DEVELOPMENT OF A MAXIMUM ENTROPY-BASED MODEL AND ALGORITHM FOR PART-OF-SPEECH TAGGING IN UZBEK TEXTS

Sharipov Maksud

Urgench State University named after Abu Rayhan Biruni, Associate Professor of the department of "Computer Science"

E-mail: maqsbek72@gmail.com

Abstract. In this work, we consider the development of a model for solving the problem of part-of-speech (POS) tagging in Uzbek language texts. Based on a morphologically annotated dataset, a Maximum Entropy (MaxEnt) probabilistic model was developed for POS tagging. Using this model, and taking into account predefined rules, an algorithm and software were created to perform POS tagging of Uzbek language texts.

Key words. POS tagging, NLP, Maximum Entropy (MaxEnt), parts of speech, probabilistic model.

Annotatsiya. Bu ishda biz o'zbek tili matnlaridagi so'zlarning turkumlarini aniqlash masalasini(POS tagging- Part of Speech Tagging) yechish uchun model ishlab chiqishni ko'rib chiqamiz. Morfologik belgilab chiqilgan dataset asosida POS tagging uchun Maximum Entropy (MaxEnt) ehtimolliliok modeli ishlab chiqildi va shu modelga asoslangan holda oldindan aniqlangan qoidalarni hisobga olib o'zbek tili matnlarini POS tagging qiluvchi algoritm va dastur ishlab chiqilgan.

Kalit so'zlar. POS tagging, NLP, Maximum Entropy (MaxEnt), so'z turkumlari, ehtimollilik modeli.

Аннотация. В данной работе рассматривается разработка модели для решения задачи разметки частей речи (POS) в текстах на узбекском языке. На основе морфологически аннотированного набора данных была разработана вероятностная модель с максимальной энтропией (MaxEnt) для разметки POS. С использованием данной модели и с учетом предопределенных правил были созданы алгоритм и программное обеспечение для выполнения разметки POS текстов на узбекском языке.

Ключевые слова. POS-тегирование, обработка естественного языка, максимальная энтропия (MaxEnt), части речи, вероятностная модель.

Introduction

Part-of-speech tagging (POS tagging) is the process of assigning grammatical labels—such as noun, verb, adjective, and others—to each word in a sentence. POS tagging is one of the fundamental tasks in the field of Natural Language Processing (NLP) and is widely used in various applications such as machine translation, sentiment analysis, and speech synthesis from text. Due to the

ISSN 3030-3907

rich morphological structure of the Uzbek language, solving this task is considered complex. While the POS tagging problem has been addressed for resource-rich languages years ago, it still remains an insufficiently solved issue for the Uzbek language.

In addressing the POS tagging problem, researchers typically employ various approaches: stochastic methods, rule-based methods grounded in linguistic rules, combinations of these two approaches, as well as artificial intelligence techniques. In the POS tagging process, a tagset (a set of tags) and a tagged corpus specific to the language are generally required. A tagset is a collection of tags that represent grammatical categories. It is used to assign the appropriate tag to each word [1,2]. For instance, Table 1 presents several words along with their corresponding proposed tags and tag descriptions.

Table 1.

Tags developed for the noun part of speech in the Uzbek language.

Tag name	Meaning	Examples	A common tag for all
BOT	Birlik ot uchun	*Mashina, kitob, olma*	NOUN
KOT	Ko‘plik otlar	*Mashinalar, kitoblar, talabalar*	
SOOT	Sodda ot	*Bino, bog‘dorchilik*	
QOT	Qo`shma ot	*Markaziy Osiyo*	
JFOT	Juft oti	*Uy-joy*	
TOT	Takroriy ot	*Tomir-tomiriga*	
ABR	Qisqartma otlar	*BMT, UNESCO*	
HOT	Hisob so‘zlar	*Gramm, metr, litr, sm, kg*	

Rule-based part-of-speech tagging algorithms primarily rely on rules written by linguistic experts.

The first rule-based taggers utilized a two-stage architecture. In the first stage, a set of possible parts of speech (POS) corresponding to the exact form of the word is assigned using a lexicon. This lexicon includes not all word forms, but only the lemmas of words. This implies that a certain level of morphological analysis is performed prior to accessing the lexicon. In the second stage, if multiple tags are associated with a word, a set of disambiguation rules is applied to select the most appropriate single tag.

This article [3] is one of the early works in natural language processing based on the MaxEnt model. The authors justify the use of the maximum entropy principle in constructing a statistical model: the model selected is the one that is consistent with known facts while remaining maximally uniform with respect to

ISSN 3030-3907

unknown aspects. The paper presents the mathematical foundations of the MaxEnt model, including the use of Lagrange multipliers and dual optimization methods for parameter estimation. It also discusses algorithms and guidelines for feature selection and application of the model to tasks such as translation, phrase ordering, and sentence segmentation.

This article [4] focuses on enhancing the stability and accuracy of MaxEnt models by introducing constraints such as uncertainty, normalization, and Bayesian priors. The paper explores optimization approaches for MaxEnt models involving complex feature structures, such as "feature forests," and connects them with constrained optimization methods like L-BFGS. The study was tested on textual data and demonstrated the effectiveness of the MaxEnt model in classification tasks.

This work addresses the problem of automatically identifying and selecting features for POS tagging. In random field models, including those based on MaxEnt, gradient-based selection algorithms were developed to determine a set of relevant features. This model is considered useful for POS tagging, sentence segmentation, and other sequence-related problems in NLP [5].

In this scientific paper, the constructed model directly applies the MaxEnt method for part-of-speech (POS) tagging. It generates a probabilistic model to assign the most appropriate tag to each word based on context-dependent features. This model has proven to be highly effective, demonstrating strong performance on datasets in CoNLL format [6].

Materials and methods. Part-of-speech tagging is the process of assigning the appropriate part-of-speech tag (POS label) to each word in a text. The input consists of a sequence of words tokenized as x1,x2,x3,…,xn and a predefined tagset; the output is a corresponding sequence of tags y1,y2,y3,…,yn. Each output element yi corresponds to a specific input element xi.

In the task of part-of-speech tagging, there are cases where ambiguity arises—for example, the word “olma” (apple / don’t take) can belong to the noun class (“Qizil olma qani?” – “Where is the red apple?”) or the verb class (“Sen kitobni olma” – “Do not take the book”). In such cases, the main goal of POS tagging is to determine the correct interpretation based on the context of the text. Traditional probabilistic models, such as Hidden Markov Models (HMM), belong to the class of generative models. These models typically define the joint probability distribution P(X,Y) for the random variables X and Y, where X(x1,x2,x3,…,xn) is the sequence of words, and Y(y1,y2,y3,…,yn) is the corresponding sequence of tags.

ISSN 3030-3907

Maximum Entropy Model. For each token x, the tag y is selected based on the following probability:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y)\right) \quad (1)$$

$$Z(x) = \sum_{y' \in Y} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x,y')\right) \quad (2)$$

Here, *x* is a token (for example, *"kitobni"*), *y* is a candidate tag (for example, "NOUN", "VERB"), $f_k(x,y) \in \{0,1\}$ is the k-*th* feature function, λ_k is the weight coefficient associated with each feature (learned by the model), and *Z(x)* is the normalization constant that sums the probabilities over all possible tags, resulting in: $\sum_{y \in Y} P(y|x) = 1$.

Let us consider the construction of feature functions:

$$f_1(x,y) = \begin{cases} 1, if \ \ x = "kitobni" \ and \ y = NOUN \\ 0, \ \ else \end{cases} \quad (3)$$

- this is a lowercase word function;

$$f_2(x,y) = \begin{cases} 1, if \ x[-2:0] = "ni" \ and \ y = NOUN \\ 0, \ \ else \end{cases} \quad (4)$$

- this is a function to check the suffix;

$$f_3(x,y) = \begin{cases} 1, \ \ if \ is_title(x) = True \ and \ y = PROPN \\ 0, \ \ else \end{cases} \quad (5)$$

- this function is a case-sensitive function;

To define individual feature functions for the model, suffixes are stored in the file suffix_features.txt, which are read for all tags and used to construct the model.

Dataset and model. Tagging is a complex task in the field of Natural Language Processing (NLP). POS (part-of-speech) tagging refers to assigning each word a tag that corresponds to its grammatical function in context. These tags represent the grammatical properties of each word. The process of tagging texts in the Uzbek language is considered highly complex. In order to tag Uzbek texts, the text must first be segmented into tokens (words), for which a tokenizer adapted to the Uzbek language was developed. Subsequently, a specialized web-based tool was created for tagging Uzbek texts, and a tag was assigned to each word. Once the texts were manually tagged, the resulting tagged dataset (corpus) underwent linguistic validation by computational linguistics experts [10]. The texts were annotated as follows:

- *Yer/NOUN yuzida/NOUN yashaydigan/VERB kishilar/NOUN hayoti/NOUN tabiatga/NOUN juda/ADV bog'langan/VERB ./PUNCT*

ISSN 3030-3907

- *Inson/NOUN o‘ziga/PRON kerakli+bo‘lgan/VERB hamma/PRON narsani/NOUN tabiatdan/NOUN oladi/VERB ./PUNCT*

The process of building a model from such a tagged corpus is carried out according to the following algorithm:

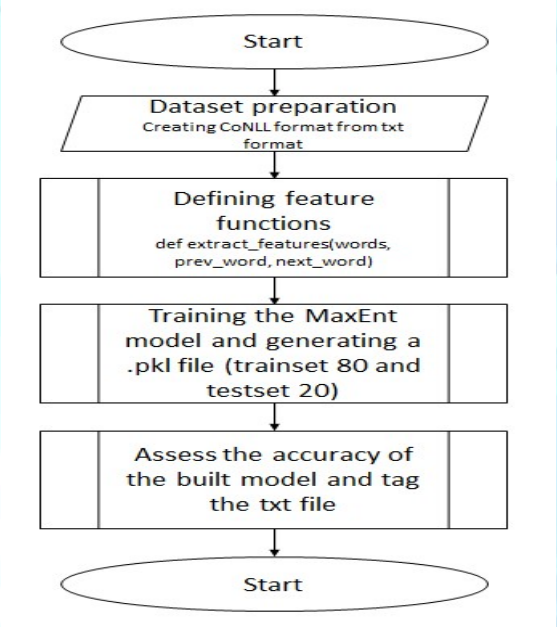


Figure 1. Uzbek word tagging algorithm based on maximum entropy.

The algorithm for building a Maximum Entropy model using a dataset annotated with morphological tags. This algorithm tokenizes sentences in Uzbek texts (written in the Latin script) and predicts the appropriate tags for each token. The constructed model has been uploaded to the following address: https://github.com/MaksudSharipov/UzbekMaxEntTagger. The model can be downloaded and tested via this link.

Table 2.

Accuracy metrics of the POS tagger by tags.

Tag	Precision	Recall	F1-score	Support
ADJ	0.933	0.761	0.839	4564
ADP	0.884	0.790	0.835	1493
ADV	0.875	0.554	0.679	1267
AUX	0.780	0.354	0.486	280
CONJ	0.971	0.935	0.953	1520
MOD	0.743	0.585	0.655	287
NOUN	0.787	0.972	0.870	16567
NUM	0.946	0.622	0.750	1047
PART	0.833	0.725	0.778	309
PRON	0.962	0.841	0.898	1959
PROPN	0.864	0.639	0.734	634

ISSN 3030-3907

PUNCT	0.946	0.988	0.966	3991
VERB	0.934	0.724	0.815	6841

As shown in Table 2, the F1-score accuracy for the AUX tag is 0.486, which is the lowest among all tags. The highest accuracy was achieved in identifying the PUNCT and CONJ tags, with scores of 0.966 and 0.953, respectively. Thus, the overall accuracy of the model is 0.858.

Conclusion. Up to the present time, due to the fact that POS tagger algorithms based on the Hidden Markov Model (HMM) have hardly been developed for Uzbek texts, the model and algorithm proposed in this study are considered highly relevant. To develop HMM-based algorithms for the POS tagging task—a natural language processing problem in the Uzbek language—a morphologically annotated dataset across various categories is required. Creating such datasets demands significant effort, and for the purposes of this study, a dataset was also developed. As a continuation of this work, the future goal is to address the POS tagging problem for Uzbek texts using artificial intelligence models in order to achieve more accurate results.

REFERENCES

1. Ekbal, A., Haque, R., & Bandyopadhyay, S. (2007). Bengali part of speech tagging using conditional random field. In Proceedings of seventh international symposium on natural language processing (SNLP2007), (pp. 131–136).
2. Brants, T. (2000). TnT: A statistical part-of-speech tagger. Sixth Applied Natural Language Processing Conference (Association for Computational Linguistics, Seattle, Washington, USA), (pp. 224–231).https://doi.org/10.3115/974147.974178.
3. Berger, Della Pietra, & Della Pietra (1996). “A Maximum Entropy Approach to NLP”. [Computational Linguistics, 22(1)]
4. Kazama et al. (2003). “Maximum Entropy Models with Inequality Constraints”. [EMNLP 2003]
5. Della Pietra et al. (1997). “Inducing Features of Random Fields”. [IEEE TPAMI].
6. Ratnaparkhi, A. (1996). “Maximum Entropy POS Tagging”. [EMNLP]
7. Cardey, S. and Greenfield, P. (2003), Disambiguating and tagging using systemic grammar, in ‘Proceedings of the 8th International Symposium on Social Communication’, pp. 559–564
8. Birocheau (2003), Etiquetage morphologique et contribution la dsambigusation automatique des ambiguts morphologiques sur un lexique anglais, PhD thesis, Thse de doctorat sous la dir. de Sylviane Cardey, Centre Tesnire, Universit de Franche-Comt, Besanon.

ISSN 3030-3907

9. Eric Brill. 1992. A Simple Rule-Based Part of Speech Tagger. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

10. Sharipov M., Mattiev J., Sobirov J., Baltayev R, Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language (2022) CEUR Workshop Proceedings, 3315, pp. 93 - 98, [Online]. Available: https://www.scopus.com/record/display.uri?eid=2-s2.0-85146120658&origin=recordpage

ISSN 3030-3907