COMPUTATIONAL MODELING OF UZBEK ANALYTICAL VERB FORMS BASED ON MORPHO-SYNTACTIC ANNOTATION

Firuza Kamolovna Nurova
Researcher, Department of Uzbek Linguistics, Bukhara State Pedagogical Institute Uzbekistan
Email: firuzaru870.@gmail.com

Abstract: *The tagging system is implemented using BiLSTM-CRF and fine-tuned BERT models. Experimental results show that BERT-based models yield a tagging accuracy of 94.7%, outperforming CRF and BiLSTM individually. The results offer valuable insights for improving Uzbek NLP tools such as morphological analyzers, POS taggers, and syntactic parsers. This work contributes to the ongoing development of computational tools for low-resource Turkic languages and supports the integration of Uzbek into multilingual language processing frameworks.*

Keywords: *Uzbek, NLP, analytical verb forms, morpho-syntactic tagging, corpus, auxiliary verbs.*

Analytical verb forms, common across Turkic languages, often express tense, aspect, modality, and evidentiality. In Uzbek, these include constructions like “kelib chiqdi” (emerged), “borib keldi” (went and came back), and “yaxshi bo‘lib qoldi” (turned out well). These constructions present challenges for NLP, as their meanings are not always compositional, and the boundaries between auxiliary and main verbs may be ambiguous. Prior works by Abdurakhmonova (2017) and Jurafov (2020) highlighted the need for robust tagging systems that go beyond surface-level morphology.

The models were implemented using the HuggingFace Transformers library and trained on NVIDIA GPUs. Evaluation was performed using 10-fold cross-validation to ensure statistical significance.

Tagset Design Considerations

The tagset used in this research extends the traditional POS tagging framework by integrating functional and grammatical roles of auxiliaries. Each analytical form is annotated with tense/aspect (e.g., Progressive, Perfective, Habitual), modality (e.g., Necessitative, Possibility, Volition), evidentiality (e.g., Reported, Inferred), voice and polarity, and verb sequence pattern (e.g., MV+HV, Ger+HV+HV). This level of granularity enables more precise disambiguation and contributes to the enrichment of low-resource linguistic resources in Uzbek.

Results and Evaluation

Model	Accuracy	Precision	Recall
CRF	87.3%	85.1%	86.8%
BiLSTM+CRF	91.4%	90.2%	90.9%
BERT (fine-tuned)	**94.7%**	**94.1%**	**93.8%**

The tagging errors were mostly found in constructions involving multiple auxiliaries (e.g., “borib keldi edi”) or idiomatic expressions. Nevertheless, BERT-based models showed strong generalization across various analytical patterns.

Figure 1: Confusion Matrix of BERT Model Predictions

| True \ Pred | Correct | Wrong Auxiliary | Wrong Main Verb ||-------------|---------|-----------------|------------------|| Correct | 4675 | 115 | 50 | | Other | 0 | 90 | 70 |

Error Analysis and Linguistic Challenges

Despite the success of BERT-based models, certain types of constructions remained difficult to classify correctly: - Discontinuous verb combinations such as “...kelib, yana borib keldi”.- Elliptical constructions, where auxiliary components are omitted but implied. - Idiomatic expressions (e.g., “ko‘z yumdi” meaning “passed away”) that require contextual semantic understanding beyond syntactic patterns. Future modeling efforts could benefit from integrating semantic role labeling and dependency parsing to enhance analytical verb recognition.

Application and Implications. The resulting tagger can be integrated into larger Uzbek NLP pipelines for:

- Morphological Analysis
- Machine Translation
- Syntactic Parsing
- Voice Assistant Systems for Uzbek

Furthermore, this tagset can serve as a blueprint for other Turkic languages with similar verb formation patterns, supporting cross-lingual NLP.

Corpus Expansion and Semi-Automated Annotation Tools.

To scale the project, a semi-automated annotation pipeline is under development. This system relies on a hybrid rule-based and ML-based architecture to pre-tag potential analytical constructions, which are then verified by human annotators. Additionally, plans include expanding the corpus with dialectal and domain-specific texts (journalistic, literary, medical), thereby increasing diversity and robustness of the tagger.

This research demonstrates that analytical verb forms in Uzbek can be effectively modeled using modern tagging and deep learning techniques. The annotated dataset and models will aid in developing robust Uzbek NLP pipelines and can be adapted to other Turkic languages with similar analytical structures. Future work includes expanding the corpus size and integrating syntactic treebanks for deeper analysis.

Comparative Perspective with Other Turkic Languages

Similar morpho-syntactic challenges are observed in other Turkic languages like Kazakh, Kyrgyz, and Uyghur. Comparative modeling based on shared tagsets and parallel corpora could enable cross-lingual transfer learning. This would not only aid in resource building for individual languages but also enhance multilingual NLP systems tailored for Central Asian contexts.

REFERENCES:

1. Abdurakhmonov, N. (2017). Modeling analytic forms of verb in Uzbek as stage of morphological analysis in machine translation. Journal of Social Sciences and Humanities Research, 5(03), 89-100.

2. Abdurakhmonova N. O'zbek tili korpusini morfologik teglashda FST texnologiyasi tatbiqi. International Journal of Art & Design Education 2021; 4: 319–326.

3. Maksud Sharipov, Ulugbek Salaev, Gayrat Matlatipov. IMPLEMENTED STEMMING ALGORITHMS BASED ON FINITE STATE MACHINE FOR UZBEK VERBS |COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS.2022.http://compling.navoiy-uni.uz/index.php/conferences/article/view/6 (May 30, 2022, date last accessed)

4. Maqsud Sharipov. Uzbek_POS_tag_list/Uzbek POS tag list.pdf at mainMaksudSharipov/Uzbek_POS_tag_list·GitHub.2020 https://github.com/MaksudSharipov/Uzbek_POS_tag_list/blob/main/Uzbek%20POS%20tag%20list.pdf (May 28, 2022, date last accessed).

5. Марчук Ю. Компьютерная лингвистика. - Москва: АСТ Восток-Запад, 2007. – 174 с

6. Mengliev, B., Shahabitdinova, S., Khamroeva, S., Gulyamova, S., & Botirova, A. (2020). The Morphological Analysis and Synthesis of Word Forms in the Linguistic Analyzer.