GALAXY INTERNATIONAL INTERDISCIPLINARY RESEARCH JOURNAL (GIIRJ)
ISSN (E): 2347-6915
Vol. 13, Issue 1 January (2025)

POS TAGGING OF LEXICAL UNITS IN THE UZBEK LANGUAGE

Zilola Karimova Dilmurod qizi
1st Year Master's Student, Computer Linguistics, Tashkent State
University of Uzbek Language and Literature named after Alisher Navoi
E-mail: karimovazilola791@gmail.com

ABSTRACT

In corpus construction, the issue of linguistic annotation is important and complex. The process of assigning identifiers to linguistic units in the text is a problem because it is related to both the rules of tagging and the regularities of the language. Tagging, particularly grammatical tagging or Part of Speech (PoS) tagging, is also a critical issue in Uzbek corpus linguistics. This article discusses the methods of POS tagging for Uzbek language texts using widely used tagsets worldwide.

Keywords: Tag, annotation, markup, tagset, NLP, corpus, PoS tagging.

INTRODUCTION

It is well-known that any corpus construction begins with the annotation of available data. The process of tagging is often referred to as annotation or markup in various literature, sometimes used synonymously. Annotation refers to general information providing linguistic or extralinguistic data about a part of the text that is not directly related to it. Annotation may include metadata and tags. Markup is considered part of the annotation process, specifically for metadata. Standard principles for markup have been developed in global practice. Tag refers to a conditional symbol or special code used to facilitate and accelerate text analysis by a computer. Tags are divided into several types: semantic tags, syntactic tags, and grammatical tags. Grammatical tags, also known as PoS (Part of Speech) tags, are widely used. PoS tagging is the task of assigning the part of speech (noun, verb, adjective, pronoun, etc.) to each word form in a given sentence. PoS tagging is one of the key tasks in Natural Language Processing (NLP), and it is an essential part of the pipeline process.

Due to the rapid development of information and communication technologies, there has been growing interest in tools for Natural Language Processing (NLP). As a result, various NLP methods and tools are being developed today. However, developing effective NLP tools for processing natural language texts requires addressing tasks such as linguistic tagging. Linguistic tagging involves linking descriptive or analytical labels to language data. Unstructured (unprocessed) data can come in the form of text, or from sources like audio, video, or physiological recordings, and it may contain various types of transcription (from phonetic features to speech structure), PoS tagging, semantic features, syntactic analysis, Named Entity Recognition (NER), semantic role labeling, time and event recognition, and syntactic chains of words.

After the development of language corpora, they need to be tagged. For example, POS tagging was used in Brown corpus texts [Maverick, 1969: 35(1)]. This research also refers to K.V. Gruch [Church, 1989], S.J. Derose [Derose, 1988: 14(1)], R. Garside [Garside, 1987], and B.B. Greene's studies [Greene, Rubin, 1971: 23]. Like the Brown corpus, corpora developed in the

1970s and 1980s are typically POS-tagged, but the lack of effective automated methods and the complexity of manual tagging made it difficult to create sufficiently large corpora that included tags for other linguistic phenomena. At the end of the 1980s, the availability of new large-scale linguistic data led to the proliferation of linguistic tagging systems. These systems are primarily based on PoS or morphosyntactic tagging, and statistical methods were developed for automatic tagging. One of the first significant efforts in this area was the Lancaster-Oslo-Bergen (LOB) corpus, which consisted of one million words and was tagged based on morphosyntactic and syntactic tags [Beale, 1985]. Based on this research, the Penn Treebank project developed a one-million-word corpus of Wall Street Journal articles [Marcus, Santorini, Marcinkiewicz, 1993], which was also tagged with PoS and syntactic tags [Marcus, Kim, Marcinkiewicz, MacIntyre, Bies, Ferguson, Katz, Schasberger, 1994].

In the 1990s, automatically tagged corpora such as the 100-million-word British National Corpus [Erjaveç, 1998], the MULTEXT multilingual corpus [Ide, Véronis, 1994: 588-592], and the PAROLE and SIMPLE corpora [Kolodnytsky, Bernsen, Dybkjær, 2004] were developed, containing data from fourteen European languages and their PoS tags.

Once a text is segmented into tokens, each token or sequence of tokens can be tagged with its part of speech (noun, verb, pronoun, etc.). This task is called PoS tagging, which is a form of morphological analysis because it helps identify the part of speech by the grammatical form attached to the root. For instance, the process of morphological tagging indicates that suffixes like “-gan” or “-di” are morphological units attached to verbs (with some exceptions). However, words without morphological (grammatical) forms are also tagged. For example, "mergan" (shooter) is tagged as a noun, and "bergan" (gave) is tagged as a verb. In these words, "-gan" is not a morphological unit but a part of the word in the first case.

In the PoS tagging task, each token or sequence of tokens in a text is assigned a PoS tag. The Uzbek language has 12 parts of speech, and their corresponding PoS tags have been identified. Various researchers have proposed different classification systems for PoS tagging. For example, in the paper "PoS Tagging in Uzbek: Problems and Proposals" by B. Elov and Sh. Hamroyeva [Elov, Hamroyeva, Abdullayeva, Uzakova, 2022: 51-68], a tag system for POS tagging Uzbek lexical units is described. Furthermore, in the paper "POS Tagging of Uzbek Texts Using Hidden Markov Models (HMM) and the Viterbi Algorithm" [Elov, Hamroyeva, Xudayberganov, Yodgorov, Yuldashev, 2023], the same tag system is employed for POS tagging Uzbek language units. Similarly, in the paper "POS Tagging and Stemming in Agglutinative Languages (with Examples from Turkish, Uighur, and Uzbek)" [Elov, Hamroyeva, Abdullayeva, Husainova, Xudayberganov, 2023: 6-39], the tag systems for agglutinative languages like Turkish, Uighur, and Uzbek are compared. The paper "Building a Syntactically Tagged Database for Simple Sentences in Uzbek" [Ramatova, 2023] proposes POS tags only for independent parts of speech in Uzbek. Below, we explain the important grammatical features in the POS tagging of Uzbek lexical units. Linguistics classifies these main categories further into smaller subcategories.

For example, when tagging the verb part of speech, it is possible to identify other grammatical features along with its verb status. The morphological (PoS) tagging of a verb is done as follows: "Og‘rib qoldi" (He got sick): [verb], [main verb], [intransitive verb], [past tense], [3rd person singular], [mood marker]. Here, the grammatical categories are fully listed, but in the

tagging process, an extended tag system is used. The extended tag system for adjectives includes distinctions like:
"yaxshi" (good): [adjective], [positive form], [simple degree], [attribute].
"tuzuk" (good): [adjective], [positive form], [simple degree], [descriptive adjective].
In conclusion, experiments on language corpora have shown that integrating base information with syntactic tasks improves PoS tagging results for morphologically rich languages, which contributes to the efficiency of solving NLP tasks. The specific features of PoS tagging tasks for Uzbek language corpora, the PoS tag categories, and the list of subcategories were presented. The classification of grammatical categories for verbs and adjectives was explained with examples. The methods of PoS tagging described in the article can be used in the development of morphological analyzers for Uzbek and in solving other complex NLP tasks.

REFERENCES

1. Maverick, G. v. (1969). Computational Analysis of Present-Day American English. Henry Kučera, W. Nelson Francis. International Journal of American Linguistics, 35(1). https://doi.org/10.1086/465045
2. Church, K. W. (1989). Stochastic parts program and noun phrase parser for unrestricted text. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2. https://doi.org/10.3115/974235.974260
3. Derose, S. J. (1988). Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 14(1).
4. Garside, R. (1987). The CLAWS word-tagging system. The Computational Analysis of English: A Corpus-Based Approach.
5. Greene, B. B., Rubin, G. M.: Automatic Grammatical Tagging of English. Brown University, Department of Linguistics (1971)
6. Beale, A. D. (1985). Grammatical analysis by computer of the Lancasteroslo/Bergen (LOB) corpus of British English texts. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1985-July. https://doi.org/10.3115/981210.981246
7. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2). https://doi.org/10.1162/coli.2010.36.1.36100
8. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The Penn Treebank: Annotating predicate argument structure. In: Proceedings of the Workshop on Human Language Technology, pp. 114–119. Association for Computational Linguistics, Stroudsburg, PA, USA (1994) Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol. I, pp. 588–592. Kyoto, Japan (1994)
9. Bakhtiyorovich, Ismonov Khurshidbek, and Ruziyev Nuriddin Mukhammadaliyevich. "Pairing, Their Own Aspects and Corresponding Methods of Work with Pairing in the AutoCAD Software." International Journal on Orange Technologies 3.12 (2021): 211-216.
10. Qizi Abduraimova, Muazzamoy Abduqodir. "PERSPECTIVE." INTERNATIONAL CONFERENCES. Vol. 1. No. 11. 2022.

11. Xurshidbek, Ismonov, Rustamov Umurzoq, and Abduraimova Muazzamoy. "Central and Parallel Projections, Orthogonal Projections, and Their Models." Educational Research in Universal Sciences 1.4 (2022): 70-81.
12. Ismonov, Xurshidbek Baxtiyorivich, and Muazzamoy Abduqodir qizi Abduraimova. "Orthogonal Projections and Their Models." Educational Research in Universal Sciences 1.3 (2022): 288-296.
13. Qizi, Abduraimova Muazzamoy Abduqodir. "Projection and Axonometry."