

# Efficient POS Tagger for Kokborok Language

<sup>[1]</sup> Khumbar Debbarma, <sup>[2]</sup> Ruman Sarkar

<sup>[1]</sup>Department of CST, TTAADC Polytechnic Institute, Tripura.

<sup>[2]</sup> Department of CST, Tripura Institute of Technology, Narsingarh.

Corresponding Author Email: <sup>[1]</sup> khum.10jan@gmail.com, <sup>[2]</sup> ruman10k@gmail.com

---

**Abstract**— *The Part of Speech (POS) tagging tries to tag each word with its correct part of speech. In this paper we discuss about the rule-based POS tagger for Kokborok, a resource poor and less digitized Indian language. We employ two machine learning algorithms for supervised approach, Naive Bayes (NB) and decision tree. Experimental results showed rule-based approach performed better than Naive Bayes and Decision tree from supervised approach giving accuracy of 79%, 70% and 71% respectively.*

**Index Terms**—Kokborok, Part of Speech Tagger, Naive Bayes, Decision Tree.

---

## I. INTRODUCTION

Kokborok is the native language of Tripura and is also spoken in the neighboring states like Assam, Manipur, Mizoram as well as the countries like Bangladesh, Myanmar etc., comprising of more than 2.5 millions of people. Kokborok belongs to the Tibeto-Burman (TB) language falling under the Sino language family of East Asia and South East Asia. Kokborok shares the genetic features of TB languages that include phonemic tone, widespread stem homophony, subject-object-verb (SOV) word order, agglutinative verb morphology, verb derivational suffixes originating from the semantic bleaching of verbs, duplication or elaboration [1].

Part-of-speech (POS) tagging is normally a sentence-based approach where each word forming a sentence is labelled or tagged with its correct part of speech. POS tagging deals with assigning a POS tag to the given surface form word. Recently, rule-based approaches, which learn symbolic tagging rules automatically from a corpus, have been reconsidered, to overcome the limitations of statistical approaches. However, in general rule-based approaches alone are not very robust and are not portable enough to be adjusted to new tag sets or new languages. Also, they usually perform no better than their statistical counterparts. Using statistical methods requires rich repertoire of data sources. Kokborok having limited resources is not suitable for all stochastic approaches.

## II. RELATED WORKS

This section comprises of some efficient POS tagging approaches: The authors in [2] present a comparative study of three efficient POS tagging techniques for Bangla language. The authors investigate the performance of n-gram, HMM and Brill's tagger with a very limited resource of an annotated corpus. In [3], a key phrase extraction system is proposed. Three lexicons are introduced for word segmentation (a verb lexicon, a functional lexicon and stop

word lexicon). Moreover, key phrases are sifted by their weighted TF-IDF values.

An evolutionary algorithm-based POS tagger is presented in [4]. The authors try to increase the accuracy of tagging by exploiting statistical measures in evaluating the solutions. A joint solution for POS tagging and dependency parsing is proposed in [5]. However, the authors present an effective pruning strategy to reduce the search space after introducing dynamic programming based decoding algorithms for their models. A novel harmony search optimization method-based POS tagging algorithm is proposed in [6]. The authors investigate the performance of HS (Harmony search) meta-heuristic against genetic algorithm and simulated annealing.

A detailed survey on Indian languages is carried out in [7]. The authors compare different taggers based on Hidden Markov Model (HMM), Maximum Entropy (ME), Support Vector Machine (SVM) and Conditional Random Field (CRF) approaches. The authors conclude that SVM based taggers achieve more accuracy in case of Malayalam language where CRF based taggers performs well for Bengali. Hammad Ali [8] propose a four level layered unsupervised POS tagger based on Baum-Welch trained HMM. The author follow the common tag set for Indian languages and IIIT tag set guidelines. The algorithm identifies a universal category comprising of 12 different categories. Second level consists of the application of disambiguation rules. The fourth level tags the multi verb words and per-forms local word grouping.

A bigram HMM based approach is proposed in [9]. The authors use an annotated corpus of 20,000 words. The authors implement the HMM approach using the well known Viterby algorithm and estimation of HMM parameter is based on Maximum likelihood method. In [10] a CRF based POS tagger for Malayalam language is proposed. The authors present their analysis based on an annotated corpus of 1028 sentences. The authors incorporate a trigram-based tagging scheme. A trigram HMM based POS tagger is investigated in [11]. Furthermore, unknown word handling is based on prefix analysis and word type analysis combined with suffix

analysis. The authors investigate their proposed algorithm with respect to Bengali, Hindi, Marathi and Telugu languages.

### III. LINGUISTIC CHARACTERISTICS OF KOKBOROK

Kokborok is classified as an agglutinative language. Traditional grammars of Kokborok recognize five parts of speech i.e., noun, adjective, pronoun, preposition and verb. A word in Kokborok consists of several morphemes with clear morpheme boundaries. For example, “phataro ta thang di” meaning “do not go outside” consist of 2 words and 5 morphemes. Two different types of morphologies are recognized for simple Kokborok words [12].

#### A. Inflectional Verb and Noun Morphology

Most verbs have a monosyllabic root, and the main method for processing verb phrases is to add suffixes to the root. Kokborok verbs always occur in bound form to which multiple affixes are added to give the tense, manner of action. The suffixes can be classified in three layers at least [13].

Monosyllabic nouns are relatively rare in Kokborok where bisyllabic formations are dominant. This is due to the widespread process of compounding, either true compounding when two lexical roots form a new word.

POS tagging of Kokborok is usually performed on a morpheme basis. Accordingly, morphological analysis is essential for POS assignment. Kokborok is a postpositional language with many kinds of noun endings and verb endings. It is these functional morphemes, rather than the order of tokens, that determine grammatical relations such as a nouns syntactic function, a verbs tense, aspect, modals, and even modifying relations between to kens.

### IV. RULE BASED KOKBOROK POS TAGGER

The rule-based POS tagging models apply a set of hand written rules and use contextual information to assign POS tags to words. These rules are often known as context frame rules. For example, a context frame rule might say something like: If a word X is segmented to corresponding unknown morpheme and morpheme with possessive feature tag unknown as noun. Rule Based Kokborok POS Tagger is composed of Tokenizer, Stemmer, Morphological Analyzer and Tag generator modules as in [14]. The input sentence is tokenized by the tokenizer module based on space in between the consecutive word. Each token is stemmed to affixes and root words using the affix and root dictionary. The morphemes are then analyzed by morphological Analyzer using morpho syntactic features and lexical rules which tags the morphemes and gives feature information. The tag generator applies morphological rules based on tags and feature information of the morphemes and tags the tokens. Below gives the algorithm for rule-based Kokborok POS Tagger.

#### Algorithm 1 POS Tagger

- 1: **procedure** TAGPOS (Input Text)
- 2: **while** until every Token is tagged **do**
- 3: **if** Prefixes and/or Suffixes are attached
- A. *then*
- 4: separate attached Prefixes and/or Suffixes by looking into the affix dictionary and retrieve the free and bound morphemes.
- 5: **if** the free morpheme does not occur in the root dictionary **then**
- 6: send the free morpheme to the complex word handler
- 7: **if** complex word handler cannot handle the provided morpheme **then**
- 8: tag it as Proper Noun.
- 9: **end if**
- 10: **end if**
- 11: **end if**
- 12: **end while**
- 13: Assign proper POS Tag to the tokens by applying the morphological rules
- 14: **end procedure**

### V. PERFORMANCE EVALUATION AND ANALYSIS

Kokborok being a highly agglutinative language requires accurate morphological analysis for proper tagging of each token. Deletion of some part of free morphemes after inflection results in unknown morphemes after segmentation. For example, Ang(I) (free morpheme) + ni (bound morpheme/suffix) → Ani(my/mine). The word Ani is segmented to unknown morpheme “A” and suffix “ni”. Analyzing unknown morphemes and tagging is a problem in itself. Difficulty in also arises from indistinct word categories.

Word having same semantics may have different part of speech depending on contextual position of the word. For example, in the phrase “buphang kuchuk” (high tree) and “abo kuchuk” (it is high) the word “kuchuk” (high) behaves as adjective and noun respectively. The rule-based Kokborok POS Tagger is evaluated using input text of 5468 Kokborok sentences having 46348212 words. Table below gives the percentage of output of POS Tagger. The Tagged words are manually checked for correctness. Incorrect tagging results from insufficient rules to handle unknown morphemes, segmentation of a part of free morpheme that is present in the affix dictionary. Context dependency of words contributes to the assignment of wrong POS tags. Unknown words and morphemes are wrongly tagged due to absence of unknown word/morpheme handler.

**Table 1.** Evaluation Results

Categories	Percentage
correctly tagged words	79%
wrongly tagged words	12%
wrongly tagged unknown words	9%

## VI. SUPERVISED APPROACH

Stochastic models are more popular than rule-based POS taggers as these are language independent and easy to use. Among the entire stochastic models, HMMs is quite popular but it requires a huge amount of annotated corpus. Simple HMMs do not work well when small amount of labelled data are used to estimate the model parameters. Incorporating diverse features in an HMM-based tagger is also difficult and complicates the smoothing typically used in such taggers [14]. Stochastic approach using CRF and SVM gave better results than rule-based approach [14]. We have used Naive Bayes and Decision tree for supervised approach

## VII. EXPERIMENTS AND EVALUATION

The corpus obtained from Kokborok magazine “Aitorma” is tagged using rule-based POS tagger and corrected manually. We use the WEKA tool, which provides these two algorithms among many other machine learning algorithms. Naive Bayes is denoted with NB and decision tree is denoted with J48. Table below presents the performance evaluation based on precision, recall and F-score measure.

**Table 2.** Evaluation Results

Method	Precision	Recall	F-score
NB	70.60	40.20	51.10
J48	71.90	51.60	60.80

## VIII. CONCLUSION AND FUTURE WORK

In this paper, we have discussed the rule-based POS tagger and statistical model. We achieved the accuracies of 79%, 70% and 71% in rule based, Naive Bayes and Decision tree-based POS taggers respectively. Since Kokborok is an agglutinative language POS tagging depends on the morphological features the rule-based POS tagger gave better results compared to supervised models.

Future work includes the development of language resources. The unknown morpheme handler and Named Entity recognition module may be included to improve the accuracy in the POS taggers. Inclusion of Rules for handling ambiguous words.

## REFERENCES

- [1] Khumbar Debbarma, Braja Gopal Patra, Dipankar Das and Sivaji Bandyopadhyay. 2012. Morphological Analyzer for Kokborok. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING 2012,
- [2] F. M. Hasan, N. UzZaman, M. Khan. 2007. Comparison of different POS Tagging Techniques (n-gram, HMM and Brills tagger) for Bangla. Advances and Innovations in Systems, Computing Sciences and Software Engineering, 121–126
- [3] X. Huang, J. Chen, P. Yan, X. Luo. 2005. Word Segmentation and POS Tagging for Chinese Keyphrase Extraction. Advanced Data Mining and Applications Lecture Notes in Computer Science, 3584:364–369 .
- [4] R. Forsati, M. Shamsfard. 2014. Hybrid PoS- tagging: A cooperation of evolutionary and statis- tical approaches. Applied Mathematical Modelling, 38(13):3193-3211
- [5] Zhenghua Li, Min Zhang, Wanxiang Che, Ting Liu, Wenliang Chen. Joint Optimization for Chinese POS Tagging and Dependency Parsing. IEEE/ACM Transactions on Audio, Speech, and Language Pro- cessing, 22(1):274 – 286.
- [6] R. Forsati, M. Shamsfard, P. Mojtahedpour.. An Efficient Meta Heuristic Algorithm for POS-tagging. Fifth International Multi-Conference on Comput- ing in the Global Information Technology (ICCGI), 2010, 93 – 98
- [7] D. Kumar, G. S. Josan. Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. International Journal of Computer Applications, 6(5):93 – 98 .
- [8] H. Ali. 2010. An Unsupervised Parts-of-Speech Tagger for the Bangla language. Department of Computer Science, University of British Columbia
- [9] S.K. Sharma, G.S. Lehal. 2011. Using Hidden Markov Model to improve the accuracy of Punjabi POS tagger. IEEE International Conference on Computer Science and Automation Engineering (CSAE), 697 –701.
- [10] V. Krishnapriya, P. Sreesh, T.R. Harithalakshmi, T.C. Archana, N. J. Vettath. 2014. Design of a POS tagger using conditional random fields for Malayalam. First International Conference on Computational Systems and Communications (ICCS), 370 –373.
- [11] K. Sarkar, V. Gayen. 2013. A Trigram HMM- Based POS Tagger for Indian Languages. Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) Advances in Intelligent Systems and Computing, 199:205–212.
- [12] Francois Jacquesson, 2003, Kokborok a short analysis, Hukumu, 10th anniversary volume, Kokborok tei Hukumu Mission, pp.109-122,
- [13] Francois Jacquesson, 2008 A Kokborok Grammer, Agartala Dialect
- [14] B. G. Patra, K. Debbarma, D. Das, S. Bandyopadhyay. 2012 Part of Speech (POS) Tagger for Kokborok. Proceedings of COLING 2012, Mumbai, December 2012., 923-932.