithm works with a dictionary containing 10 thousand manually annotated sentences, which include more than 58 thousand words. Words are annotated according to their belonging to a particular part of speech, the type of named entity (person, organization, time, etc.), and whether it is (part of) an allusion. In addition, during working process algorithm also uses 2nd dictionary, which consist of 80 thousand manually annotated words (most of them are roots).

The sentences in the corpus were manually tagged using the BIOES method[17], where each word or phrase is assigned one of three tags:

- B-ALL (Begin-Allusion): the beginning of the allusion,
- I-ALL (Inside-Allusion): the inner part of the allusion,
- E-ALL (Ending-Allusion): the final part of the allusion,
- S-ALL (Single-Allusion): the allusion, which consist of single word,
- (Outside): other words not related to the allusion.

For example: "Bu hikoya ham bir Layli-Majnun qissasiga o'xshaydi" (This story is also similar to the story of Layli and Majnun).

The marking will occur in the following order: Bu [O] hikoya [O] ham [O] bir [O] Layli [B-ALL] - Majnun [I-ALL] qissasiga [E-ALL] ukhshaydi [O]

In this example, "Layli-Mazhnun qissasiga" (the story of Layli and Majnun) is marked as an allusion, where "B-ALL" indicates the beginning of the allusion, "I-ALL" is the continuation of the allusion, "E-ALL" is the end of the allusion, and "O" indicates words that are not related to the allusion.

The algorithm works as follows:

1. The text to be analyzed is fed to the algorithm.
2. The text is broken down into an array of sentences, and then each sentence is broken down into an array of words.
3. Each word in the text is compared with the dictionary. Example: If the dictionary contains the allusion "Farhad and Shirin", then when the word "Farhad" is found, the algorithm checks whether it is followed by the word "and", and then "Shirin" to confirm the presence of the allusion. The process continues until all the words in the allusion are found in the correct order.
4. If the word is not found in the dictionary, morphological analysis is performed to determine the stem of the word and its grammatical characteristics (e.g. root, prefixes, suffixes). After the analysis, the word is re-searched in the dictionary, taking into account possible morphological changes. If the word is not found, it is marked as "unrecognized".
5. If an allusion is successfully identified, each word in it is marked by part of speech (e.g. noun, verb, adjective) and type of named entity (e.g. person, place, organization). Example: "Farhad and Shirin" - "Farhad" (name, person), "and" (conjunction, is not a named entity), "Shirin" (name, person).

6. The result is a list of detected allusions and their classification, as well as the part of speech and type of named entity that each word belongs to (as in the previous step).

The algorithm's operation diagram is shown in Fig. 1.

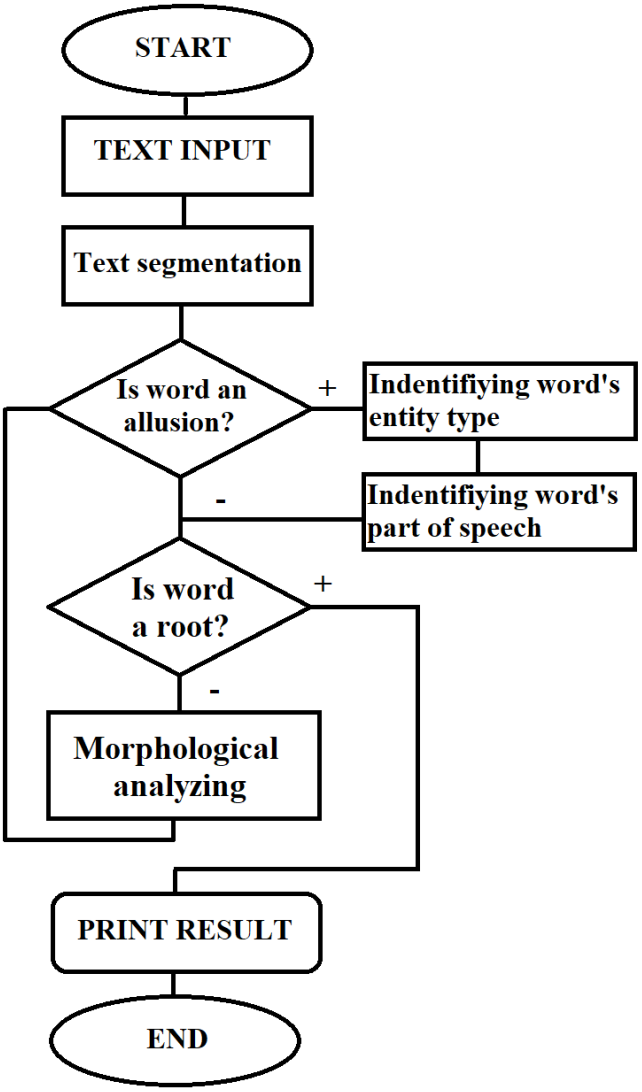

Fig. 1. Scheme of algorithm's work.

V. Testing and Results of the Algorithm

The authors conducted a series of experiments to determine the efficiency of the algorithm, in particular, by preparing two datasets, each of which contained allusions of different topics. The first dataset contains allusions from cultural, poetic and historical topics. The second dataset contains allusions from topics such as jurisprudence, politics and healthcare. It should be noted that the language model was trained for the topics specified in the first dataset. However, the authors were interested in testing the algorithm for other topics. Each dataset contains 100 sentences, each of which contains 1 allusion.

The metrics chosen to evaluate the efficiency of the algorithm were accuracy and recall. Accuracy in this case is the ratio of the number of correctly identified allusions to the number of identified allusions. Meanwhile, recall is the ratio of identified allusions to the number of all allusions in the text.

As a result of testing, the algorithm was able to identify allusions in the first dataset with high accuracy, achieving 100% accuracy and 100% recall. The explanation for such a high accuracy and completeness rate is that all the words and phrases are present in the first dictionary that contains these allusions. Moreover, when analyzing the second dataset, the algorithm showed less positive results, in particular, achieving 81% accuracy and 98% recall. The result of such a decline was a change in the composition of sentences, namely, the allusions themselves.

The second dataset contains allusions of a completely different nature. For example, for the field of jurisprudence: "Haqiqat ko'rinishidan 'Hammuropi kodeksi' yoliga o'tmoqda".

Translation: "Truth moves along the path of the 'Code of Hammurabi'".

Explanation: An allusion to the "Code of Hammurabi" implies an appeal to ancient principles of justice and fairness, emphasizing strict and possibly outdated methods of law enforcement or decision-making.

Although, as you can see, the algorithm was still able to detect most of the allusions. Although, thanks to different of the sentences, algorithm made a mistake during detecting words like allusions. For example: "Men kecha 'Layli-Mazhnun qissasi' nomli shirinliklar do'konini topdim".

Translation: "Yesterday I found a shop, which's name 'Layli-Mazhnun qissasi'".

In dictionary 'Layli-Mazhnun qissasi' phrase is saved as allusion, though in the sentence above it is not about allusion, it is just name of the shop. Neverthless, it should be noted, that this is exception, and this kind exceptional cases are not so much.

More details on the testing results can be found in Table I.

TABLE I. Result of Model's Testing

Number of dataset	Correctly detected allusions / Detected allusions	Precision	Identified allusions / total allusions	Recall
1	100 / 100	100%	100 / 100	100%
2	81 / 98	82,6%	98 / 100	98%

VI. Conclusion

As part of the study, the authors developed an algorithm for identifying allusions in Uzbek texts. The proposed solution is based on dictionary-based approach, which uses two big dictionaries to correctly analyze and recognize allusions in data. The first dataset built from 10,000 sentences, which consist of over 58 thousand manually annotated words. In addition, each word marked by named entity type, part of speech and allusion. As for structure of tagging words and sentences, authors used the BIOES approach, which has proven itself in many similar tasks.

Moreover, authors created second dictionary, which has over 80 thousand wordforms, most of which are root words, and they are annotated by part of speech. In case the algorithm does not find words from first dictionary, the morphological analysis is launched for undetected word thanks to using this dictionary. After stemmatization the word is searched againg from the first dictionary. This approach provides additional opportunity to find necessary elements from text.

Besides, authors conducted an experiment to test the algorithm, where two datasets were used, each of which has sentences and allusions on a specific topic. As a result of testing, the algorithm showed the highest indicator with a sample consisting of sentences and allusions on cultural, historical and poetic topics. At the same time, the accuracy reached 100%, and the recall 100%. Meanwhile, when analyzing the second dataset, the indicators dropped to 81%

accuracy and 98% recall, which is explained by a completely different topic of the content.

In addition, the authors conducted a comparative analysis of existing works, where the main advantages and disadvantages of each similar work were identified.

In the future, it is planned to increase the volume of the corpus for training the model in other topics, for example, in the field of healthcare, law, etc. Moreover, it is planned to use neural network architectures, in particular LSTM, etc.

REFERENCES

[1] T.Saleem, "A Comparative Study of Allusions in the Poetry of English Poet John Milton and Persian Poet Hafiz Sherazi", Journal of Education and Practice, vol. 6, issue 7, 2015.

[2] F. Rakhmatov, "Linguoculturological and semantic features of poetic terms in English and Uzbek languages", Science and Innovation International scientific journal, vol. 4, 2022.

[3] M. Galieva, "Conceptual essence of allusio in the fictional text", Bulletin of Science and Practice, vol. 6, issue 11, 2020.

[4] 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.

[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. Xoshimova, "Comparative analysis of allusions in two languages (Uzbek and English)", Academicia Globe: Inderscience Research, vol. 2, issue 6, 2021.

[7] 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.

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

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

[10] D. Mengliev, V. Barakhnin, M. Eshkulov, B. Palvanov, N. Abdurakhmonova, S. Khamraeva, "Dictionary-Based Medical Text Analysis in Uzbek: Overcoming the Low-Resource Challenge", IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine, pp. 85-89, September 2023.

[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] 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] M. Sharipov, J. Mattiev, J. Sobirov, R. Boltayev, "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.

[15] B. Elov, Sh. Khamroeva, R. Alayev, Z. Khusainova, U. Yodgorov, "Methods of processing the uzbek language corpus texts", International Journal of Open Information Technologies, vol. 11, no. 12, 2023.

[16] 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.

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