Identification of POS Tag for Khasi Language based on Hidden Markov Model POS Tagger

Sunita Warjri[1], Partha Pakray[2], Saralin Lyngdoh[1], Arnab Kumar Maji[1]

North Eastern Hill University Shillong, Meghalaya[1] &
National Institute of Technology Silchar, Assam[2]
India
{sunitawarjri,parthapakray,saralyngdoh,arnab.maji}@gmail.com

Abstract Computational Linguistic (CL) becomes an essential and important amenity in the present scenarios, as many different technologies are involved in making machines to understand human languages. Khasi is the language which is spoken in Meghalaya, India. Many Indian languages have been researched in different fields of Natural Language Processing (NLP), whereas Khasi lacks substantial research from the NLP perspectives. Therefore, in this paper, taking POS tagging as one of the key aspects of NLP, we present POS tagger based on Hidden Markov Model (HMM) for Khasi language. In this present preliminary stage of building NLP system for Khasi, with the analyses of the categories and structures of the words is started. Therefore, we have designed specific POS tagsets to categories Khasi words and vocabularies. Then, the POS system based on HMM is trained by using Khasi words which have been tagged manually using the designed tagsets. As ambiguity is one of the main challenges in POS tagging in Khasi, we anticipated difficulties in tagging. However, by running with the first few sets of data in the experimental data by using the HMM tagger we found out that the result yielded by this model is 76.70% of accurate.

Keywords: Natural Language Processing (NLP), Computational Linguistic, Part of speech (POS), POS tagger, Hidden Markov Model (HMM).

1 Introduction

Natural Language Processing (NLP) deals with the inter-relation and inter-communication between the computer and natural human language by combining the technology of artificial intelligent and computer science. The most important part of any NLP task is the issue of understanding the natural language. Application of NLP helps machines to learn, read and understand the human language, by simulating the human ability of understanding the language and by combining the technology of computational linguistics, computer science and artificial intelligence. The most basic and important starting level of NLP for any language is POS tagging. POS in language processing is the aspect that deals with the identification of grammatical class of each word in a given sentence. POS is used in many fields of NLP such as Semantic Disambiguation, Phrase identification (chunking), Named Entity Recognition, Information Extraction, Parsing, etc. [2, 17]. The task of creating POS Tagger involves many stages such as: building tagsets, creating dictionary, considering the rules of the context and also checking the inflexions, dependent anomalies of the particular language, etc.

Though, substantial work has already been carried out in different fields of NLP for Indian Language, Khasi lacks such study from the NLP perspective. Other Indian languages like Hindi, Bengali, Assamese, Manipuri, Marathi, Tamil, etc. have already been employed in Computational linguistic. In this paper we present POS tagger for Khasi language. Khasi Language is an official language of the state Meghalaya, in North East India [16]. The name 'Khasi' classify both the tribe and as well as the language. Khasi Language is spoken in the Khasi Hills district of the state Meghalaya, India. Also, this language is spoken in the border area of Assam -Meghalaya as well as India-Bangladesh border. Khasi is a part of Mon-Khmer family which is branch of Austro-Asiatic, Southeast of Asia. There have been very limited research works in computational linguistics with regard to Khasi Language.

To perform research work on NLP there is a need of corpus in Khasi Language. Dataset or Corpus is basically a large and extensive collection of texts or words, which are used for analysis of any Natural Language. Corpus is an essential component for any Natural Language Processing research. Therefore, in this paper we describe tagger for Khasi Language based on supervised system HMM trained model.

The paper is organized as follows: Sect. 2 describes related works on POS Tagging; Sect. 3 describes Methodology used in HMM based POS Tagger system; Sect. 4 describes experimental results ; Sect. 5 Conclusions and some future perspectives.

2 The Literature Review

In this section, the existing relevant works in the POS tagging of different languages are presented. There are several existing research works on POS tagging on different languages such as Indian, English, German, Spanish, etc. whereby many researchers have also proposed different methods for POS tagging and show the achieved results.

A Hidden Markov Model (HMM) based Part of Speech (POS) Tagger for Hindi language as discussed in [5].Indian Language (IL) POS tag set have been employed for the system. In the experimental result the HHM POS tagger acquires accuracy of 92%.

A POS tagging for Manipuri language demonstrated 69% of accuracy by using morphology driven POS tagger in [12]. This POS tagger uses 10917 unique words and three dictionaries consisting of prefixes, suffixes and root words and also information with respect to the text content.

Part-of-speech (POS) Tagger for Malayalam Language using supervised learning based on Support Vector Machine (SVM)as discussed in [1]. For the tagger, tag-set consisting 29 tags was developed. The corpus with dataset consisting 180,000 words, the system achieved 94% of accuracy.

Part-of-Speech Tagging for Marathi Language discussed in [8] using Training data of 576 words with tag-set of 9 tags. Tokenization, Morphological analysis and Disambiguation process have been carried out for POS tagging. The author concluded with the attained accuracy of 78.82% by the system.

Kokborok language based on rule based, Conditional Random Field (CRF) and Support Vector Machines (SVM) for Part of Speech (POS) Tagger discussed in [9]. A tagged dataset of 42,537 words with 26 tag were used. The POS taggers methods attains the accuracies of 69% for rule based, 81.67% CRF based and 84.46% for SVM.

In [10]has discussed POS Tagging and Chunking using Conditional Random Fields and Transformation. The POS tagger accuracy achieved for CRF and TBL of about 77.37% for Telugu, 78.66% for Hindi, and 76.08% for Bengali and the chunker performance accuracy of 79.15% for Telugu, 80.97% for Hindi and 82.74% for Bengali respectively.

Part-of-Speech (POS) tagger for Bengali language in [3] has been reported. Tagging based on Hidden Markov Model (HMM) and Maximum Entropy (ME) stochastic taggers has been discussed. Study is conducted to improve the efficiency and performance of the tagger by using a morphological analyzer. An accuracy of 76.8% has been reported. The author reported to achieved good performance with the suffix information and morphological restriction on the grammatical categories for the supervised learning model.

In the paper [4] the POS tagger based on SVM and HMM for Bengali has been proposed. The author show the result as the accuracy of 79.6% for the manually checked corpus consisting 0.128 million words. The result of developed POS taggers were given as the accuracies of 85.56% for HMM, and 91.23% for SVM.

Part of Speech tagging for Assamese was reported in paper [11]. The system was design based on Hidden Markov Model approach (HMM). With tagsets consisting 172 tags and corpus consisting 10000 words which were manually tagged for training the system. The accuracy of 87% was achieved by the HMM POS tagger.

For khasi language the author in [13] had introduce tagsets consisting 61 tags. In the paper [14] the author had also introduce Morphological Analyzer for Khasi language. Using morphotactic rules the author had used dictionary consisting 8000 words. The analyzing system use based word of the word classes with grammatical relationship of subject verb object. For deriving the morphotactic rules the prefixes, infixes and suffixes of the words had been made used.

3 Methodology for HMM POS tagger

In this paper, the POS tagger for Khasi language based on Hidden Markov Model (HMM) as supervised learning has been employed. In the subsection below, we present the discussion about the methods that have been carried out in this work for building the supervised POS Tagger:

3.1 Tag Sets

Tag is the label that is used to describe a grammatical class of information (these are: nouns, verbs, pronouns, prepositions, and so on). For example: NN could represent the Noun class, JJ the Adjective class, PRP the Pronoun class, etc. Each language has a different pattern, frequency, and speaking style. Thus, the grammatical class is also different for different languages. Therefore, for this work we have designed tagset consisting 54 tags for identifying the grammatical class or Part-of-Speech (POS) of Khasi Language [15] as listed in the table below.

Table 1: POS tagset for Khasi Language [15].

No.	Tag	Description	No.	Tag	Description
1	PPN	Proper nouns	32	3PSM	3rd Person singular Masculine gender
2	CLN	Collective nouns	33	3PPG	3rd Person plural common Gender
3	CMN	Common nouns	34	3PSG	3rd Person singular common Gender
4	MTN	Material nouns	35	VPT	Verb, present tense
5	ABN	Abstract nouns	36	VPP	Verb, present progressive participle
6	RFP	Reflexive Pronoun	37	VST	Verb, past tense
7	EM	Emphatic Pronoun	38	VSP	Verb, past perfective participle
8	RLP	Relative Pronouns	39	VFT	Verb, future tense
9	INP	Interrogative Pronouns	40	Mod	Modalities
10	DMP	Demonstrative Pronouns	41	Neg	Negation
11	POP	Possessive Pronoun	42	CLF	Classifier
12	CAV	Causative Verb	43	COC	Coordinating conjunction
13	TRV	Transitive verb	44	SUC	Subordinating conjunction
14	ITV	Intransitive verb	45	CRC	Correlative conjunction
14	DTV	Ditransitive verb	46	CN	Cardinal Number
15	ADJ	Adjective	47	ON	Ordinal number
16	CMA	Comparative Adjective marker	48	QNT	Quantifiers
17	SPA	Superlative Adjective marker	49	CO	Copula
18	AD	Adverb	50	InP	Infinitive Participle
19	ADT	Adverb of Time	51	PaV	Passive Voice
20	ADM	Adverb of Manner	52	COM	Complementizer
22	ADP	Adverb of Place	53	FR	Foreign words
23	ADF	Adverb of frequency	54	SYM	Symbols
24	ADD	Adverb of degree			
25	IN	Preposition			
26	1PSG	1st Person singular common gender			
27	1PPG	1st Person plural common gender			
28	2PG	2ndPerson singular/plural common gender			
29	2PF	2ndPerson singular/plural Feminine gender			
30	2PM	2ndPerson singular/plural Masculine gender			
31	3PSF	3rd Person singular Feminine gender			

For more details regarding the designed tag-sets of Khasi language can be found in [15] respectively.

3.2 Data Sets or Corpus building

In linguistics, the corpus is the large number of data or written texts, which is collected for analyzing in computational linguistic. It is found that no such Khasi corpus is available till date. Therefore, in this work the corpus has been built-in Khasi language based on context, by collecting the written text from online Khasi newspaper. These raw texts are then tag appropriately to each word by using the tagset respectively. Tagging to words at this stage has been done manually. Therefore, we were able to create 7,500 words in our data set.

The corpus has been built painstakingly with care so that it can efficiently handle the problem of Ambiguity. The dataset have also been built under the observation and validation of a linguistic expert from North Eastern Hill University, Department of Linguistic, Shillong, Meghalaya, India. Then, this dataset has been used for training and testing the tagger.

3.3 POS Identification based on HMM for Khasi language

This subsection presents the steps for POS tagger based on HMM that has been carried out in this work. Part-of-speech tagging is a sequence classification problem. The main objective of this supervised learning HMM trained model is to give the most i.e. the maximum probable tag y as outputs for the given word x, as shown in equation (1).

$$f(x) = arg \max_{y \in Y} P(y|x) \tag{1}$$

The following are the elements consist in the HMM POS tagger system and the steps followed for calculating the efficiency of the tagger.
1. A finite set of words, $W = \{w_1, w_2, ..., w_n\}$
2. A finite tags, $T = \{t_1, t_2 ..., t_n\}$
3. n is the number in length.

Therefore, from our objective, to find the optimal tags sequence $\hat{t^n}$ we have equation (3):

$$\hat{t^n} = arg \max_{t^n} P(t^n|w^n) \tag{2}$$

$$\hat{t^n} = arg \max_{t^n} P(t^n)P(w^n|t^n) \tag{3}$$

NOTE: Prior probability -> $P(t^n)$ and Likelihood probability -> $P(w^n|t^n)$.

4.Two properties of assumptions are considered in HMM based POS taggers, shown in equation (4) and (5):

i.The probability of bi–gram assumption.

$$P(t^n) = P(t_1)P(t_2|t_1)P(t_3|t_2)P(t_4|t_3)...P(t_n|t_{(n-1)}) \approx \prod_{i=1}^{n} P(t_i|t_{(i-1})) \tag{4}$$

ii.The assumption of Likelihood probability.

$$P(w^n|t^n) = P(w_1|t_1)P(w_2|t_2)P(w_3|t_3)...P(w_n|t_n) \approx \prod_{i=1}^{n} P(w_i|t_i) \tag{5}$$

5.The POS Tagger use equation (6) for estimating the most probable sequence of tag.

$$(t^n) = arg \max_{t^n} P(t^n|w^n) \approx arg \max_{t^n} \prod_{i=1}^{n} P(t_i|t_(i-1))P(w_i|t_i) \tag{6}$$

6. Tag Transition probabilities $P(t_i|t_{i-1})$ defining the probability of going from tag t_{i-1} to tag t_i , as shown in equation (7).

$$P(t_i|t_{i-1}) = \frac{Count(t_{i-1}, t_i)}{Count(t_{i-1})} \tag{7}$$

7. Word Emission probabilities $P(w_i|t_i)$ defining the probability of emitting word w_i in tag t_i shown in equation (8).

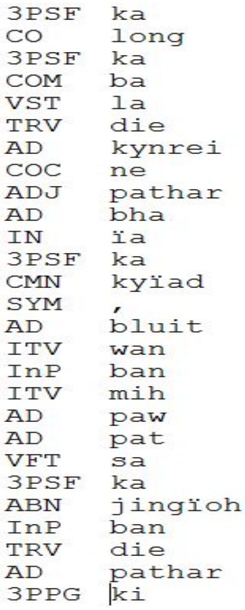
$$P(w_i|t_i) = \frac{Count(t_i, w_i)}{Count(t_i)} \tag{8}$$

For more details on supervised learning system based on HMM POS tagger and its assumption considered can be found in [7] respectively.

4 Experimental Results

4.1 Corpus

As discussed in the methodology, the corpus has been manually designed and check by the linguistic expert. In this work dataset or corpus of around of 7,500 words has been used for training and 312 words for testing the HMM based POS Tagger. The corpus of Khasi language for this research has been collected from the online Khasi newspaper from [6], and the data collected comprise of the political news and article news had been used in the corpus. The table below shows some sample dataset that are manually tagged using the respective tagset.



```
3PSF  ka
CO    long
3PSF  ka
COM   ba
VST   la
TRV   die
AD    kynrei
COC   ne
ADJ   pathar
AD    bha
IN    ïa
3PSF  ka
CMN   kyïad
SYM   ,
AD    bluit
ITV   wan
InP   ban
ITV   mih
AD    paw
AD    pat
VFT   sa
3PSF  ka
ABN   jingïoh
InP   ban
TRV   die
AD    pathar
3PPG  ki
```

Fig. 1: Manually tagged dataset for trainning

4.2 Discussion on some challenges

During the collection of data and creating the dataset, we encountered some challenges; two main challenges in building the dataset for Khasi discussed in this paper are orthography and ambiguity. In orthography, the major problem is the spelling consistency. Spellings in Khasi have not been fully standardized. Different authors spell differently for the same words and that many words that are spelt alike have different meanings and are pronounced differently in different context. Whereas ambiguities are found in categorizing words that are spelt and pronounced alike but differ in categories when they are used in a sentence. Assigning labels or tags to each word of a given sentence is a difficult task, because there are words that represent more than one grammatical part of speech. The challenge of POS tagging is the 'Ambiguity'. Some other structural are also there, but are kept out of this paper as they are not within the scope of the paper.

Keeping in mind these problems, we tried to consider by checking and accounting these problems wisely and built the dataset accordingly. The assumption that is considered for the word categories is based on the root category and prefix information. Therefore accordingly, we have designed the data sets to account these problems. Below are some of the examples cited, based on the challenges which are mentioned above:

For example: Orthography problem -

It is found that the orthography is very complex problem in Khasi language, as in some context words are different compare with the other; like the word *ia* in some context it is written as *ïa*. There are many such words that are spell and written different by different people, some more of those words are like: *ïadei*, *ia dei*, *Khamtam*, *Kham tam*, *eiei*, *ei ei*, *ïatreilang*, *ia trei lang*, *watla*, *wat la*, and so no.

For example: Ambiguity problem -

The table below show some of the ambiguous words of Khasi language:

Table 2: Some of the Khasi ambiguous words.

Khasi word	**Meaning**	**Part-Of-Speech(POS) class**
Kot	book	noun
Kot	reach	verb
Kam	work	noun
Kam	pace	verb
lum	hill	noun
lum	collect	verb
bah	sir	noun
bah	enough	adjective
mar	as soon as	adverb
mar	material	noun
tam	pick	verb
tam	over	adjective
kham	more	adjective
kham	hold	verb

Therefore according to our need of the data we have been considered and solve the problem.

4.3 Result

Using this HMM POS tagger based on the supervised learning method for the Khasi language, with the corpus of 7,500 words the system yield 76.70% of accuracy as a performance. Due to the unavailability of dataset or corpus and we had to make our own dataset, the accuracy of the tagging can be improved further by creating more data in the corpus. The table below show the comparison result with the other language using HMM approach.

Table 3: Comparison with others HMM POS tagger.

Sl. no.	Paper Title	Language Used	Approach	corpus training	Accuracy
1	HMM based pos tagger for Hindi.	Hindi	Hidden Markov Model	-24 tags -3,58,288 words	92%.
2	Development of Marathi Part of Speech Tagger Using Statistical Approach.	Marathi	Unigram, Bigram, Trigram and HMM Methods	-26 lexical tags -1, 95,647 words	Unigram-77.38%, Bigram-90.30%, Trigram-91.46% and HMM- 93.82%
3	Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario Approach.	Bengali	Maximum Entropy (ME) and HMM Methods	-45,000 words	76.8%
4	Web-based Bengali News Corpus for Lexicon Development and POS Tagging.	Bengali	HMM and SVM Methods	- 0.128 million words	SVM - 1.23% HMM- 85.56%
5	Part of Speech Tagger for Assamese Text.	Assamese	HMM Method	- 10000 words	87%
6	Identification of POS Tag for Khasi Language based on Hidden Markov Model POS Tagger (Our proposed method).	Khasi	HMM Method	- 7,500 words	76.70%

4.4 Result Analysis

Apart from the correct tagged words, in the experimental result some errors have also been analyses. It is found that some words are tagged incorrectly with the tag that does not belong to the respective word. As we have tag manually the Khasi words with respect to the content of context, therefore some words are tagged wrongly by the system due to ambiguity problem.

For example the word:
bha are tagged as ADJ or AD.
Namar are tagged as COC or SUC.
Hapdeng are tagged as IN or ADP.

In the result it is found that for the words: *ia* it is tagged 12 times as InP and it is tagged 2 times as IN, *bah* is tagged as CMN 3 times and 1 time as ADJ, *dang* is tagged 1 time as AD and 1 time as VPP. Due to the present of ambiguous words in the annotated corpus it reduces the result accuracy. Therefore, with more annotated data there is high chance that the system will improve the result.

Conclusion and Future Works

As very limited works have been done for Khasi language in NLP till date, and tagging from the semantic and technical problems cited above have not been discussed at all, this work culminated as paper that addressed these issues from a larger perspective of NLP. The problems and issues being raised in this paper does not solve all the problems encounter in the POS tagging of Khasi, therefore, there is a future scope of accounting the other problems in future research. Therefore, we aim to improve the result of the system by introducing more data in the corpus, for both training and testing. We also aim to develop some syntactic rules for Khasi language and to employ them in POS tagger for good results. This will help us to evaluate good performance of the POS tagger of Khasi language.

Acknowledgement

Authors would like to acknowledge the Centre for Natural Language Processing, National Institute of Technology Silchar, 788010, Assam, India for this work.

References

1. Antony, P., Mohan, S.P., Soman, K.: Svm based part of speech tagger for malayalam. In: Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference on. pp. 339–341. IEEE (2010)
2. Chowdhury, G.G.: Natural language processing. Annual review of information science and technology 37(1), 51–89 (2003)
3. Dandapat, S., Sarkar, S., Basu, A.: 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. pp. 221–224. Association for Computational Linguistics (2007)
4. Ekbal, A., Bandyopadhyay, S.: Web-based bengali news corpus for lexicon development and pos tagging. Polibits (37), 21–30 (2008)
5. Joshi, N., Darbari, H., Mathur, I.: Hmm based pos tagger for hindi. In: Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013). pp. 341–349 (2013)
6. Mawphor: Mawphor (2017), `https://www.mawphor.com/index.php/`, [Online; accessed 07-Nov-2017]
7. Pakray, P., Majumder, G., Pathak, A.: An hmm based pos tagger for pos tagging of code-mixed indian social media text. In: Annual Convention of the Computer Society of India. pp. 495–504. Springer (2018)
8. Patil, H., Patil, A., Pawar, B.: Part-of-speech tagger for marathi language using limited training corpora. In: IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT (4). pp. 33–37. Citeseer (2014)
9. Patra, B.G., Debbarma, K., Das, D., Bandyopadhyay, S.: Part of speech (pos) tagger for kokborok. Proceedings of COLING 2012: Posters pp. 923–932 (2012)
10. PVS, A., Karthik, G.: Part-of-speech tagging and chunking using conditional random fields and transformation based learning. Shallow Parsing for South Asian Languages 21, 21–24 (2007)
11. Saharia, N., Das, D., Sharma, U., Kalita, J.: Part of speech tagger for assamese text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 33–36. Association for Computational Linguistics (2009)

12. Singh, T.D., Bandyopadhyay, S.: Morphology driven manipuri pos tagger. In: Proceedings of the IJCNLP-08 Workshop on NLP for less privileged languages. pp. 91–97 (2008)
13. Tham, M.J.: Design considerations for developing a parts-of-speech tagset for khasi. In: Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on. pp. 277–280. IEEE (2012)
14. Tham, M.J.: Preliminary investigation of a morphological analyzer and generator for khasi. In: Emerging Trends and Applications in Computer Science (ICETACS), 2013 1st International Conference on. pp. 256–259. IEEE (2013)
15. Warjri, S., Pakray, P., Lyngdoh, S., Kumar Maji, A.: Khasi language as dominant part-of-speech (pos) ascendant in nlp. International Journal of Computational Intelligence & IoT 1(1), 109–115 (2018)
16. Wikipedia contributors: Khasi language — Wikipedia, the free encyclopedia (2018), `https://en.wikipedia.org/w/index.php?title=Khasi_language&oldid=838847215`, [Online; accessed 02-Feb-2018]
17. Wilks, Y., Stevenson, M.: The grammar of sense: Using 6 part-of-speech tags as a first step in semantic disambiguation. Natural Language Engineering 4(2), 135–143 (1998)