

UzbekVerbDetection

Rule-based Detection of Verb in Uzbek Texts

LREC-COLING 2024 22-24 MAY

AUTHORS

Sharipov Maksud

Kuriyozov Elmurod

Yuldashov Ollabergan

Sobirov Ogabek



Table of contents

01

Introduction

Verb detection is a crucial task in NLP

02

Related work

Other works in the field of NLP

03

Methodology

We propose a rule-based approach for verb detection in Uzbek

04

Step-by-step

Step-by-step detection algorithm

05

Analysis

Our approach achieved an F1-score of 0.97.

06

Conclusions


Future work can explore the use of machine learning approaches



01

Introduction

Verb detection is a crucial task in NLP

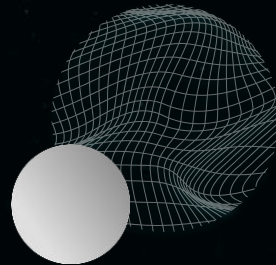


Purpose statement

~30,000,000

The Uzbek language is spoken by over 30 million people worldwide, making it one of the most widely spoken languages in Central Asia.

Uzbekistan, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, and Afghanistan





Uzbek is a Turkic language spoken primarily in Uzbekistan

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.

Navoi is a great Uzbek poet, a representative of the Uzbek literature which is called Chagatai literature in the West.



02

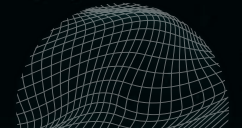
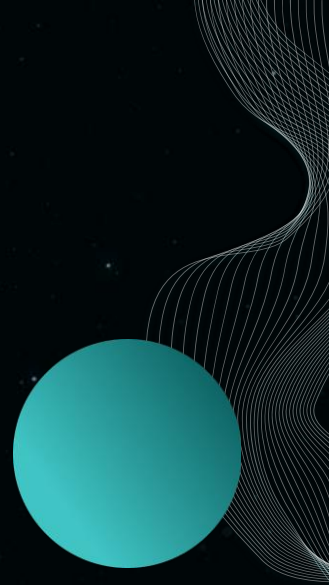
Related work

Other works in the field of NLP



Literature review

- 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*.
- Bakaev, I., & Shafiyev, T. (2020). Morphemic analysis of Uzbek nouns with Finite State Techniques. *Journal of Physics: Conference Series*, 1546, 12076.
- 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.
- Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: The rule-based POS tagger for Uzbek language.




A decorative graphic on the left side of the slide, consisting of a white grid pattern that curves and warps, with a solid teal circle positioned in the upper left area.

03

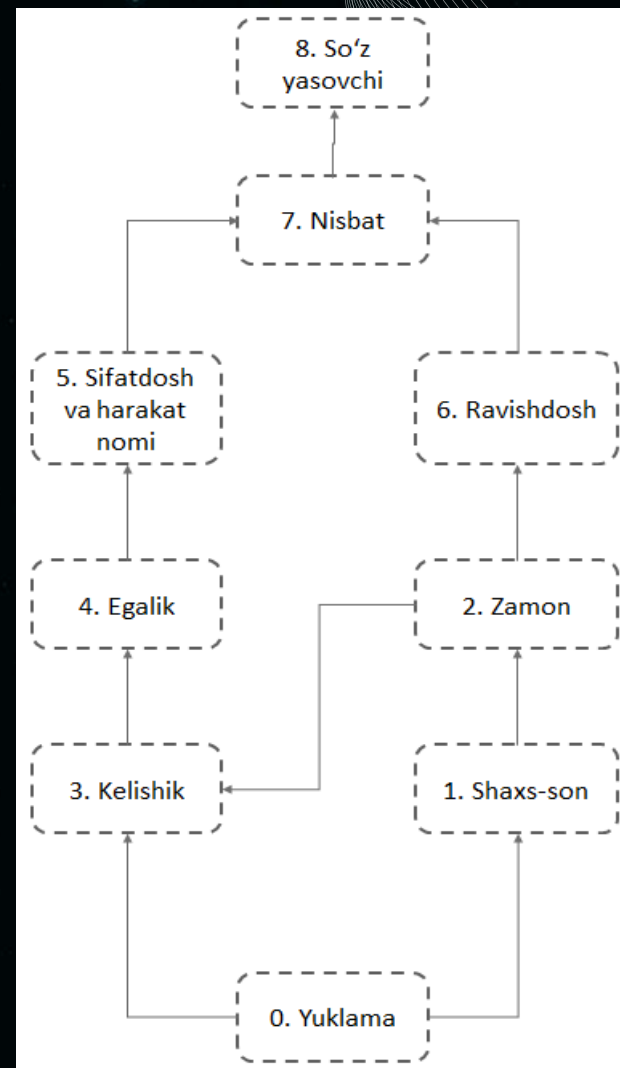
Methodology

We propose a rule-based approach for verb
detection in Uzbek

A decorative graphic on the right side of the slide, consisting of a white grid pattern that curves and warps, with a solid purple circle positioned in the upper right area.

Methodology

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 *soʻz yasovchi* (word-formative).



A decorative graphic on the left side of the slide, consisting of a white grid pattern that curves and warps, with a solid teal circle positioned in the upper left quadrant.

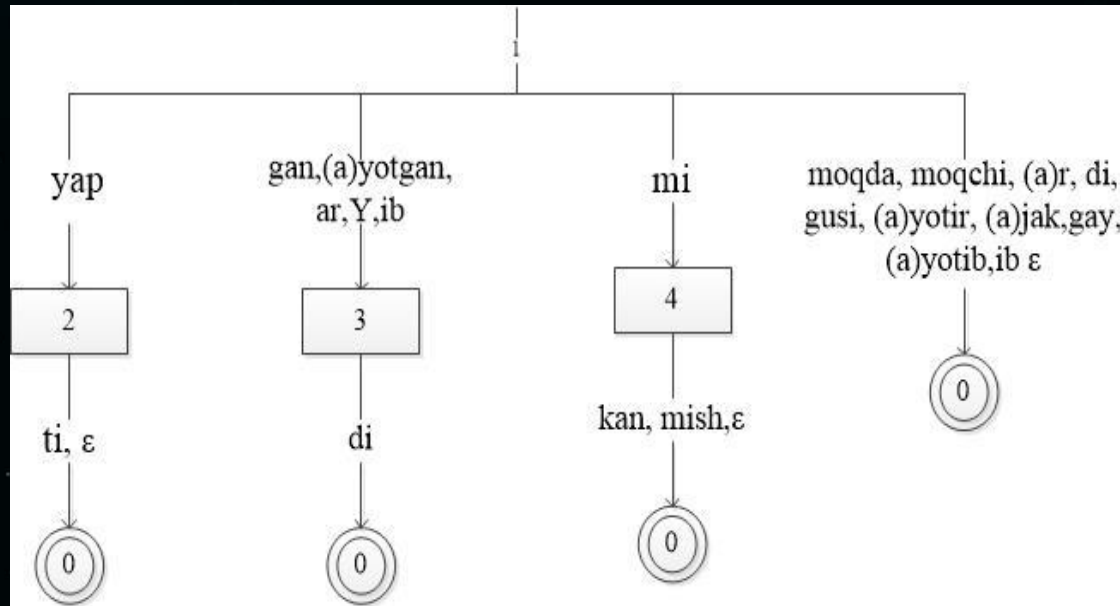
04

Step-by-step

Step-by-step detection algorithm

A decorative graphic on the right side of the slide, consisting of a white grid pattern that curves and warps, with a solid purple circle positioned in the upper right quadrant.

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

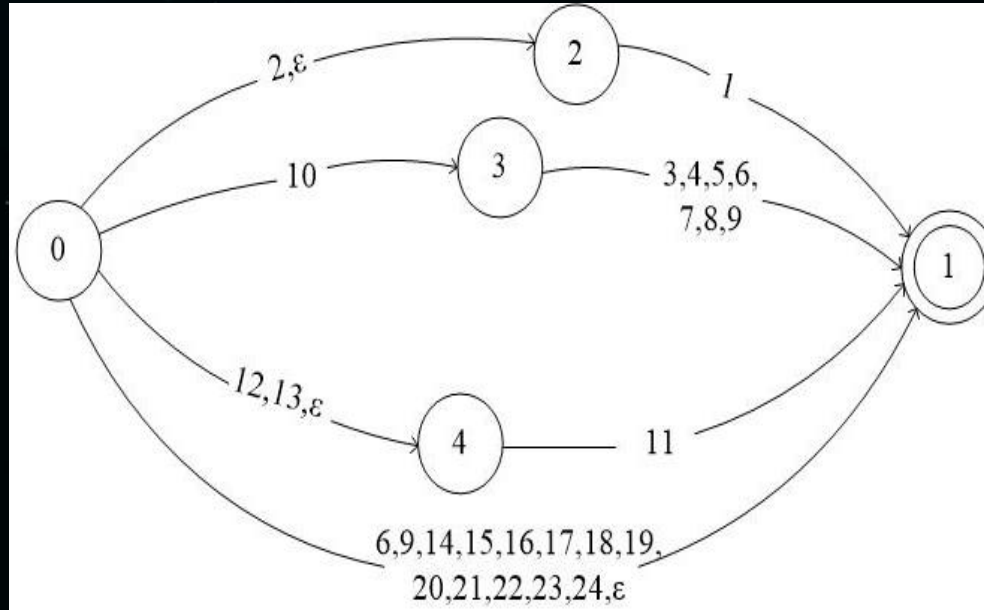


Step 2. Sorting out relative affixes.

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



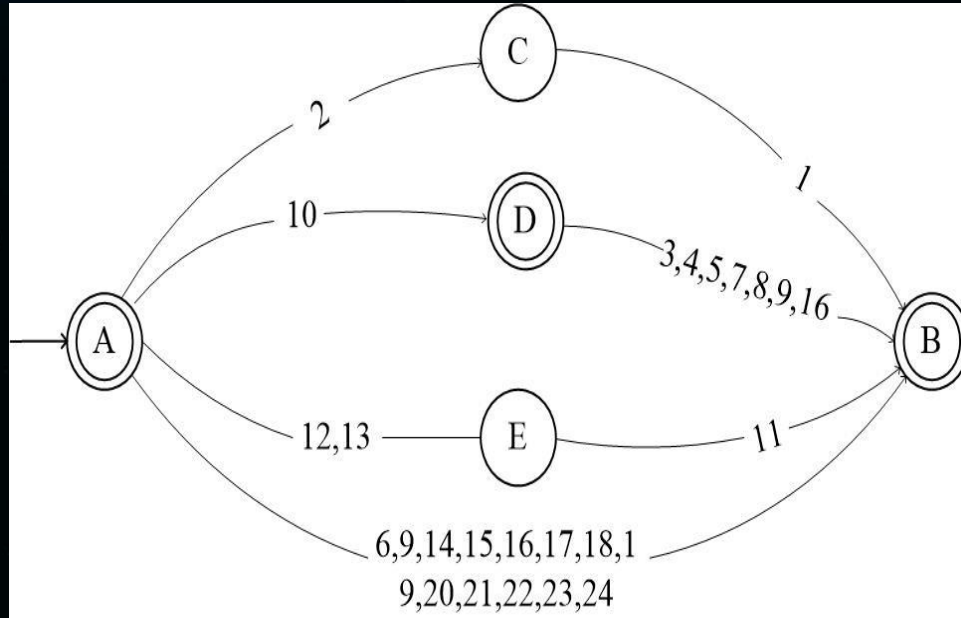
Step 3. Searching for relative is a non-deterministic finite automaton (NFA).



**Step 4. Transition
from non-
deterministic
finite automaton
(NFA) to
deterministic
finite automaton
(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

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





05

Analysis

Our approach achieved an F1-score of 0.97.



Results and Discussion



25



~25,000



0,97

№	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

25	Zoology	1012	212	4	1	795	0.99
TOTAL:		25142	4528	133	101	20380	0.97


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.

A decorative graphic on the left side of the slide, featuring a white wireframe grid that curves and warps, with a solid teal circle positioned in the upper left quadrant.

06

Conclusions

Future work can explore the use of machine learning approaches

A decorative graphic on the right side of the slide, featuring a white wireframe grid that curves and warps, with a solid purple circle positioned in the upper right quadrant.

Conclusions and Future work

01

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.

03

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.

02

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

04

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.

A decorative graphic on the left side of the slide. It features a white wireframe grid that is distorted into a wavy, undulating shape. A solid teal circle is positioned in the upper-left quadrant of this grid.

THANKS!

QUESTIONS?

A decorative graphic on the right side of the slide. It features a white wireframe grid that is distorted into a wavy, undulating shape. A solid purple circle is positioned in the upper-right quadrant of this grid.