Part of Speech (POS) Tagger for Kokborok

Braja Gopal Patra[1] Khumbar Debbarma[2] Dipankar Das[3] Sivaji Bandyopadhyay[1]
(1) Department of Compute Science & Engineering, Jadavpur University, Kolkata, India
(2) Department of Compute Science & Engineering, TIT, Agartala, India
(3) Department of Compute Science & Engineering, NIT Meghalaya, Shillong, India
brajagopal.cse@gmail.com, khum_10jan@yahoo.co.in,
dipankar.dipnil2005@gmail.com,sivaji_cse_ju@yahoo.com

ABSTRACT

The Part of Speech (POS) tagging refers to the process of assigning appropriate lexical category to individual word in a sentence of a natural language. This paper describes the development of a POS tagger using rule based and supervised methods in Kokborok, a resource constrained and less computerized Indian language. In case of rule based POS tagging, we took the help of a morphological analyzer while for supervised methods, we employed two machine learning classifiers, Conditional Random Field (CRF) and Support Vector Machines (SVM). A total of 42,537 words were POS tagged. Manual checking achieves the accuracies of 70% and 84% in case of rule based and supervised POS tagging, respectively.

KEYWORDS : Kokborok, POS Tagger, Suffix, Prefix, CRF, SVM, Morph analyser.

1 Introduction

From the very beginning, POS tagging has been playing its significant roles in several Natural Language Processing (NLP) applications such as chunking, parsing, developing Information Extraction systems, semantic processing, Question Answering (QA), Summarization, Event Tracking etc. To the best of our knowledge, no prior work on POS tagging has been done for Kokborok except the development of a stemmer (Patra et al., 2012). Thus, in this paper, we have basically described the development of a POS tagger in Kokborok, a less privileged native language of the Borok people of Tripura, a state in North Eastern part of India. Kokborok is also spoken by neighboring states such as Assam, Manipur, Mizoram and the countries like Bangladesh, Myanmar etc. The language comprises of more than 2.5 millions of people[1] and belongs to Tibeto-Burman (TB) language family. It has several unique features if compared with other South-Asian Tibeto-Burman languages. Kokborok literatures were written in Koloma or Swithaih borok script which suffered massive destruction. Overall, the Kokborok language is very scientific and the people use a script similar to Roman script to project the tonal effect. As the language follows the Subject-Object-Verb (SOV) pattern and its agglutinative verb morphology is enriched by the Indo-Aryan languages of Sanskrit origin. The affixes play an important role in framing the structure of the language, e.g., prefixing, suffixing and compounding form new words in this language. In case of compound words, some infixing are also seen where no specific demarcation and morphology is found. Mainly, the root words appear in bounded forms and are joined together to form the compound words.

In general, the POS tagger for the natural languages are developed using linguistic rules, probabilistic models and combination of both. To the best of our knowledge, the POS tag set is not available in Kokborok as no prior work has been carried out in this language. Thus, we prepared a POS tag set by ourselves with the help of linguists by considering different characteristics of the similar Indian languages.

Several POS taggers have been developed in different languages using both rule based and statistical methods. Different approaches to POS tagging for English have already been developed such as Transformation based error-driven learning (Brill, 1995), Decision tree (Black et al., 1992), Hidden Markov Model (Cutting et al., 1992), Maximum Entropy model (Ratnaparkhi, 1996) etc. It was also found that in a practical Part-of-Speech Tagger (Cutting et al., 1992), the accuracy exceeds 96%.

The rule based systems require handcrafted rules and are typically not very robust (Brill, 1992). POS tagger in different Indian languages such as in Hindi (Dalal et al., 2007; Shrivastav et al., 2006; Singh et al., 2006), Bengali (Dandapat et al., 2007; Ekbal et al., 2007; Ekbal and Bandyopadhyay, 2008a), and Manipuri (Kishorjit et al., 2011; Singh and Bandyopadhyay 2008; Singh et al., 2008) etc. have also been developed using both rule based and machine learning approaches. In case of rule based POS Tagging, we considered the help of three dictionaries, namely prefix, suffix and root dictionary. It is also observed that the Probabilistic models have been widely used in POS tagging as they are simple to use and language independent (Dandapat et al., 2007). Among the probabilistic models, Hidden Markov Models (HMMs) are quite popular but it performs poor when less tagged data is used to estimate the parameters of the model. Due to the scarcity of POS tagged corpus in Kokborok, among different machine learning algorithms,

[1] http://tripura.nic.in

we have used only CRF and SVM to accomplish the POS tagging task. CRF is a widely used probabilistic framework for sequence labelling tasks. In our case, we observed that the accuracies achieved in the rule based POS tagger is less than the CRF based POS tagger whereas the accuracy of CRF based POS tagger is less than SVM based POS tagger.

The rest of the paper is organized in the following manner. Section 2 gives a brief discussion about word features in Kokborok whereas Section 3 details about resources preparation. Section 4 describes the implementation of rule based POS tagger and Section 5 gives the detail study of Machine learning algorithms, feature selection, implementation and their results while the conclusion is drawn at the end.

2 Word Features in Kokborok

In general, Kokborok possesses unique features like agglutination and compounding. Specially, it has both free and bound root words and has more numbers of bound root words compared to English. In Kokborok, the inflections play the major role and almost all verbs and many of noun root words are bound. It is found that the free root words are nouns, pronouns, some adjectives, numerals etc. The compound words are formed by joining multiple root words affixed with multiple suffixes or prefixes. It is found by the linguistic observations that we can classify the Kokborok words into following seven categories as given below.

i) Only root word (RW). For e.g., Naithok (beautiful)
ii) Root words (RW) having a prefix (P). For e.g., **Bu**pha (my father)
iii) Root words having a suffix (S). For e.g., Braja**no** (to Braja)
iv) P+RW+S. For e.g. **Bu**kumui**ni** (His/Her Brother In Law's)
v) P+RW+S+S… For e.g., Ma(P)+thang (to go)+lai(S)+nai(S)→**Ma**thang**lainai**(need to go)
vi) RW+RW… For e.g., Khwn (Flower)+Lwng(Garden)→Khwmlwng(Flowergarden)
vii) RW+S+RW+S. For e.g., Hui(RW)(to hide)+jak(S)+hui(RW)+jak(S)+wi(S) → Hujakhujakwi (Without Being Seen)

We observed that there is less number of free root words. In Kokborok, affixes are of two types, i.e. derivational affixes and inflectional affixes (Debbarma et al., 2012). In Kokborok, the prefixes are very limited in numbers, generally inflectional and do not change the syntactic category when added to a root word but the suffixes are of both inflectional and derivational. A total of 19 prefixes and 72 suffixes are found in Kokborok.

3 Resource Preparation

In the following sections, we have discussed about the basic requirements of our experiments. The first section discusses about the dictionaries used in the experiments and their formats and in the final section, we have presented the POS tagset for Kokborok which is used for our experiments.

3.1 Dictionaries

We used three dictionaries namely prefix, suffix and root. Prefix and suffix dictionaries contain the list of prefixes and suffixes along with the word features like TAM (Tense, Aspect and Modality), gender, number and person etc. Root dictionary is a bilingual dictionary containing

1895 root words. The format of root dictionary is <root><lexical category><English meaning>. This bilingual dictionary is used for testing of the POS tagger.

3.2 The Tagset

The Kokborok language is one of the agglutinative languages in India and its word formation technique is quite different from other Indian languages. Thus, the POS tagset for Kokborok has been developed keeping the similarity of the POS tagset with other Indian languages[2] in mind. The POS tagset used in this task is given below in Table 1.

POS	**Types/ Tag**	**Examples**
Noun	Proper (NNP), Common (NNC), Verbal (NNV)	Aguli, yachakrai (All names), Chwla(boy), bwrwi(girl), khaina(to do), phaina(to come)
Pronoun	Personal (PRP)	Ang(I), Nwng(you), Bo(He/she), Ani(my)
Adjective	JJ	Naithok(beautiful), kwchwng(bright)
Determiner	Singular (DTS), Plural (DTP)	Khoroksa(a), Joto(all), bebak(every)
Predeterminer	PDT	Aa(that), o(this)
Conjunction	CC	Bai(and), tei(or)
Verb	Root (VB), Present (VBP), Past (VBD), Gerund (VBG), Progression (PROG), Future (VBF)	Cha (to eat), khai (to do), Chao (eat), khaio (do), Chakha (ate), phaikha (came), Chawi (eating), khaiwi (doing), Tongo (is/am/are), tongmani (was/were), Chanai(will eat), khainai (will do)
Inflectors	*D	O (to), Rok([charai(child)rok]-children
Quantifiers	QF	Kisa(less), kwbang(more)
Cardinal	CD	Sa(one), nwi(two)
Adverb	RB	Twrwk(slow), dakti(fast)
Interjection	UH	Bah(wao), uh(huh)
Indeclinable	ID	Haiphano(still), Abonibagwi(that’s why)
Onomatopes	ON	Sini-sini, sek-sek,sep-sep
Question Words	QW	boh(which), sabo(who), Saboni(whose)
Compound word	CW	
Unknown	UNK	
Symbol	SYM	`,~,@,#,$,%,^,&,*,_,+,-,=,<,>,.,’, etc.

Table 1 – POS Tagset for Kokborok.

4 Rule Based POS Tagger

In case of rule based POS tagger, the basic POS tags are assigned to each of the words in a natural language sentence using the morphological rules. The descriptions of the different modules as shown in Figure.1 are as follows:

- **Tokenizer**: Based on the space in between consecutive words, each word of a sentence is separated or tokenized.

[2] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

- **Stemmer (Patra et al., 2012):** It identifies the prefixes and suffixes using the affix dictionaries and finds the root words.
- **Morphological Analyzer & Tag generator:** Different analysis on the stemmed words and suffixes are performed using the lexical rules and morpho-syntactic features. Then, the POS tags are assigned to the words based on the tagset and morphology rules.
- **Dictionary:** Prefix, suffix and root dictionaries are described in Section 3.
- **Morpho syntactic Rules:** These are the heuristic rules based morphological characteristics of the words. For e.g., VB + kha (suffix) = VBD, VB + o(suffix)=VBP etc.

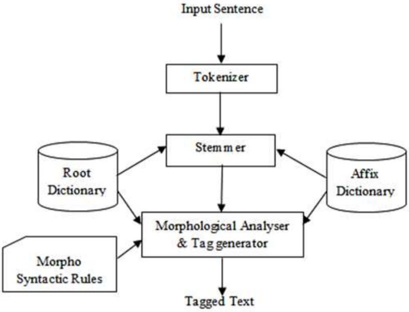

FIGURE 1 –System Diagram of Rule based Morphology driven POS Tagger.

4.1 Algorithm

1. Give input text to the tokenizer module.
2. Repeat step 3 and 4 until each token is tagged.
3. Check for prefixes and suffixes and separate them with the help of affix dictionaries and check if the stemmed word occurs in the root dictionary or not. The words which are not stemmed are sent to the complex word handler module.
4. The complex words are stemmed separately, if these words are not stemmed by complex word handler and tag them as the Named Entities (NEs).
5. Apply the morphological rules on the affixes and root words for identifying the POS tag of the words according to the output of the morphological analyzer.

4.2 Evaluation and Result Discussion

In Kokborok, word categories are not distinct; all the verbs are under the bound categories whereas another problem is to classify basic root forms according to their word classes as the distinction between noun and adjectives is often vague while the distinction between the noun and verb classes is relatively clear. It is found that distinction between a noun and an adjective becomes unclear because structurally a word may be a noun but contextually it is an adjective. For e.g., Uttor Bharato watwi kwbang wakha (“North” “India” “lots” “rain” “happened”). Here north is an adjective where as in the sentence, “Abo uttor” (that is north) the word ‘uttor’ is a noun. Thus, the word ‘uttor’ may be an adjective or a noun but the POS of the word in lexicon is noun there by making it difficult to extract the exact POS for the word appearing in various sentences.

The assumption made for the word categories depends upon the root category and affix information that are available from the dictionaries. Further a part of root may also be a prefix which leads to wrong tagging. It is found that the verb morphology is more complex than that of noun. When multiple suffixes added to a verb, it's difficult to find the POS category of the word as the specific rules are not available. The input of 2525 Kokborok sentences of 42537 words was supplied to the tagger . Sometimes, two words get fused to form a complete word and handling such collocations is difficult. Table 2 shows the percentage of tagging output based on the actual and correctly tagged words. There are some unknown words which could not be tagged based on rules available. Due to the unavailability of root dictionary, the performance of POS tagger was reduced effectively. A word can be easily formed by affixation or compounding in Kokborok, so the number of unknown words are relatively large. The accuracy of the tagging can be further improved by introducing more numbers of linguistic rules and adding more root words to the dictionary.

Items	**Percentage**
Correctly tagged words	70%
Wrongly tagged words	22%
Wrongly tagged unknown words	8%

TABLE 2 – Results of the Rule Based POS Tagger.

5 Stochastic POS Taggers

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 (Ekbal and Bandyopadhyay, 2008b). Thus, we have used Conditional Random Fields (CRF) (Lafferty et al., 2001) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995) frameworks to develop Stochastic POS taggers for the resource constrained Kokborok language.

5.1 Feature Selection

Feature selection plays important role in CRF based machine learning framework. The main features for POS tagging are selected based on the different combinations of available words and tags. As the Kokborok is one of the highly inflected and agglutinative Indian languages, the suffix and prefix features are the effective features in POS tagging task. We have considered different combinations of features to get the best feature set for POS tagging task. Following are the sample and the details of the set of features that have been included in the above list for POS tagging in Kokborok:

F={$w_{(i-m)}$,$w_{(i-m+1)}$,……$w_{(i-1)}$, w_i, $w_{(i+1)}$,……..$w_{(i+n)}$, |prefix|=n, |suffix|=n, Context word feature, Digit information, Symbol, Length of the word, Frequent word}

Word suffix: Kokborok is highly inflected language. So, the word suffix information is one of the most important features as it is very helpful to identify the POS classes. This feature can be used in two different ways. The first way is to check whether a word has a suffix or not. If yes, then set the suffix feature 1 else set 0. The second way is to check whether a suffix is changing the POS class of the root word. If yes, then set change POS feature 1 else set 0.

Word prefix: Word prefix information is also helpful to identify the POS class of the word. This feature has been introduced with the observation that the words of the same category POS tags contain some common prefix. This feature has been used in a similar way as word suffixes.

Context word Feature: The immediate previous and next word of a particular word can also be used as feature, i.e., the surrounding words can play an important role in deciding the POS tag of the current word.

Digit information: If any word consists of any digit, then set the digit feature to 1 otherwise 0. It helps to identify the QF (Quantifier) tag.

Symbol: If the token consists of symbols like (%, $,. etc.), then set the symbol feature to 1, otherwise set it to 0. This helps to identify the SYM tag.

Length of a word: It is found that length of a word is an effective feature in deciding POS tag of the word (Singh et al., 2008). If the length of a word is four or less, set the length word feature to 1, otherwise set it as 0. The motivation of using this feature is to distinguish the Personal pronoun from the nouns. We observed that words of very short length are generally Personal pronoun.

Frequent Word: A list for frequently occurring word is prepared for the training corpus. The words that occur more than 10 times in the entire training corpus are considered as the frequent words. The feature for the frequent word is set to 1 if they are in the list else set it as 0. This has been observed that frequently occurring words are rarely proper nouns.

5.2 Evaluation

For applying the statistical models in Kokborok, we required huge amount of annotated corpus in order to achieve good result. But, Kokborok is less computerized language and the corpora for training and testing were not available. During the manually annotation, we faced the problems due to agglutinative structure of the Kokborok language.

5.2.1 Experimental Results of CRF

We have conducted several experiments by considering the different combination of features to find out the best combination of features and feature templates. From the analysis, we observed that our proposed features as mentioned in Section 5.1 give the best results for testing purpose.

We have designed three types of modules based on the CRF Frameworks. The first module makes use of simple contextual features (i.e. CRF), whereas the second module uses the information of affixes along with contextual information (i.e. CRF+suf.). In order to increase the accuracy of the system, we have integrated the morphological information with the model (i.e. CRF + suf. +MA_F). The tagging accuracy of the CRF based POS tagging model has been evaluated as the ratio of correctly tagged words with respect to the total numbers of words. We have trained the system on different data size and the result is shown in Table 3.

The above experiment leads us to the following observations that the use of suffix information plays an important role in achieving the accuracy of the system, especially when the training data is less. Furthermore, the morphology of the word gives significant improvement in the accuracy over the CRF and CRF+suf models.

It was found that the CRF based POS tagger performs far better than the morphology driven POS tagger and has less computational complexity. We have also conducted the experiments with large number of features but, the inclusion of the features decreases the accuracy. It is found that large number of features works well when large amount of annotated corpus is available for training. The other reason was the biasness of noun tags in the corpus.

	10K	**20K**	**40K**
CRF baseline model	59.67	63.51	65.72
CRF + suf.	67.23	73.57	76.25
CRF + suf. + MA_F	74.57	79.53	81.67
SVM baseline model	60.51	64.26	68.32
SVM + suf.	69.38	72.66	76.97
SVM + suf. + MA_F	75.52	80.47	84.46

TABLE 3 – Tagging Accuracies In %age With Different Template For CRF & SVM.

5.2.2 Experimental Results of SVM

Same training set which was used for CRF is also used for SVM based experiments. We also conducted several experiments considering the different combination of features to find out the best combination of features and feature templates. From the analysis, we found that the similar features of CRF also produced the best results for testing of SVM based POS Tagger.

We have also conducted several experiments for the various polynomial kernel functions and found that the system is giving the best result for the second degree kernel functions. It has been also observed that the pair wise multi-class decision strategy performs better than the than the one-vs.-rest strategy. The models described here are simple and quite good for automatic POS tagging even less amount of tagged corpus was available. The best performance is achieved when suffix information and morphological information is added to the system.

SVM performs far better than the CRF based POS tagger. The performance in SVM can be improved significantly by including the language specific resources such as lexicon and inflection lists. It is found that a Named Entity Recognizer (NER) and a Multiword Identification Systems are necessary to reduce the large number of errors that involve proper nouns and different multiword expressions. The experiments of SVMs are also conducted on same type of data set and same features as shown in Table 3.

Conclusion and Future works

In this paper, we have described the development of POS taggers using both rule based and statistical models. We achieved the accuracies of 69%, 81.67% and 84.46% in rule based, CRF based and SVM based POS taggers, respectively with respect to 26 different POS tags.

Future work includes the development of language specific resources such as lexicon and inflection lists. The Named Entity recognition module may be included to improve the accuracy in the POS taggers. Some language specific rules should be implemented to handle the Complex words in rule based POS tagger. Other experiments like voting technique for two or more models may be an interesting research direction.

References

Black, E., Jelinek, F., Lafferty, J., Mercer, R., and Roukos, S. (1992). Decision tree models applied to the labeling of text with parts-of-speech. In Proceedings of the DARPA Speech and Natural Language Workshop, pages 117-121.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing, pages 224-231, Association for Computational Linguistics.

Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language, pages 112-116, Association for Computational Linguistics.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics, 21(4):543-565.

Carlos, C. S., Choudhury, M., and Dandapat, S. (2009). Large-coverage root lexicon extraction for Hindi. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 121-129, Association for Computational Linguistics.

Choudhury, S., Singh, L., Borgohain, S., and Das, P. (2004). Morphological Analyzer for Manipuri: Design and Implementation. Applied Computing, 123-129.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3): 273-297.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In Proceedings of the third conference on Applied natural language processing, pages 133-140. Association for Computational Linguistics.

Debbarma, Binoy and Debbarma, Bijesh (2001). Kokborok Terminology P-I, II, III, English-Kokborok-Bengali. Language Wing, Education Dept., TTAADC, Khumulwng, Tripura.

Debbarma, K., Patra, B. G., Debbarma, S., Kumari, L., and Purkayastha, B. S. (2012). Morphological analysis of Kokborok for universal networking language dictionary. In Proceedings of First International Conference on Recent Advances in Information Technology, pages 474-477. IEEE.

Dalal, A., Nagaraj, K., Swant, U., Shelke, S., and Bhattacharyya, P. (2007). Building feature rich pos tagger for morphologically rich languages: Experience in Hindi. In Proceedings of ICON.

Dandapat, S., Sarkar, S., and Basu, A. (2007). Automatic Part-of-Speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 221-224. Association for Computational Linguistics.

Ekbal, A., and Bandyopadhyay, S. (2008a). Part of speech tagging in Bengali using Support Vector Machine. In proceedings of the International Conference on Information Technology, 2008. ICIT'08, pages 106-111. IEEE.

Ekbal, A., and Bandyopadhyay, S. (2008). Web-based Bengali News Corpus for lexicon Development and POS tagging. POLIBITS, ISSN 1870, 9044(37):20-29.

Ekbal, A., Haque, R., and Bandyopadhyay, S. (2007). Bengali Part of Speech Tagging using

Conditional Random Field. In Proceedings of Seventh International Symposium on Natural Language Processing (SNLP2007), pages 131-136.

Kishorjit, N., Laishram, J., Haobam, V., Soibam, A., Longjam, N., Lourembam, S. and Bandyopadhyay, S. (2009). Unsupervised POS Tagging for Manipuri Text. In Reso-illusion 2009, MIT, Imphal, India.

Kishorjit, N., Salam, B., Romina, M., Chanu, N. M., and Bandyopadhyay, S. (2011). A Light Weight Manipuri Stemmer. In The Proceedings of National Conference on Indian Language Computing (NCILC), Chochin, India.

Kumar, D., and Josan, G. S. (2010). Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. International Journal of Computer Applications IJCA, 6(5):1-9.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 282–289.

Patra, B. G., Debbarma, K., Debabarma, S., Das, D., Das, A. and Bandyopadhyay, S. (2012). A light Weight Stemmer for Kokborok. In Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012), Yuan Ze University, Chung-Li, Taiwan, pages 318-325.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Proceedings of the conference on empirical methods in natural language processing, volume 1, pages 133-142.

Shrivastav, M., Melz, R., Singh, S., Gupta, K. and Bhattacharyya, P. (2006). Conditional Random Field Based POS Tagger for Hindi. In Proceedings of the MSPIL, pages 63-68.

Singh, S., Gupta, K., Shrivastava, M., and Bhattacharyya, P. (2006). Morphological richness offsets resource demand-experiences in constructing a POS tagger for Hindi. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 779-786.

Singh, T. D. and Bandyopadhyay, S. (2005). Manipuri Morphological Analyzer. In Proceedings of the Platinum Jubilee International Conference of LSI, University of Hyderabad, India.

Singh, T. D., and Bandyopadhyay, S. (2008). Morphology driven Manipuri POS tagger. IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91-98, IIIT, Hyderabad, India.

Singh, T. D., Ekbal, A., and Bandyopadhyay, S. (2008). Manipuri POS tagging using CRF and SVM: A language independent approach. In proceeding of 6th International conference on Natural Language Processing (ICON-2008), pages 240-245.