A Hybrid POS Tagger for Khasi, an Under Resourced Language

Medari Janai Tham
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India

***Abstract*—Khasi is an Austro-Asiatic language spoken mainly in the state of Meghalaya, India, and can be considered as an under resourced and under studied language from the natural language processing perspective. Part-of-speech (POS) tagging is one of the major initial requirements in any natural language processing tasks where part of speech is assigned automatically to each word in a sentence. Therefore, it is only natural to initiate the development of a POS tagger for Khasi and this paper presents the construction of a Hybrid POS tagger for Khasi. The tagger is developed to address the tagging errors of a Khasi Hidden Markov Model (HMM) POS tagger by integrating conditional random fields (CRF). This integration incorporates language features which are otherwise not feasible in an HMM POS tagger. The results of the Hybrid Khasi tagger have shown significant improvement in the tagger's accuracy as well as substantially reducing most of the tagging confusion of the HMM POS tagger.**

***Keywords*—Khasi corpus; BIS tagset; Khasi POS tagger; Conditional Random Fields (CRF); Hidden Markov Model (HMM)**

I. Introduction

Part-of-speech (POS) tagging is the process of automatically assigning a part of speech to each word present in a sentence. It differs from a morphological analyzer, which gives a detailed analysis of a word such as root word, multiple parts of speech (if any), etc., by assigning a part of speech to the word depending on its context. These parts of speech are assigned from a specific list of POS tags called a tagset, applicable to the language at hand. The annotated corpus and tagset utilized in this work are described in [1]. The tagset was formulated according to the Bureau of Indian Standards (BIS) guidelines and referred to as the Khasi BIS tagset. In this paper, the Hidden Markov Model (HMM) approach has been incorporated in the development of a Khasi POS tagger and a ten-fold cross-validation has been carried out to rigorously test the performance of the tagger. To address the tagging errors of the Khasi HMM tagger, conditional random fields (CRF) have been integrated. The CRF approach has shown its capability in resolving issues in various natural language processing tasks [2], [3], [4], and integrating CRF allows the inclusion of features in a sentence such as capitalization, prefixes which are prevalent in Khasi, part of speech tag of the previous word, and context words. This leads to the development of a Hybrid POS tagger for Khasi with improved performance and the details of the tasks undertaken are given in the sections below. The background work is discussed in Section II, which briefly introduces Khasi and POS tagging approaches. Section III contains a description of the resources utilized in this work, while the construction of the Khasi HMM POS tagger is described in Section IV. The integration of CRF in developing a Hybrid Khasi POS tagger is presented in Section V, and finally the conclusion of the paper is given in Section VI.

II. Background Work

A. A Brief Overview of the Khasi Language

Khasi belongs to the Austro-Asiatic language family and is categorized under the Mon-Khmer branch. It is a language spoken by the Khasi tribe who mainly inhabits the state of Meghalaya in India. As per the 2011 census of government of India, there are about 1.4 million speakers of the language in the state. Khasi is an analytic and isolating language, devoid of inflection, but typical of its Mon-Khmer features, it demonstrates simple derivational morphology contributing to the partial agglutination present in the language [5], [6]. Derivational morphology occurs when affixes attached themselves to a word base and they are easily distinguished from any given word. Another Mon-Khmer characteristic is that the word order is subject verb object (SVO). Khasi is written in the Latin script comprising of 23 letters with the exclusion of the letters `c`, `f`, `q`, `v`, `x`, `z` and the inclusion of the diacritic letters `ï` and `ñ`, and the diagraph `ng` [1].

B. Part-of-Speech Tagging Approaches

India's rich language diversity can be understood by the presence of five language families namely Indo-Aryan, Dravidian, Austro-Asiatic, Tibeto-Burmese, and Semito–Hamitic. The reported accuracies for POS taggers for Hindi, a morphologically rich language and one of India's official languages, are 87.55% on a rule-based tagger [7], 93.45% accuracy using a small-sized training corpus of 15,562 words aided with an extensive morphological analyzer and a massive lexicon [8], and 93.12% using HMM on corpus size of 66,900 words [9]. A trend observed across POS taggers for Indian languages is that stochastic taggers have to deal with the availability of only small-sized training data. In the Khasi language scenario, two HMM POS taggers have been reported for Khasi but trained and tested on two independent data sets and tagsets. The first HMM POS tagger trained on a dataset of 86,087 tokens using the Khasi BIS tagset of 33 tags provided an accuracy of 95.68% [1]. The second HMM tagger was trained on 7,500 words with a custom-made tagset of 54 tags reporting an accuracy of 76.7% [10]. However, both taggers reported accuracy performing only in a single run on their respective test data.

III. Resources utilized – Khasi Corpus and Khasi BIS Tagset

The present available Khasi corpus comprises of Khasi literature containing 4,386 sentences and 94,651 tokens [1]. Excluding the punctuations, there are 83,312 tokens and 5,465 word types. Text segmentation has been performed on the corpus to visibly identify characters, words, and sentences. Written Khasi is very similar to written English because of the usage of the Latin script and the use of whitespace for marking word boundaries. Each sentence in the corpus is written in one line and marked with an end of sentence marker such as the period (.), the question mark (?), or the exclamation mark (!). Each token in a sentence is separated by a whitespace. Punctuations are also considered as tokens, except for two punctuations- the apostrophe (') which is part of a contracted word and the hyphen (-) which is part of a compound word. The data has been annotated with the Khasi BIS tagset containing 33 tags [1].

Corpus analysis revealed that 10.9% of the word types are multifunctional. The abbreviations used are in accordance with the Leipzig glossing rules except when clearly specified[1]. Table I shows the frequency of the most common words occurring more than five hundred times in the corpus. The statistics show that these 15 most frequent words account to 34.7% of the word tokens in the corpus. If the most frequent words occurring 100 times or more are taken into account, it amounts to 55.8% of the tokens in the corpus. However, from Table II we can see that approximately 47.7% of the word types occur only once in the corpus. These statistics are in line with what is reported by Manning and Shütze [11] about the difficulty in predicting the behavior of words even with the availability of a larger and bigger corpus.

Another phenomenon related to natural language data is the Zipfian distribution. When frequencies (f) of different word types is calculated and ranked in order of occurrences, then according to Zipf's law, $f \propto \frac{1}{r}$ and we can also say that there is a constant *k* where f * r=k. Drawing this information from the training corpus, the extracted values and their respective calculations are shown in Table III. Based on this extraction, along with the usage of logarithmic scales, the graph of Fig. 1 shows the plot of the rank of word type on the X-axis versus the frequency of the respective word type on the Y-axis. The double line graph shows the ranks and frequencies of the words in the corpus, and the straight line shows Zipf's predicted value for k = 10000. The graph seems to approximately hold Zipf's law, except for very low rank and high rank words.

Table I also reveals other suspected language phenomena. For instance, pronouns such as *ka*, *u*, *i*, and *ki* can have other functions and various researchers [12], [13], [14], [15] have referred to them as articles, pronominal markers, noun gender markers, subject enclitic, and so on. The frequencies indicate that the third personal pronouns- *ka* 'singular feminine', *i* 'singular neutral' (183 occurrences as a pronominal marker versus 93 occurrences as a pronoun), and *ki* 'plural'- are more likely to have a sense of pronominal markers (tagged as PR_PRP_M) than of personal pronouns (tagged as PR_PRP). However, *u* "singular masculine" is more likely to be a personal pronoun than a pronominal marker. Since all animate and inanimate objects in Khasi have gender, and given that the Khasi tribe follows a matrilineal system, Khasi corpus analysis indicates that there are more objects tagged as feminine than masculine. The feminine pronominal marker *ka* is approximately 2 times more than the masculine pronominal marker *u* (Table I).

TABLE I. Most Common Word Frequency

Sl. No	Word	Frequency	Frequency of Parts of Speech
1.	ka 3SF	7,212	pronominal marker= 4,946; pronoun= 2,263; others= 3
2.	u 3SM	4,332	pronominal marker= 2,040; pronoun= 2,292
3.	ki 3PL	4,272	pronominal marker= 2,515; pronoun= 1,757
4.	la AUX	3,214	auxiliary verb= 2,590; possessive particle= 506; subordinating conjunction= 69 ; others= 49
5.	ïa ACC	2,864	preposition= 2,852; verb=12
6.	bad 'and'	2,276	coordinating conjunction= 2,166; preposition=110
7.	ha DAT	1,540	preposition= 1,534; others= 6
8.	ba	1,452	subordinating conjunction= 1,443; coordinating conjunction = 6; others=3
9.	ban INF	1,346	infinitive= 1,311; preposition=19; others=16
10.	jong GEN	1,102	preposition=1,096; others=6
11.	nga 1SN	753	pronoun = 752; noun = 1
12.	da INS	725	preposition= 415; auxiliary verb= 308; verb=2
13.	kaba 3SREL	615	relative pronoun
14.	na 'from'	576	Preposition
15.	don COP	545	auxiliary verb = 527; proper noun=18

TABLE II. Frequency of Frequency

Frequency	Frequency of Frequency
1	2608
2	820
3	397
4	221
5	188
6	145
7	108
8	94
9-10	111
11-30	432
31-100	240
>100	101

[1] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

TABLE III. EXPERIMENTAL CALCULATION OF ZIPF'S LAW ON THE CORPUS

Word	Frequency (f)	Rank	f*r	Word	Frequency (f)	Rank	f*r
ka	7212	1	7212	phin	108	100	10800
u	4332	2	8664	yn	55	200	11000
ki	4272	3	12816	iarap	35	300	10500
la	3214	4	12856	jumai	25	400	10000
ia	2864	5	14320	phareng	19	500	9500
jong	1102	10	11020	pdeng	15	600	9000
long	414	20	8280	synshar	12	700	8400
sa	321	30	9630	nguh	10	800	8000
haba	273	40	10920	jingdum	8	900	7200
iing	229	50	11450	jingkylli	7	1000	7000
bha	180	60	10800	syndakor	3	2000	6000
ri	150	70	10500	b.ed.	1	3000	3000
shong	138	80	11040	keep	1	4000	4000
ngam	126	90	11340	satdam	1	5000	5000

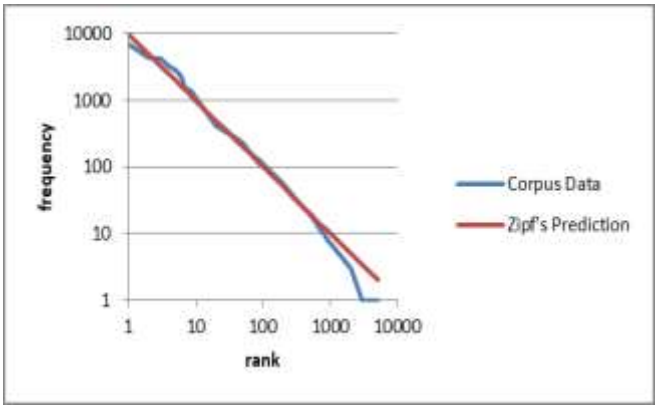

Fig. 1. Zipf's Law: The Red Line shows Relationship between Rank and Frequency Predicted by Zipf for k = 10000, that is f*r = 10000. The Blue Line Corresponds to the Ranks and Frequencies of the Words in the Khasi Corpus. Logarithmic Scales are used for Both Rank and Frequency.

Finally, Table IV highlights the 10 most common tags in the training corpus excluding the punctuation tag. The most common tag is the common noun (N_NN). Remarkably, the fact that Khasi is known to be rich in adverbs is also reflected by its usage in the corpus and its position in the table (RB is the fifth most common tag out of 33 tags).

TABLE IV. MOST FREQUENT TAGS

Rank	Tags	Frequency
1	N_NN	13025
2	V_VM	11193
3	PR_PRP_M	9708
4	PR_PRP	8355
5	RB	7490
6	IN	7373
7	V_VAUX	4970
8	CC_CCD	3057
9	JJ	2348
10	CC_CCS	2225

IV. APPLYING THE HIDDEN MARKOV MODEL FOR POS TAGGING

Given a Khasi sentence of n words $W = w_1\, w_2 \ldots w_n$, we have to assign the best possible tag sequence $T = t_1\, t_2 \ldots t_n$, to the given sentence. Here, t_i is a tag from the BIS tagset for.

Khasi and assigned to word w_i, where $1 \leq i \geq n$. A Khasi sentence tagged using the Khasi BIS tagset is as follows:

Kane/DM_DMD ka/PR_PRP_M shnong/N_NN ka/PR_PRP long/V_VAUX halor/N_NST u/PR_PRP_M lum/N_NN u/PR_PRP baitynnat/JJ shikatdei/RB eh/RP_INTF ./RD_PUNC.

'This village is on a very beautiful hill.'

As proposed by Brants [16], a second-order Markov model along with additional tags t_{-1}, t_0, and t_{n+1} for beginning and end of sentence indicators is incorporated in part of speech tagging for Khasi as follows:

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

To handle data sparsity in (1), he suggested linear interpolation of unigrams, bigrams, and trigrams. Hence, the probability is recalculated as in (2), where the λs are evaluated using deleted interpolation and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

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

A. Integrating Khasi Morphology to Handle unknown Words

As mentioned in Tham [1], Khasi affixes are easily detectable, especially the prefixes which play a major role in Khasi derivational morphology. There is a consistent pattern of Khasi words with prefixes such as *jing-*, *nong-* and *maw-* mapping to common nouns (N_NN), and prefixes such as *pyn-* and *ïa-* (excluding the preposition *ïa*) mapping to verbs (V_VM). To estimate the probability of unknown words having these features, words in the Khasi corpus having prefixes *jing-*, *nong-* , *maw-*, *pyn-*, and *ïa-* (excluding preposition) are mapped to pseudowords _JING_, _NONG_, _MAW_, _PYN_ and _IA_ respectively. To handle unknown

words which do not have the above-mentioned prefixes, low frequency words in the training data are mapped to pseudoword _UNK_. As suggested by Manning and Shütze [11], words occurring only once in the corpus are treated as rare words or out-of vocabulary items, and hence can be mapped to pseudoword _UNK_. They have stated that these words, correspondingly known as hapax legomena, tend to comprise half of the word types, but only a fraction of the tokens in the corpus. Hence, these words will not significantly affect the model. The same phenomenon is likewise observed in Khasi, where such words comprise 47.7% of the word types but only 0.03% of the tokens in the corpus. Therefore, low frequency is taken to be less than or equal to a selected value of γ, and in this tagger γ=1. After the mappings are done, the HMM parameters are evaluated as mentioned earlier where the pseudowords _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.

B. Testing Results and Error Analysis of HMM POS Tagger

The corpus comprising of 94,651 tokens is used for training and testing a baseline tagger, a Natural Language Toolkit (NLTK) tagger [17], and an HMM POS tagger. The mappings mentioned in Section A are incorporated in the baseline tagger and HMM POS tagger. A baseline tagger employed here is a tagger that tags the most probable tag to each word in the test data as put forward by Jurafsky and Martin [18]. Unigram, bigram, trigram taggers, etc., are also provided in NLTK. In the case of the NLTK tagger, it integrates a trigram tagger which backs off to a bigram tagger, the bigram tagger which backs off to a unigram tagger, and the unigram tagger which backs off to a Khasi regular expression. The Khasi regular expression tagger incorporates Khasi morphology, tagging 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). Hapax legomena words not containing the mentioned prefixes are preprocessed and mapped to pseudoword _UNK_. The results of all the three taggers using ten-fold cross-validation are given in Table V, with the HMM POS tagger giving a relatively good performance of 93.39% accuracy.

A confusion matrix of the HMM POS tagger, shown in Table VI, is used in analyzing the errors during the HMM POS tagging, with the values reflecting errors occurring at 0.5% and above (i.e., an average frequency of 3 and above). The rows in Table VI indicate the correct tags, the columns indicate the HMM tagger's predicted tags, and each cell indicates the percentage of the tagging error. The most common error which is difficult to disambiguate is when proper nouns are tagged as common nouns and vice versa, accounting to 12% of the errors. Here, the HMM tagger has not been able to take into consideration the capitalization feature of proper nouns. A brief discussion on some of the tagging errors is given as follows:

TABLE V. KHASI POS TAGGER RESULTS USING TEN-FOLD CROSS VALIDATION

Tagger	**Accuracy**
Baseline Tagger	84.05%
NLTK Tagger	87.58%
HMM POS Tagger	**93.39%**

Verb Noun / Noun Verb confusion.

An interesting phenomenon that mainly contributes to the collective occurrence of 20% of the errors is when pronouns are tagged as pronominal markers, the words following them are inadvertently tagged as nouns rather than verbs. When pronominal markers are tagged as pronouns, the words following them are likewise tagged as verbs rather than nouns. For example:

U/PR_PRP lum/V_VM ïa/IN u/PR_PRP_M soh/N_NN ka/PR_PRP_M jingtrei/N_NN shitom/JJ jong/IN u/PR_PRP ./RD_PUNC.

'He reaps the fruit of his hard work.'

In the sentence above, the verb *lum* 'reap' is incorrectly tagged as a noun *lum* 'hill/mountain' in the sentence given below. This is because the preceding word *u* was incorrectly tagged as a pronominal marker rather than a personal pronoun.

U/PR_PRP_M* [2] lum/N_NN* ïa/IN u/PR_PRP_M soh/N_NN ka/PR_PRP_M jingtrei/N_NN shitom/JJ jong/IN u/PR_PRP ./RD_PUNC.

Another example where a noun is incorrectly tagged as a verb is shown in the sentences below:

Baroh/JJ ki/PR_PRP_M diengsohniamtra/N_NN sawdong/RB ïa/IN u/PR_PRP ki/PR_PRP don/V_VAUX tang/RB kawei/QT_QTC ban/V_VAUX_VINF iathuh/V_VM ,/RD_PUNC ka/PR_PRP_M jingjot/N_NN bad/CC_CCD ka/PR_PRP_M jinglehnohei/N_NN ./RD_PUNC.

'All the orange trees have only one thing to tell, tales of destruction and lost.'

Baroh/JJ ki/PR_PRP* diengsohniamtra/V_VM* sawdong/RB ïa/IN u/PR_PRP ki/PR_PRP don/V_VAUX tang/RB kawei/QT_QTC ban/V_VAUX_VINF iathuh/V_VM ,/RD_PUNC ka/PR_PRP_M jingjot/N_NN bad/CC_CCD ka/PR_PRP_M jinglehnohei/N_NN ./RD_PUNC.

Apart from what is described above, other instances when a noun is erroneously tagged as a verb are when the tagger cannot distinguish a verb functioning as a noun.

Adjective Verb / Verb Adjective confusion

Various researchers have put forward their views on the existence of adjectives in Khasi due to their syntactic similarity with verbs [12], [19], [20]. The confusion matrix also indicates that the tagger has tagged some adjectives as verbs, especially when the words follow a subordinating conjunction *ba* or the auxiliary word *la*.

[2] *Incorrect tag

TABLE VI. CONFUSION MATRIX IN % OF KHASI HMM POS TAGGER

	N_NN	PR_PRP	V_VM	PR_PRP_M	N_NNP	RB	JJ	RD_ECH	RD_RDF	CC_CCD	RP_RPD	QT_QTC	IN	PR_PRF
N_NNP	12.7		1.3			0.8			0.6					
V_VM	5.3					2.9	1.4	0.8						
PR_PRP				5.0				0.5						
N_NN			4.9		3.7	3.3	1.6	1.5	1.2			0.7		
PR_PRP_M		4.8						1.0						
JJ	3.3		1.9			2.7						0.5		
RB	3.0		3.1				1.9							
IN						0.8		0.6		1.2				
QT_QTC	1.0													
PR_PRF											0.9			
RD_PUNC								0.6						
RP_INJ						0.6								
CC_CCD								0.5					0.5	
CC_CCS						0.5								
RP_RPD														0.5
RD_RDF	0.5													

Ka/PR_PRP_M khmat/N_NN jong/IN ka/PR_PRP ka/PR_PRP la/V_VAUX stem/JJ blaid/RB blaid/RB ./RD_PUNC

'Her eyes have turned yellowish.'

Ka/PR_PRP_M khmat/N_NN jong/IN ka/PR_PRP ka/PR_PRP la/V_VAUX stem/V_VM* blaid/RB blaid/RB ./RD_PUNC

Likewise, there are some instances where a verb has also been tagged as an adjective. Here, *bah* 'carry on the back/shoulders' has been erroneously tagged as an adjective.

Nga/PR_PRP la/V_VAUX kit/V_VM, /RD_PUNC nga/PR_PRP la/V_VAUX bah/V_VM, /RD_PUNC bad/CC_CCD nga/PR_PRP la/V_VAUX bysa/V_VM la/V_VAUX btiah/V_VM ïa/IN phi/PR_PRP baroh/JJ ./RD_PUNC

'I bore the burden and raised all of you.'

Nga/PR_PRP la/V_VAUX kit/V_VM, /RD_PUNC nga/PR_PRP la/V_VAUX bah/JJ*, /RD_PUNC bad/CC_CCD nga/PR_PRP la/V_VAUX bysa/V_VM la/V_VAUX btiah/V_VM ïa/IN phi/PR_PRP baroh/JJ ./RD_PUNC

Noun Adverb confusion

When a noun is tagged as an adverb, it is because the noun follows the verb without a preceding pronominal marker. This is an example where the mandatory pronominal marker is dropped before the noun *Sein Iong* 'black snake'.

U/PR_PRP la/V_VAUX kylla/V_VM Sein/N_NN Iong/N_NN ./RD_PUNC.

'He turned into a black snake.'

U/PR_PRP la/V_VAUX kylla/V_VM Sein/RB* Iong/RB* ./RD_PUNC.

Adjective Noun / Noun Adjective confusion

When an adjective is tagged as a noun, it is more likely that the tagger cannot differentiate a compound noun from a noun having a qualifying adjective. This is mainly because in most cases adjectives follow the nouns they qualify. Here the adjective *badon baem* 'well-to-do' is tagged as a compound noun.

Ka/PR_PRP pher/JJ na/IN kiwei/JJ pat/RP_RPD ki/PR_PRP_M khun/N_NN badon/JJ baem/JJ kiba/PR_PRL nga/PR_PRP la/V_VAUX iakynduh/V_VM ./RD_PUNC.

'She is different from all the well-to-do kids that I have met.'

Ka/PR_PRP pher/JJ na/IN kiwei/JJ pat/RP_RPD ki/PR_PRP_M khun/N_NN badon/N_NN* baem/N_NN* kiba/PR_PRL nga/PR_PRP la/V_VAUX iakynduh/V_VM ./RD_PUNC

Nouns are erroneously tagged as adjectives in compound nouns, or when an adjective or a verb semantically functions as a noun, as seen in the sentence below. Here *u rit u riat* 'low-class people' is confused as *rit* and *ria*, which means 'small' in another sense of the words. Additionally, the preceding word being tagged as a pronoun rather than a pronominal marker adds to the confusion.

Mynta/RB ,/RD_PUNC u/PR_PRP_M rit/N_NN u/PR_PRP_M ria/N_NN u/PR_PRP shu/RB pyrta/V_VM shla/RB ./RD_PUNC.

'Now, the low-class people are just shouting angrily.'

Mynta/RB ,/RD_PUNC u/PR_PRP* rit/JJ* u/PR_PRP* ria/JJ* u/PR_PRP shu/RB pyrta/V_VM shla/RB ./RD_PUNC.

The above discussion indicates that one way of addressing the existing confusion is to consider the properties or attributes of words such as capitalization, next occurring word, and others. These considerations are presented in the next section.

V. A Hybrid Khasi POS Tagger to Address Tagging Errors

To reduce the errors present in the HMM POS tagger output, the errors identified in Section IV B need to be addressed.

To do so, the sklearn-crfsuite[3] has been engaged as a means to achieve this purpose. Sklearn-crfsuite is a thin python-crfsuite wrapper which provides a fast implementation of conditional random fields (CRF). Unlike HMM, CRFs allow the inclusion of features that are non-independent and varied in depth even on the same observation [21]. Using CRF, given a sentence $X = x_1\, x_2 \ldots x_T$, the conditional probability of the tag sequence $Y = y_1\, y_2 \ldots y_T$ is given by:

$$P(y|x) = \frac{1}{Z(x)} * \exp\left(\sum_{t=1}^{T} \sum_k \theta_k . f_k\ (y_{t-1}, y_t, x)\right) \quad (3)$$

where $Z(x) = \sum_T exp\ (\sum_{t=1}^{T} \sum_k \theta_k . f_k\ (y_{t-1}, y_t, x))$ is the normalization factor, θ_k is the weight and $f_k(y_{t-1}, y_t, x)$ is the feature function. Implementing POS tagging in sklearn-crfsuite permits the possibility to include as many word features as possible to aid the tagging process. The word features included for Khasi are capitalization, prefixes (prevalent in Khasi, unlike suffixation in English) of length >=2 and length<=4, current word under consideration, previous word, next word, and whether a word begins or ends a sentence. An additional feature that can be included is the previous tag of a word. In the interface provided by sklearn-crfsuite the features are extracted from the training data and from the test data. It expects the training data to contain annotated data, i.e., the words and their respective POS tags. This will enable sklearn-crfsuite to extract all the specified features and learn the tagging process. However, when tagging the test data, the problem arises during feature extraction from the test data. In the provided interface, all the above-mentioned features are possible to extract from the test data except the previous word tag feature. This feature is not available in the test data because it contains only sentences where the respective words are yet to be tag. To overcome this problem, the output of the Khasi HMM POS tagger is used as input to the Khasi CRF POS tagger. This enables the previous tag feature to be easily extracted from the tagged output of the HMM tagger.

[3] https://sklearn-crfsuite.readthedocs.io/en/latest/

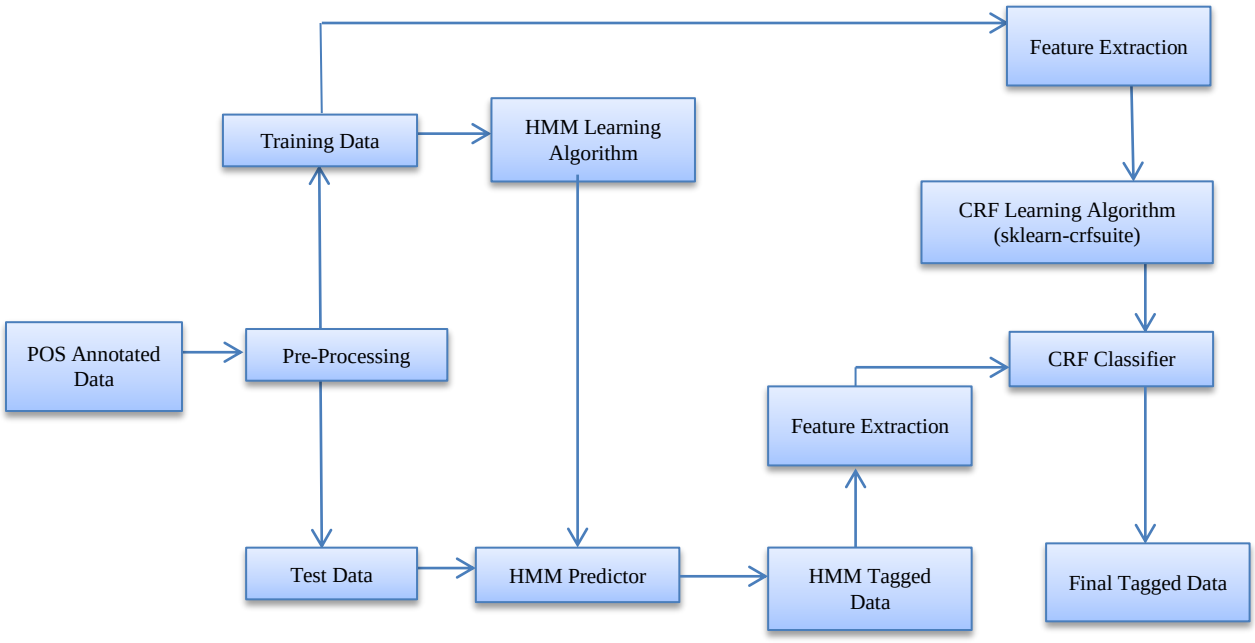

Fig. 2. Block Diagram of the Hybrid POS Tagger.

Fig. 2 shows the block diagram of the implementation. The features mentioned above are included in the CRF tagger. The CRF POS tagger is then trained on the same training data used by the HMM POS tagger. To ensure consistency, ten-fold cross-validation is undertaken for training and testing the CRF tagger. During tagging, the word features are extracted from the test data and the previous tag feature is provided by the output of the HMM tagger. In doing so, training on 4k sentences and tagging on 10% of the training sentences, an average tagging accuracy of **95.29**% is achieved with an average improvement of 1.9% over the performance of the HMM tagger shown earlier in Table V.

VI. Evaluation of Hybrid POS Tagger

Table VII shows the average precision, recall, and F-measure of both the HMM POS Tagger and the Hybrid POS Tagger. The F-measure of the Hybrid POS Tagger shows a significant improvement over the F-measure of the HMM POS Tagger in 23 tags out of 32 tags, with one tag RD_UNK (for unknown tags) giving 0 F-measure in both taggers. This may be attributed that RD_UNK occurs exactly once in the corpus. Since the corpus captures only prose genre, it may be the factor where the frequency of symbols in the corpus is only 4. This may be the reason that the Hybrid POS Tagger has failed to predict the symbol tag by giving an F-measure of 0 for RD_SYM symbol tag. Fig. 3 shows the graphical comparison of the average confusion frequency among the tags between the HMM tagger and the Hybrid tagger, arranged in descending order of the HMM tagger confusion percentage. Table VIII shows the percentage of reduction or increase in confusion in the Hybrid POS tagger from the HMM POS tagger. The rows in Table VIII indicate the correct tags, the columns indicate the Hybrid tagger's predicted tags, and each cell indicates the percentage increase or reduction in tagging error from the HMM tagger. The biggest improvement of the Hybrid tagger is the 100% elimination of all the tags confused as echo tags (RD_ECH) shown earlier in Table VI.

Another significant improvement is in its ability to disambiguate proper nouns (N_NNP) from common nouns (N_NN) with an 89% reduction of confusion; a trait where CRF classifiers are good at capturing word features such as capitalization. The same is observed in noun and foreign word confusion (N_NN, RD_RDF), noun and adverb confusion (N_NN, RB), and the coordinating conjunction and preposition confusion (CC_CCD, IN); all of them over the 80% confusion reduction. Overall, it is clear that the Hybrid tagger has reduced most of the confusion mentioned in Section IV-B. Interestingly, even the Hybrid tagger has a problem in disambiguating adjectives from verbs, a language phenomenon debated by researchers [11], [13], [19], [20]. In this category, the confusion has not reduced but increased by 65%.The other two tags that showed a relatively small increase in confusion are the adverb and noun (RB, N_NN) -confusion by 10%- and the interjection and adverb (RP_INJ, RB)- confusion by 8%. The remaining three confusing tags showed a 2% or less increase in confusion. All the most common confusion tags that have an average frequency of 3 or more are indicated in Table VIII.

TABLE VII. AVERAGE PRECISION, RECALL, AND F-MEASURE OF BOTH TAGGERS

Sl. No	POS Tags	**HMM POS Tagger**			**Hybrid POS Tagger**		
		Precision	Recall	F-measure	Precision	Recall	F-measure
1	CC_CCD	95.95	97	96.46	96.13	98.29	97.18
2	CC_CCS	96.99	95.68	96.31	98.66	96.71	97.66
3	DM_DMD	100	96.03	97.93	98.55	97.16	97.79
4	IN	98.44	96.3	97.35	98.64	97.76	98.2
5	JJ	82.6	75.16	78.54	89.03	78	83.01
6	N_NN	87.38	91.69	89.45	92.75	93.57	93.15
7	N_NNP	75.97	54.72	61.89	88.15	94.83	91.2
8	N_NST	99.57	96.12	97.76	97.66	91.19	94.25
9	PR_PRF	73.02	67.64	68.88	82.65	72.03	74.5
10	PR_PRI	91.58	90.04	90.52	95.5	92.4	93.71
11	PR_PRL	99.67	98.03	98.83	99.8	98.79	99.29
12	PR_PRP	96.02	95.53	95.75	96.58	95.96	96.26
13	PR_PRP_AUX	98.43	99.59	99	97.94	99.35	98.63
14	PR_PRP_M	96.64	96.26	96.44	96.21	97.44	96.81
15	PR_PRQ	92.56	75.23	82.01	96.24	77.69	85.06
16	QT_QTC	85.37	83.43	83.93	96.25	88.46	92.02
17	QT_QTF	97.79	97.15	97.43	96.86	96.82	96.76
18	QT_QTO	83.17	78.22	78.64	97.15	64.71	74.01
19	RB	88.29	91.06	89.64	91.53	93.41	92.46
20	RD_ECH	21.28	22.5	13.88	30	13.34	17.53
21	RD_PUNC	89.98	89.6	89.79	99.99	99.88	99.93
22	RD_RDF	38.51	66.51	41.35	36.55	47.1	33.44
23	RD_SYM	10	10	10	0	0	0
24	RD_UNK	0	0	0	0	0	0
25	RP_CL	98	98.23	97.96	100	98.75	99.34
26	RP_INJ	71.04	71.95	70.87	76.71	73.46	74.36
27	RP_INTF	95.05	95.67	95.18	94.82	99.51	96.95
28	RP_NEG	99.73	98.88	99.3	100	98.84	99.42
29	RP_POS	89.46	92.87	90.92	89.31	97.07	92.61
30	RP_RPD	94.16	92.09	93.04	94.38	93.82	93.95
31	V_VAUX	97.08	97.31	97.19	97.52	98.08	97.8
32	V_VAUX_VINF	98.25	98.24	98.24	97.91	99.45	98.67
33	V_VM	92.71	92.86	92.76	93.22	94.38	93.78

TABLE VIII. REDUCTION % IN CONFUSION OF HYBRID POS TAGGER

	N_NN	PR_PRP	V_VM	PR_PRP_M	N_NNP	RB	JJ	RD_ECH	RD_RDF	CC_CCD	RP_RPD	QT_QTC	IN	PR_PRF
N_NNP	88.9		68.2			84.9			85.4					
V_VM	1.7					42	23	100						
PR_PRP				2.2#				100						
N_NN			0.6		15.7	13.6	37	100	86.1			87.5		
PR_PRP_M		18.6						100						
JJ	21.8		65.3#			31						76.5		
RB	10.3#		43.7				51.2							
IN						47.1		100		1.3#				
QT_QTC	48.2													
PR_PRF											1.8#			
RD_PUNC								100						
RP_INJ						8.3#								
CC_CCD								100					82.9	
CC_CCS						71								
RP_RPD														25.8
RD_RDF	3.1													
	#Increase % in confusion													

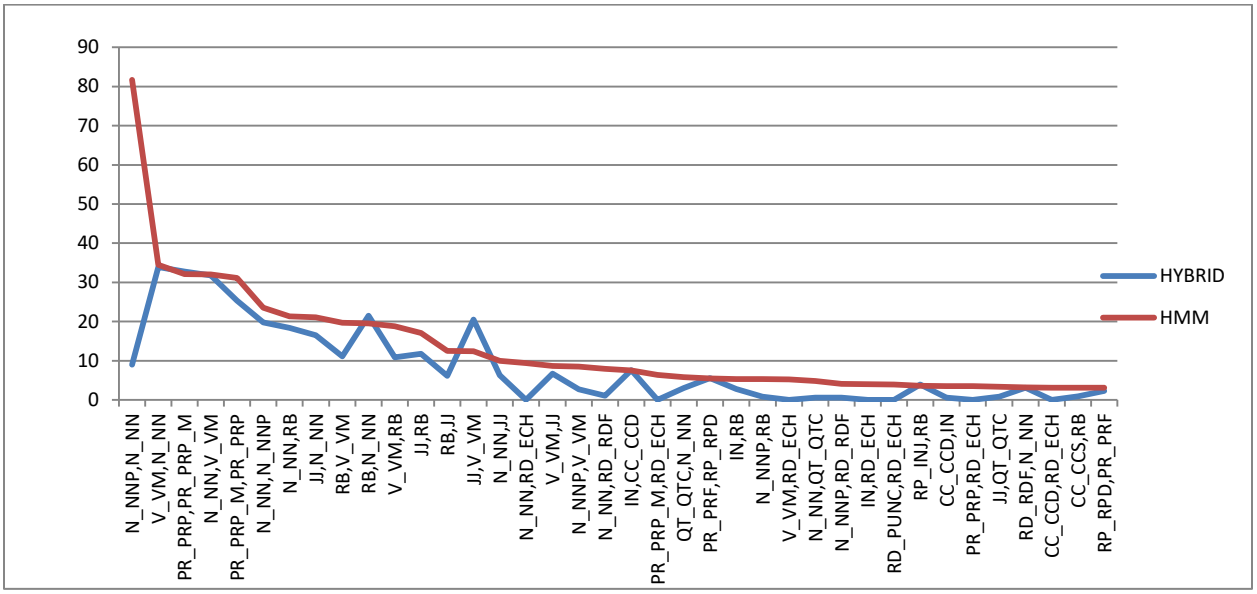

Fig. 3. The Graph Shows the Tag Confusion between the HMM POS Tagger and the Hybrid POS Tagger where Label x,y means Tag x is Confused as y.

VII. Conclusion

Although the present annotated 90k corpus available for Khasi is relatively small, nevertheless experiments with automatic tagging using HMM along with the BIS tagset for Khasi have shown performance that does not lack behind reported performance in other languages. As shown in this paper, addressing the tagging errors of the HMM POS tagger by coupling it with the sklearn-crfsuite, a fast implementation of CRF has given a Hybrid POS tagger for Khasi with an improved accuracy of 95.29%.

The results are very promising for an under resourced language such as Khasi. Apart from the concerns regarding the performance of the tagger, Khasi POS tagger development has also highlighted issues that were often raised in the literature of Khasi language. Does the Hybrid tagger's confusion between verbs and adjectives imply that the confused adjectives are actually attributive verbs? However, to answer this question, further investigation in this direction is still needed. Finally, the next step is to include a wider range of genres for the corpus, which hopefully, with the current POS tagger in place, will ease the development towards a larger size annotated corpus.

References

[1] M. J. Tham, "Challenges and issues in developing an annotated corpus and HMM POS tagger for Khasi", Proceedings of the 15th International Conference on Natural Language Processing (1CON 2018), Patiala, Punjab, India, pp .15-18, December 2018.

[2] K. Darwish et al. "Multi-dialect Arabic POS tagging: a CRF approach", Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018.

[3] W. Khan, et al. "Urdu part of speech tagging using conditional random fields", Language Resources and Evaluation, 53, pp. 331–362. 2018.

[4] S. Song, N. Zhang, and H. Huang, "Named entity recognition based on conditional random fields", Cluster Comput, 2017.

[5] K. S. Nagaraja, Word formation in Khasi, Bulletin of the Deccan College Research Institute 60/61:387-417, 2000.

[6] B. War, Ki sawa bad ki dur jong ka ktien Khasi, 2nd edn. Ri-Ia-dor, Shillong, Meghalaya, 2011.

[7] N. Garg, V. Goyal, and S.Preet, "Rule based Hindi part of speech tagger", Proceedings of COLING 2012, Mumbai, India, pp. 163-174, 2012.

[8] S. Singh, K. Gupta, M. Shrivastava, and P. Bhattacharyya. "Morphological richness offsets resource demand – experiences in constructing a POS Tagger for Hindi", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia, pp. 779-786, 2006.

[9] M. Shrivastava, P. Bhattacharyya, "Hindi POS tagger using naive stemming: harnessing morphological information without extensive linguistic knowledge", Proceedings of the International Conference on NLP (ICON 08). Pune, India, 2008.

[10] S. Warjri, P. Pakray, S. Lyngdoh, and A. K. Maji, "Identification of POS tag for Khasi language based on Hidden Markov Model POS tagger". Computación y Sistemas, 23(3):795-802, 2019.

[11] C. D. Manning, and H. Schütze, Foundations of statistical natural language processing. MIT press, 1999.

[12] M. B. Jyrwa, A descriptive study of the noun phrase in Khasi. Dissertation, North Eastern Hill University, 1989.

[13] K. S. Nagaraja, Khasi a descriptive analysis, Deccan College Post-Graduate & Research Institute, Pune, India, 1985.

[14] L. Rabel, Gender in Khasi nouns, Mon Khmer Studies 6:247-272, 1977.

[15] H. Roberts, A grammar of the Khasi language, Mittal Publications, New Delhi, India , 2005.

[16] T. Brants, "TnT-A statistical part of speech tagger", Proceedings of the Sixth Conference on Applied Natural Language Processing, Seattle, Washington, USA, pp. 224-231, 2000.

[17] S. Bird, E. Klein, and E. Loper, Natural language processing with python. O'Reilly Media Inc, CA, 2009.

[18] D. Jurafsky, and J. H. Martin, Speech and language processing. An introduction to natural language processing, Computational Linguistics, and Speech Recognition, 2nd edn. Pearson India Education, Noida, 2009.

[19] L. Rabel, Khasi a language of Assam. Baton Rouge: Louisiana State University Press, 1961.

[20] I. M. Simon, "Some observations on the adjectives in Khasi", Khasi Studies 1(3), 1987.

[21] J. Lafferty, A. McCallum, and F. C Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", 2001.