ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 11, Issue 11, November 2024

Punctuation Prediction for the Uzbek Language

Maksud Sharipov Siddiqovich, Hushudbek Adinaev Saylboyevich

PhD. Professor, Department of Computer Science, Urgench State University, Urgench, Uzbekistan.
Assistant Professor, Department of Information technologies, Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Urgench, Uzbekistan.

ABSTRACT: Punctuation prediction plays an important role in natural language processing and written text understanding, a model that can predict the correct punctuation in a number of tasks such as text preprocessing, spell checking, grammar checking, information retrieval, etc. provides the need. The task of predicting the correct punctuation is context dependent, making language-independent generic punctuation tools irrelevant to the task. Although the idea of creating such a tool has already been implemented for many languages, Uzbek is one of the low-resource languages, and to our knowledge, punctuation prediction algorithms for Uzbek texts have yet to be developed. not released. This article proposes a rule-based algorithm and model for predicting punctuation marks in Uzbek texts. Although the main contribution of this paper is a rule-based algorithm for predicting the correct or incorrect placement of full stop (period), commas, question marks, hyphen, colon and exclamation mark in Uzbek text, the authors also present the analysis results on a corpus with different domains.

I. INTRODUCTION

The problem of determining whether periods and commas are correctly or incorrectly placed in texts (punctuation analysis) is one of the problems of natural language processing (NLP). Punctuation analysis is used to solve problems such as machine translation, text analysis, semantic analysis, and syntactic analysis in natural language processing. In the Uzbek language, like many other languages in general, a period (or a full stop) is one of the most used and oldest punctuation marks (.); It appeared as a writing symbol in Arabic texts in the 11th century. Until the 19th century, its function and use were completely different from that of the present punctuation mark. The use of the dot as a symbolic sign goes back to ancient Arabic sources. It has been used as a punctuation mark in Uzbek since the second half of the 19th century. The comma mark has been used as a punctuation mark in Western Europe since the 15th century. We find it in Uzbek texts from the beginning of the 20th century. The comma (,) is one of the most used punctuation marks. The place of use, the form was different in different eras and writings of different languages. In the Uzbek language, the comma is also called "inverted comma", "pesh", "half stop", "half pause". Comma was originally used for a short pause, and later its scope and function expanded.

Many scientists of the world have been engaged in the analysis of punctuation marks, including, until recently, punctuation marks were ignored by most researchers of theoretical and computational linguistics. This is due to the lack of a concise, formal framework for this abstract problem. However, if we remember that punctuation is an orthographic component of the written language, we can see that research on punctuation makes sense. Accordingly, in the last decade, interest in the topic has increased, because it was understood that it is absolutely impossible to understand and process the written language more fully without taking into account punctuation marks. Although punctuation was originally invented as a means of reflecting intonation in written text, it is now a separate system of linguistic punctuation [1].

In this paper, we present a rule-based approach for punctuation analysis and prediction, focusing only on period and comma marks, with wider coverage is coming in the upcoming works. More than 30 rules have been created to analyze and predict the placements of two above-mentioned punctuation marks in Uzbek texts, with its analysis on an Uzbek raw texts corpus with 30K word balanced over 30 fields.

We also present the experiment results of the created model's performance, with accuracy results over all field areas the used Uzbek corpus covers. The experiment results indicate that the created rule-based model is accurate enough to be used in practical applications, also leaving gap for future works like modern approaches with machine learning and deep learning models.

All the rules presented as part of the methodology in this paper, as well as the Python codes used for the punctuation analysis and the experiments are openly available at the project www.punktuatsiya.uz/ repository[1].

The Uzbek Language

Uzbek, a Turkic language (Karluk branch), is the most-spoken language in Central Asia in terms of number of people who use it for communication. Being an official language of Uzbekistan, and second language in neighboring Central Asian countries, Uzbek is a vital language with more than 32 million native speakers. The Turkic literary language Chagatai, employed during the Timurid dynasty, is its ancestor. Uzbek has adopted Persian, Arabic, and Russian lexicon due to historical and cultural influences. After independence in the early 1990s, the language uses the Latin alphabet, with old Cyrillic form still as a popular choice. Uzbek's rich language structure comes from its agglutinative nature, which adds affixes to express meaning. Beyond its language importance, Uzbek is a cultural link that reflects Uzbekistan's unique past and historical connections[2].

II.RELATED WORK

One of the early approaches to analyze and detect punctuation irregularities in the text has been discussed in the case of the Danish language [2], where the authors describe research using the Brill tagger to study the detection of irregular commas in Danish. The proposed model was trained on a 600,000-word part-of-speech labeled corpus. The system is designed by inserting random commas into the text, which are marked as false, while the original comma is marked as true.

Further punctuation analysis models and algorithms were developed for languages of the Indo-European group [1] where the created models not only make use of linguistic rules, but also train machine learning algorithms to successfully analyze the placements of various punctuation marks.

The need for correctly placed punctuation marks grew over time, making it a crucial NLP task as part of a text pre-processing, since the wrong use of a punctuation, especially comma, can change the entire meaning of a sentence, for content-rich languages like Uzbek [3].

One of the very first mentions of punctuation analysis in a form of tagging for the Uzbek language appears in the research from Uzbek Part-of-speech (POS) tagging algorithm [4], where the research work proposes a correct handling and tagging of punctuations in an Uzbek raw text, while lacking the ability to predict them.

The importance of correct placements of punctuation is further needed in other downstream tasks, in the case of Uzbek, it has been researched with several NLP tasks which mention the punctuation analysis, especially the tasks like morphological analyzers, text summarizers benefit a lot from correctly analyzed punctuations [5], [6].

Recent works in Uzbek NLP. Considering the low-resource nature of the Uzbek language, it is a good practice to address some of the vital advancements in the creation of the NLP resources and tools for the language. Uzbek NLP has seen a sharp rise in the NLP research, taking not only itself to the level of well-resourced languages, but also taking the minority language spoken only inside Uzbekistan - Karakalpak to a new level. One of the prominent instances to the case is a recent work that aims at solving the problem of automatic extraction of stop words and text summarization for the low-resource Karakalpak language, which is spoken by about two million people in Uzbekistan [7].

Furthermore, another useful research work is about building a lemmatization algorithm for the Uzbek language. The main goal of the work is to remove affixes of words in the Uzbek language using a finite state machine and to determine the lemma of the word [8].

Since Uzbek is an agglutinative language, many words are formed by adding adverbs, and the variety of adverbs also makes a large dictionary on its own. For this reason, it is difficult to find the root of the word. The methodology proposed to perform the core of Uzbek words by the method of separating affixes, without including a database of simple word forms in the Uzbek language. Word affixes are divided into fifteen classes and constructed as Finite State Machines (FSM) for each class according to morphological rules [8]. The field of sentiment analysis within the framework of the Uzbek language still remains an understudied field of science, despite the progress with researches from above-mentioned works. Despite its importance, the paucity of existing research may be due to limited resources that have prevented in-depth research [9], [10].

[1] https://punktuatsiya.uz/

[2] More about the Uzbek language: https://en.wikipedia.org/wiki/Uzbek_language

III.METHODOLOGY

Since this work only focuses on two punctuation marks for the start: period and comma marks (with a plan to widen the marks' list in the upcoming works), this section will be dedicated to show only some limited number of rules to analyze the correct placement and/or generation of both two marks. The full list is presented in the project repository.

The algorithm behind the created model works as a Python library, becoming one piece with an Uzbek morphological analyzer from [6] as well as the part-of-speech data of word tokens obtained from the Apertium monolingual package for Uzbek[3].

Application of the Period Mark.

There are more than 20 rules that help for the period mark correct placement and generation if missing, and we mention some of them below:

1. General rules that apply to many languages that use period for shortened versions of statuses/degrees (***"Mr.", "Mrs.", "Ph.D.", "M.Sc."***, etc.) or the first letter of aliases (***"I. A. Karimov", "Sh. M. Mirziyoyev"***, etc.);
2. *Before (possible) sentences beginning with the conjunctions like* ***"ammo, lekin, biroq, chunki, shuning uchun, go'yo"*** *("but, however, hence, therefore, as if").*
   Ex.: "Bobo dehqon yerga urug'ni ekish bilan band. ***Chunki*** *dalalarda ish qizg'in" (Elder farmer is busy planting seeds in the ground.* ***Because*** *the work in the fields is intense);*
3. In Uzbek, there are shapes of verbs that add its tense and meaning to the main verb like ***"edi, ekan, emish,…"*** *(was, has been, is told so,…)*, and they come at the end of a sentence, followed by a period mark;
4. There is a form of an Uzbek sentence that ends with the word ***"emas"*** *(not so)* which means the negation of the action or state. This word usually means the end of a sentence, preceding a period mark; *Ex.: "Bu ruchka meniki* ***emas****." (This pen is not mine);*
5. In Uzbek, generally, a sentence ends with adverbs [11];

Application of the Comma Mark.

There are more than a dozen rules of the comma, and we will mention some of them below:

1. To separate the sentence fragments that come after words such as *"****ha, yo'q, rahmat, xo'sh, qani, xayr, ofarin, salom****" (yes, no, thank you, okay, come on, goodbye, well done, hello)* that express confirmation, emphasis, denial and similar meanings, a comma is inserted. Ex.: ma'nolarni bildiruvchi kabi so'zlardan keyin kelgan bo'laklarni ulardan ajratish uchun ham vergul qo'yiladi. *(Yes, being vigilant is the sacred duty of every citizen to the Motherland.);*
2. A comma is placed at the beginning of the second part of connected sentences: ***"yo-yo; na-na; dam-dam; goh-goh, …"*** *(either-or, neither-nor, as-as, …)*. Ex.: "*Yo biz boraylik, yo siz keeling.*"*(Either we will go or you will)*
3. In conjunctions related to opposition or negation, a comma is placed before the conjunctions ***"ammo, lekin, biroq"*** *(but, however, hence)*. Ex.: "*Kechasi yana qor yog'gan, biroq havo unchalik sovuq emas edi.*" *(It snowed again last night, but it wasn't too cold.);*
4. A comma is used when the clauses in the following cases are connected with each other using the following conjunctions *"****chunki, negaki, toki, go'yo****" (because, why, so that, as if)* and auxiliary devices ***"shuning uchun, shu sababli, shu tufayli, shu bois"*** *(therefore, therefore, because of this, therefore)*. Ex.: *Oyna opa xatni oxirigacha o'qiy olmadi, chunki hovlining eshigini kimdir taqillata boshladi. (Sister Oyna could not read the letter to the end, because someone started knocking on the door)*;
5. No commas are placed between the conjunctions connected by means of ***"va, ham, hamda, yoki"*** *(and, also, and, or)* if it comes without repetition.[11]

The process is straightforward, with the core process takes place only at the stage where rules apply for periods and commas. Moreover, the second stage where the tokenization takes place is ambiguous, and in that stage a text is ot only tokenized into words, but also each token is sent to a morphological analyzer, which the returning result will accompany each token as a meta info, so the punctuation rules can judge the nature of each token when analyzing.

The stages of the algorithm are also presented in a pseudo-language at
ALGORITHM 1.

[3] Apertium Uzbek monolingual package for tagging and morphological analysis: https://github.com/apertium/apertium-uzb

IV.EXPERIMENTS AND RESULTS

As an experiment, an Uzbek corpus with 30000 words, balanced over 30 different fields/areas has been chosen, giving about 1000 words in each field to analyze the performance of our model.

The corpus we processed contained 2766 period marks, totaling 6221 punctuation marks in the corpus. The raw text corpus was converted into an annotated dataset using a simple conversion method where each document was kept as it is with its punctuations for evaluation, and the same document was cleared from punctuation, removing periods and commas, used for feeding the text into the model for punctuation prediction.

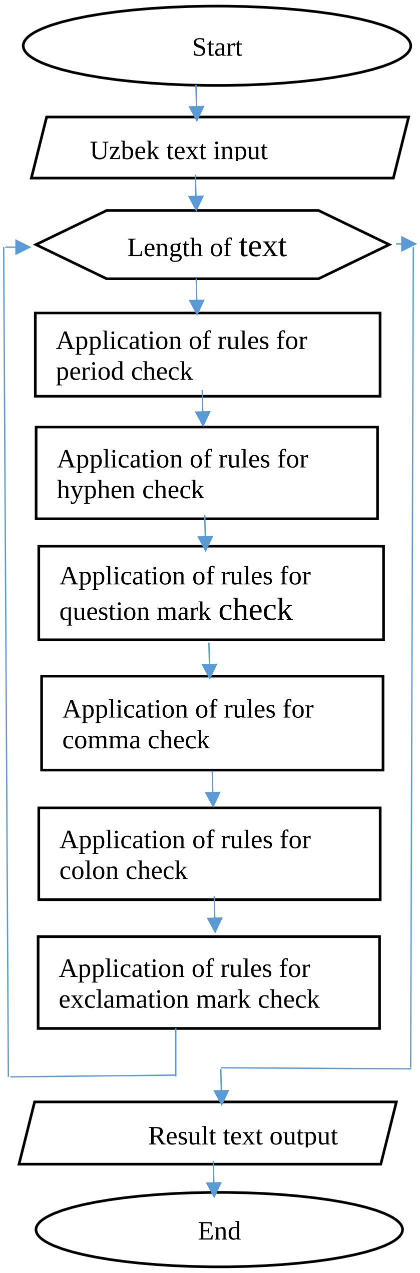
Start
Uzbek text input
Length of text
Application of rules for period check
Application of rules for hyphen check
Application of rules for question mark check
Application of rules for comma check
Application of rules for colon check
Application of rules for exclamation mark check
Result text output
End

Fig. 1. Stages of rule-based punctuation analysis/correction and prediction algorithm.

ALGORITHM 1. Algorithmic representation of the proposed rule-based punctuation analysis/prediction method.

```
procedure RuleBasedPunctuationGeneration(inputText):
    // Start the program
    initialize program
    // Tokenization process
    tokens = tokenize(inputText)
    // Loop through each token in the text
    for each token in tokens:
        // Consult the part program for periods
        if isCorrectPeriodPlacement(token):
            // Period placed correctly
            return to_text(tokens)
        else:
            // Period placed incorrectly, apply correction
           // Consult the part program for commas
        if isCorrectCommaPlacement(token):
            // Comma placed correctly
            continue
        else:
            // Comma placed incorrectly, apply correction
   return to_text(tokens)
…
  // Program completion
end procedure
```

The overall accuracy of the created rule-based model for punctuation prediction was almost 84% percent, showing prominent result.

Detailed reports of the estimation accuracy results for each area/region of the period in the texts are given in Table I. We determine the results of the remaining punctuation marks in the same way.

TABLE I. To check the correctness of the developed algorithm, we calculated the results of identifying Uzbek texts in a corpus of 30 categories or 30,000 words. The results are presented in the table below.

№	Area/Field	Number of period	Period		Not period		F1 score
			Period	Not period	Period	Not period	
1	Literature	83	43	7	7	26	0,86
2	Anatomy	92	46	9	6	31	0,86
3	Biology	79	40	4	11	24	0,84
4	Botany	86	44	4	7	31	0,89
5	Religion	92	48	9	11	24	0,83
6	World	98	49	7	7	35	0,88
7	Physics	78	40	9	10	19	0,81
8	Geography	99	50	6	8	35	0,88
9	Stories	87	44	3	5	33	0,91
10	Law	78	42	9	6	21	0,85
11	IT	97	49	10	4	34	0,88
12	Economy	89	40	4	12	33	0,83
13	Sociology	88	44	6	7	31	0,87
14	Chemistry	101	51	5	9	36	0,88
15	Cinema	95	45	9	9	32	0,83
16	Culture	105	52	11	13	29	0,81
17	Math	98	49	9	7	33	0,86
18	Language	87	48	11	5	23	0,86

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 11, Issue 11, November 2024

19	Agriculture	98	38	9	13	38	0,78
20	Media	87	48	6	8	25	0,87
21	Art	89	30	11	9	39	0,75
22	Politics	97	49	9	9	30	0,84
23	Sport	101	55	7	4	32	0,91
24	Treatment	110	56	9	11	34	0,85
25	Education	96	59	6	12	19	0,87
26	History	86	46	9	9	22	0,84
27	Technology	88	44	9	9	26	0,83
28	Medicine	99	52	9	8	30	0,86
29	War	86	45	9	8	24	0,84
30	Zoology	97	48	11	11	27	0,81
Total/Average:		**2766**	**1394**	**236**	**260**	**876**	**0.85**

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 2).

		Target true class Period	Not period
Predicted class	Period	True positives (TP)	False positives (FP)
	Not period	False Negatives (FN)	True Negatives (TN)

Figure 2. F1-score confusion matrix

The analysis of the results table indicates that although the rule-based model has performed relatively close among fields, there is a remarkable difference between certain fields that needs to mention. For instance, the model has performed exceptionally well in the fields of Sport and Stories, all with higher than 90% accuracy, whereas the performance in the fields of Agriculture and Art is seen less than 80% accuracy. This drastic performance difference between the fields can be explained by several factors, such as the source of the text collection on the chosen Uzbek corpus, sentence structures, terminology, as well as written level of difficulty/sentence complexity.

V.DISCUSSION

Some challenges faced when dealing with the Uzbek punctuation analysis, especially its complex nature and linguistic intricacies of Uzbek text highlight the necessity for very complex models. Following points are worth addressing in the scope of punctuation analysis for Uzbek:

- ***Agglutinative nature of Uzbek.*** Rule-based punctuation prediction gets even more complex in agglutinative Uzbek. This characteristic creates complicated words with several affixes, making sentence analysis difficult. Hence, the model may have trouble distinguishing clause and sentence boundaries, affecting punctuation prediction. Precise morphological analysis model for the language helps mitigate the difficulty regarding this matter;
- ***Diacritics/Digraph Ambiguities.*** Uzbek alphabet contains diacritics and digraphs, such as letters "o‘" and "g‘", which may cause model ambiguities. The characters used to represent linguistic sounds may be misinterpreted for punctuation marks. Misspelling the second parts of those digraphs for apostrophes or commas can cause punctuation errors. To avoid this, diacritic handling and tokenization could be modified to distinguish linguistic aspects from punctuation.

- ***Addressing Data Sparsity and Model Generalization***. A robust rule-based punctuation prediction model for Uzbek must overcome data sparsity and ensure generalization. The model struggles to learn varied linguistic patterns due to limited Uzbek punctuation datasets. More data collection and/or data augmentation or transfer learning methods could improve model generalization. Adding context-aware features and improving rule sets based on linguistic peculiarities may improve model performance in more textual scenarios.

VI.CONCLUSION

Since punctuation analysis algorithms and models for the Uzbek language texts have not been developed so far, the algorithm proposed in this work opens a path to explore more ways to approach the task. In this work, a rule-based algorithm and model for punctuation analysis and generation for period and comma marks has been proposed, with openly-available linguistic rules and the code used for the experiments. The analysis and generation of correct punctuation placements highly demand on the syntactic and morphologic characteristic of the language, especially if the language is a highly agglutinative one, like in our case with the Uzbek language. Our results from an experiment with an Uzbek corpus, which consists of 30K words from 30 different fields show that th newly created punctuation analysis/correction model gives almost 84% accuracy rate, meaning that the model, with a little more rule addition, is enough to be used for practical applications.

As a future work, more accurate rule-based model creation, also explore more developed approach to the task of punctuation generation, including not only the above-mentioned two punctuation marks, but also all the available punctuations, and apply various machine learning and deep-learning techniques to enhance the model performance. Especially, use of a large-size monolingual Transformer-based language model for Uzbek to fine-tune for the punctuation prediction task is underway.

ACKNOWLEDGMENT

This publication has been produced within the framework of the Grant “REP-25112021/113 - UzUDT: Universal Dependencies Treebank and parser for natural language processing on the Uzbek Language”, funded under the MUNIS Project, hosted by The National University of Uzbekistan named after Mirzo Ulugbek, supported by the World Bank and the Government of the Republic of Uzbekistan. The statements do not necessarily reflect the official position of the World Bank and the Government of the Republic of Uzbekistan.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] M. Bayraktar, B. Say, and V. Akman, “An analysis of English punctuation: the special case of comma,” *International Journal of Corpus Linguistics*, vol. 3, Jul. 1998, doi: 10.1075/ijcl.3.1.03bay.

[2] D. Hardt, “Comma checking in Danish,” 2001.

[3] U. Salaev, E. Kuriyozov, and C. Gomez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022, pp. 199 – 206. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138700420&partnerID=40&md5=bf476cd74317f06577dd0548c5c600d6

[4] A. M. Abdurashetona and I. O. Ismailovich, “Methods of Tagging Part of Speech of Uzbek Language,” in *Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021*, 2021, pp. 82 – 85. doi: 10.1109/UBMK52708.2021.9558900.

[5] A. M. Abdurashetona and U. Mokhiyakon, “Software Features and Linguistic Features of Uzbek Synonymizer,” in *Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022*, 2022, pp. 171 – 175. doi: 10.1109/UBMK55850.2022.9919447.

[6] B. Mengliyev, S. Shahabitdinova, S. Khamroeva, S. Gulyamova, and A. Botirova, “The morphological analysis and synthesis of word forms in the linguistic analyzer,” *Journal of Language and Linguistic Studies*, vol. 17, no. 1, pp. 558 – 564, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103797274&partnerID=40&md5=8a4052419f721c3c734f8fe1c48984ec

[7] K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of Karakalpak language stop words,” *Data Brief*, vol. 48, 2023, doi: 10.1016/j.dib.2023.109111.

[8] M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, ‘Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language’, *CEUR Workshop Proceedings*, vol. 3315, pp. 93–98, 2022.

[9] D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, and O. Alloyorov, “Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language,” Jan. 2023, pp. 1720–1724. doi: 10.1109/APEIE59731.2023.10347765.

[10] K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of stopwords extracted from Uzbek texts”, *Data in Brief*, vol. 43, 2022.

[11] M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.