International Journal of Speech Technology (2021) 24:853–864
https://doi.org/10.1007/s10772-021-09860-w

Part-of-speech (POS) tagging using conditional random field (CRF) model for Khasi corpora

Sunita Warjri[1] · Partha Pakray[2] · Saralin A. Lyngdoh[3] · Arnab Kumar Maji[1]

Received: 1 October 2020 / Accepted: 21 May 2021 / Published online: 4 June 2021

Abstract
Khasi is a language that belongs to the Mon-Khmer language of the Austroasiatic group. Khasi language is spoken by the indigenous people of the state of Meghalaya in India. This paper presents a work on Part-of-speech (POS) tagging for the Khasi language by using the Conditional Random Field (CRF) method. The main significance of this work, is to experiment with the CRF model for PoS tagging in the Khasi language. This method produces a reliable agreement on the features of the language. POS tagging for Khasi is essential for creating lemmatizers which are used to lessen a word to its root structure and the POS corpus or dataset can be used in other NLP applications. In this research work, we have designed a tag set and POS tagging corpus. Khasi does not have any standard POS corpus. Therefore, we have to build a Khasi corpus that consists of around 71,000 tokens. After feeding the Khasi corpus to the CRF model for learning, the system yields a testing accuracy of 92.12% and an F1-score of 0.91. The result is compared with few other state-of-art techniques. It is observed that our approach produces promising results in comparison with other techniques. In future, we will increase the size of the Khasi POS corpus.

Keywords Natural language processing (NLP) · Parts of speech (POS) tagging · Khasi language · Lexical morphology · Conditional random field (CRF) model · Khasi corpus

1 Introduction

An area that deals with the interaction between the human communication and the machine is called Natural Language Processing (NLP). NLP or Computational Language (CL) uses Artificial Intelligence (AI) techniques of computer science in order to make the machine understand the human natural language. NLP is having huge domains of applications and many languages have been introduced to it's different fields.

In this paper, our main motivation is to discuss the findings based on implementation of Conditional Random Field (CRF) model for Part-of-Speech (POS) tagging in Khasi language by using the designed Khasi POS corpus (Warjri, 2020). POS is one of the initial field and strong backbone of NLP. POS tagging is about identifying the grammatical class or grammatical property of each word. POS tagging is used to label each word in a sentence with its appropriate grammatical syntactic category. In language processing, POS tagging is one of the most important foundation. Different languages may have a different kind of syntax/semantics while writing or speaking. Hence the grammatical structure of different languages varies widely in different contexts. Thus, POS or grammatical class of the language is a challenging task.

POS tagging is very useful in performing NLP operation as a sequencing problem, especially for deep parsing of text. POS as a sequence tagging is found to be very helpful and powerful in order to achieve good system

✉ Arnab Kumar Maji
arnab.maji@gmail.com

Sunita Warjri
sunitawarjri@gmail.com

Partha Pakray
parthapakray@gmail.com

Saralin A. Lyngdoh
saralyngdoh@gmail.com

[1] Department of Information Technology, North-Eastern Hill University, Shillong, Meghalaya, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Silchar, Assam, India

[3] Department of Linguistics, North-Eastern Hill University, Shillong, Meghalaya, India

performance. In many other fields of NLP, such as Named Entity Recognition (NER), Question Answering, Music modeling, Machine Translation, Speech Recognition, and Information Retrieval, POS can be used. Many researchers have already performed works in many different languages concerning many fields of NLP. To tag POS on the natural texts, most of the researcher uses different methods such as the stochastic technique, rule-based technique i.e based on the language context rule, or by combining both the technique (Brants, 2000; Merialdo, 1994; Ekbal et al., 2007; Cutting et al., 1992). In POS tagging, we usually need a tagset, POS corpus of a particular language. A set of tags is usually called a tag set. Tagset consists of tags or labels that describe a grammatical class. Tagset is used by assigning each word in sentences with its corresponding tag or label. As an example, in Table 1, few words along with their corresponding tags and description of the tags, is shown. In the example, the word Cat is tagged as CMN that represent or describe "Common Noun".

The main objective of the work is to use Conditional Random Field (CRF) for POS tagging in Khasi language using the designed corpus. A brief comparison with few other existing work in Khasi language is also introduced in this work. As Khasi is a very low resourced Indian language, there are very few works available in this domain. We have used our designed Khasi POS corpus (Warjri, 2020), as there is a paucity of standard Khasi corpus. POS is a prepossessing technique for any work related to Natural Language Processing. As it is found that CRF based technique can produce promising accuracy, our work will really have benefits in all NLP related works in this low resourced Khasi language. The motivation of this research is to include Khasi language and its resources in Computational Linguistics research. The paper is organized as follows: Sect. 2 describes related works on POS Tagging for Indian languages; Sect. 3 describes the CRF approach; Sect. 4 describes Khasi POS Tagging methodology using CRF model; Sect. 5 shows the experimental results ; Sect. 7 consists of Conclusions and some future perspectives of the work. Finally the paper ends with a brief discussion about the work.

Table 1 Example of POS tagging

Word	Tag	Description
Cat	CMN	Common noun
John	PPN	Proper noun
Climb	VB	Verb
Happy	ADJ	Adjective

2 Existing literature on conditional random field

CRF approach is used widely in many Indian languages for POS tagging. In this section, a brief discussion is presented on different Indian languages, where CRF approach is used for POS tagging. The summary of all the existing works on relevant domain is presented in Table 2.

1. Bengali: Bengali or Bangla is an Indo-Aryan language, that is spoken in South Asia by the Bengalis. The language is spoken mostly in a large part of India such as the states like West Bengal, Tripura, Assam. This language is also widely spoken in Bangladesh (Wikipedia contributors, 2020a). The paper (Ekbal et al., 2007) reported the work on POS tagging for the Bengali language by using the statistical model like the Conditional Random Fields (CRFs). In this work, the tagset consisted of 26 different tags. For predicting the POS classes, the designed tagger utilized the features and information from the context. The system was trained with 72,341 words and tested with 20k words. Using the statistical model, it produced an accuracy of 90.30%. It was also reported that usages of different kinds of word suffixes, named entity recognizer and the lexicon, proved to be very effective in handling unknown words.
2. Assamese: Assamese is an official language and spoken by the people of Assam, India. Assamese is an Indo-Aryan language and it performs as a common language to the nearby state of Assam. This lingua franca is served in the region of Nagaland and Arunachal Pradesh (Wikipedia contributors, 2020b). In paper (Barman et al., 2013), POS tagging for the Assamese language was discussed using the CRF model and Transformation Base Learning (TBL) approach . It was reported that the annotated corpus was not available; therefore an annotated corpus was prepared for the experiment. Using the created corpus that is manually tagged the result obtained are 87.17% for TBL and 67.73% for CRF. Experimenting with both the taggers i.e. CRF and TBL, it was found that the performance of CRF was not satisfactory in comparison with TBL, as CRF couldn't extract the suffixes, prefixes, and other morphological information.
3. Gujarati: Gujarati is an Indo-Aryan language that is spoken in Gujarat, India (Wikipedia contributors, 2020c). In paper (Patel & Gali, 2008), POS tagging is performed using the CRF machine learning model. The manually tagged corpus was used which consists of 600 sentences. Corpus was designed by using 26 tags. For training and testing the system 10,000 and

Table 2 POS tagging on different Indian languages using the CRF model

Sl. no.	Language	Technique/s	Data & Tagsets	Results & Accuracy	References	Year
1	Hindi	CRF++	21,000 words	82.67%	Agarwal and Mani (2006)	2006
2	Bengali	CRF	tags Train data-72,341 words Test data-20k words	90.30%	Ekbal et al. (2007)	2007
3	Hindi	CRF	Train data-21470 words Test data-2924 words	78.66%	Pvs and Karthik (2007)	2007
4	Gujarati	CRF	26 tags 600 sentences	92%	Patel and Gali (2008)	2008
5	Manipuri	CRF and SVM	26 tags 63,200 tokens	72.04% for CRF 74.38% for SVM	Singh and Ekbal (2008)	2008
6	Tamil	CRF	36,000 sentences	F-score of 0.88 (for 18345 words), 0.89 (for 19834 words), and 0.89 (for 18907 words)	Pandian and Geetha (2009)	2009
7	Kannada	HMM and CRF	Train data-51269 words Test data-2932 words	84.54% for CRF 79.90% for HMM	BR and Ramakanth Kumar (2012)	2012
8	Assamese	CRF model and Transformation Base Learning (TBL) approach	–	87.17% for TBL 67.73% for CRF	Barman et al. (2013)	2013
9	Kashmiri	CRF	30,000 words	81.10%	Ahmad and Syam (2014)	2014
10	Malayalam	CRF	100 tags 36,315 words	85.7%	Krishnapriya et al. (2014)	2014
11	Hindi	SVM & CRF++	90k tokens	82 to 86.7% for CRF++ 88 to 93.7% for SVM	Ojha et al. (2015)	2015
12	Kannada	CRF	36 tags 80,000 words	92.94%	Pallavi and Pillai (2016)	2016
13	Punjabi	CRF	36 tags 38k to 42k of words	Precision of 98.9% for Articles, 98.1% for News, 99.6% for stories, 98.6% for Novel, 98.9% EBook and Recall score of 100% for all data	Sharma (2016)	2016
14	Kannada	CRF	19 tags 3000 sentences	96.86%	Suraksha et al. (2017)	2017
15	Odia	CRF	600k tokens	94.11%	Behera (2017)	2017
16	Khasi	HMM tagger NLTK Bigram NLTK Trigram	Train data-86,087 words Test data-8,565 words	88.23% 88.64% 95.68%	Tham (2018)	2018
17	Khasi	HMM tagger	54 tags 7812 tokens	76.70%	Warjri et al. (2019)	2019
18	Urdu	CRF	BJ dataset CLE dataset	F-measure of 86.99% for CLE dataset 93.56% for BJ dataset	Khan et al. (2019)	2019
19	Khasi	CRF tagger	53 tags 41000 tokens	0.922(Precision), 0.922(Recall), & 0.921(F-measure)	Warjri et al. (2011)	2021

5,000 words were used respectively. Using the corpus 92% accuracy was achieved by the system. During the experiment, it was observed that the performance of system accuracy could increase if the rules for the language could be framed.

4. Kannada: Kannada language is spoken at Karnataka, India. Kanarese or Kannada is a Dravidian language (Wikipedia contributors, 2020d). POS tagging is performed for the Kannada language in paper (BR and Ramakanth Kumar 2012) using Hidden Markov Model (HMM) and CRF. The systems were trained with corpus data consisting of 51269 words and for testing 2932 tokens were used. Using the corpus, the systems produce accuracies of 84.54% for CRF and 79.9% for HMM. In another paper (Pallavi & Pillai, 2016) also POS tagging system was discussed for Kannada language using CRF. For this work, 36 tags from Technology Development for Indian Languages (TDIL) were used for annotating the corpus. The corpus consists of 80,000 words. Out of 80,000 words, 64,000 words were used for training and 16,000 for testing. Based on linguistic information such as number, symbol, foreign word, png marker, prefix, tense marker, suffix, and punctuation, 12 features were extracted and trained.

Also different linguistic rules are formed based on word information and permutation. Using the corpus with the features and rules, the CRF system produces an accuracy of 92.94%. Parsing and POS tagging in Kannada language, were also discussed in the paper (Suraksha et al., 2017). The CRF model is implemented for the purpose of parsing and POS tagging. The corpus is built using 19 tags. The corpus consists of 3000 sentences, out of which 2500 sentences were used for training and 500 sentences for testing the system. Using the corpus in the CRF system, an accuracy of 96.86% was achieved.

5. Kashmiri: Kashmiri is a language spoken by people of Kashmir, India. Though Urdu is an official language of Jammu and Kashmir, Kashmiri belongs to the Dardic language of Indo-Aryan language family (Wikipedia contributors, 2020e). In the paper, Ahmad and Syam (2014) POS tagging for Kashmiri language was discussed. CRF model was used for this purpose. This was the first attempt for Kashmiri towards the NLP approach. The corpus of 30,000 words was built. The corpus was divided into sets of training and testing data for checking the accuracy of the system. One of the set consisted of training data of 27,000 words and testing data of 3,000 words yielded an accuracy of 81.10%.
6. Malayalam: Malayalam is a language spoken at Kerala, India. Malayalam belongs to the Dravidian language. Malayalam derives from mala and alam. Which means "mountain" for mala, and alam meaning "region" or "ship". The Malayalam language, on the other hand, is also called as Wikipedia contributors (2020f). In paper (Krishnapriya et al., 2014), POS tagging for Malayalam was discussed using the CRF method. The corpus for this work consisted of 36,315 words with 100 tags. In their experiment, they also used bigram and trigram approach. Out of 3026 sentences from the corpus, 2/3 were used for training the system and 1/3 sentences were used for testing. Using the corpus an accuracy of 85.7% was achieved as a testing result.
7. Manipuri: Manipuri language is also known as Meitei. Meitei belongs to the Sino-Tibetian language. This language is spoken in Manipur, India. This language is found to be generally spoken in North-East India apart from Assamese and Bengali (Wikipedia contributors, 2020g). The paper (Singh & Ekbal, 2008), had discussed about POS tagging using CRF and Support Vector Machines (SVM) model for Manipuri language. In this work, the corpus was annotated using a tagset, consisting of 26 different tags. The corpus consists of 63,200 tokens. The system was trained with 39449 tokens. The features such as word-level and contextual information were also trained in the model. For testing the systems 8672 tokens were used. As a results, 72.04% accuracy was yielded for CRF and 74.38% accuracy for SVM.
8. Odia: Odia is an official language, spoken by the people of Orissa, India. Odia language belongs to the Indo-Arian language. Odia is also the second official language of the state Jharkhand, India (Wikipedia contributors, 2020h). The paper (Behera, 2017) discussed POS tagging for Odia language using the CRF model. For this work, corpus of around 600k tokens were annotated using Bureau of Indian Standards (BIS) tagset. Out of 2,36,793 tokens used in the work, the system had been trained and tested with 1,28,646 tokens. With the training data, an accuracy of 94.11% was achieved.
9. Punjabi: The Punjabi is a local language of the people of Punjab region, India. This language is widely spoken in India and Pakistan, and also in other counties such as Bangladesh, Malaysia, Australia, and the United State. The Punjabi language belongs to the Indo-Aryan language. Punjabi can be written in Gurmukhī script and Shahmukhī script (Wikipedia contributors, 2020i). In the paper (Sharma, 2016), the POS tagger for the Punjabi language was developed using a hybridized approach of CRF and rule-based. The corpus used for the system consisted of around 38k to 42k number of words using TDIL (Technical Development of Indian Languages). Tagset consisted of 36 different tags. The designed corpus data consisted of Articles, News, Stories, Novel, and E-Book. For training the system from the corpus, 2/3 of the sentences were used and 1/3 sentences were used for testing the system. Using the toolkit CRF++, the proposed approach produced the following results. A recall of 100% for Articles, News, Stories, Novel, and EBook. The Precision of 98.9% for Articles, 98.1% for News, 99.6% for Stories, 98.6% for Novel, and 98.9% for EBook.
10. Tamil: Tamil is spoken by people of the state Tamil Nadu, India. The Tamil language belongs to the Dravidian language. The Tamil language is also spoken in its nearby state of Tamil Nadu such as Karnataka, Kerala, Telangana, Andhra Pradesh and also at Andaman and Nicobar Islands. This language is also spoken by some people of Sri Lanka and Singapore (Wikipedia contributors, 2020j). The paper (Pandian & Geetha, 2009), discussed POS tagging and chunking for Tamil language using the CRF model. For this work, semi-automatic POS tagged corpus consisting of 36,000 sentences was used. The corpus was also manually verified. The corpus was used to train the system, for both chunking and POS tagging. Using the corpus, comparison was made between the baseline CRF and the

modified CRF model. For comparing the performance, three sets of testing data comprising of 18,345, 19,834 and 18,907 words and 6342, 6834 and 6521 chunks were considered. The F-scores for baseline CRF were 0.84, 0.86, and 0.85 respectively for the testing size of 18345, 19834 and 18907 words. The F-scores for the modified CRF model were 0.88, 0.89, and 0.89 respectively for the testing data comprising of 18345, 19834 and 18907 words. F-scores for Chunk tag by baseline CRF model were 0.79, 0.77, and 0.78 respectively for the testing data size of 6342, 6834 and 6521 chunks. F-scores for Chunk tag by modified CRF model were 0.83, 0.84, and 0.85 respectively for the testing size of 6342, 6834 and 6521 chunks.

11. Hindi: Hindi is an official language and spoken by many people in India, it is an Indo-Aryan language. This language is also spoken in other countries such as Singapore, South Africa, and Nepal. Hindi is written using the Devanagari script. The government of India uses the Hindi script as an official language apart from English (Wikipedia contributors, 2020k). The paper (Agarwal & Mani, 2006) presented Hindi language using the CRF approach for part-of-speech tagger and chunker. CRF++ tool had been used in this work that provide more information for training the POS tags. In this works 21,000 words were used for training the system. For Evaluation conll evaluation script had been taken. The accuracy achieved were 82.67% for POS tagging and 90.89% for chunking. The paper (Ojha et al., 2015) had also presented POS tagging methodology using SVM and CRF++ models for Hindi, Odia, and Bhojpuri languages. BIS tags were used for annotating the corpus. The systems were trained with 90k tokens and tested with 2k tokens. The accuracies achieved were as follows. In case of CRF based methodology, the result was in the range of 82 to 86.7%, whereas in case of SVM, accuracy obtained was in the span of 88 to 93.7%.

    In paper (Pvs & Karthik, 2007) POS tagging and Chunking discussion was presented for Hindi, Telugu, and Bengali languages. Training data for the work were of the following sizes: 21470 for Hindi, 21425 for Telugu, and 20397 for Bengali. Testing data consisted of 4924 Hindi words, 5225 Bengali words, and 5193 Telugu words. 78.66%, 76.08%, and 77.37% was obtained for Hindi, Bengali, and Telugu languages respectively. In case of chunking, the accuracy was observed 80.97% for Hindi, 82.74% for Bengali, and 79.15% for Telugu.

12. Urdu: Urdu is spoken mostly by the people of Pakistan. The Urdu language belongs to the Indo-Aryan language. Urdu is also recognized as one of the official languages in India. Many parts of India that speak Urdu are Uttar Pradesh, West Bengal, Telangana, and Jharkhand (Wikipedia contributors, 2020l). The paper (Khan et al., 2019) discussed POS tagging for Urdu language using the CRF approach. For this work, two types of data-sets were used the Bushra Jawaid (BJ) dataset (Jawaid et al., 2014) and Center for Language Engineering (CLE) dataset (CLE, 2020). It was reported that using the CRF model it improved the F-measure to 86.99% for the CLE dataset and 93.56% for the BJ dataset.

13. Khasi: Khasi is a Mon-Khmer language of the Austro-Asiatic language family, spoken by the native people of the state Meghalaya in the North-Eastern part of India. The Khasi language is also spoken by many people in the border area of Assam and Meghalaya. Also, it is spoken in the border area of India and Bangladesh (Wikipedia contributors, 2020m). In the fields of NLP there are some of the existing work on Khasi language. Some of the PoS tagging work on Khasi are discussed briefly:

In the paper (Tham, 2018), Khasi POS tagging based on HMM tagger has been discussed. The designed corpus consists of 86,087 tokens. In this same paper, the NLTK tool tagger had also discussed that has been applied to the same corpus (Tham, 2018). Accuracy's of 86.76% for BaselineTagger, 88.23% for NLTK Bigram Tagger, 88.64% for NLTK Trigram tagger, 89.7% for NLTK Tagger, and 95.68% for HMM POS Tagger.

In the paper (Warjri et al., 2019), the Khasi POS tagger had been developed based on the HMM method. In this work, the Khasi tagset was designed using 54 tags. The corpus was manually tagged using the tagset. The Khasi lexicon consists of around 7,500 tokens for training and for testing data of 312 words were used. Using the manually tagged Khasi lexicon on the proposed HMM-based POS Tagger, 76.70% of testing accuracy was achieved. In paper (Warjri et al., 2011), a CRF POS tagger is briefly discussed for Khasi. The experiment are done on 41000 tokens with 53 tags. The result achieved are 0.922 for Precision, 0.922 for Recall, & 0.921 for F-measure.

A summary of all the existing CRF based related works on PoS Tagging for Indian languages is presented in Table 2 .

3 Conditional random field (CRF) approach

In this paper, the POS tagger for Khasi language based on Conditional Random Field (CRF) approach is discussed. In this section, a brief discussion is presented on Conditional Random Field (CRF).

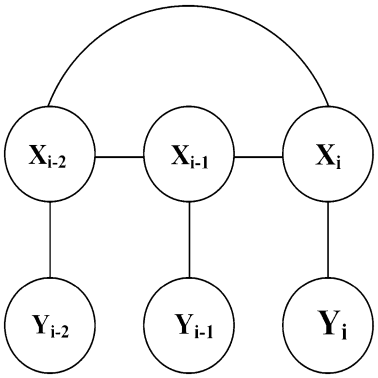

Fig. 1 Graphical structure of CRF model

Conditional random field (CRF) is a technique used in data sequencing problem. CRF is a probabilistic technique that is used for labeling and sampling (Lafferty et al., 2001). CRF is basically an undirected graphical model with X & Y nodes. Graphical structure of CRF is represented in Fig. 1, which have nodes of chain structure looks. In Fig. 1, X is taken as sequence of words, $X(X_1X_2 \ldots X_n)$ and Y represent the tag or the label of a particular word, $Y(Y_1Y_2 \ldots Y_n)$.

The main aim is to find the probable tag or label "Y" for the word sequence "X", based on the maximum conditional probability $P(Y|X)$.

The conditional probability chain structure of the CRF is represented as follows:

$$P(Y|X) = \frac{1}{Z_x} exp[\sum i \sum j \lambda j f_j(y_{i-1}, y_i, X, i)] \quad (1)$$

where λ is the weight concerning the distinct features f within the training stage, which is defined by the user.

Estimation of weight is done by maximum likelihood, where $f_j(y_{i-1}, y_i, X, i)$ denotes function for feature and Z_x is used for normalization.

Z_x can be expressed as follows:

$$Z_x = \sum y \epsilon Y exp[\sum i \sum j \lambda j f_j(y_{i-1}, y_i, X, i)] \quad (2)$$

The normalization Z_x probability of all state sequences are summed in such a manner that it becomes 1 (one). CRF is a model that depends on the given set of features and is encoded to the conditional probability. For determining the most probable sequence label from the given data one can use the equation below:

$$Y^* = argymaxP(Y|X) \quad (3)$$

3.1 Necessities of CRF

There are several existing models for sequencing problem or pattern recognition. The most promising models amongst them are regular expressions and graph-based models. The Conditional Random Fields (CRF) model is a popular and better prospective approach used for recognition of entity or sequencing problems. It uses both regular expression and graph for the purpose. The advantage of using CRF in POS Tagging is, it uses a property of an undirected graph-based model that can examine the words which are before the entity and also after the entity. To incorporate local features in a log-linear model, is one of the best characteristic of the CRF model (Pallavi & Pillai, 2016; Khan et al., 2019; Pandian & Geetha, 2009; Zhang et al., 2008).

4 Khasi POS tagging methodology using CRF approach

In this section, we present the different steps which have been followed for Khasi Part-of-Speech (POS) Tagging (KPOST). Figure 2 below represents the KPOST architecture using the CRF model. The subsections below represent brief discussions on the model architecture.

4.1 Tagset

Tagset used in this research work is unique in nature. There are significant differences in the tagsets in comparison with the tagset proposed in paper (Tham, 2018) for Khasi POS tagging. The tagset used here consists of 54 tags, out of which we have: Noun, Adjectives, Verbs, Pronouns,

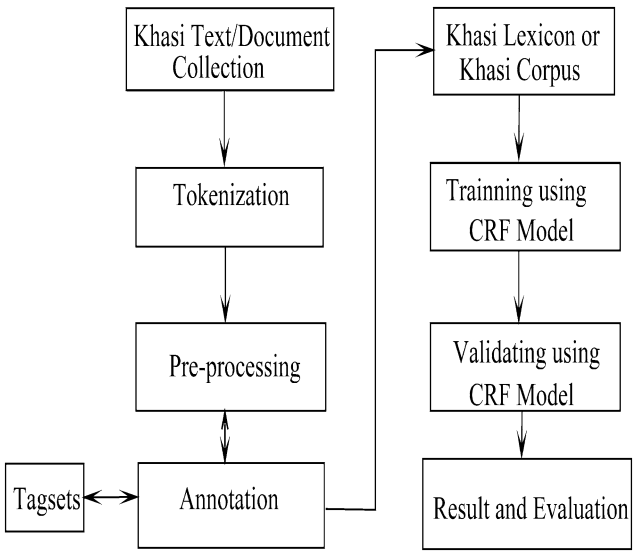

Fig. 2 Graphical structure of CRF model

Adverbs, and Prepositions as primary classes of part-of-speech. For secondary classes of part-of-speech, we have Tense, Aspects, Modality, Person, Number and Gender markers, Conjunctions, Quantifiers, Number, Copula, Passive Voice, and Emphatic. Therefore, we have altogether 54 tags which is used for annotating the Khasi data. For more details on the designed Khasi tagset that is used in this research work can be found in the paper (Warjri et al., 2018).

4.2 Pre-processing

Pre-processing or preliminary processing helps to clean the collected data or text for more clarity. Pre-processing includes removing or separating punctuation marks and symbols from the context. In Khasi language context, orthography and ambiguity are the main problems which needs to be addressed. Therefore in this work, we have to dealt with most of these problems by manually correcting them. Some of the Khasi words are morphologically analyzed manually for more clarity. Several words are shortened in many Khasi script or context. As an example, the word like *"ban"* is shortened to *"ba yn"* or *"ba n"*. In the same way, we have also splitted some other Khasi words like *ngan, ngin, kan, kin,* into *nga n, ngi n, ka n, ki n*. Some words that are orthographically splitted or combined, are also processed accordingly using python script. In our work, most of the word processing are done manually due to the ambiguity problem.

4.3 Annotation

Annotation is a challenging task. Annotation in NLP for POS corpus means assigning a tag to each word in a context. These tags symbolized the grammatical property of each word. Annotation process of Khasi text is found to be very complex. We have manually annotated the tag for each collected Khasi raw data. For annotation, one need to tokenize the collected text. For tokenization, we have used a Python program to split each word column-wise and then we have performed some pre-processing task manually. Then each word is annotated. After manually annotating the Khasi text, the corpus is verified by the linguistic expert from the Department of Linguistics, North-Eastern Hill University, Shillong. For any incorrectly tagged words, the correction of the tags is done manually.

4.4 Khasi Corpus

Khasi Corpus is the output after annotation of the raw Khasi text using a tagset. Therefore in this work, we have collected raw data from Mawphor (2017). These raw data are then manually annotated using tagset (Warjri et al., 2018), as discussed in the sub-section 4.3. Therefore, we have formed a Khasi lexicon or Khasi corpus for POS tagging purposes, which consists of around 71,000 words along with their tag properties. This annotated Khasi POS corpus is used for training and validating the CRF model. Some of our designed Khasi POS corpus can be found online in Warjri (2020).

4.5 Testing and validating data

In this work, we have used the designed Khasi corpus for training and validating the CRF model. For training, 80% of data is used from the corpus and 20% of it is used for validating purposes.

4.5.1 Feature function for CRF model

For identification of Khasi POS tags, feeding features to the CRF model is an important task. The main component of the CRF model is word features. These features are used to extract linguistic hidden information based on the contextual words of a particular language. Therefore some of the functions are created to extract the features. In this work, the features that are used, are capitalized words, the first and last word of the sentence, whole words that are capitalized, the first three or four prefix of the word, words that have symbols such as hyphen, numbers, and the word that has numbers and alphabets together.

5 Experimental results

In this subsection, a brief discussion is presented based on the state-of-art results and the experimental results obtained in this work. The designed Khasi POS corpus is experimented using the CRF model. To perform experiment using this model, training and validating data is divided with the ratio of 80:20 i.e. 80% of data are used for training purpose and rest are used for the purpose of validation. Table 3 represents a comparison between the results obtained for some

Table 3 Validating result achieved using state-of-art and proposed CRF method for the Khasi Lexicon

Sl. no.	Khasi Corpus	Technique	Accuracy
1.	40,800 tokens	NLTK Bi-gram	72.08%
	40,800 tokens	NLTK Tri-gram	75.15%
	40,800 tokens	combining (Bigram + Trigram)	79.35%
	40,800 tokens	CRF (Proposed work)	90.10%
2.	71,000 tokens	NLTK Bi-gram	84.11%
	71,000 tokens	NLTK Tri-gram	86.07%
	71,000 tokens	combining (Bigram + Trigram)	88.16%
	71,000 tokens	CRF (Proposed work)	92.12%

state-of-art technique and our proposed technique, using our designed Khasi POS corpus. In the same maaner,for state-of-art technique also, the training and validating data is divided with the ratio of 80:20. Therefore, the system yields an accurate validating result. From the result, we can observe that proposed CRF technique out performs the state-of-art method.

Figure 3, shows the nemenyi test which is a statistical posthoc test for comparison of the result achieved on 71,000 tokens over the four techniques as shown in Table 3. In the figure, the techniques are identified as **1** for Bigram, **2** for Trigram, **3** for Bigram+Trigram, and **4** for the CRF.

Table 4, presents the comparison result of the proposed CRF model for POS tagging in Khasi and with the other Indian languages that have used the CRF model for POS tagging.

In the comparison Table 4, we can see that the achieved result and the dataset that is used during the experiment of

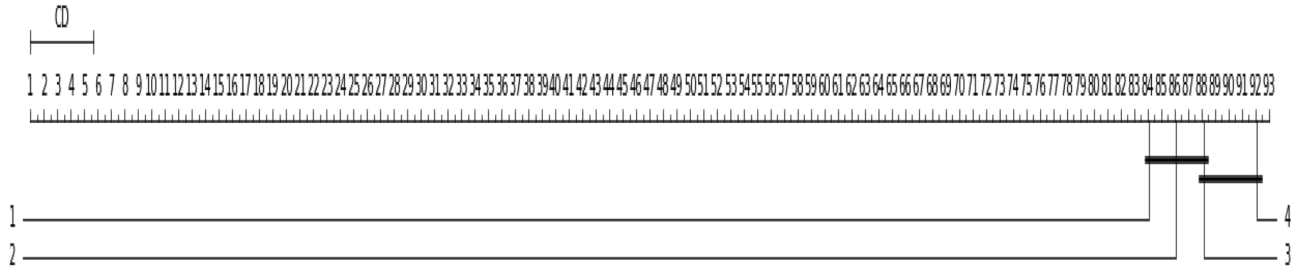

Fig. 3 Graphical structure using Nemeniyi Test on the results validity

Table 4 Comparison with different existing CRF POS tagger for others Indian languages

Sl. no.	Language	Technique/s	Dataset	Results & Accuracy
1	Hindi (Agarwal & Man,i 2006)	CRF++	21,000 words	82.67%
2	Bengali (Ekbal et al., 2007)	CRF	26 tags Train data-72,341 words Test data-20k words	90.30%
3	Hindi (Pvs & Karthik, 2007)	CRF	Train data-21470 words Test data-2924 words	78.66%
4	Gujarati (Patel & Gali ,2008)	CRF	26 tags, 600 sentences	92%
5	Manipuri (Singh & Ekbal, 2008)	CRF	26 tags 63,200 tokens	72.04%
6	Tamil (Pandian & Geetha, 2009)	CRF	36,000 sentences	F-score of 0.88 (for 18345 words), 0.89 (for 19834 words), & 0.89 (for 18907 words)
7	Kannada (BR & Ramakanth Kumar, 2012)	CRF	Train data-51269 words Test data-2932 words	84.54%
8	Assamese (Barman et al., 2013)	CRF	-	67.73%
9	Kashmiri (Ahmad & Syam, 2014)	CRF	30,000 words	81.10%
10	Malayalam (Krishnapriya et al., 2014)	CRF	100 tags 36,315 words	85.7%
11	Hindi (Ojha et al., 2015)	CRF++	90k tokens	82 to 86.7%
12	Kannada (Pallavi & Pilla,i 2016)	CRF	36 tags 80,000 words	92.94%
13	Punjabi (Sharma, 2016)	CRF	36 tags 38k to 42k words	Precision of 98.9%, 98.1%,99.6%, 98.6%,98.9% for Articles,News, stories,Novel, & EBook respectively & Recall score of 100% for all data
14	Kannada (Suraksha et al., 2017)	CRF	19 tags, 3000 sentences	96.86%
15	Odia (Behera, 2017)	CRF	600k tokens	94.11%
16	Urdu (Khan et al., 2019)	CRF	BJ dataset CLE dataset	F-measure of 86.99% for CLE dataset 93.56% for BJ dataset
17	Khasi (Warjri et al., 2011)	CRF	53 tags & 41000 tokens	0.922(Precision), 0.922(Recall), & 0.921(F-measure)
18	Khasi (Proposed work in this paper)	CRF	54 tags & 71000 tokens	92.12% (Accuracy) 0.92(Precision), 0.92(Recall), & 0.91(F-measure)

the CRF model by different languages. From the table, we can see that there is some language that has achieved good accuracy. But in comparison to the Khasi language, this is the extension work of the paper (Warjri et al., 2011) to investigate POS tagging using the CRF model with the designed Khasi POS corpus.

The Table 5, shows the average Precision, Recall, and F1-score yielded by the system for our proposed work. Figure 4, represents the graphical outline of the F1-score for each POS tag when the Khasi POS corpus consisting of 71,000 tokens is feet to the CRF model.

For calculating the Precision, Recall and F1-score the following formula is applied.

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Table 5 Average precision, recall and F1-score of proposed CRF for Khasi Lexicon

Sl.no.	Khasi Corpus	Precision	Recall	F1-score
1.	71,000 tokens	0.92	0.92	0.91

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - score = \frac{2 * (precision * recall)}{precision + recall} \tag{6}$$

TP-True Positives, FP-False Positives, and FN-False Negatives.

Similarly, Table 6 represents the result of the Precision, Recall and F1-score for each POS tag.

6 Performance analysis

We have also used some of the testing data to cross-validate the system. From the testing result it can be observed that the system has automatically tagged most of the Khasi words correctly. But, apart from the correctly tagged Khasi words, we have analysed some errors. From the result, it is found that some words are tagged incorrectly. This is due to the ambiguity problem. When we have trained the system with around 33000 tokens the system have more tagging error as shown in Table 7. In Table 7, the column *Khasi words* represents the words which are not tagged.

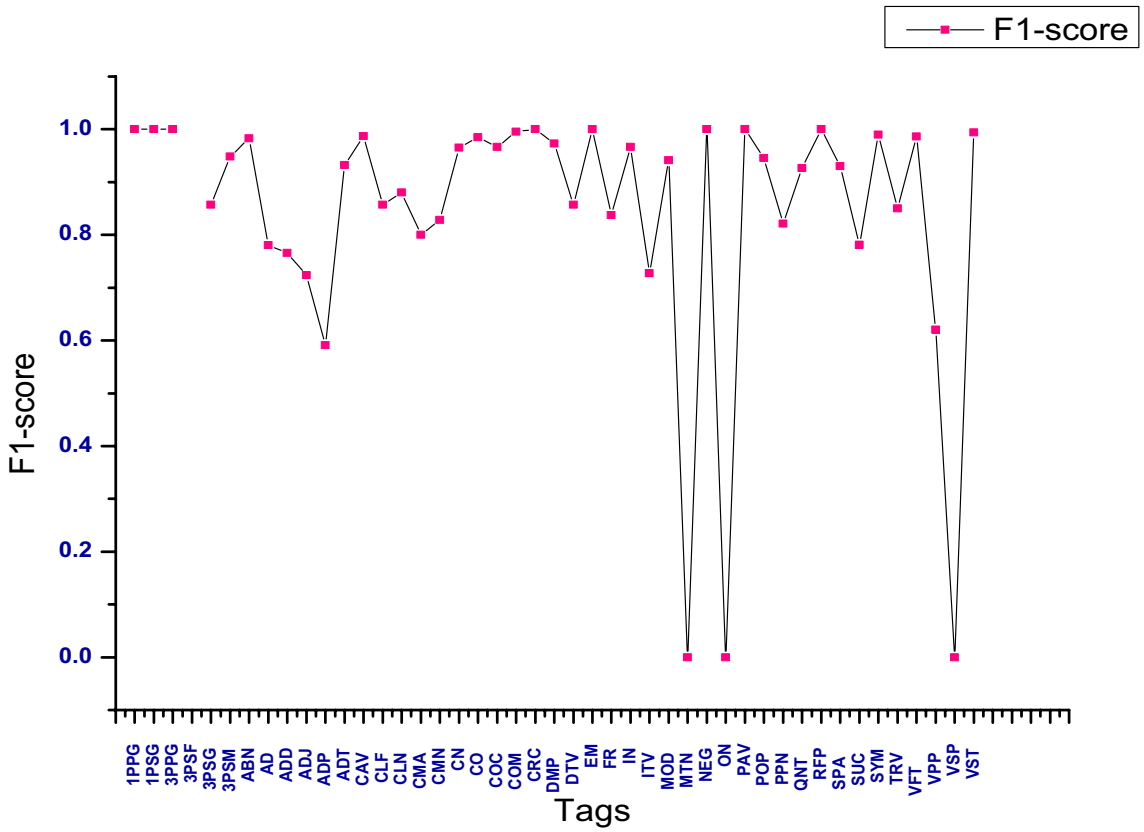

Fig. 4 Graphical representation of F1-score on individual POS tags

Table 6 The average Precision, Recall, & F1-score of each POS tags

Sl. no.	Tags	Precision	Recall	F-score
1	ABN	0.990	0.977	0.984
2	ITV	0.813	0.590	0.684
3	ADT	0.816	0.764	0.789
4	TRV	0.780	0.781	0.780
5	3PSF	0.988	0.983	0.985
6	AD	0.739	0.718	0.728
7	IN	0.935	0.961	0.948
8	COM	0.991	0.995	0.993
9	CMN	0.853	0.880	0.866
10	3PPG	0.978	0.994	0.986
11	POP	0.955	0.959	0.957
12	VFT	0.989	0.948	0.968
13	ADD	0.699	0.782	0.738
14	SPA	0.604	0.941	0.736
15	CLN	0.932	0.911	0.921
16	FR	0.906	0.913	0.910
17	DMP	0.796	0.950	0.866
18	CO	0.907	0.988	0.946
19	QNT	0.642	0.779	0.704
20	SUC	0.812	0.657	0.726
21	CAV	0.976	0.970	0.973
22	VST	0.945	0.973	0.959
23	SYM	0.988	0.996	0.992
24	ADJ	0.750	0.661	0.702
25	PPN	0.884	0.892	0.888
26	COC	0.939	0.958	0.948
27	3PSM	0.983	0.971	0.977
28	EM	0.991	0.966	0.979
29	CN	0.958	0.923	0.940
30	1PPG	0.976	0.976	0.976
31	ADP	0.564	0.393	0.463
32	DTV	0.811	0.682	0.741
33	NEG	0.954	0.945	0.949
34	1PSG	0.864	1.000	0.927
35	PAV	0.750	0.600	0.667
36	3PSG	0.875	1.000	0.933
37	CLF	0.963	0.963	0.963
38	RFP	0.667	0.571	0.615
39	MOD	0.958	0.821	0.885
40	CRC	1.000	0.851	0.920
41	VPP	0.647	0.957	0.772
42	INP	0.600	0.107	0.182
43	XX	1.000	0.250	0.400
44	ON	1.000	0.750	0.857
45	CMA	0.600	0.375	0.462
46	MTN	0.667	0.143	0.235
47	IM	0.000	0.000	0.000
48	IDP	0.000	0.000	0.000
49	2PG	0.333	1.000	0.500
50	VSP	0.000	0.000	0.000
51	ADF	0.000	0.000	0.000
52	RLP	0.000	0.000	0.000
53	ADM	0.000	0.000	0.000
54	VPT	0.000	0.000	0.000

Table 7 System performance based on training of 33,000 tokens

Khasi words	Predicted tag	Correct tag
Mynta	CMN	ADT
yn	VFT	VFT
kynduh	CMN	TRV
ka	3PSF	3PSF
NPP	FR	FR
UDP	FR	FR
halor	IN	AD
ka	3PSF	3PSF
shuki	3PSF	MTN
CEM	FR	FR
Shillong	PPN	PPN
Lber	PPN	PPN
03	CN	CN

Table 8 System performance based on training of 58,000 tokens

Khasi word	Predicted tag	Correct tag
Mynta	ADT	ADT
yn	VFT	VFT
kynduh	TRV	TRV
ka	3PSF	3PSF
NPP	FR	FR
UDP	FR	FR
halor	AD	AD
ka	3PSF	3PSF
shuki	CMN	MTN
CEM	FR	FR
Shillong	PPN	PPN
Lber	PPN	PPN
03	CN	CN

Predicted tag are the tags produced for each word by the CRF system and *Correct tag* are the tag which is supposed to be tagged for a particular word. It is found that out of 13, only four predicted tags are incorrect.

But as the Khasi corpus data is increased, the CRF system performs more accurately. Table 8 shows the result when the corpus size consists of around 58,000 tokens. It

can be easily found that except one words, all other words are tagged correctly.

We have also experimented with the system, by feeding the whole designed Khasi corpus (i.e. 71,000 tokens). By training the 71K tokens, we have tested more raw Khasi text or untagged Khasi words. The CRF system mostly yields the correct tag for the given Khasi words. But for few words, the system have also predicted wrong tags due to ambiguity and unknown words problems.

The discussion above presents the reactivity of the research as a threat to validity. With the participation of the raw testing data in the experiment, we can see the behaviors of the system and its result. Giving the untagged data to the CRF system when trained with 33000 tokens, as a result, the reaction of the system is that it could produce the tag to a particular word. But for some words, the POS system has tagged wrongly by the system as shown in Table 7. Again, as the Khasi corpus size is increased to 58,000 tokens we can observe the reaction of the system, the result is more accurate and the system can also tag most of the given words correctly as shown in Table 8.

7 Conclusion and future works

In this paper, Part-of-speech (POS) tagging using Conditional Random Field (CRF) on Khasi language has been discussed. This work is certainly a resource task on Khasi language towards the Natural Language Processing aspect. In this research, the CRF model is fed with the designed Khasi POS corpora, which consists of around 71,000 tokens. The training and validating data are divided into a ratio of 80:20. The result is also compared with some state-of-art techniques. It is observed that, the proposed approach achieves superior accuracy than other state-of-art techniques. Though, some words were tagged wrongly by the system due to ambiguity problem, especially for unknown Khasi words. To solve the said problem and to make the POS tagging system more efficient, more tagged Khasi data are needed. Therefore in our future work, we would like to collect more Khasi data and tag each word appropriately. We will also experiment with the designed Khasi corpus on other models.

8 Discussion

The main objective of this paper is to use the Conditional Random Field (CRF) method for POS tagging on the designed Khasi POS corpus. A few set of tagged Khasi data are available at Warjri (2020). A brief comparison with some state-of-art methodologies for the same Khasi corpus is also introduced in this work. In this research work, the most challenging task is to annotate the Khasi text. We have created a corpus consisting of 71,000 tokens for this experiment. As Khasi is a very low resourced Indian language, there are very few works available in this domain. POS is a prepossessing technique for any work related to Natural Language Processing. As it is found that CRF based technique can produce promising accuracy, our proposed system can automatically tag the Khasi words more efficiently. But some words or unknown words are tagged wrongly by the system due to ambiguities, which is the main limitation of the work. Therefore, to solve this problem, the Khasi corpus size may be increased and can also be used for parsing purposes. This research work will have benefits in all NLP related works in this low resource Khasi language.

Acknowledgements Authors would like to thanks and acknowledge the Government of India, Ministry of Science & Technology, Department of Science & Technology (DST), KIRAN Division, Technology Bhavan, New Delhi for the financial assistance (Grant: DST/WOS-B/2018/1216/ETD/Sunita(G)) during the study.

References

Agarwal, H., & Mani, A. (2006). Part of speech tagging and chunking with conditional random fields. In the Proceedings of NWAI workshop.

Ahmad, A., & Syam, B. (2014). Kashmir part of speech tagger using CRF. *Computer Science*, *3*(3), 3.

Barman, A. K., Sarmah, J., & Sarma, S. K. (2013). POS Tagging of Assamese language and performance analysis of CRF++ and fnTBL approaches. In: 2013 UKSim 15th international conference on computer modelling and simulation IEEE, pp. 476–479. Retrieved from https://doi.org/10.1109/UKSim.2013.91.

Behera, P. (2017). An experiment with the CRF++ parts of speech (POS) tagger for Odia. *Language in India*, *17*(1), 18.

Brants, T. (2000). TnT: A statistical part-of-speech tagger. *Sixth Applied Natural Language Processing Conference (Association for Computational Linguistics, Seattle, Washington, USA)*, (pp. 224–231). https://doi.org/10.3115/974147.974178.

Br, S., & Ramakanth Kumar, P. (2012). Kannada part-of-speech tagging with probabilistic classifiers. *International Journal of Computer Applications*, *48*(17), 26.

CLE. (2020). Center for language engineering. Retrieved January 12, 2020, from https://www.cle.org.pk/

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. *Third Conference on Applied Natural Language Processing*, *6*, 133–140.

Ekbal, A., Haque, R., & Bandyopadhyay, S. (2007). Bengali part of speech tagging using conditional random field. In Proceedings of seventh international symposium on natural language processing (SNLP2007), (pp. 131–136).

Jawaid, B., Kamran, A., & Bojar, O. (2014). A tagged corpus and a tagger for Urdu. *LREC*, *2*, 2938–2943.

Khan, W., Daud, A., Nasir, J. A., Amjad, T., Arafat, S., Aljohani, N., et al. (2019). Urdu part of speech tagging using conditional random fields. *Language Resources and Evaluation*, *53*(3), 331.

Krishnapriya, V., Sreesha, P., Harithalakshmi, T., Archana, T., & Vettath, J. N. (2014). Design of a POS tagger using conditional random fields for Malayalam. In 2014 first international conference

on computational systems and communications (ICCSC) IEEE, (pp. 370–373). https://doi.org/10.1109/COMPSC.2014.7032680.
Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In proceedings of the eighteenth international conference on machine learning, (pp. 282–289)
Mawphor. (2017). Mawphor. Retrieved November 2017, June 2019, from https://www.mawphor.com/index.php/
Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, *20*(2), 155.
Ojha, A. K., Behera, P., Singh, S., & Jha, G. N. (2015). Training & evaluation of pos taggers in indo-aryan languages: A case of Hindi, Odia and Bhojpuri. In the proceedings of 7th language & technology conference: human language technologies as a challenge for computer science and linguistics, (pp. 524–529).
Pallavi, K., & Pillai, A. S. (2016). Kannpos-Kannada parts of speech tagger using conditional random fields. In: Emerging research in computing, information, communication and applications. Springer (pp. 479–491). Retrieved from https://doi.org/10.1007/978-81-322-2553-9_43.
Pandian, S. L., & Geetha, T. V. (2009). CRF models for Tamil part of speech tagging and chunking. In W. Li & D. Mollá-Aliod (Eds.), *Computer processing of oriental languages. Language technology for the knowledge-based economy* (pp. 11–22). Berlin: Springer.
Patel, C., & Gali, K. (2008). Part-of-speech tagging for Gujarati using conditional random fields. In: Proceedings of the IJCNLP-08 workshop on NLP for less privileged languages. Retrieved from https://www.aclweb.org/anthology/I08-3019.
Pvs, A., & Karthik, G. (2007). Part-of-speech tagging and chunking using conditional random fields and transformation based learning. *Shallow Parsing for South Asian Languages*, *21*, 21.
Sharma, S. K. (2016). Assigning the correct word class to Punjabi unknown words using CRF. *International Journal of Computer Applications*, *142*, 14. https://doi.org/10.5120/ijca2016909684.
Singh, T. D., & Ekbal, A. (2008). Manipuri POS tagging using CRF and SVM: A language independent approach. In proceeding of 6th international conference on natural language processing (ICON-2008), (pp. 240–245)
Suraksha, N., Reshma, K., & Kumar, K. S. (2017). Part-of-speech tagging and parsing of Kannada text using Conditional Random Fields (CRFs). In: 2017 international conference on intelligent computing and control (I2C2) IEEE, (pp. 1–5). Retrieved from https://doi.org/10.1109/I2C2.2017.8321833.
Tham, M. J. (2018). Challenges and Issues in Developing an Annotated Corpus and HMM POS Tagger for Khasi. In the 15th international conference on natural language processing, (pp. 10–19).
Warjri, S. (2020). Khasi corpus. Retrieved from https://github.com/sunitawarjri/Khasi-Corpus/blob/master/Khasi%20Corpus.txt.
Warjri, S., Pakray, P., Lyngdoh, S., & Maji, A. K. (2021). Adopting conditional random field (CRF) for Khasi part-of-speech tagging (KPOST). In proceedings of the international conference on computing and communication systems: I3CS 2020, NEHU, Shillong, India, vol. 170 (Springer Nature), vol. 170, p. 75.
Warjri, S., Pakray, P., Lyngdoh, S., & Kumar Maji, A. (2018). Khasi language as dominant part-of-speech (POS) ascendant in NLP. *International Journal of Computational Intelligence & IoT*, *1*(1), 109.
Warjri, S., Pakray, P., Lyngdoh, S., & Maji, A. K. (2019). Identification of POS Tag for Khasi Language based on Hidden Markov Model POS Tagger. *Computación y Sistemas*, *23*(3), 795. https://doi.org/10.13053/CyS-23-3-3248.
Wikipedia contributors. (2020a). Bengali language: Wikipedia, the free encyclopedia. Retrieved February 02, 2020, from https://en.wikipedia.org/w/index.php?title=Bengali-language&oldid=941772762.
Wikipedia contributors. (2020b). Assamese language: Wikipedia, the free encyclopedia. Retrieved February 02, 2020, from https://en.wikipedia.org/w/index.php?title=Assamese-language&oldid=939154061.
Wikipedia contributors. (2020c). Gujarati language: Wikipedia, the free encyclopedia. Retrieved February 03, 2020, from https://en.wikipedia.org/w/index.php?title=Gujarati-language&oldid=942374083
Wikipedia contributors. (2020d). Kannada: Wikipedia, the free encyclopedia. Retrieved February 05, 2020, from https://en.wikipedia.org/w/index.php?title=Kannada&oldid=942703407.
Wikipedia contributors. (2020e). Kashmiri language: Wikipedia, the free encyclopedia. Retrieved February 04, 2020, from https://en.wikipedia.org/w/index.php?title=Kashmiri-language&oldid=942627183.
Wikipedia contributors. (2020f). Malayalam: Wikipedia, the free encyclopedia. Retrieved February 03, 2020 from https://en.wikipedia.org/w/index.php?title=Malayalam&oldid=941882964
Wikipedia contributors. (2020g). Meitei language: Wikipedia, the free encyclopedia. Retrieved February 02, 2020, from https://en.wikipedia.org/w/index.php?title=Meitei-language&oldid=936096557
Wikipedia contributors. (2020h). Odia language: Wikipedia, the free encyclopedia. Retrieved February 03, 2020, from https://en.wikipedia.org/w/index.php?title=Odia-language&oldid=941768688
Wikipedia contributors. (2020i). Punjabi language: Wikipedia, the free encyclopedia. Retrieved February 02, 2020, from https://en.wikipedia.org/w/index.php?title=Punjabi-language&oldid=941520253
Wikipedia contributors. (2020j). Tamil language: Wikipedia, the free encyclopedia. Retrieved February 04, 2020, from https://en.wikipedia.org/w/index.php?title=Tamil_language&oldid=941234813
Wikipedia contributors. (2020k). Hindi: Wikipedia, the free encyclopedia. Retrieved February 03, 2020, from https://en.wikipedia.org/w/index.php?title=Hindi&oldid=942598408.
Wikipedia contributors. (2020l). Urdu: Wikipedia, the free encyclopedia. Retrieved February 02, 2020, from https://en.wikipedia.org/w/index.php?title=Urdu&oldid=942705946
Wikipedia contributors. (2020m). Khasi: Wikipedia, the free encyclopedia. Retrieved January 15, 2020, from https://en.wikipedia.org/w/index.php?title=Khasi_language&oldid=914412473
Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., & Wang, B. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169.