IEEE Turkey Section

UBMK'2021

Bildiriler Kitabı
Proceedings

6. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

6th International Conference on Computer Science and Engineering

15-16-17 Eylül (September) 2021 Ankara- Turkey

Methods of Tagging Part of Speech of Uzbek Language

Abjalova Manzura Abdurashetona
Uzbek Language and Literature
Tashkent Stata University named after Alisher Navoi
Tashkent, Uzbekistan
abjalova.manzura@gmail.com

Iskandarov Otabek Ismailovich
Tashkent State Polytechnic University named after Islam Karimov
Tashkent, Uzbekistan
bekor_ok@mail.ru

***Abstract* - As in all other fields, linguistics is accelerating the process of adapting to digital technologies. Consequently, it is important to process the traditional linguistic norms of natural language for computer programs and information systems. One such important task in NLP is tagging parts of speech. Part-of-speech tagging (abbreviation: (POS tagging or PoS tagging or POST) in Russian “частеречная разметка”) is a stage of automatic text processing, the function of it which is a series of words (forms) used in the text and it is to determine grammatical features. With this function, POS-tagging is one of the first steps in automatic text analysis. The article discusses the need for tagging parts of speech, tagging methods.**

***Keywords* – natural language processing, tagging, tag, word group, formal language, pragmatic feature, corpus, polysemy, homonymy, PoS-tagging.**

I. INTRODUCTION

It is important to understand the difference in the use of a word in different contexts when defining the category (categories) of each word in a language according to its semantic, syntactic, morphological, and word-formation features, such as noun, adjective, number, verb, adverb and pronoun. There are cases when part of speech of the word cannot be defined without analyzing its semantics (especially multi-functional and polysemantic words), or even contextual pragmatics. In computer programs, the creation of stages of semantic and pragmatic analysis that ensure their perfection is a more complex process. Here is another examplee of this idea of the pragmatic nature of the language is given by the word “long” in Uzbek “uzoq” (Table-1):

In corpus linguistics, the grouping of words, the tagging of parts of speech and grammatical categories, in order to avoid ambiguity and remove homonymy, is based not only on their lexical form, but also on their expression in a sentence (on a paragraph, in a phrase) and on their semantic relationship with other words. identification of sentences members tags is a complex process [15]. Because it is impossible distinguish all Uzbek words for 12 parts of speech. As it is known there are 12 parts of speech in the Uzbek language. A word can be polyfunctional depending on the state of its realization in the sentence structure and the semantic valence of the N-gramm words. [Abjalova, 2020: 73-77] E.g. The sentences classified according to the parts of speech the first word “sick” is answering to the question “who?” and it is noun, in the second one (it is answering the question What kind of?) it is an adjective “It was brought the sick to the hospital” and “It was brought sick person to the hospital” (how? answers the question) is a word in the function of the category of quality. Out of 11,000 borrowed words in the Uzbek Explanatory Dictionary, 66 such polyfunctional words were identified [14].

TABLE I. TABLE-1: PRAGMATIC NATURE OF THE LANGUAGE

№	**The concept of the word “long” in the context (in Uzbek)**	**English translation**	**Pragmatic meaning**	**In English**
1	*xolislik*	impartiality	*Bu voqeadan uzoqda*	This is far from the case
2	*ayrilish*	leave	*uzoqlashishdi*	Moved away
3	*ko‘plik*	plural	*Uzoq (ko‘p) kutdim*	I waited a long time
4	*yuzakilik*	superficiality	*Uzoqdan yondashdim*	I approached from a distance
5	*uzunlik*	length	*Uzoq-uzoq karvonlar*	Long caravans
6	*vaqt*	time	*Uzoq o’yladi*	He thought for a long time
7	*qarindosh(lik) munosabati*	relationship	*Uzoq qarindosh*	A distant relative
8	*kelajak*	future	*Uzoqni o’ylamoq*	Think long and hard
9	*tarix*	history	*Uzoq yillar*	Many years
10	*masofa*	distance	*Uzoqda*	Away

The fact that a word is categorized according to its context does not make it possible to set a common parameter for word group tags. Thus, this indicates that it is not possible to manually perform ST tags for the case. The emergence of new contexts and the appearance of neologisms in language also indicate the continuity of the tagging process. Therefore, PoS correction in the language body relies on machine coding.

II. The Need to Tag Word Groups

Before discussing PoS tagging methods, let's first look at why PoS tagging in Uzbek NLP is necessary and where they are used (Figure 1).

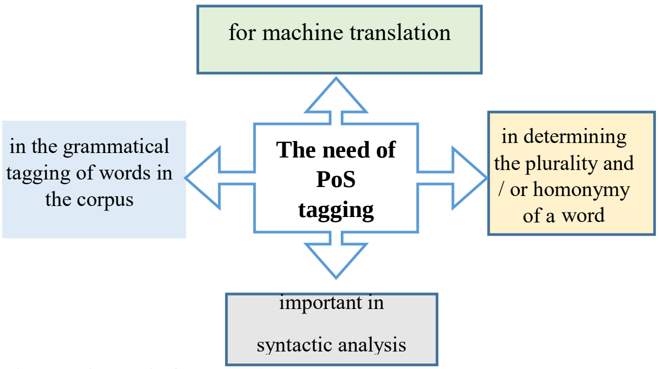
for machine translation
in the grammatical tagging of words in the corpus
The need of PoS tagging
in determining the plurality and / or homonymy of a word
important in syntactic analysis

Fig. 1. The need of PoS tagging

Most importantly, PoS tags are the most essential linguistic elements for (Natural Language Processing (NLP)), so PoS tagging is done as a prerequisite to simplify various problems in NLP.

In a universal linguistic information system (corpus), which includes the phonetic, lexical, morphological, syntactic, and semantic levels of a language, it is also necessary to tag word groups in order to perform morphological analysis. This will allow the reader to present the results of the correct analysis of any word (form) in the language corpus. In the absence of a language corpus, many languages refer to morphological dictionaries that are included in the linguistic support in the process of word variation and assimilation. [Abjalova, 2020: 38-39, 155].

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 for the basis of corpus and linguistic computer programs. There are words that do not have a part of the speech sign or that the contextual meaning of the sentence confuses the reader to determine his parts 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 in 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 the sentences. In some places it appears as a personal pronoun, and in some cases it is clearly visible as an adverb.

In this case, the parts of speech of the word is determined by the categorical characteristics of the parts of speech. They are four [22]: semantic, syntactic, morphological, and word-formation features.

It is known that in 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, 4 word groups are formed: noun, verb, adjective, adverb. Among the identified constructive affixes (337: noun-forming suffixes are 114, verb-forming suffixes are 58, 117 adjective-forming suffixes, 48 adverb-forming suffixes) [Abjalova, 2020: 122-123] is the affix “*-an*” to form the adverb. Based on this parameter, it can be concluded that adding the suffix “-an” to the word “shaxs” (“person”) belonging to a noun forms a derivative adverb: *shaxs (Ot) ∪{-an} => shaxsan. (person (noun) ∪ {-ly} = > personally.*

In general, the word is interpreted as a very complex phenomenon, emphasizing both the unity of language and the unity of speech. The equivalence of the unit of language with the unit of speech is mainly observed in the variable categories. The nature of the polyfunctional, ambiguous, and homonymous words we know requires serious research and practical observation.

In the sentence “*It's better to read a book than to blame your work in impossibility*” (Figure 2) you can see that the members of the sentence are identified as a result of tagging words to parts of speech, and the parts of speech of the homonymous word “pesh” are determined by the sequence of words (*pesh qilmoq* – verb (noun+Verb)); *o‘qiganing pesh* (verb+adj) and their location (at the end of the sentence predicative => adjective).

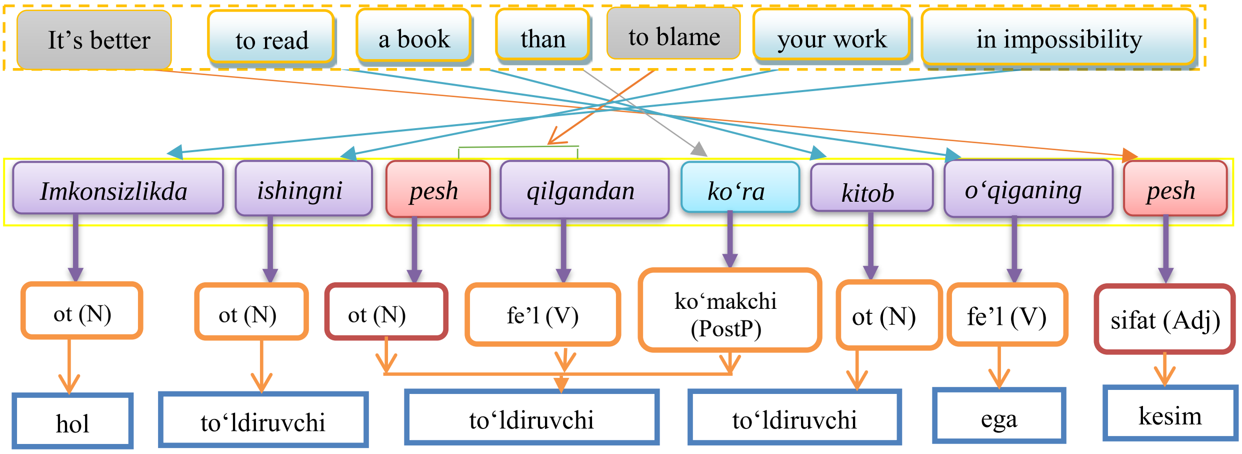

Fig. 2. Parsing of sentense in Uzbek language

A part of speech can does perform different syntactic functions in any structure of a phrase and a sentence (In Figure 3, “*pesh qilgandan ko‘ra*” (“than to blame”) is a combination of three words referring to the three parts of speech, an indirect object). Such words are named polyfunctional. But each part of speech performs a basic, primary syntactic function. The primary syntactic function comes from the lexical meaning of a part of speech and is embodied in the form of a specific transposition of this meaning. [Ganiyeva, 2019: 277].

It is not enough to include a list of words and their parts of speeches in the linguistic database for PoS tagging. The loss of consistency, as in the case of the definition of the above phrase, or the finding of a set of polyfunctional, homonymous or polysemous words in a sentence, encourages even an expert linguist to think and search. Also, many words in the Uzbek language are not definitely the exact part of speech. Taking into account the aforementioned problems in the Uzbek language when processing natural language for creating computer programs, several methods of tagging parts of speech are used.

III. METHODS OF TAGGING WORD GROUPS IN UZBEK NLP

On the following methods (algorithms) [16, 18, 19] is based on the tagging of parts of speech in the NLP of the Uzbek language:

- rule-based method
- stochastic (or statistical) method.

Rule-based PoS tags. One of the oldest tagging methods is POS-tagging based on these rules. The Brill method is mainly useful for tagging it. [Brill, 1992]. Rule-based taggers use a dictionary or vocabulary of a language to tag each word. If a word (multifunctional, homonymous, plural) has multiple tags, then rule-based taggers use handwritten rules to correctly identify the part-of-speech tag of a word in a sentence. Giving more specific tags can also be done by defining the linguistic features of a word based on rules by analyzing the words before and after it. For example, the linguistic unit that comes after the noun in the accusative case is a word from the category of nouns with the possessive suffix: “*me**ning** kitob**im***” (*my book*), *“aka**mning** uy**i**” (my brother’s house), “Salima**ning** ko‘ylag**i**” (Salima’s shirt).* Therefore, in this case, the noun is defined by the noun in the possessive case that comes before the word being defined. Let's look at an example from English: if the preceding word is an article, then the word following it is a lexical unit of the noun category. For example, *an egg, a book, the train, the windows.*

In such cases, the tagging of parts of speech words in the Uzbek language are encoded in the form of the following rules:

1. Rules based on linguistic norms. Hundreds of rules based on the spelling rules of the language are formed as a base of general, special and exception rules. [Abjalova, 2020].

2. Contextual template rules, that are, the regular use of a word with a figurative meaning in a sentence in a connotative sense, are stored in the program memory, as a result of which the ambiguities associated with that connotative word are eliminated in subsequent processes.

According to the rule-based method, word grouping was done in two stages [16]:

In the first stage, PoS tagger was based on dictionaries (explanatory, morphological or spelling). With the help of dictionaries, the parts of speech (categories) of each word were determined.

In the second stage, a lists of polyfunctional or homonymous words were written by hand, and a long list of rules was developed to define the function of such words in a sentence.

As a result, you can determine the parts of speech for each word in the sentence (Figure 4). This is important when defining polysemous, polyfunctional and homonymous words.

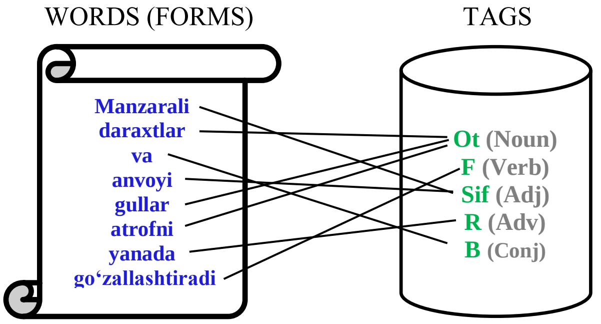

Fig. 4. Assigning part-of-speech tagging to each word in a sentence.

A striking example of the method of automatic generation of rules is the method of American linguist Eric Brill [Brill, 1995]. The method of work is as follows:

1. Get started: Each word should be associated with the most frequently used tag of that word. Unknown words are treated as noun phrases. At this stage, not only the learning process begins, but also the method of eliminating homonyms [2].

2. Create a change (rework) rule for a common errors.

3. Repeat the second step until the desired minimum error is reached.

Features of rule-based ST tags in Uzbek NLP

Rule-based PoS tags have the following features:

- These tags are based on knowledge.
- Rules are created manually.
- Information is coded in the form of rules.
- The rules will be limited. For digital technology, infinity represents abstraction and combinations and symbols such as “...” (many points), “*etc.*” indicate the ambiguity of the list, not the length. Therefore, in computer linguistics, such ambiguities are not allowed, but the nature and characteristics of each linguistic unit must be clearly stated in the software database.
- Language modeling is based on the rules of tagging

Stochastic tagging method. This method is based on frequency or probability (statistics). Therefore, in some sources it is explained as a statistical or probabilistic method. [16, 19, 20, 21]. As can be seen, the following methods are used for PoS tags in simple stochastic tagging: the frequency approach, the probability of a sequence of tags or the n-gram method, and the hidden Markov model and model modification.

Features of the method of stochastic tagging of word groups in Uzbek NLP

Stochastic PoS-tegs have the following features:

- This tagging is based on the probability of the tags being applied in series.
- Educational corpus required.
- There is no possibility for words that do not exist in the corpus.
- Except the educational corpus, other types of language corpuss can be used.
- The simplest PoS tagging method, because this method selects the actively used serial tags in the language corpus.

IV. PRACTICAL RESULT

As a result of several years of research and practical efforts in the years 2020-2021 at the Tashkent State University of Uzbek language and literature, in collaboration with the Department of Information Technology, Applied Linguistics and didactics in the practical part of the project AM-FZ-201908172 named after “Creation of the educational corpus of the Uzbek language” created a NATIONAL CORPUS and EDUCATIONAL CORPUS OF THE UZBEK LANGUAGE. Today, the corpus has a mosfoanalyzer (automatic morphological analysis), a synonymizer (a program for introducing synonyms into a search word) [4], as well as the ability to divide a word into syllables, provide explanations (s), and specify antonyms [3].

V. CONCLUSION

In conclusion, the creation of algorithms for tagging word groups in natural language processing (NLP) is a prerequisite for automated analysis, morphoanalysis, and translation programs, resulting in a high-quality grammatical analysis of texts. Today, rule-based and stochastic methods are used hybridly in neural network-based artificial intelligence systems.

Linguistic knowledge is the most important source of information for computer programs and language corporations. If the formation of a linguistic database in a linguistic processor is based on the ability to use language knowledge in everyday life, and if such situations are translated into literary language norms, the analysis capabilities of computer programs will be perfected at the expert level. Therefore, linguistic competence and verbal competence are constantly interdependent. The practical significance of digital technology programs and systems is further enhanced when the definition parts of speech of particular word (mainly polyfunctional, polysemantic, homonymous) is based on the pragmatic and contextual meanings and additional semantics of the word, as in the case of the word “long” at the beginning.

In general, if today's linguistic knowledge is improved based on the practical use of the language and based on such theoretical resources to create a formal Uzbek language for computer programs and systems, in the future the basis for the emergence of electronic systems.

REFERENCES

[1] Abjalova M. Linguistic modules of editing and analysis programs. [text] : Monograph / M.A. Abjalova. – Tashkent: Nodirabegim, 2020. – 176 p.

[2] Abjalova M., Yuldashev A. Methods for Determining Homonyms in Linguistic Systems // ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 11, Issue 2, February 2021. Impact Factor: SJIF 2021 = 7.492 (https://saarj.com). ISSN: 2249-7137. DOI: 10.5958/2249-7137.2021.00522.X

[3] Abjalova M. Possibilities of Searching Words on the Basis of Lexicography in the National Corpus of the Uzbek Language // Kompyuter lingvistikasi: muammo, yechim, istiqbollar. Respublika I Ilmiy-texnik konferensiya to‘plami. – Toshkent: ToshDO‘TAU. – B. 12-17.

[4] Abjalova M. The issue of creating a synonymizer in the national corpus of the Uzbek language // O‘zbek Milliy va ta’limiy korpuslarining yaratishning nazariy hamda amaliy masalalari. Xalqaro ilmiy-amaliy konferensiya to‘plami. – Toshkent: ToshDO‘TAU. – B. 38-40. DOI: doi.org/10.52773/CL:PSP/vol_1_issue_1/A2

[5] Brill E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging // Computational Linguistics. Vol. 21. – P. 543-565. http://acl.ldc.upenn.edu/J/J95/J95-4004.pdf

[6] Brill E. 1992. A simple rule-based part of speech tagger //Proceedings of ANLC. – P. 154.

[7] Baum, L. E.; Sell, G. R. 1968. Growth transformations for functions on manifolds. Pacific Journal of Mathematics. 27 (2) – P. 211-227.

[8] Ganiyeva, Dildora. 2019. Мазмуний синкретизм ва полифункционаллик: NamDU ilmiy axborotnomasi – Nauchniy vestnik NamGU 6-son: 275-278.

[9] Rizayev S. O‘zbek tilshunosligida lingvostatistika asoslari. – Toshkent: Fan, 2006. – B. 18.

[10] Thad Starner, Alex Pentland. 1995. Real-Time American Sign Language Visual Recognition From Video Using Hidden Markov Models. Master's Thesis, MIT, Program in Media Arts

[11] *Li, N; Stephens, M (December 2003). “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data”. Genetics. 165 (4): 2213-2233. doi:10.1093/genetics/165.4.2213.*

[12] *Ernst, Jason; Kellis, Manolis (March 2012). “ChromHMM: automating chromatin-state discovery and characterization”. Nature Methods. 9 (3): 215–216. doi:10.1038/nmeth.1906. PMC* 3577932. *PMID 22373907*

[13] *Qurbonova M., Abjalova M. va boshq.* O‘zbek tili o‘zlashma so‘zlarining urg‘uli lug‘ati. [Matn]: o‘quv-uslubiy lug‘at / M.Qurbonova, M.Abjalova, N.Axmedova, R.To‘laboyeva. – Toshkent: Nodirabegim, 2021. – 988 b.

[14] https://en.wikipedia.org/wiki/Hidden_Markov_model

[15] https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-markov-model-953d45338f24/

[16] https://uzjournals.edu.uz/namdu/vol1/iss6/46

[17] https://coderlessons.com/tutorials/akademicheskii/obrabotka-estestvennogo-iazyka/pometka-chasti-rechi-pos

[18] https://habr.com/ru/post/125988/

[19] https://ru.wikipedia.org/wiki/Частеречная_разметка

[20] https://en.wikipedia.org/wiki/Part-of-speech_tagging#:~:text=In%20corpus%20linguistics%2C%20part%2Dof,its%20definition%20and%20its%20context.

[21] https://ru.wikipedia.org/wiki/Часть_речи

[22] http://uzschoolcorpara.uz/uz/Dictionary (Uzbek language educational corpus)