Natural Language Processing: comparing rule-based and stochastic part-of-speech tagging methods in the Uzbek language

Mr. Sardorbek Turdimurodov

SCRUM master
BAI
turdimurodovs33@gmail.com

Abstract:
This paper explores different approaches to Part-of-Speech (POS) tagging and their application to the Uzbek language. Although NLP progress has improved tagging in high-resource languages, Uzbek still lacks annotated corpora. This study compares rule-based and stochastic approaches and suggests ways to improve tagging accuracy for low-resource languages.

1. INTRODUCTION

Natural language processing (NLP) is the processing of natural language information by a computers. NLP is one of the main subfields of AI, and its uses are really important for AI's wide usages. Major processing tasks in an NLP system include speech recognition, text classification, natural language understanding, and natural language generation. The main examples of NLP can be seen on Alexa, Yandex, and all other chatbots. NLP does the following jobs: coreference resolution, named entity recognition, part-of-speech tagging, and word sense disambiguation. Thus, POS tagging is the process of determining which part of speech a word or piece of text is, based on its use and context.

Despite recent advances in Natural Language Processing, low-resource languages like Uzbek have not benefited equally from large datasets or pre-trained models. Therefore, foundational, linguistic methods such as rule-based and stochastic tagging remain critical for language modelling, corpus creation, and syntactic analysis. This paper compares these two approaches to evaluate their effectiveness for Uzbek text processing.

This research's aims are as follows:

To understand the current methods of POS tagging;

To explain how such methods work with Uzbek language;

To give further suggestions for the improvement of POS tagging for low-resource languages such as Uzbek.

The research questions guiding this study are as follows:

What are the methods of tagging, and what is the difference?

How to improve the accuracy of taggers, and what are the challenges?

Why don't current improvements in NLP help current taggers to work with low-resource languages?

2. LITERATURE REVIEW

2.1 About POS tagging

Part-of-speech Tagging, also called grammatical tagging, is the process of determining which part of speech a word or piece of text is, based on its use and context. For example, part-of-speech identifies "make" as a verb in "I can make a paper plane" and as a noun in "What make of car do you own?" [1] Parts of Speech (PoS) Tagging is a fundamental task in natural language processing (NLP), and it plays an important role in various NLP applications, including machine translation, sentiment analysis and information retrieval, by bridging the gap between human language and machine understanding. [2]

Parts of Speech: These are categories like nouns, verbs, adjectives, adverbs, etc that define the role of a word in a sentence.

Tagging: The process of assigning a specific part-of-speech label to each word in a sentence.

Corpus: A large collection of text data used to train POS taggers. [3]
Consider the sentence: "Tez ari meni qo'limdan chaqdi chaqdi."
After performing POS tagging, we get:

- **Tez** — adverb (ADV)
- **ari** — noun (NN)
- **meni** — pronoun (PRP)
- **qo'limdan** — noun (NN) *(possessive + ablative)*
- **chaqdi** — verb (VBD)
- **chaqdi** — verb (VBD)

2.2 Workflow of POS tag set:
Tokenisation: The text is divided into individual tokens, representing words or subwords;
Loading a Language Model: By using tools which are already trained to recognize the grammatical structures and rules, POS tagging is used;
Text processing: Standardising the text's format. Lowercasing all the words and removing unnecessary content;
Linguistic analysis: understands the grammatical role of each token and analyses sentences' syntactic structure;
POS tagging: Based on the role and context, each word is assigned a specific part-of-speech label;
Result Evaluation: any misclassifications are identified and corrected, preparing data for usage. [4]

2.3 Methods of POS tagging
1. **Manual Tagging**: This means having people versed in syntax rules applying a tag to every and each word in a phrase.

- This is the time-consuming, old-school, non-automated method. Reminds you of homework? Yeah… But it is also the basis for the third and fourth ways.

1. **Rule-Based Tagging**: The first automated way to do tagging. Consists of a series of rules (if the preceding word is an article and the succeeding word is a noun, then it is an adjective…). Has to be done by a specialist and can easily get complicated (far more complicated than the Stemmer we built).

- The position of "Most famous and widely used Rule Based Tagger" is usually attributed to E. Brill's Tagger.

1. **Stochastic/Probabilistic Methods**: Automated ways to assign a PoS to a word based on the probability that a word belongs to a particular tag or based on the probability of a word being a tag based on a sequence of preceding/succeeding words. These are the preferred, most used and most successful methods so far. They are also the simpler ones to implement (given that you already have pre-annotated samples — a corpus).

- Among these methods, there could be defined two types of automated probabilistic methods: the Discriminative Probabilistic Classifiers (examples are Logistic Regression, SVM's and conditional random fields — CRF's) and the generative probabilistic classifiers (examples are Naive Bayes and Hidden Markov models — HMMs).

4. **Deep Learning Methods**: Methods that use deep learning techniques to infer PoS tags. So far, these methods have not been shown to be superior to stochastic/probabilistic methods in PoS tagging — they are, at most, at the same level of accuracy — at the cost of more complexity/training time. Today, some consider PoS tagging a solved problem. Some closed context cases achieve 99% accuracy for the tags, and the gold standard for Penn Treebank has been kept at above a 97.6 F1 score since 2002 in the ACL (Association for Computer Linguistics) gold standard records. [5]

2.4 Uzbek language
There are more than 50,000 lexemes in the Uzbek language, and it is very important to determine the part of speech of each lexeme as a basis for corpus and linguistic computer programs. There are words that do not have a clear part of speech marker or whose contextual meaning within a sentence makes it difficult for the reader to determine their part of speech. For example: "... test sinovlaridan o'tkazish yuzasidan shaxsan javobgarligi belgilab qo'yilsin", "Shaxsan o'zim keldim", "Shaxsan bajardim", "Bular hammasi lotincha yoki

lotinchaga yaqin so‘zlar. Men, shaxsan, shunday deb bilaman.” (A. Qahhor, *Adabiyot muallimi*). “I came in personally”, “I did it personally”, “These are all Latin words or close to Latin. I personally know that.” (A. Qahhor, *Literature Teacher*).

It is difficult to determine the part of speech of the word “shaxsan” (personally) in these sentences. In some contexts, it appears as a personal pronoun, while in others it is clearly used as an adverb.

In such cases, the part of speech of a word is determined by the categorical characteristics of the parts of speech. These are four [22]: semantic, syntactic, morphological, and word-formation features.

It is known that in the Uzbek language there are 12 word groups (independent word groups: noun, verb, adjective, adverb, numeral, pronoun; auxiliary word groups: conjunction, postposition, particle; separate word groups: modal, interjection, imitative words). As a result of the addition of word-forming suffixes, four word groups are formed: noun, verb, adjective, and adverb. Among the identified constructive affixes (337 in total: 114 noun-forming suffixes, 58 verb-forming suffixes, 117 adjective-forming suffixes, and 48 adverb-forming suffixes) [Abjalova, 2020: 122–123], one such affix is “-an”, which forms adverbs. Based on this, it can be concluded that adding the suffix “-an” to the noun “shaxs” (“person”) forms a derivative adverb:

shaxs (noun) ∪ {-an} => shaxsan

(person (noun) ∪ {-ly} => personally).

2.5 Uzbek POS taggers

The workability and accuraccu of Uzbek POS taggers was determined after creating a rule-based POS tagger. This tagger was created for testing purposes only.

It runs as follows:

1. Using words that are already imputed on the tag set:

Id	Tag	Meaning	Uzbek examples
1	NOUN	OT (*Noun*)	olma (*apple*)
2	VERB	FE’L (*Verb*)	yugurmoq (*run*)
3	ADJ	SIFAT (*Adjective*)	ko‘p (*many/much*)
4	NUM	SON (*Numeral*)	besh (*five*)
5	ADV	RAVISH (*Adverb*)	tez (*fast*)
6	PRON	OLMOSH (*Pronoun*)	bu (*this*)
7	AUX	KO‘MAKCHI (*Auxiliary*)	bilan (*with*)
8	CONJ	BOG‘LOVCHI (*Conjunction*)	va (*and*)
9	PART	YUKLAMA (*Particle*)	faqat (*only*)
10	MOD	MODAL (*Modal*)	darhaqiqat (*actually*)
11	IMIT	TAQLID (*Imitation*)	kuk-kuk (*imitation of a hen*)
12	INTJ	UNDOV (*Interjection*)	hoorah! (*when you win*)

2. Using the rules that apply for specific set of words, making it more commonly applied for larger numbers of words:

```
# --- FE'L ---
  # O'tgan zamon
  elif soz_lower.endswith(("di", "gan", "ibdi", "mish")):
      return "FE'L (o'tgan zamon)"
  # Hozirgi davomiy zamon
  elif soz_lower.endswith(("yapti", "yapman", "yapsan", "moqda")):
      return "FE'L (hozirgi davomiy zamon)"
  # Kelasi zamon
  elif soz_lower.endswith(("adi", "ar", "moqchi", "ajak", "ajakman")):
      return "FE'L (kelasi zamon)"
  # Buyruq mayli
  elif soz_lower.endswith(("gin", "ing", "sin", "aylik")):
      return "FE'L (buyruq mayli)"
```

The accuracy of the created POS tagger was between 45% and 50%. But this accuracy can be improved through the implementation of further rules and word sets.
It's hard to get solid accuracy as Uzbek language, with over 40.000.000 users, has wide grammar and word range, as mentioned earlier.
Currently, there is no publicly available Uzbek POS-tagged corpus, which makes the development and evaluation of taggers challenging. Therefore, this research will focus on constructing an initial Uzbek tag set and manually annotating a small corpus to enable further experiments and model training.

3. METHODOLOGY

The proposed research aim is achieved by comparing developed POS tag sets with the one that is in the process of being created. The first aim is achieved by learning developed tag sets like Hidden Markov, the Brown Corpus and others. The second aim is achieved by learning how Uzbek linguistics work. The third aim's main intention is currently on the process of being fulfilled, as it'll be done after the creation of the tag set for the Uzbek language. The accuracy and the usage of current Uzbek tagger was determined with the created POS tag set. Thus, the accuracy cannot be determined without the tag set which should include all Uzbek grammars and Uzbek words

4. CONCLUSION AND FUTURE WORK

This research emphasises the importance of developing effective Part-of-Speech (POS) tagging methods for the Uzbek language, a low-resource yet morphologically rich language. Through the comparison of rule-based and stochastic approaches, this study highlights that traditional linguistic methods still play a crucial role in building foundational tools for Uzbek NLP. Accurate POS tagging is essential not only for syntactic and semantic analysis but also for creating robust language models, machine translation systems, and linguistic corpora.
While the current paper focuses on reviewing existing methods and their applicability to Uzbek, future work will involve the design and implementation of a full Uzbek POS tagger. This upcoming stage will include constructing a balanced and annotated corpus, defining an optimised tag set based on linguistic and morphological features, and testing hybrid approaches that combine rule-based precision with probabilistic or deep learning models. The goal is to achieve a high-performance POS tagger that can serve as a foundation for future Uzbek NLP tasks such as dependency parsing, named entity recognition, and text classification.
Ultimately, this line of research contributes to expanding digital linguistic resources for under-represented languages and strengthens the foundation for broader AI applications in Central Asian languages.

BIBLIOGRAPHY:

1. IBM. (2023). *What is Natural Language Processing (NLP)?* IBM Think. Retrieved from https://www.ibm.com/think/topics/natural-language-processing
2. GeeksforGeeks. (2023). *Part-of-Speech (POS) Tagging in NLP*. Retrieved from https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/
3. Abjalova, G. (2020). *Uzbek Language Morphology.* Tashkent State University of Oriental Studies Press.
4. ACL Anthology. (2002). *POS Tagging Benchmarks and Evaluation (Penn Treebank).* Association for Computational Linguistics.
5. Medium Analytics Vidhya. (2021). *Part-of-Speech Tagging: What, When, Why, and How.* Retrieved from https://medium.com/analytics-vidhya/part-of-speech-tagging-what-when-why-and-how-9d250e634df6
6. Uzbek School Corpora. (n.d.). *Educational Corpus Dictionary (Uzbek Language).*
7. Retrieved from http://uzschoolcorpara.uz/uz/Dictionary%20(Uzbek%20language%20educational%20corpus)