UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts

[1]Sharipov Maksud, [2]Kuriyozov Elmurod, [1]Yuldashov Ollabergan, [1]Sobirov Ogabek
[1]Urgench State University, Department of Computer Science, 220100, Urgench, Uzbekistan
[2]Universidade da Coruña, CITIC; Campus de Elviña, A Coruña, 15071, Spain
[1]{m.sharipov, elmurod1202, yuldoshev.o, sobirov.o}@urdu.uz
[2]e.kuriyozov@udc.es

Abstract

Verb detection is a fundamental task in natural language processing that involves identifying the action or state expressed by a verb in a sentence. However, in the Uzbek language, verb detection is challenging due to the complexity of its morphology and the agglutinative nature of the language. This paper proposes a rule-based approach for verb detection in Uzbek texts based on affixes/suffixes. Our method is based on a set of rules that capture the morphological patterns of verb forms in the Uzbek language. We evaluate the proposed approach on a dataset of Uzbek texts and report an F1-score of 0.97, which outperforms existing methods for verb detection in Uzbek language. Our results suggest that rule-based approaches can be effective for verb detection in Uzbek texts and have potential applications in various natural language processing tasks.

Keywords: Verbs, Affixes, Suffixes, FSM, Rule-based, NLP, Uzbek NLP, Verb detection

1. Introduction

Verb detection is a crucial task in natural language processing that involves identifying the action or state expressed by a verb in a sentence. In the Uzbek language, verb detection poses several challenges due to the complex morphology and agglutinative nature of the language. The Uzbek language is spoken by over 30 million people worldwide, making it one of the most widely spoken languages in Central Asia. Thus, the development of effective methods for verb detection in Uzbek texts is of great importance for various natural language processing tasks, such as machine translation, text summarization, and sentiment analysis.

Uzbek language. Uzbek is a Turkic language spoken primarily in Uzbekistan and is also spoken by significant communities in neighbouring countries such as Afghanistan, Kazakhstan, Kyrgyzstan, and Tajikistan (Allaberdiev et al., 2024). It is an agglutinative language, meaning that it uses affixes and suffixes to convey meaning and grammatical information. The Uzbek language has a rich history, with its origins tracing back to the Chagatai language of the Turkic Khaganate. Over time, it has been influenced by various languages such as Persian, Arabic, and Russian. The Uzbek language has its unique script known as the Latin-based Uzbek alphabet, which was adopted in the 1990s to replace the Cyrillic-based script used during the Soviet era. Uzbek is an important language for Central Asia and has become a significant language for international trade and diplomacy in the region.

2. Related work

Several studies have been conducted on verb detection in the Uzbek language, using various approaches such as machine learning, rule-based, and hybrid methods. For instance, the paper discusses the important significance that morpheme analysis plays in the modelling of Uzbek's grammatical categories for parts of speech in machine translation. The article demonstrates the modelling of grammatical categories based on forms, the boundary of syntactic attitudes, and the combinations of affixes in verb forms (Abdurakhmonov, 2017). The article presents a sophisticated web application created for the morphological analysis of Uzbek language words. The idea behind the online application is word form generation and stem analysis in the Uzbek language (Mengliev et al., 2021). However, to the best of our knowledge, no study has focused on the use of rules based on affixes/suffixes for verb detection in Uzbek texts.

2.1 Other works in the field of NLP

The methodology is proposed for the stemming of the Uzbek words with an affix stripping approach not including any database of the normal word forms of the Uzbek language (Sharipov & Yuldashov, 2022). According to Uzbek, an agglutinative language can be designed with finite state machines (FSMs)[1] (Sharipov & Salaev, 2022). This newly presented dataset and tagger tool can be used for a variety of natural language processing tasks such as language modelling, machine translation, and text-to-speech synthesis (Sharipov et al., 2022). The main purpose of the work is to remove affixes of words in the Uzbek language utilizing the finite state machine and to identify a lemma of the word (Sharipov et al., 2022; Sharipov & Sobirov, 2022). The article deals with the automatic deletion of stop words for the Uzbek text and, if necessary, its return to the original text (Madatov et al., 2022, 2023). To our knowledge, the majority of human language processing technologies for low-resource languages don't have well-established linguistic resources for the development of sentiment analysis applications (Matlatipov et al., 2022). Machine transliteration, as defined in this paper, is a process of automatically transforming the written script of words from a source alphabet into words of another target alphabet within the same

[1] FSM - is a mathematical model of computation.

LREC-COLING 2024, pages 17343–17347

language, while preserving their meaning, as well as pronunciation (U. Salaev et al., 2022). Since verbs are the heart and soul of English sentences, it is crucial to correctly identify English verb grammatical problems (He & others, 2021). As in all other fields, linguistics is accelerating the process of adapting to digital technologies. As a result, computer programs and information systems must process the traditional linguistic conventions of natural language. Tagging speech segments is one of these crucial NLP tasks (Abdurashetona & Ismailovich, 2021). The article created a mathematical model to represent the morphemic description of the Uzbek language's noun structure. All word-forming suffixes, affixes, and word forms that fall under the lexical category of nouns are taken into account by the proposed model (Bakaev & Shafiyev, 2020). The relationship between the author's corpus and lexicography, the function of concordance in the growth of the author's lexicography, the significance of the author's language dictionary, and the role of the dictionary article in the case design will all be examined in this article (Khamroeva, 2019). The unique methodologies for the Bengali morphological synthesis of the verb, pronoun, and noun systems have been given in this study. The systems explained here are platform-neutral and have been implemented in Java. Many randomly chosen words were used to test the systems' performance, and the results showed that they performed fairly accurately overall (Bhattacharya, 2023).

3. Methodology

We propose a rule-based approach for verb detection in Uzbek texts based on affixes/suffixes. Our approach relies on a set of rules that capture the morphological patterns of verb forms in Uzbek language. We identified the most common suffixes and prefixes used for verbs in the Uzbek language and developed a set of rules that specify the conditions for identifying a verb based on these affixes/suffixes. The rules were implemented using Python programming language and the Natural Language Toolkit (NLTK) library. Uzbek verbs are complex (Mattiev et al., 2023), so we have created a special database for Uzbek verbs. Verbs are one of the Uzbek language's most important and complex parts. Therefore, the verb affects and controls the entire sentence. This work creates an opportunity for new scientific research among the NLP models and software tools being created for the Uzbek language.

- Root verb – *tup fe'l*
- Artificial verb – *yasama fe'l*

When identifying verbs from Uzbek texts, we have created a database of root verbs, because the root verb is not formed with suffixes, so it cannot be identified using FSMs. To identify artificial verbs, we create separate FSMs for *yuklama* (particle), *shaxs-son* (person-number), *zamon* (tense), *ravishdosh* (adverb), *sifatdosh va harakat nomi* (adjective and action noun), *egalik* (possessive), *kelishik* (agreement), *nisbat* (relative) and *so'z yasovchi* (word-formative).

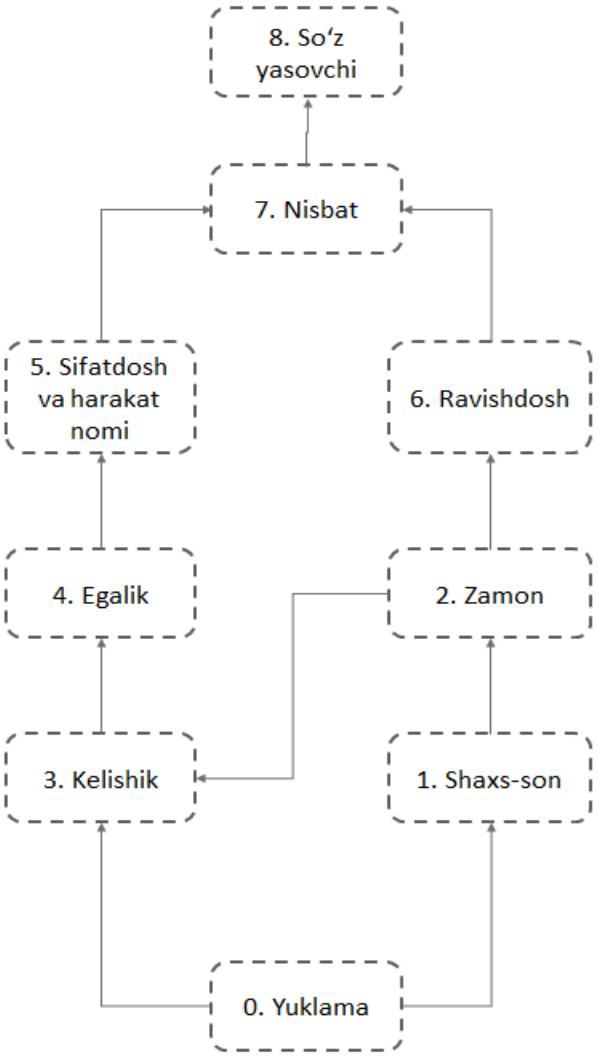

Figure 1: FSM to identify the verb with affixes

3.1 Step-by-step detection algorithm

Verb detection by affixes is shown in Figure 1, where there are 9 FSMs. Creating each FSM consists of 5 steps. Now let's create an FSM that determines the verb for relative (7 in Figure 1).

Step 1. Create FSM (left to right) to search for relative.

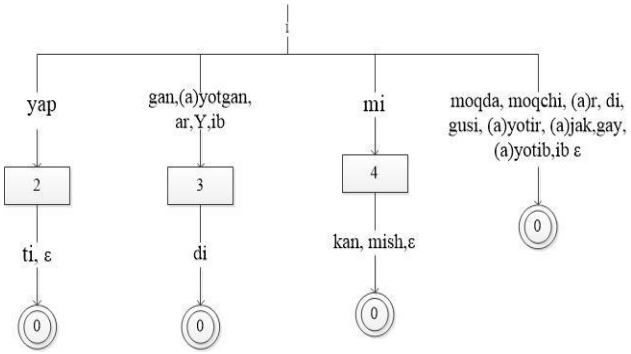

Figure 2. Relative FSM (left to right)

Step 2. Sorting out relative affixes.

1. -yap	13. -kan
2. -ti	14. -mish
3. -gan	15. -moqda
4. -yotgan	16. -moqchi
5. -ayotgan	17. -gusi
6. -r	18. -yotr
7. -ar	19. -ayotr
8. -a	20. -jak
9. -y	21. -ajak
10. -ib	22. -gay
11. -di	23. -yotib
12. -mi	24. -ayotib

Table 1. Teble of relative affixes

Step 3. Searching for relative is a non-deterministic finite automaton (NFA).

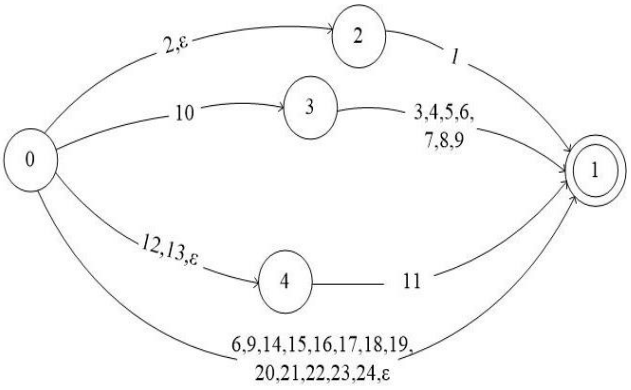

Figure 3. NFA

Step 4. Transition from non-deterministic finite automata (NFA) to deterministic finite automata (DFA).

A={0,1,2,3,4} “1,2” : T={1,2,3} B “8,9,11,12” : T={1,2,3,6} C “20” : T={1,3,4} D “3,4,5,6,7,10,13-19” : T={1} E “19” : T={1} E
B={1,2,3} “1,2” : T={1,2} F “3-19” : T={1} E
C={1,2,3,6} “1,2” : T={1,2,5} I “3-19” : T={1} E
D={1,3,4} “1,2” : T={2,3} G “8,9,11,12” : T={1} B “17” : T={3} H
F={1,2} “1-19” : T={1} E I={1,2,5} “1-19” : T={1} E
G={2,3} “1,2” : T={1,2} F “3-19” : T={1} E
H={3} “1,2” : T={2} J
J={2} “1-19” : T={1} E

Table 2. From NFA to DFA

Step 5. Create FSM (right to left) to search for ralative.

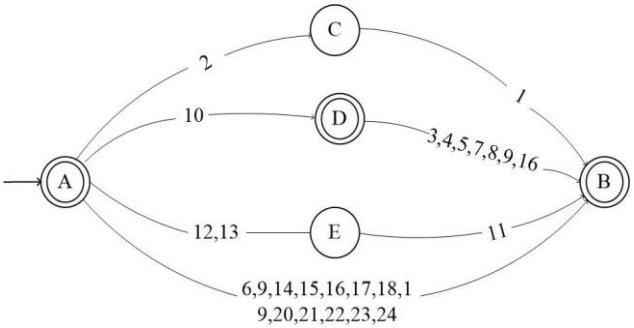

Figure 4. FSM (left to right)

Also, we found such exceptions that words that do not belong to the verb word classes can also find affixes in this FSM. Because *sifatdosh va harakat nomi* (adjectives and action noun) verbs are the forms adapted to the noun stem, they receive syntactic and lexical suffixes of the noun.

For instance, *kitob-imiz(4)-ni(3)-mi(0)* here,

- *-mi* ∈ 0.Yuklama
- *-ni* ∈ 3.Kelishik
- *-imiz* ∈ 4.Egalik

Below are more such exceptions:

1. *qalam-imiz-ni-mi* (0,3,4)
2. *bir-ga-miz-a* (0,1,3)
3. *o‘yin-i-ni* (3,4)
4. *qizil-mi* (0)
5. *maktab-ga* (3)

…

4. Results and Discussion

Our approach achieved an F1-score of 0.97. Our approach also achieved higher precision and recall values than the previous method. The results demonstrate the effectiveness of our rule-based approach for verb detection in Uzbek texts based on affixes/suffixes.

The results suggest that rule-based approaches can be effective for verb detection in Uzbek texts. Our

approach based on affixes/suffixes outperformed the previous rule-based method based on suffixes and prefixes. The proposed approach has potential applications in various natural language processing tasks, such as text classification, named entity recognition, and sentiment analysis, where verb detection plays a crucial role. However, the proposed approach has some limitations (U. I. Salaev et al., 2023). Firstly, it heavily relies on the morphological patterns of verbs and may not work well for irregular verbs or rare verbs that do not follow the typical morphological patterns. Secondly, the proposed approach may not generalise well to other agglutinative languages with different morphological patterns.

To check the accuracy of the developed algorithm, we calculated the results of identifying Uzbek verbs in a corpus of 25 categories or 25,000 words. The results are presented in the table below.

№	File name	Number of words	Verb		Not verb		F1 score
			Verb	Not verb	Verb	Not verb	
1	Biology	1003	164	4	0	835	0.99
2	Literature	999	236	11	7	745	0.96
3	Anatomy	1021	184	2	2	833	0.99
4	Botany	1012	177	6	0	829	0.98
5	History of religion	1015	162	3	2	848	0.98
6	World	1006	179	1	8	818	0.98
7	Physics	1000	238	11	11	740	0.96
8	Geography	977	167	7	3	800	0.97
9	Law	989	147	4	10	828	0.95
10	Informatics	1005	250	4	2	749	0.99
11	Economy	1027	194	4	4	825	0.98
12	Society	1003	172	1	1	829	0.99
13	Chemistry	995	162	11	5	817	0.95
14	Culture	1000	162	4	3	831	0.98
15	Mathematics	999	192	2	14	791	0.96
16	Music	1000	159	4	4	833	0.98
17	Mother tongue	1012	177	17	2	816	0.95
18	Agriculture	1006	154	10	2	840	0.96
19	Legislation	1000	144	1	0	855	0.99
20	Politics	1030	152	8	1	869	0.97
21	Sports	1008	169	5	2	832	0.98
22	History	1005	171	5	3	826	0.98
23	Technology	1005	239	4	9	753	0.97
24	Medicine	1013	165	0	5	843	0.99
25	Zoology	1012	212	4	1	795	0.99
TOTAL:		**25142**	**4528**	**133**	**101**	**20380**	**0.97**

Table 3: F1-score results

In our model, we used F1 estimation because class was not balanced. The F1 estimate was calculated using the following formula:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

Here, TP, FP and FN are given in the Confusion matrix (Figure 5).

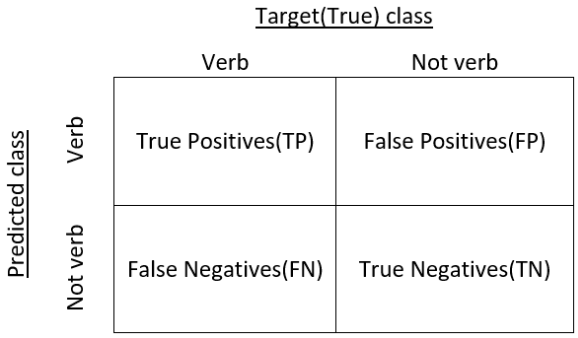

Figure 5. F1-score confusion matrix

5. Conclusion and Future work

In this paper, we proposed a rule-based approach for verb detection in Uzbek texts based on affixes/suffixes. The proposed approach outperformed existing methods for verb detection in Uzbek language and demonstrated the potential of rule-based approaches for natural language processing tasks in Uzbek language. The results suggest that morphological patterns of verbs can be effectively captured using rule-based approaches, which can be applied to other agglutinative languages with similar morphological patterns.

Future work can explore the use of machine learning approaches, such as deep learning and transfer learning, for verb detection in Uzbek language. Such methods can be used to capture more complex patterns and improve the generalization capability of the model. Moreover, the proposed rule-based approach can be extended to handle other parts of speech, such as nouns and adjectives, which also exhibit complex morphological patterns in Uzbek language. Finally, the proposed approach can be

integrated into existing natural language processing pipelines for the Uzbek language and evaluated in real-world applications, such as machine translation and text summarization.

6. Bibliographical References

Abdurakhmonov, N. (2017). Modeling analytic forms of verb in Uzbek as stage of morphological analysis in machine translation. *Journal of Social Sciences and Humanities Research*, *5*(03), 89–100.

Abdurashetona, A. M., & Ismailovich, I. O. (2021). Methods of Tagging Part of Speech of Uzbek Language. *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 82–85.

Allaberdiev, B., Matlatipov, G., Kuriyozov, E., & Rakhmonov, Z. (2024). Parallel texts dataset for Uzbek-Kazakh machine translation. *Data in Brief*, 110–194. https://doi.org/https://doi.org/10.1016/j.dib.2024.110194

Bakaev, I., & Shafiyev, T. (2020). Morphemic analysis of Uzbek nouns with Finite State Techniques. *Journal of Physics: Conference Series*, *1546*, 12076. https://doi.org/10.1088/1742-6596/1546/1/012076

Bhattacharya, S. (2023). *Inflectional morphology synthesis for bengali noun, pronoun and verb systems*.

He, Y., & others. (2021). Automatic Detection of Grammatical Errors in English Verbs Based on RNN Algorithm: Auxiliary Objectives for Neural Error Detection Models. *Computational Intelligence and Neuroscience*, *2021*.

Khamroeva, S. (2019). The author's lexicography and author's corpus approach. *International Journal of Applied Research*, 26–29.

Madatov, K., Bekchanov, S., & Vičič, J. (2022). Accuracy of the Uzbek Stop Words Detection: a Case Study on "School Corpus." *CEUR Workshop Proceedings*, *3315*, 107 – 115.

Madatov, K., Bekchanov, S., & Vičič, J. (2023). Automatic Detection of Stop Words for Texts in Uzbek Language. *Informatica*, *47*(2).

Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek Sentiment Analysis Based on Local Restaurant Reviews. *CEUR Workshop Proceedings*, *3315*, 126–136. www.scopus.com

Mattiev, J., Salaev, U., & Kavsek, B. (2023). Word Game Modeling Using Character-Level N-Gram and Statistics. *Mathematics*, *11*(6), 1380.

Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. *Applied Sciences*, *11*(19), 9117.

Salaev, U. I., Kuriyozov, E. R., & Matlatipov, G. R. (2023). Design and Implementation of a Tool for Extracting Uzbek Syllables. *Proceedings of the 2023 IEEE 16th International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering, APEIE 2023*, 1750 – 1755. https://doi.org/10.1109/APEIE59731.2023.10347773

Salaev, U., Kuriyozov, E., & Gómez-Rodríguez, C. (2022). A Machine Transliteration Tool Between Uzbek Alphabets. *CEUR Workshop Proceedings*, *3315*, 42 – 50. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146119140&partnerID=40&md5=be670d829670d883b2f8326559ce954a

Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language. *CEUR Workshop Proceedings*, *3315*, 93 – 98. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146120658&partnerID=40&md5=dfdea47deba344b3df4ed2372856847b

Sharipov, M., & Sobirov, O. (2022). Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language. *CEUR Workshop Proceedings*, *3315*, 154 – 159.

Sharipov, M., & Yuldashov, O. (2022). UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. *ArXiv Preprint ArXiv:2210.16011*.