AUTOMATIC IDENTIFICATION OF SENTENCE COMPONENTS IN THE UZBEK LANGUAGE USING THE HIDDEN MARKOV MODEL

Nuriddinova Gulirano Baxtiyor qizi
Tashkent State University of Uzbek
language and literature Tashkent, Uzbekistan
guliranonuriddinova@gmail.com
https://doi.org/10.5281/zenodo.20063714

Abstract. This article investigates the application of the Hidden Markov Model (HMM) and the Viterbi algorithm in the automatic syntactic analysis of the Uzbek language. Uzbek belongs to the agglutinative group of the Turkic language family and has a distinctive syntactic structure. In this study, sentence components are labeled using the BIO (Begin, Inside, Outside) tagging scheme within a syntactic analysis system, and the statistical probabilities of the model are examined. The results demonstrate the effectiveness as well as the limitations of HMM in the syntactic analysis of the Uzbek language.

Keywords: NLP, HMM, Viterbi algorithm, BIO chunking, syntactic analysis, Uzbek language, machine learning.

Introduction. In recent years, applied linguistics has entered a new stage of development through the integration of artificial intelligence technologies. The problem of modeling, in particular, requires the adoption of new approaches and the integration of methods from machine learning and deep learning. Modeling sentence components is one of the most fundamental and complex tasks in natural language processing. In general, modeling involves studying not the object itself, but its representation or model [5]. Within this context, the automatic identification of sentence components is considered one of the most challenging problems. In agglutinative languages such as Uzbek, the rich morphological structure of word forms introduces both advantages and significant challenges in identifying sentence components [1].

On the Concept of the Hidden Markov Model. The Hidden Markov Model (HMM) is one of the classical and effective methods for modeling sequential data, based on the statistical relationships between hidden states (sentence components) and observed events (words) [2]. HMM represents the dependency between hidden states and observable outputs. The model can be described as follows: Hidden refers to states that cannot be directly observed; Markov indicates that each state depends only on the previous state (the Markov property); and Model denotes a mathematical representation of these relationships. In the field of natural language processing (NLP), HMM is widely applied in tasks such as part-of-speech (POS) tagging, named entity recognition (NER), speech recognition, text segmentation, and syntactic analysis. An HMM consists of three main components:

1.Initial probability (π) – represents the probability distribution of the initial hidden states at the beginning of the sequence.

2.Transition probability (A) – defines the probability of transitioning from one hidden state to another.

3.Emission probability (B) – indicates the probability of observing a particular output given a specific hidden state. HMM formulation:

$$P(O,S)=\pi_{s_1}\prod_{t=2}^{T}a_{S_{t-1}S_t}\prod_{t=1}^{T}b_{S_t}(o_t)$$

Here: O — observed words; S — hidden states (sentence components).

Application of HMM in syntactic analysis works as follows:

Example: “Talaba kitobni diqqat bilan o‘qidi”

Observed sequence: Talaba / kitobni / diqqat / bilan / o‘qidi
Hidden states: subject / object / adverbial modifier / predicate

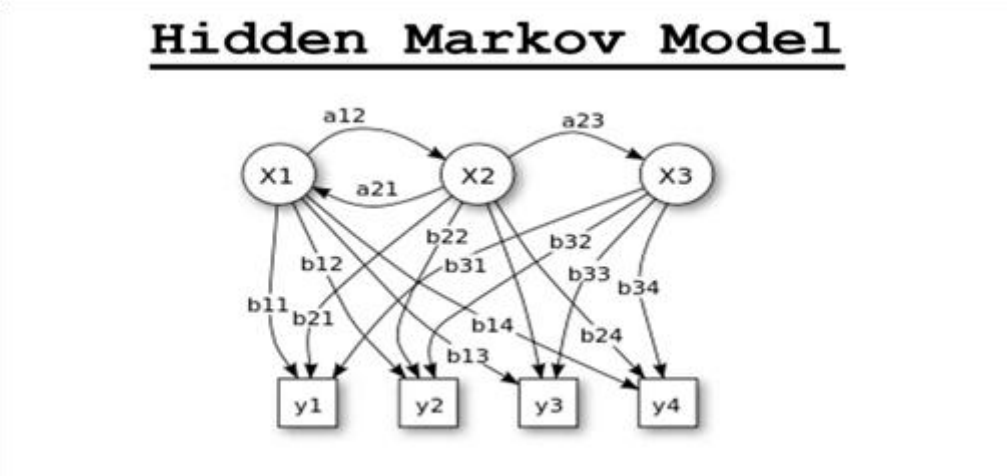

Figure 1. Graphical representation of the Hidden Markov Model

Viterbi Algorithm. The Viterbi algorithm is used to find the most probable syntactic sequence in a Hidden Markov Model (HMM). It computes the highest probability path using the following mathematical formulation:

$$v_t(j) = \max_{i=1} N\left[v_{t-1}(i) * a_{ij} \right] * b_j(o_t)$$

In this context: a_{ij} — transition between states; $b_j(o_t)$ — emission probability of a word [4].

Sentence	Parts of sentence	HMM Tag
Talaba	ega	B-SUBJ
kitobni	to‘ldiruvchi	B-OBJ
diqqat	hol	B-AdvC
bilan	_	I-AdvC
o‘qidi	kesim	B-PRED

Table 1. Data annotated using the BIO chunking method.

For the study, 500 sentences were annotated using the BIO chunking scheme as shown in Table 1 and were used to train the HMM model. The capabilities of the model were evaluated, and general conclusions were drawn.

The Hidden Markov Model demonstrated good performance in sequence labeling tasks and showed the ability to work effectively even with a limited amount of data. It was also able to correctly identify sentence components consistent with the valency of predicates. However, in the Uzbek language, the presence of complex compound sentences and frequent cases where dependent sentence elements are located far from each other led to relatively lower performance of the model. The accuracy of the model for sentence components can be observed in Table 2:

Sentence constituents	Model accuracy (%)	Comment
Ega	80-82%	Ko‘pincha gap boshida keladi.
Kesim	85-90%	Kesim aniq morfologik ko‘rsatkichlar bilan va odatda gap so‘nggida kelgani uchun yuqori aniqlikda.
Aniqlovchi	75-85%	Egaga bog‘langan holatda ajratib oladi, ba’zan hol bilan almashtirish ehtimoli bor.
To’ldiruvchi	80%	Aniq morfologik qo‘shimchalar bilan kelib obyektni ifodalagani uchun yuqori aniqlikda.
Hol	70-80%	Murakkab birikma holatlari ko‘p uchraydi, aniqlovchi bilan almashtirish o‘rinlari mavjud.

Table 2. Accuracy metrics of sentence components in the HMM model

In conclusion, the Hidden Markov Model can be effectively applied to the Uzbek language and is capable of producing good results even on small corpora. It is a computationally efficient and relatively simple model to implement. However, since it is based on statistical computations, it does not achieve high performance in identifying context-dependent sentence components. In future work, higher performance can be achieved by integrating the HMM with deep learning models such as BiLSTM [6].

References:

1.Madatov A. “Morphological Analysis of the Uzbek Language for NLP Tasks.” Journal of Central Asian Studies, 2023.
2.Jurafsky D., Martin J. H. Speech and Language Processing. Stanford University, 2021.
3.Ziyayev A. “BIO Tagging Approach in Uzbek Syntax.” Journal of Computational Linguistics, 2022.
4.Rabiner L. R. “A Tutorial on Hidden Markov Models.” Proceedings of the IEEE, 1989.
5.Po‘latov A. Computational Linguistics. Tashkent, 2009.
6.Sultonov B. “Probabilistic Models for Sentence Component Identification in Uzbek.” Tashkent: Proceedings of UzMU, 2023.
7.Manning C. D., Schutze H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
8.Sayfullayeva R. Modern Uzbek Literary Language. Tashkent: O‘qituvchi, 2007.

9.Nurmonov A. System Linguistics and Its Foundations. Tashkent: Uzbek National Encyclopedia, 2010.
10.G‘ulomov A., Asqarova M. Modern Uzbek Language: Sentence Components and Their Functions. Tashkent: Fan Publishing, 1985..