Challenges and Issues in Developing an Annotated Corpus and HMM POS Tagger for Khasi

Medari Janai Tham
Computer Science & Engineering and Information Technology
Assam Don Bosco University
medaritham16@gmail.com

Abstract

An attempt has been made to annotate a Khasi corpus with Part-of-Speech (POS) tags, using the Bureau of Indian Standards (BIS) POS tagset prepared by the POS Tag Standardization Committee of the Department of Information Technology (DIT), New Delhi, India for annotating Indian language corpora. This is the first initiative taken for Khasi- an understudied and under-resourced language, in developing an annotated corpus and POS tagger essential for language technology. This article highlights the challenges and issues that surfaced during annotation, and the decisions that were taken when tagging features characteristic of Khasi that are absent from mainstream Indian languages. A Hidden Markov Model (HMM) POS tagger is then constructed, taking into consideration the information provided by the morphological features of the Khasi language. The results of training and testing the Khasi HMM POS tagger are compared with the results of a Khasi baseline tagger, and a Khasi tagger constructed using Natural Language Toolkit (NLTK).

1 Introduction

Construction of resources is necessary for natural language processing and this article describes the process initiated in the development of an annotated corpus and POS tagger for Khasi, which are basic resources required for natural language applications such as parsing, information retrieval, question and answering, etc.

Standard guidelines in annotating text corpora are essential when an attempt is made to annotate a corpus from scratch as is the case with Khasi. The benefits of annotating the corpus using the prescribed standard such as Bureau of Indian Standards (BIS) (Chaudhary et al., 2010) for Indian languages will facilitate the corpus in interlinguistic analysis and study. The annotated Khasi corpus has been constructed from a collection of Khasi literature of prose and fiction genre and it comprises of 3,984 sentences which include 86,087 tokens out of which 75,736 are tokens excluding punctuation and 5,313 word types. The applied BIS tagset for Khasi is given in Table 1 and Table 2. The questions and issues that emerged when annotating the corpus and their proposed suggestions are discussed in section 6. The construction and analysis of the Khasi HMM POS tagger and the comparison of its results with the Khasi baseline tagger and the Khasi NLTK tagger are given in section 7.

2 Related Work

POS tagging is the process of automatically assigning a part of speech to each word present in a corpus. These part of speech tags are assigned from a specific tagset applicable to the language. Current POS tagging accuracy is about 96%-97% for languages such as English, French, etc. (Güngör, 2010). Approaches to tagging algorithms are either rule-based taggers or stochastic taggers. The most widely used tagger in rule-based tagging is the Transformation Based Learning (Brill, 1995) often called the "Brill tagger". This approach also uses machine learning to learn the rules form the data and achieved 96.6% accuracy when trained and tested on the WSJ corpus. On the other hand, stochastic taggers utilize the availability of lexicons and corpora and one such learning approach is the Hidden Markov Model (HMM) which has obtained high accuracies in POS tagging. For example, the most available tagger and highly accurate is the TnT tagger (Brants, 2000). Its influence comes from its sensitive dealing with unknown words and smoothing. Another HMM tagger is the HunPos trigram tagger (Halácsy et al., 2007) which unlike TnT, provided mechanisms where a language morphological features can be tweaked into the

Proc. of ICON-2018, Patiala, India. December 2018 ©2018 NLPAI, pages 10–19

tagger and achieved 98.24% accuracy for Hungarian when compared to TnT's 97.42% on the same corpus. However, the TnT tagger remains one of best performing taggers across different languages (Plank et al., 2016).

According to the 2001 Indian census, the language families present in India are Indo-Aryan, Dravidian, Austro-Asiatic, Tibeto-Burmese and Semito–Hamitic. Among these language families, Indo-Aryan and Dravidian are the two major family groups of India comprising approximately 97% of India's population. A recurring pattern with stochastic POS taggers developed for Indian languages, is that they have to content with small size training data and language specific tagsets. Reported tagging accuracies for Indian languages range from 69%-96% and some of the POS taggers developed for Indo-Aryan (Hindi), Tibeto Burman (Manipuri and Kokborok) and Dravidian (Tamil) families are as follows.

Apart from English, Hindi is the official language of India. It is a morphologically rich language, and one POS tagger (Singh et al., 2006) developed for Hindi has taken advantage of this feature to compensate the lack of annotated corpora by utilizing extensive morphological analysis along with a high-coverage lexicon and decision tree based learning algorithm where the size of the corpus used is 15,562 words, and achieved POS tagging accuracy of 93.45%. Another POS tagger (Shrivastava and Bhattacharyya, 2008) for Hindi that does away with the need of a morphological analyzer and structured lexicon, uses the HMM approach where a list of all possible suffixes in Hindi is employed to perform stemming on a corpus of 66,900 words and achieved 93.12% accuracy. On the other hand a rule based Hindi POS tagger (Garg et al., 2012) reported an accuracy of 87.55%.

In the absence of tagged corpora, a morphologically driven POS Tagger for Manipuri (Singh and Bandyopadhyay, 2008) achieved 69% accuracy tested on 3,784 sentences. A Manipuri POS tagger (Singh et al., 2008) using condition random field and support vector machine trained on 39,449 tokens and tested on 8,672 tokens reported 72.04% and 74.38% accuracies respectively. Another condition random field Manipuri POS tagger (Nongmeikapam and Bandyopadhyay, 2012) used for transliterating from Bengali script to the Meitei Mayek script achieved a precision of 74.31%, a recall of 80.20% and an F-measure of 77.14%. POS taggers developed for Kokborok, another resource constrained language (Patra et al., 2012), include rule based tagger with 69% accuracy and stochastic taggers using condition random field and support vector machine with 81.67% and 84.46% accuracies respectively.

POS taggers for Tamil include rule based (Selvam and Natarajan, 2009) with 85.56% accuracy, and a morpheme based language model (Pandian and Geetha, 2008) involving 35 tags and a test set of 43,678 words with 95.92% accuracy.

3 Concise Overview of BIS

The BIS standard has been prepared to work for languages even beyond Indo-Aryan and Dravidian families and the guidelines have been formulated taking into account existing tagsets designed under various projects such as the Indian Language Machine Translation (ILMT) POS tagset (Bharati et al., 2006), Microsoft Research India Indian Language POS (Baskaran et al., 2008) and others. Taking into consideration the existence of various language families in India, the tagset has been designed to be all accommodating. The annotation follows a layered approach where the linguistics features can be incorporated in layers such as morphology in one layer, part of speech in another layer, syntactic analysis in another layer and the others in different layers, and within each layer there is a hierarchy of categories. Extensibility is a key feature of the tagset where a top category or a sub category can be added to the existing hierarchy if the language under question requires one. On the other hand, a tag may not be utilized if it is not required even if it exist in the BIS tagset. The POS tagging has to be carried out on text that have been pre-processed, where each token in the corpus is a single lexical item and any morphological analysis required should have been processed by a morphological analyzer. In total, the tagset has 11 top level categories with very few categories having two levels of subtypes, reflecting the coarse nature of the tagset.

4 Brief Introduction to Khasi Language

Khasi is classified under the Mon-Khmer branch of the Austro-Asiatic language family (Diffloth, 2018). It is the associate official language of the state of Meghalaya, India and according to the

2011 Indian census there are approximately 1.4 million speakers in Meghalaya and Assam, placing it less than 1 percent of India's population. Khasi is an analytic and non-inflectional language exhibiting derivational morphology which contributes to the partial agglutinative behavior of the language (Nagaraja, 2000).

Khasi is written in the Latin script comprising of 23 letters where the letters *c*, *f*, *q*, *v*, *x*, *z* have been removed with the addition of the diacritic letters *ï* and *ñ* and the diagraph *ng* which is adopted as a single letter.[1]

5 Khasi Corpus Construction

A corpus is designed to represent a particular natural language or language variety by virtue of the range of text included and the sampling from each text used in collecting the data contained in the corpus (Xiao, 2010). Due to the unavailability of any corpus in Khasi, the required corpus has to be built from scratch which consumes time and effort. The data collected for the current corpus are samples from the prose and fiction genre of Khasi literature that are prescribed for studies in higher secondary, graduation, and post-graduation. The selection of Khasi literature is compelled by the fact that though newspapers are easily available online, they are not accepted by language experts as a representation of the language because of the lack of consensus on how the language should be written in terms of its grammar and orthography. On the other hand, it is also observed that in most instances, the written literature does not conform to any single standard when it comes to orthography even within text written by the same author. To cite a few examples the preposition *ïa* (to) is written as *ïa* or *ia*, where the word is written with the letter *i* with diaeresis or without it. Other examples are the words *duai* (pray) where it is also written as *duwai*, *mynmied* (night) which it is also written as *mynmiet*, etc. Another category of a nominal that do not follow a uniform orthography are doublets. These are two nouns that occur together having the same semantics and are often used more for their stylistic value. A few examples are *ki-mrad ki-mreng* (animals), *ki khun-ki kti* (children) , *u-kñi-u-kpa* (ancestor) where the hyphen (-) is used according to the author's style.

In analyzing natural language in digital format it is necessary that the characters, words and sentences are clearly identified before any natural language processing task can be carried out. This process of dividing a text document into words and sentences is called text segmentation. Khasi utilizes the Latin script for writing and like English the whitespace is used to marked word boundaries. The data for analysis is pre-processed manually where each word is separated by a space and each sentence is marked with an end of sentence marker such as a period (.), a question mark (?) or an exclamation mark (!). Thus the words identified are also called tokens and these tokens include punctuations. This implies that the punctuations are not attached with a word but are delimited with a whitespace. The only exception is the use of apostrophes (') and the hyphens (-). The apostrophe is used to marked contractions such as *bar'bor* (everytime), and the hyphen to form compound words such as *Khasi-Khara* (the Khasis); the reduplicated forms often used with adverbs such as *khah-khah* (regularly) where these punctuations are also part of the tokens. The corpus is then manually tagged using the BIS tagset shown in Table 1 and Table 2.

6 Annotating Khasi using BIS tagset

This section discusses the challenges and the issues faced when tagging Khasi and the decisions that were taken on encountering features prevalent in the language. The grammatical characteristics of the language taken into consideration are with references to the works of various contributors on the Khasi language (Rabel, 1961; Bars, 1973; Henderson, 1976; Nagaraja, 1985; Jyrwa, 1989; Roberts, 2005; War, 2011)

6.1 Personal Pronouns

The structure of personal pronouns in Khasi is simple except in the case of third person singular and plural forms. Apart from their basic functionality, third person singular and plural personal pronouns such as: /i/ 'singular, neutral', /u/ 'singular, masculine' /ka/ 'singular, feminine' and /ki/ 'common, plural' also function as number and/or gender markers. The personal pronoun *i* when used, indicates reverence or refers to diminutive objects. They are also described as articles, determiners, gender indicators and pronominal markers. It is mandatory that every noun in Khasi is preceded by pronominal markers

[1] http://www.ciil-lisindia.net/Khasi/Khasi.html

		Categories				
Sl. No	Top Level	Subtype (Level 1)	Subtype (Level 2)	Label	Annotation Convention	Example(s)
1	Noun			N	N	
1.1		Common		NN	N_NN	jingsuk ‘peace’ ksew ‘dog’
1.2		Proper		NNP	N_NNP	Melam, Shillong
1.3		Nloc		NST	N_NST	sha-lor LOC-top ‘on top’
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR_PRP	nga 1S ‘I’
2.1.1			Pronominal	PRP_M	PR_PRP_M	ka kot PM book ‘a/the book’
2.1.2			Auxiliary	AUX	PR_PRP_AUX	nga-n 1S-FUT ‘I will’
2.2		Reflexive		PRF	PR_PRF	lade ‘self’
2.3		Relative		PRL	PR_PRL	u-ba 3SM-that ‘he that’
2.4		Wh-word		PRQ	PR_PRQ	u-ei 3SM-who ‘who’
2.5		Indefinite		PRI	PR_PRI	ka-no ka-no 3SF-whoever 3SF-whoever ‘whoever’
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM_DMD	ka-ta 3SF-out of sight ‘that’
4	Verb			V	V	
4.1		Main		VM	V_VM	bam ‘eat’
4.2		Auxiliary		VAUX	V_VAUX	lah ‘can’
4.2.1			Infinitive	VINF	V_VAUX_VINF	ban ‘to’
5	Adjective			JJ	JJ	bakhraw ‘great’

Table 1: Khasi BIS Tagset

(PM) which are third person personal pronouns. Exceptions where the pronominal marker is dropped are in vocative sentences, optionally in locative phrases where inanimate nouns are used,

	Categories					
Sl. No	Top Level	Subtype (Level 1)	Subtype (Level 2)	Label	Annotation Convention	Example(s)
6	Adverb			RB	RB	suki-suki ‘slowly-slowly’
7	Conjunction			CC	CC	
7.1		Coordinating		CCD	CC_CCD	bad ‘and’
7.2		Subordinating		CCS	CC_CCS	namar ‘because’
8	Particles			RP	RP	
8.1		Default		RPD	RP_RPD	noh PRT
8.2		Classifier		CL	RP_CL	tylli
8.3		Interjection		INJ	RP_INJ	wa, ada
8.4		Intensifier		INTF	RP_INTF	shuh, eh
8.5		Negation		NEG	RP_NEG	ki-m 3PL-will not ‘they will not’
8.6		Possessive		POS	RP_POS	la POS
9	Quantifiers			QT	QT	
9.1		General		QTF	QT_QTF	shi ‘one’
9.2		Cardinals		QTC	QT_QTC	wei ‘one’
9.3		Ordinals		QTO	QT_QTO	banyngkong ‘first’
10	Residuals			RD	RD	
10.1		Foreign		RDF	RD_RDF	a word not written in Khasi
10.2		Symbols		SYM	RD_SYM	#, $
10.3		Punctuation		PUNC	RD_PUNC	; ,
10.4		Unknown		UNK	RD_UNK	
10.5		Echowords		ECH	RD_ECH	lyngaiň
11		Preposition		IN	IN	na ‘from’

Table 2: Khasi BIS Tagset cont...

and when nouns immediately follow a verb- they blend with the verb and cease to be nouns (Jyrwa, 1989). Another functionality of these pronominal markers is their occurrence before a verb (also called subject enclitic (Jyrwa, 1989)) indicating subject verb agreement and highlighted in bold in the example below.

ka	Iba	**ka**	ai	ka	kot
PM	Iba	PM	give	PM	book

‘Iba gave the book’

Ideally, tagging them as pronominal markers will be appropriate in highlighting the fact that they stand in agreement with the head noun, but there are instances where their occurrences can also be quite far from the head noun which is not feasible for machine learning purposes in disambiguating them from personal pronouns. Therefore, the new tag PR_PRP_M representing a pronominal marker is applied only to pronouns occurring before a noun only and not for subject enclitic. The existing personal pronoun tag PR_PRP has been maintained for subject enclitic.

Another problem is the personal pronouns attached with the suffix -n or -m such as *ngan* (I will), *ngam* (I will not), etc. For instance *ngan* indicates tense equivalent to English (will) and (shall) and *ngam* indicates tense and negation (will not) and (shall not). As per BIS guidelines, any morphological analysis required must have been carried out before tagging, such that each token is a lexical item and requires no further processing. If morphological analysis is applied to these pronouns, we now have *ngan* (I will) mapping to *nga* 'first person, singular' and *yn* (will) an auxiliary verb. *Ngam* on the other hand will have two mappings-- a) (I will not) mapping to 'first person, singular' and *ym* (will not) an auxiliary verb and b) *ngam* (drown) which is a verb. Since this analysis is applicable to a finite number of words, a morphological analyzer is not employed, and specifically when these words function as pronouns in the corpus, they are given a newly created tag PR_PRP_AUX which is a sub-type of the personal pronoun category. It may be mentioned that these words do not function as pronominal markers.

6.2 Multi-functionality of la

The word *la* in Khasi can function as a past tense marker or an auxiliary verb or particle or a possessive particle or a subordinating conjunction. When *la* functions as an auxiliary verb or a past tense marker, it has been tagged as an auxiliary verb V_VAUX because their occurrences in sentences are syntactically similar. The BIS tagset has provisions for subordinating conjunction and particle but not for possessive marker. While tagging, the tags applied for subordinating conjunction and particle are CC_CCS and RP_RPD respectively. Again, keeping in mind BIS extensibility feature, a new sub-type of the particle category RP_POS is created to accommodate *la* functioning as a possessive particle.

6.3 Tagging of Adverbs

The BIS tagset specifies that only manner adverbs should be tagged as RB such as *ïaid suki suki* (walk slowly). It appears that no BIS tag is appropriate for adverbs such as *ruh* (also), *ju* (in the habit of, used to), etc. In order to maintain minimalism of new tags in the BIS tagset, in the present corpus any occurrences of such words are still tagged as RB.

6.4 Absence of Prepositions in BIS Tagset

BIS has incorporated postpositions with the tag PSP which is a prevalent feature in the Indo-Aryan and Dravidian families but absent in Khasi. Khasi utilizes prepositions and in order to accommodate them, a new top level category is constructed with tag IN.

6.5 Nouns of Location Space and Time (Nloc)

The BIS tagset clearly states that only a finite number of nouns of location, space and time that can also function as postpositions are tagged as N_NST. The question was, whether this category is also applicable to Khasi or not. From the literature and the data in the corpus, it came to attention that a certain group of words in Khasi can function as a noun or a preposition or as an adverb depicting the behavior mentioned in the BIS specification. These are compound words comprising of a preposition (ha/na/sha) and a bound or a free element. These words are also referred as prepositional adverbials such as *halor* (on top), *sharum* (downwards/south), etc. The conclusion that was brought forward in the BIS tagset specification is to facilitate machine learning and simultaneously avoid confusion in annotation- therefore in the present corpus they have been uniformly tagged as N_NST irrespective of their function.

6.6 Tagging Compound Words and Imitative

Compounds in Khasi are primarily formed when a space or a hyphen separates the elements of the compound word, or they are collocated. For example, *khia thew* (graceful), *bai-sngi* (wage) and *metbneng* (planets). The compound word that is written as a single word or where the elements of the compound are separated by hyphens is tagged by taking its grammatical function in the sentence.

7 Applying the Hidden Markov Model for POS tagging

7.1 POS Tagging

During POS tagging, each word in the corpus is automatically tagged with its part of speech. Therefore, given an input string of words and a tagset the output of a POS tagger should be the best possible tag for each word. For example, using the BIS tags for Khasi from Table 1 and Table 2, a sentence in Khasi is tagged as follows.

Tiap\RB tang\RB shu\RB poi\V_VM ha\IN bri\N_NN ,\RD_PUNC u\PR_PRP_M slap\N_NN u\PR_PRP sdang\V_VM hap\V_VM .\RD_PUNC

'Immediately when he reached the field the rain started falling'

7.2 Hidden Markov Model Approach

Given a tagset, in this instance the BIS tagset in Table 1 and Table 2, and a sentence of n words W= $w_1,w_2,...w_n$, the POS tagger has to find the sequence T= t_1, $t_2...t_n$, where T is a set of tags from the tagset that satisfies the following equation.

$$argmax_T \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1} \dots t_{i-k}) \qquad (1)$$

In other words the best possible tag sequence is a sequence that maximize the lexical P(W|T) and transition P(T) probabilities. Since the tags are hidden and only the words are observed we have a hidden Markov model where states represent the tags and the outputs are the observed words. In the lines of Brants (2000) TnT tagger, a second order Markov model is used where k=2 in equation 1 and adding tags t_{-1}, t_0, and t_{n+1} for beginning of sentence and end of sentence markers. Equation 1 is now calculated as follows.

$$argmax_T (\prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1},t_{i-2}))P(t_{n+1}|t_n) \qquad (2)$$

Using an annotated corpus, the probabilities in equation 2 are estimated using the maximum likelihood estimation.

$$P(w_i|t_i) = \frac{f(w_i,t_i)}{f(t_i,)} \qquad (3)$$

$$P(t_i|t_{i-2},t_{i-1}) = \frac{f(t_{i-2},t_{i-1},t_i)}{f(t_{i-2},t_{i-1})} \qquad (4)$$

where f(w,t) is the number of occurrences of words w with tag t and $f(t_1,t_2,...t_m)$ is the number of occurrences of the tag sequence $t_1,t_2,...t_m$.

We can compute equation 2 for each possible tag sequence of length n and then take the sequence with the highest probability. However the complexity of this algorithm is exponential to the number of words. An efficient algorithm operating in linear time is the Viterbi (Rabiner, 1989) algorithm which is used here to determine the optimal sub paths rather than keeping track of all paths during execution. The trigram tagger given in equation 2 has one problem and that is data sparsity. Any trigram instance in the test set may not have occurred in the training set implying that equation 4 will give zero probability and in turn give rise to zero probability tag sequences. Considering **N** as the total number of tokens in the training corpus, from equation 4 the maximum likelihood estimation can be calculated as follows

$$Trigram\ \hat{P}(t_i|t_{i-2},t_{i-1}) = \frac{f(t_{i-2},t_{i-1},t_i)}{f(t_{i-2},t_{i-1})} \qquad (5)$$

$$Bigram\ \hat{P}(t_i|t_{i-1}) = \frac{f(t_{i-1},t_i)}{f(,t_{i-1})} \qquad (6)$$

$$Unigram\ \hat{P}(t_i) = \frac{f(t_i)}{N} \qquad (7)$$

As suggested in Jurafsky and Martin (2009), linear interpolation can be used and we now estimate the probability as

$$P(t_i|t_{i-2},t_{i-1}) = \lambda_3\hat{P}(t_i|t_{i-2},t_{i-1}) + \lambda_2\hat{P}(t_i|t_{i-1}) + \lambda_1\hat{P}(t_i) \qquad (8)$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$

In order to approximate the value of λ Brants (2000) version of deleted interpolation is used for setting the λ's.

7.3 Using Morphology in Handling Unknown Words

As mentioned in section 4, Khasi exhibits derivational morphology in the form of agglutination by adding affixes to word base to derive other words. These affixes can be easily separated from the root and the focus here are on the prefixes attached to Khasi nouns and verbs. Khasi words reveal that words with prefixes such as jing-, nong- and maw- always map to common nouns (N_NN). Words with prefixes such as pyn- and ïa- excluding the preposition ïa, always map to verbs (V_VM). It may be noted that *pynban* (cause to press) which is a verb can also function as an adverb (nonetheless).

In the training and test data, the words having prefixes jing- are mapped to pseudo-word _JING_, nong- to pseudo-word _NONG_, maw- to pseudo-word _MAW_, pyn- are mapped to pseudo-word _PYN_ and ïa- excluding preposition ïa, are mapped to pseudo-word _IA_.

This mapping is carried out for data in the training set and in the test set, to estimate the probabilities of unknown words having these prefixes.

In order to handle unknown words not having the above mentioned prefixes, low frequency words in the training data are mapped to pseudo-word _UNK_. Similarly, words in the test set that were unseen in the training data are also mapped to pseudo-word _UNK_. Since the corpus size is relatively small, it is observed that words occurring only once in the training set account to 49.1% of the training data. Therefore low frequency is taken to be less than or equal to a selected value γ and in this tagger γ=1.

After the mappings are done, the HMM parameters are evaluated as mentioned earlier where the pseudo-words _JING_, _PYN_, _NONG_, _IA_ and _UNK_ are treated like regular words. This mapping is carried out to ensure that the probability of P(w_i|t_i) is never zero.

7.4 Testing and Evaluation

The corpus has been divided into training set and test set. The training set consists of 3,984 sentences comprising of 86,087 tokens and 5,313 word types. The test set consists of 402 sentences which include 8,565 tokens and 1,110 word types. The test set is a sample from a book not included in the training set.

The data has been tested using a baseline tagger, an NLTK tagger, and the HMM POS Tagger and the results are shown in Table 3. As proposed by Jurafsky and Martin (2009), the baseline tagger tags the words in the test data with their most frequent tag obtained from the training data.

NLTK (Bird et al., 2009) also provides taggers such as the trigram tagger, bigram tagger, default tagger and regular expression tagger. Taking into account the morphological features of Khasi mentioned in section 7.3, an NLTK tagger for Khasi was constructed where an NLTK trigram tagger backs off to a bigram tagger, the bigram tagger backs off to a unigram tagger and the unigram tagger backs off to a Khasi regular expression tagger. The Khasi regular expression tagger tags words with prefixes jing-, nong-, and maw- as common nouns (N_NN), words with prefixes pyn- and ïa- as verbs (V_VM) and defaults to the most common tag which is the common noun (N_NN). Words having frequency less than or equal to 1 in the training data and unseen words in the test data are also mapped to the pseudo-word _UNK_ to handle unknown words. However, the words having the above mentioned prefixes are not mapped to _UNK_ since the tagger eventually backs off to the Khasi regular expression tagger. Additionally, Table 3 also highlights results of the NLTK bigram tagger which backs off to a unigram tagger and an NLTK trigram tagger which backs off to a bigram tagger.

	Accuracy
Baseline Tagger	86.76%
NLTK Bigram Tagger	88.23%
NLTK Trigram tagger	88.64%
NLTK Tagger	89.7%
HMM POS Tagger	95.68%

Table 3: Results

	RB	**V_VM**	**N_NN**	**PR_PRP**	**PR_PRP_M**
N_NN	6.2	3.8			
V_VM	3.2		4.9		
N_NNP			17.6		
PR_PRP					3.8
PR_PRP_M				2.7	

Table 4: Confusion Matrix

7.5 Some Common Tagging Errors

The confusion matrix in Table 4 highlights in percentage some of the common tagging errors present in the tagger. The most common and difficult to disambiguate is when proper nouns are tagged as common nouns, and when nouns follow verbs- the tagger tags them as adverbs. Another case when verbs are tagged as nouns and vice versa are often the case of pronouns tagged as pronominal markers and vice-versa as mentioned in section 6.1.

8 Conclusion

Developing language technology tools for an under-resourced language such as Khasi has been challenging and simultaneously exhilarating to discover the nitty-gritty of the language in the way studies such as this one exposes. The performance of the HMM tagger conditioned with the features intrinsic in the language has shown that it also provides good performance as reported in the literature relating to HMM POS taggers. This work, being a new initiative, annotating the corpus and developing the tagger, is limited by available resources; however, increasing the size of the annotated corpus for further analysis will be a good step forward.

Acknowledgement

This work was sponsored as a minor project by the University Grants Commission (UGC), India Ref.No. F.5-123/2014-15/MRP/NERO/1860.

References

E. Bars. 1973. *Khasi English Dictionary*. Shillong, Meghalaya: Don Bosco.

Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharyya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L, and KVS Subbarao. 2008. A Common Parts-of-Speech Tagset Framework for Indian Languages. *In Proceedings of LREC 2008,* (pp. 1331-1337). Marrakech, Morocco: European Language Resources Association. Retrieved from http://www.aclweb.org/anthology/I08-7013

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2006. *AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages*. Hyderabad: Language Technologies Research Center, IIIT Hyderabad. Retrieved from https://researchweb.iiit.ac.in/~rashid.ahmedpg08/ilmtdocs/chunk-pos-ann-guidelines-15-Dec-06.pdf

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. CA: O'Reilly Media Inc.

Thorstens Brants. 2000. TnT-A statistical part of speech tagger. *In Proceedings of the Sixth Conference on Applied Natural Language Processing* (pp. 224-231). Seattle, Washington. doi:10.3115/974147.974178

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing. A case study in part-of-speech-tagging. *Computational Linguistics*. *21(*4), 543-565. Retrieved from http://www.aclweb.org/anthology/J95-4004

Narayan Chaudhary, Pinkey Nainwani, Ritesh Kumar, and Esha Banerjee. 2010. ILCI Parts of Speech (PoS) Annotation Guidelines. Unpublished Manuscript, Indian Languages Corpora Initiative, Jawaharlal Nehru University, New Delhi, India.

Gérard Diffloth. 2018. Austro-Asiatic languages. In *Encylopaedia Britannica*. Retrieved from https://www.britannica.com/topic/Austroasiatic-languages.

Navneet Garg, Vishal Goyal, and Suman Preet. 2012. Rule based Hindi part of speech tagger. *In Proceedings of COLING*. (pp 163-174). Mumbai. Retrieved from http://www.aclweb.org/anthology/C12-3021

Tunga Güngör. Part-of-Speech Tagging. In Nitin Indurkhya & Fred J. Damerau (Eds). *Handbook of Natural Language Processing* (2nd ed., pp. 205—235). NY: Chapman & Hall/CRC, CRC Press.

Péter Halácsy, András Kornai and Csaba Oravecz. 2007. HunPos – An open source trigram tagger. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 209-212). Retrieved from https://www.researchgate.net/publication/228524009_HunPos_an_open_source_trigram_tagger

Eugénie J. A. Henderson. 1976. Vestiges of morphology in modern standard Khasi. *Oceanic Linguistics Special Publications*. *13,* 477-522. Retrieved from https://www.jstor.org/stable/20019169

Daniel Jurafsky, and James H. Martin. 2009. Part-of-Speech Tagging. *Speech and Language Processing*. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (2nd ed, pp. 125-174). Noida: Pearson India Education

Mumtaz Bory Jyrwa. 1989. *A Descriptive study of the Noun Phrase in Khasi* (Doctoral dissertation). Retrieved from http://shodhganga.inflibnet.ac.in/handle/10603/61398

K. S. Nagaraja. 1985. *Khasi a Descriptive Analysis* Pune, India. Deccan College Post-Graduate & Research Institute.

K. S. Nagaraja. 2000. Word Formation in Khasi. *Bulletin of the Deccan College Research Institute*, 60/61, 387-417. Retrieved from http://www.jstor.org/stable/42936628

Kishorjit Nongmeikapam, and Sivaji Bandyopadhyay. 2012. A Transliteration of CRF Based Manipuri POS Tagging. *In Proceedings of 2nd International Conference on Communication, Computing & Security (ICCCS)*. *6,*(pp. 582-589). Elsevier. doi: org/10.1016/j.protcy.2012.10.070

S. Laksmana Pandian, and T.V. Geetha. 2008. Morpheme based Language Model for Part of

Speech tagging. *POLIBITS*. *38,*19-25. Retrieved from http://www.scielo.org.mx/pdf/poli/n38/n38a3.pdf

Braja Gopal Patra, Khumbar Debbarma, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Part of Speech (POS) Tagger for Kokborok. *In Proceedings of COLING 2012. Posters.* (pp. 923-932). Mumbai, India. Retrieved from http://www.aclweb.org/anthology/C12-2090.

Barbara Plank, Anders Søgaard, A. and Yoav Goldberg, Y. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 412-418). Berlin, Germany. Retrieved from http://www.aclweb.org/anthology/P16-2067

Lili Rabel. 1961. *Khasi a language of Assam*. Baton Rouge: Louisiana State University Press.

Lawrence R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *In Proceedings of the IEEE*. *77*(2), 257-285. doi: 10.1109/5.18626

H. Roberts. 2005. *A Grammar of the Khasi Language.* New Delhi, India: Mittal Publications (Original work published 1891)

M. Selvam, and A.M. Natarajan. 2009. Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques. *International Journal of Computers*. *4*(3), 357-367. Retrieved from https://pdfs.semanticscholar.org/ffb7/494c35b766d0cac1a87298ea4f9bf00f5ad2.pdf

Manish Shrivastava, and Pushpak Bhattacharyya. 2008. Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. *International Conference on NLP (ICON 08)*. Pune, India. Retrieved from https://www.cse.iitb.ac.in/~pb/papers/icon08-hindi-pos-tagger.pdf

Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. 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*. (pp. 779-786). Sydney, Australia. Retrieved from https://www.cse.iitb.ac.in/~pb/papers/ACL-2006-Hindi-POS-Tagging.pdf

Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2008. Morphology Driven Manipuri POS Tagger. *In Proceedings of IJCNLP*. (pp. 91-97). Hyderabad, India. Retrieved from http://www.aclweb.org/anthology/I08-3015

Thoudam Doren Singh, Asif Ekbal, and Sivaji Bandyopadhyay. 2008. Manipuri POS tagging using CRF and SVM: A language independent approach. *In Proceedings of 6th ICON* (pp. 240-245). Pune, India. Retrieved from http://www.academia.edu/1150196/Manipuri_POS_Tagging_using_CRF_and_SVM_A_Language_Independent_Approach

Badaplin War. 2011. *Ki Sawa bad Ki Dur Jong Ktien Khasi* (2nd ed). Shillong. Meghalaya: Ri-Ia-dor.

Richard Xiao. 2010. Corpus Creation. In Nitin Indurkhya & Fred J. Damerau (Eds). *Handbook of Natural Language Processing* (2nd ed., pp. 147—165). NY: Chapman & Hall/CRC, CRC Press.