Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Part-of-speech Tagging for Nagamese Language using CRF

Alovi Shohe
Dept. of Information Technology
Nagaland University,
Kohima Campus Meriema,
Nagaland, India

Chonglio Khiamungam
Dept. of Information Technology
Nagaland University,
Kohima Campus Meriema,
Nagaland, India

Dr. Teisovi Angami
Dept. of Information Technology
Nagaland University,
Kohima Campus Meriema,
Nagaland, India

***Abstract* - This paper investigates part-of-speech tagging, an important task in Natural Language Processing (NLP) for the Nagamese language. The Nagamese language, a.k.a. Naga Pidgin, is an Assamese-lexified Creole language developed primarily as a means of communication in trade between the Nagas and people from Assam in northeast India. A substantial amount of work in part-of-speech-tagging has been done for resource-rich languages like English, Hindi, etc. However, no work has been done in the Nagamese language. To the best of our knowledge, this is the first attempt at part-of-speech tagging for the Nagamese Language. The aim of this work is to identify the part-of-speech for a given sentence in the Nagamese language. An annotated corpus of 16,112 tokens is created and applied machine learning technique known as Conditional Random Fields (CRF). Using CRF, an overall tagging accuracy of 85.70%; precision, recall of 86%, and f1-score of 85% is achieved.**

Keywords - Nagamese, NLP, part-of-speech, machine learning, CRF.

I. INTRODUCTION

Nagamese (Naga Pidgin) is an important Creole language of Nagaland located in the North-Eastern part of India. Apart from the tribal languages spoken, it is used as a common language across the entire state of Nagaland. Nagamese is an Assamese-lexified (Assam is an Indian state bordering Nagaland) creole language developed primarily as a means of communication in trade between the Nagas and people from Assam. It is widely used in mass media in the news, radio stations, state-government media, etc. The Nagamese language is a resource-poor language, and therefore, it is a challenge to create resources for applying various language processing tasks.

Part-of-speech (POS) tagging involves labeling each word in a sentence with its appropriate part of speech.

Example: Itu/ADJECTIVE dikhikena/VERB Isor/NOUN khusi/ADJECTIVE lagise/VERB ./SYM (God was pleased with what He saw.)

POS tagging is not an easy task due to the dependence of POS tags on contextual information.

In this work, a POS tagger is built for the Nagamese Language using Conditional Random Fields (CRF) which is a machine-learning technique. The main contribution of our work is the identification of the POS tagset, the creation of an annotated corpus of 16,115 tokens, and its evaluation using CRF. A discussion on the error analysis of the tagging performance is also presented.

The rest of the paper is organized as follows: Section II gives an introduction to the Nagamese Language, Section III gives an overview of the related works, Section IV gives a description of the POS tagset, Section V discusses the methodology, Section VI reports and discusses the experimental results and Section VII draws the conclusion and discusses future works.

II. THE NAGAMESE LANGUAGE

This section provides an overview of the character set, syllabic pattern, and grammar for the Nagamese Language.

A. Character Set for Nagamese Language

The Nagamese language has 28 phonemes, comprising of 6 vowels and 22 consonants.

Vowels: i, u, e, ə, o, a

Consonants:*p, t, c, k, b, d, j, g, p^h, t^h, c^h, k^h, m, n, ṅ, s, š, h, r, I, w, y*

A sentence in Nagamese is given below:

"Moy dos baje pora yeti ase."
(I am waiting here from 10 o 'clock.)

B. Syllabic Pattern

As found in the work of Sreedhar [1], a word in the Nagamese language may consist of one or more syllables ranging up to a maximum of four syllables. The entire mono-syllabic words in this language could be sub-grouped into six classes, which when put in a schematic formula would be:

$$(C)(C)V(C)(C)^2$$

The only limitation in the operation of the above formula is that V cannot occur alone. A disyllabic word in the Nagamese language cannot consist of just two vowels alone. The structure of the disyllabic words in this language can be broadly sub-grouped into two which is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

shown here:

1) V (C)(C)(C)V (C)
2) (C)CV (C)(C)CV (C)(C) or (C)CV (C)(C)V (C)(C)

The trisyllabic words in the Nagamese language could also be broadly sub-grouped into two sub-types. These are:

1) V (C)(C)CV (C)(C)CV (C)
2) (C)CV (C)(C)V (C)(C)(C)V (C)

There are few words in the Nagamese language that have tetra syllables. The syllabic structure of the tetrasyllabic words in this language could be schematically presented as follows:

(C)V (C)CV CV (C)CV (C)

There are no pentasyllabic words in the Nagamese language unless one takes clear compound words.

B. Grammar

The detailed grammar of Nagamese can be found in the works of Sreedhar [1], Baishya [2] and Bhattacharjya [3]. Some example sentences in Nagamese are given below:-

1) Sualitu gor bitorte ase (the girl is inside the house)
2) Moy dos baje pora yeti ase (I am waiting here from 10 o'clock)
3) Syama joldi kitab porise (Shyama read the book quickly)

III. RELATED WORKS

Since Nagamese is an Assamese-lexified creole language, some of the works done in the Assamese Language are presented here.

Saharia et al. [4] worked on developing a part of speech (POS) tagger for Assamese using the Hidden Markov Model (HMM), in which a tagset of 172 tags was developed and performed morphological analysis to determine the probable tags for the unknown words. On a manually tagged corpus of 10k words for training, an accuracy of 87% on the test data was achieved.

Pathak et al. [5] worked on a Deep Learning (DL)-based POS tagger for Assamese, using several pre-trained word embeddings, and was able to attain an F1 score of 86.52%.

Phukan et al. [6] applied character-level Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) to part-of-speech (POS) tagging for the Assamese language. The annotated dataset uses the LDCIL Assamese tagset and contains 60,000 words. The LSTM model achieved an accuracy of 92.80%, whereas the BLSTM model achieved an accuracy of 93.36%.

Pathak et al. [7] proposed a BiLSTM-CRF architecture for Assamese POS tagging using a corpus of 404k tokens. They used word embeddings for its implementation, the two top POS tagging models achieving F1 scores of 0.746 and 0.745. Also, a rule-based approach was developed achieving an F1 score of 0.85. Subsequently, the DL-based taggers were integrated with rule-based to achieve an F1 score of 0.925.

Deka et al. [8] compared tagging performances of Conditional Random Field and Trigrams'nTag for POS tagging for Assamese Language.

Phukan et al. [9] explored long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) for POS tagging for Assamese, on a corpus comprising 50k words, achieving an accuracy of 91.20% for LSTM and 91.72% for BiLSTM.

Talukdar et al. [10] used Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU) to develop a POS tagger for Assamese. The tagset used the Universal Parts of Speech (UPoS) tagset on a dataset of 30k words, achieving F1 scores of 94.01 and 94.56 for RNN and GRU.

IV. POS TAGSET

From the original 36 tags employed in the Penn Treebank Tagset, the tags for the Nagamese POS tagger were reduced to 14 Tags. To differentiate Foreign words from Nagamese words a 'FW' tag was introduced into the tag set to be used for the Nagamese tagger. The tagset is shown in Table I.

TABLE I. POS TAGSET

Sl. no.	category	tag
1	Adjective	ADJ
2	Adverb	ADV
3	Conjunction	CONJ
4	Complementizers	CMP
5	Determinant	DET
6	Postposition/Preposition	PP
7	Interjection	INTJ
8	Noun	N
9	Pronoun	PN
10	Quantifier	QN
11	Verb	V
12	Foreign Word	FW
13	Symbol	SYM
14	Unknown	UNK
15	Numeral	NUM

V. METHODOLOGY

This section provides details of the dataset creation and the Machine Learning Model, i.e, Conditional Random Fields (CRF), which has been used to build the POS tagger for Nagamese.

A. Dataset Creation

The Nagamese Corpus, which contains approximately 26,000 words, was created by collecting articles from a local newspaper, 'Nagamese Khobor' (www.nagamesekhobor.com). The 'Nagamese khobor' newspaper contains a variety of content related to current state affairs, sports, etc. Based on the word frequency, a word cloud is shown in Fig. 1.

Random publications of the Nagamese newspaper and bible phrases were collected, from which various articles were extracted to obtain a mixed corpus. The corpus was manually annotated by one annotator who is a native speaker of Nagamese. We manually annotated a corpus of 16,115 tokens,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

the tag frequencies shown in Table. II. To validate the dataset, another annotator was employed to manually tag 1,864 tokens. Out of these, 125 tokens were disagreed and 102 were Foreign words. Hence, including foreign words the disagreement is 6.7%, and excluding these foreign words, the disagreement is 1.23%.

Sample of the annotated dataset is shown below:-
Titia/ADV Isor/N koise/V ,/SYM "/SYM Ujala/N hobole/V dibi/V ./SYM "/SYM Aru/CONJ Ujala/N hoise/V ./SYM Itu/ADJ dikhikena/V Isor/N khusi/ADJ lagise/V ./SYM

TABLE II. TAG FREQUENCIES

Sl. no.	tag	frequency
1	ADJ	1507
2	ADV	709
3	CONJ	591
4	CMP	35
5	DET	132
6	PP	2418
7	INTJ	65
8	N	1804
9	PN	1141
10	QN	84
11	V	1678
12	FW	3744
13	SYM	1830
14	UNK	143
15	NUM	234

B. Conditional Random fields Model

Conditional random fields (CRFs) fall into the sequence modeling family and are a class of statistical modelling methods often applied in pattern recognition and machine learning and used for structured prediction. Whereas a discrete classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account; e.g., the neighboring tokens information in a sequence. CRFs are a type of discriminative undirected probabilistic graphical model. It overcomes the problem in Maximum Entropy Markov Models (MEMM) known as the "Label bias" problem. Fig. 2 shows the linear chain CRF model.

We have implemented CRF using the *sklearn-crfsuite* library.

The features used are:-
-the current word
-whether it is the first word in the sentence
-whether it is the last word in the sentence
-whether the word is capitalized
-whether the word is in lowercase
-prefix details up to length 3
-suffix details up to length 3
-previous word
-next word
-contains hyphen
-is numeric
-contains upper case inside word

A sample of the generated features is given below:-
'has hyphen': False, 'is first': True, 'suffix-4': 'itia', 'is numeric': False, 'prefix-3': 'Tit', 'word': 'Titia', 'suffix-1': 'a', 'is capitalized': True, 'next word': 'Isor', 'prefix-1': 'T', 'is all caps': False, 'prev word': '', 'is all lower': False, 'is last': False, 'suffix-2': 'ia', 'suffix-3': 'tia', 'capitals inside': False, 'prefix-2': 'Ti',....

For Gradient descent, the L-BFGS method has been used, 100 iterations have been used for the optimization algorithm, and to avoid the overfitting problem, L1 and L2 regularizers have been employed.

VI. RESULTS AND DISCUSSIONS

This section presents the results and discusses the results of the tagging. The annotated dataset comprises a total of 16,115 tokens (749 sentences). The training: test split used is 70:30%. Table III reports the results of the POS tagging. We obtain an overall tagging accuracy of 85.70%; precision, and recall of 86%, and f1-score of 85%. An error analysis for each tag in the POS tagset is reported in the form of a confusion matrix as shown in Fig. 3.

The accuracy measures used for the model's performance are precision, recall, and f1-score.

Precision: Precision measures how many of the positive predictions made by the model are actually correct, expressed mathematically as:
$$Precision = True\ Positive/(True\ Positive + False\ Positive) \quad (1)$$

Recall: Recall measures the completeness of positive predictions, expressed mathematically as:
$$Recall = True\ Positive/(True\ Positive + False\ Negative) \quad (2)$$

F1-Score: F1-Score is the harmonic mean between recall and precision and tells us how precise and robust our classifier is, expressed mathematically as:
$$F1\text{-}Score = 2 * (Precision * Recall)/(Precision + Recall) \quad (3)$$

From the experiment conducted as shown in Table III, we obtained a precision of 1.0 for CMP. The lowest precision for N with 0.70. The highest recall for SYM with 0.98 and the lowest recall for UNK with 0.21. The highest f1-score for 0.99 for SYM, and the lowest f1-score for UNK with 0.33.

The misclassified cases are listed below:-
i. ADJ tagged as ADV (10), CONJ (1), FW (11), N (12), PN (3), PP (15), UNK (2), and V (15).
ii. ADV tagged as ADJ (31), DET (2), FW (6), N (2), PN (4), PP (9), and V (5).
iii. CMP tagged as ADJ (1), FW (1), N(1), and CONJ (3).
iv. CONJ tagged as ADJ (7), DET (1), FW (2), N (4), PN (1), PP (5), and V (1).
v. DET tagged as ADJ (6), ADV (4), PN (1), and V (9).

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Figure 1. Nagamese wordcloud based on word frequency

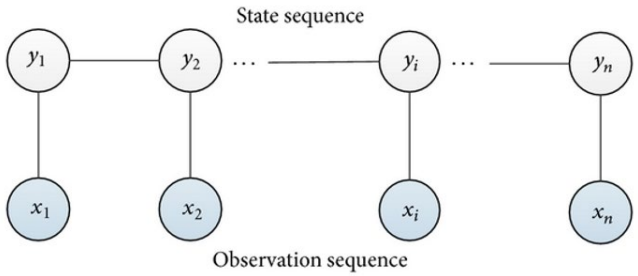

Figure 2. Linear-chain CRF model

TABLE III.POS TAGGING RESULTS

Tag	precision	recall	f1-score	support
ADJ	0.80	0.84	0.82	424
ADV	0.74	0.69	0.71	189
CMP	1.00	0.57	0.72	23
DET	0.89	0.61	0.72	51
FW	0.90	0.91	0.90	1317
INTJ	0.73	0.67	0.70	33
N	0.70	0.69	0.70	480
NUM	0.99	0.92	0.95	109
PN	0.89	0.90	0.90	321
PP	0.85	0.90	0.88	728
QN	0.80	0.70	0.74	23
SYM	0.99	0.98	0.99	524
UNK	0.77	0.21	0.33	82
V	0.77	0.88	0.82	407
avg / total	**0.86**	**0.86**	**0.85**	**4877**

vi. FW tagged as ADJ (6), ADV (4), CONJ (1), N (81), NUM (1), PN (1), PP (10), SYM (1), and V (13).
vii. INTJ tagged as ADV (3), FW (2), N (2), PP(1), and V (3).
viii. N tagged as ADJ (11), ADV (1), FW (76), PN (9), PP(26), UNK (1), and V (23).
ix. NUM tagged as FW (5), N (2), and SYM (2).
x. PN tagged as ADJ (1), ADV (2), FW (1), N (12), PP(10), QN (4), and UNK (1).
xi. PP tagged as ADJ (5), ADV (16), CONJ (1), FW (9), N (7), PN(14), and V (21).
xii. QN tagged as FW (1), N (4), PN (2), QN (16), and UNK (17).
xiii. SYM tagged as FW (8).
xiv. UNK tagged as ADJ (10), ADV (5), CONJ (1), FW (4), INTJ (5), N (4), PN (1), PP (16), and V (19).
xv. V tagged as ADJ (10), ADV (1), DET (1), FW (8), INTJ (3), N (9), PP (16), and UNK (1).
Note: The figure in bracket indicates the number of tokens.

Fig. 4 shows the top likely and unlikely transitions from one tag to the other tag. The top likely transition is from UNK to UNK and the top unlikely transition is from PP to NUM.

VII. CONCLUSION

In this paper, a part-of-speech tagger has been built for Nagamese Language using machine learning techniques by creating a POS-tagged Nagamese corpora of size 16,115 tokens. Using Conditional Random Fields (CRF), an overall tagging accuracy of 85.70%; precision, and recall of 86%, and an f1-score of 85% is obtained. The main contribution of this work is the creation of the POS annotated corpus and the evaluation of the annotated corpus. Some of the limitations of this work are the number of tags used and the size of the dataset.

Future works include:- i) Increasing the number of tags in the tagset, ii) Increasing the size of the tagged corpus, iii) Using the developed POS tagger to build other applications such as sentiment analysis, machine translation, etc for the Nagamese Language, and iv) exploring transfer learning from

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

the Assamese Language, etc.

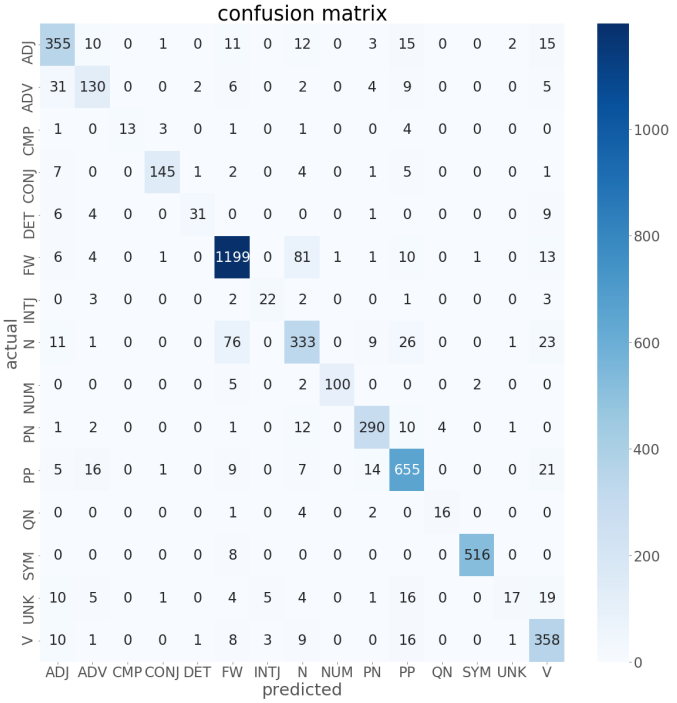

Figure 3. Confusion matrix for POS tagging

```
Top likely transitions:
UNK     -> UNK      3.402883
DET     -> ADV      2.928561
FW      -> FW       2.832922
N       -> PP       1.604715
PP      -> FW       1.582258
ADJ     -> V        1.567247
N       -> N        1.542058
P       -> ADV      1.470858
PN      -> FW       1.428755
N       -> V        1.220998
FW      -> SYM      1.166838
ADJ     -> UNK      1.152150
N       -> ADV      1.137048
CONJ    -> DET      1.118412
FW      -> V        1.096356
V       -> V        1.048017
PP      -> V        1.047655
ADJ     -> N        1.040519
FW      -> NUM      1.034534
FW      -> PP       0.998293

Top unlikely transitions:
PP      -> NUM      -0.880825
UNK     -> SYM      -0.889409
CONJ    -> PP       -0.932160
FW      -> INTJ     -0.951261
N       -> CMP      -0.997205
V       -> NUM      -1.026676
PN      -> DET      -1.063194
CONJ    -> SYM      -1.064076
NUM     -> PN       -1.069370
PN      -> SYM      -1.074111
ADJ     -> NUM      -1.171198
NUM     -> ADJ      -1.180422
SYM     -> QN       -1.183569
SYM     -> ADV      -1.239226
V       -> FW       -1.241763
SYM     -> SYM      -1.403862
PP      -> UNK      -1.414294
ADJ     -> CONJ     -1.459882
SYM     -> V        -1.650301
CONJ    -> QN       -3.594794
```

Figure 4. Top likely and unlikely transitions

REFERENCES

[1] Sreedhar, M. V. ”Standardized grammar of naga pidgin”. CIIL, Mysore, 1985.
[2] Baishya, A. K. ”The structure of Nagamese the contact language of Nagaland”. PhD diss., Assam University Silchar, 2003.
[3] Bhattacharjya, D. ”The genesis and development of Nagamese: Its social history and linguistic structure”. City University of New York, 2001.
[4] Saharia, N., Das, D., Sharma, U., and Kalita, J. ”Part of speech tagger for assamese text”. Proceedings of the ACL-IJCNLP 2009 conference short papers, pp. 33–36, 2009.
[5] Pathak, D., Nandi, S., and Sarmah, P. ”Aspos: Assamese part of speech tagger using deep learning approach”. 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1–8, 2022.
[6] Phukan, R., Baruah, N., Sarma, S. K., and Konwar, D. ”Parts-of-speech tagger in assamese using lstm and bi-lstm”. International Conference on Advances in Data-driven Computing and Intelligent Systems, Springer, pp. 19–31, 2023.
[7] Pathak, D., Nandi, S., and Sarmah, P. ”Part-of-speech tagger for Assamese using ensembling approach”. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 22, No. 10, pp. 1–22, 2023.
[8] Deka, R. R., Kalita, S., Kashyap, K., Bhuyan, M. P., and Sarma, S. K. ”A study of t’nt and crf based approach for pos tagging in Assamese language”. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), IEEE, pp. 600–604, 2020.
[9] Phukan, R., Baruah, N., Sarma, S. K., and Konwar, D. ”Exploring character-level deep learning models for pos tagging in assamese language”. Procedia Computer Science, Vol. 235, pp. 1467–1476, 2024.
[10] Talukdar, K. and Sarma, S. K. ”Deep learning based part-of-speech tagging for assamese using rnn and gru”. Procedia Computer Science, Vol. 235, pp. 1707–1712, 2024