ISSN: 2096-3246
Volume 54, Issue 02, November, 2022

PARTS OF SPEECH TAGGING OF THE NYISHI LANGUAGE USING HMM

Joyir Siram [1], Koj Sambyo [2], Achyuth Sarkar [3]
Department of CSE, National Institute of Technology, Yupia, Arunachal Pradesh, India
Joyir1020@gmail.com,sambyo.koj@gmail.com,achyuthit@gmail.com

Abstract: A natural language is one that humans speak, write, or sign for everyday communication, in contrast to formal languages. Natural language processing refers to the computational processes needed to allow a computer to process information using natural language. Nyishi part-ofspeech tagging is more challenging to answer than the English equivalent because it must be combined with the word identification problem. A POS Tagger assigns the appropriate tag, such as a noun, adjective, verb, or adverb, to each word of the input sentence. We incorporate Penn Treebank's tag set concept of word tagging format. A POS tagger's Tag set and Disambiguation Rules are essential components. The lack of a corpus for computational processing makes POS tagging for the Nyishi language challenging. Here, we discuss our work on first-order, fully linked hidden Markov models-based Nyishi part-of-speech tagging. For training and testing, a corpus of about 30,000 Nyishi characters is used. A Viterbi-based word identification algorithm divides an article into clauses and subsequently into words. The following experimental findings are presented for various testing scenarios 89% of the words in the testing data can be accurately tagged by the system.

Keywords: part of speech (POS), POS tagging, Nyishi language, NLP, HMM, tag set.

1. Introduction:

Natural language processing (NLP) [1] combines phi logy and computer technology to study the rules, the arrangements of the language and create a well-informed system capable of apprehension, analysis and elicitation of the meaning from text and speech. NLP has lengthened so many languages with its applications and many sorts of researches have passed in various unlike fields of NLP. POS tagging is the preliminary stage to figure out the Natural language processing as it labels the words with their appropriate Part of Speech grammatically. POS tagging [2] process consists of several phases for the system to describe the specified tag for each word. The task of producing a POS tagger includes many phases, for instance - building a tag set, creating a dictionary, aggregating a corpus, viewing the principles of the context and also to ascertain the inflections and subject anomalies of the given language. Applications for natural language processing (NLP)[1], such as speech recognition, text-to-speech, information retrieval, and machine translation systems, can considerably profit from part-of-speech tagged corpora. In-depth research and testing have been done on automatic part-of-speech tagging for European languages. Many language processing tasks for Western and South Asian languages have been carried out throughout the years. However, the Nyishi language receives relatively little attention. However, due to a variety of factors, automatic Nyishi part-of-speech tagging technology is still in its infancy:

i. The Nyishi language has many words that were not previously acquainted to them or utilized in their context. As a result, the dictionary now uses the sentence that best describes many of these new words.

ii. Fully automatic word segmentation is not possible.

iii. There is no well-defined tag set for the Nyishi part of speech. iv. It is challenging to find Nyishi vocabularies that include all parts of speech.

Nyishi part-of-speech tagging [3] is especially challenging due to these interrelated issues. There is a lot of interest in automating the process because manually tagging words with part-of-speech identifiers is expensive, timeconsuming, and difficult. Word ambiguity and words outside of one's vocabulary are the key issues with designing an accurate automatic part-of-speech tagging. Word ambiguity describes how different words behave in various contexts. But there hasn't been any extensive research on employing a Hidden Markov Model to construct POS tagger Nyishi language (HMM) it is first of its kind. The established probabilistic technique for automatic POS tagger is the Hidden Markov Model. The HMM method has been used by several languages to create an automatic POS tagger. It has been demonstrated that a POS tagger using the HMM method has a faster running time than any other probabilistic method. In this paper, we describe our efforts to create a part-of-speech tagger for the Nyishi language based on HMM.

2. Previous Works:

In many NLP tasks, such as question-answering, parsing, and machine translation, Part of Speech tagging is crucial. It is described as the process of giving each word in a phrase a label that indicates where that word falls within the predetermined syntax system for that particular language.

A wide variety of taggers have been created over the past 20 years, particularly for European languages with large corpora like English, Czech, and Turkish. Early PoS taggers for English include the Brill, TnT, and Claw. We discovered different methods being employed in various situations and languages. Information theory methods are frequently used to train, classify, and assign PoS tags. Examples include maximum entropy models, Hidden Markov Models, support vector machines, conditional random fields, artificial neural networks and decision trees.

Saharia et al., [4] give an overview of the research on Assamese POS tagging using the well-known Hidden Markov Model in this publication. We develop a tag set with 172 tags in cooperation with linguists in the absence of a well-defined acceptable tag set. For effective labeling, the authors examine relevant Assamese language issues. To find potential tags for unidentified words, the researchers employ conventional morphological analysis. With a manually labeled training corpus of over 10,000 words, the authors achieve a tagging accuracy of about 87 percent for test inputs.

The author has to include a tag set with 61 tags for the Khasi language in []. For the Khasi language, where the researcher employed a lexicon of 8,000 words, the author of the study included a morphological analyzer. The word class-based word analysis system was based on the subject-verb-object grammatical relationships. The prefixes, infixes, and suffixes of each word were taken into consideration when determining the morph tactic rules.

Joshi et al., [5] they used a Hidden Markov Model (HMM) for the Hindi language that was POS-tagged. A new tag set was employed for the system as Indian language (IL) POS. And using the contextual information in the text, they were able to distinguish between the right word-tag pairings and achieve an accuracy of 92.13% on test data.

Patra et al., [6] the development of well tagged corpora is necessary to apply machine learning to less computerized languages. Researchers developed POS taggers for Korborok, a less-privileged language, in this study using Conditional Random Field and Support Vector Machine. We manually annotated 42,537 tokens from written texts with a POS tag set that includes 26 tags for Indian languages. The POS taggers use a variety of contextual and orthographic word-level factors. These qualities transcend linguistic boundaries and apply to many languages. POS taggers have utilized 39449 and 8672 tokens, respectively, for training and testing. According to evaluation results, the CRF and SVM's respective accuracy rates were 72.04 percent and 74.38 percent.

Singh et al., [7] used a morphological analyzer to help with rule-based POS tagging, and we employed Conditional Random Field (CRF) and Support Vector Machines as two machine learning classifiers for supervised methods (SVM). The total number of POS-tagged words was 42,537. The accuracy rates for manual verification are 70 and 84 percent, respectively, for rule-based and supervised POS tagging.

Sunita et al., [8] the study used NLTK and hidden HMM techniques to examine POS for Khasi. In this analysis, 86,087 tokens from a corpus were employed. The accuracy of the HMM approach was 95.68 percent, whereas the accuracy of the NLTK tool was 89.7 percent. The POS tagger for the Khasi language that uses HMM was detailed in the paper. 53 tags were used to annotate the Khasi corpus; there are 7500 tokens in the train data and 312 words in the test data. Therefore, the system has a 76.70 percent accuracy rate.

Jayawera et al., [9] primarily focused on creating a tagging system to support computational linguistics analysis for the Sinhala language. In this study, they discuss the POS tagging method they created, which is an application of the stochastic model method built on HMM. The aforementioned model has a corresponding algorithm. The model was tested against a Sinhala text corpus containing 90551 words and 2754 phrases. The tagger provided more than 90% accuracy for known terms; however the system is not yet performing well for texts with unknown words.

3. The Nyishi Language:

There hasn't been much study of the Nyishi language. The roman script has been adopted and accepted for use by the phonological system of the Nyishi language. As can be seen in Fig. 1 below, the Nyishi language has a total of twentyeight alphabets, with 18 consonants, 7 vowels, 2 clusters, and 1 glottal.

Consonants	**Vowels**	**Cluster**	**Glottal**
B,C,D,F,G,H,J,K, L, M,N,P,R,S,T,X,Y, Z	A,E,I,O,U,V, W	Ng, Ny	Q

Fig 1: The alphabets of the Nyishi language (table drawn by Author)

A strong grammatical foundation exists in the Nyishi language. Unlike the third person nouns, which appear to be gender-differentiated in Nyishi, all third person pronouns lack gender. However, it should be noted that the majority of Nyishi nouns are disyllabic whereas the majority of verbs are monosyllabic. Men and women are both gendered by a marker. The Nyishi sentences generally follow the SOV (Subject-Object-Verb) word order. While there are numerous possible arrangements in the language, the verb often comes after the other sentence parts. In the Nyishi language, a word's meaning can be changed by modifying the tone, pitch, and shape of its syllables. The tonal aspect of the language is a barrier for computational linguistics because there aren't any tonal symbols that are routinely used to represent all the different tones in the language.

4. System description:

This section discusses the several stages of the proposed system's development, comprising data collection, preprocessing, tokenization, tag set, corpus development, and the HMM models.

4.1 Data collection:

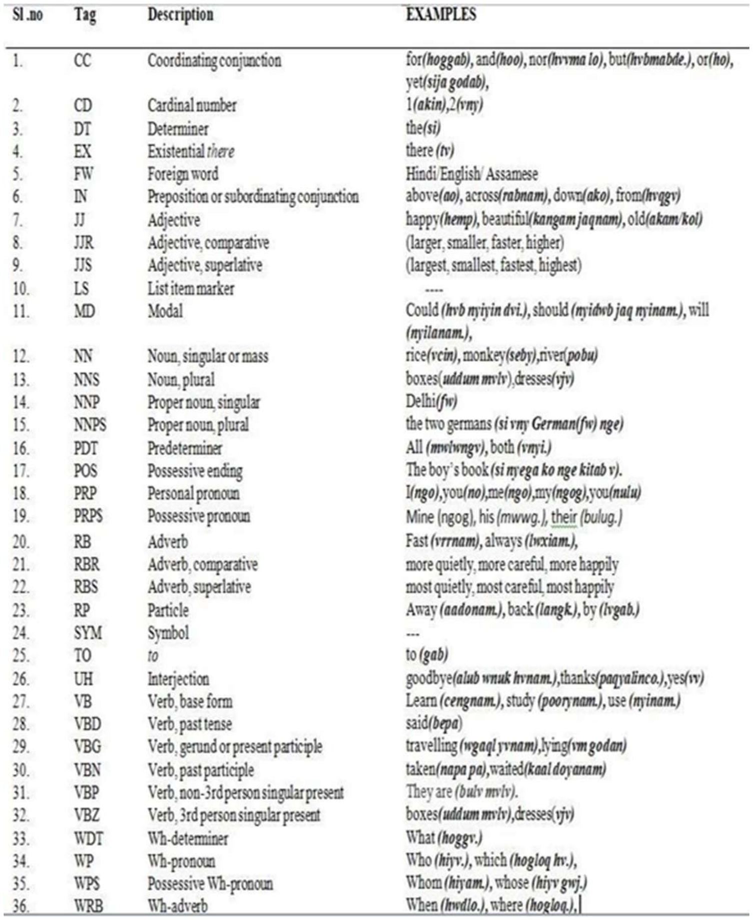
The collection of text or speech written in a certain language and situated in a specific situation is known as a corpus in linguistics. And these written texts have the appropriate tag sets applied, which will be utilized subsequently for training. The most crucial task for any tagger to complete is corpus building. It was a first for the Nyishi language that texts were manually tagged and a corpus was created for the current endeavor. The majority of the data was gathered via the limited Nyishi language dictionaries and older literature. It was intended for the corpus, a collection of data, to be this way so that it could handle ambiguity problems well. This corpus will later be used to train the tagger.

4.2 Preprocessing:

To take advantage of the disparate writing styles of the many authors, additional processing of the collected raw text is necessary. The majority of them are due to general proficiency in grammar. Data in the numeric form is necessary for machine learning, as is well known. In order to convert text into a numeric vector, we essentially used encoding techniques (Bag of Word, Bi-gram, n-gram). However, cleaning text data prior to encoding is necessary.

4.3 Tag set:

Selecting an appropriate tag set is crucial to the task of POS tagging. A tag set is a compilation of all the tags used to indicate the specifics of a language's grammatical structure. The pattern of tag sets in each language differs. A huge tag set will make it more difficult to improve the efficiency of the final model as well as to construct a corpus. As there is no tag set for the Nyishi language, creating a useful tag set becomes more challenging. So, using the Penn Treebank Tag set as a guide, we create a POS tag set for the Nyishi language. The Penn Treebank 36 tag set for recognized the
Part-Of-Speech (POS) of Nyishi language as shown in table
2.



Sl .no	Tag	Description	EXAMPLES
1.	CC	Coordinating conjunction	for(hoggab), and(hoo), nor(hvvma lo), but(hvbmabde.), or(ho), yet(sija godab),
2.	CD	Cardinal number	1(akin),2(vny)
3.	DT	Determiner	the(si)
4.	EX	Existential *there*	there (tv)
5.	FW	Foreign word	Hindi/English/ Assamese
6.	IN	Preposition or subordinating conjunction	above(ao), across(rabnam), down(ako), from(hvggv)
7.	JJ	Adjective	happy(hemp), beautiful(kangam jaqnam), old(akam/kol)
8.	JJR	Adjective, comparative	(larger, smaller, faster, higher)
9.	JJS	Adjective, superlative	(largest, smallest, fastest, highest)
10.	LS	List item marker	----
11.	MD	Modal	Could (hvb nyiyin dvi.), should (nyidwb jaq nyinam.), will (nyilanam.),
12.	NN	Noun, singular or mass	rice(vcin), monkey(seby),river(pobu)
13.	NNS	Noun, plural	boxes(uddum mvlv),dresses(vjv)
14.	NNP	Proper noun, singular	Delhi(fw)
15.	NNPS	Proper noun, plural	the two germans (si vny German(fw) nge)
16.	PDT	Predeterminer	All (mwlwngv), both (vnyi.)
17.	POS	Possessive ending	The boy's book (si nyega ko nge kitab v).
18.	PRP	Personal pronoun	I(ngo),you(no),me(ngo),my(ngog),you(nulu)
19.	PRPS	Possessive pronoun	Mine (ngog), his (mwwg.), their (bulug.)
20.	RB	Adverb	Fast (vrrnam), always (lwxiam.),
21.	RBR	Adverb, comparative	more quietly, more careful, more happily
22.	RBS	Adverb, superlative	most quietly, most careful, most happily
23.	RP	Particle	Away (aadonam.), back (langk.), by (lvgab.)
24.	SYM	Symbol	---
25.	TO	*to*	to (gab)
26.	UH	Interjection	goodbye(alub wnuk hvnam.),thanks(paqyalinco.),yes(vv)
27.	VB	Verb, base form	Learn (cengnam.), study (poorynam.), use (nyinam.)
28.	VBD	Verb, past tense	said(bepa)
29.	VBG	Verb, gerund or present participle	travelling (wgaql yvnam),lying(vm godan)
30.	VBN	Verb, past participle	taken(napa pa),waited(kaal doyanam)
31.	VBP	Verb, non-3rd person singular present	They are (bulv mvlv).
32.	VBZ	Verb, 3rd person singular present	boxes(uddum mvlv),dresses(vjv)
33.	WDT	Wh-determiner	What (hoggv.)
34.	WP	Wh-pronoun	Who (hiyv.), which (hogloq hv.),
35.	WPS	Possessive Wh-pronoun	Whom (hiyam.), whose (hiyv gwj.)
36.	WRB	Wh-adverb	When (hwdlo.), where (hogloq.),

Table2: Tag set for Nyishi language

4.4 Tokenization:

The process of tokenization includes dividing the raw text into manageable chunks. The original text is tokenized into tokens, which are words and sentences. Understanding the function of part-of-speech tagging is made easier by these symbols. Tokenization assists in understanding the meaning of the text by looking at the word order in the text.

4.5 Design of Nyishi POS tagger:

The figure below demonstrates the user interface architecture of the POS tagger developed for the Nyishi language. The user caters the input to the Nyishi language. The tagger tokenizes the input data and work on it. Then, the input text is sent to the respective algorithm and process the tagged output. Finally, we get the final POS tagged output.

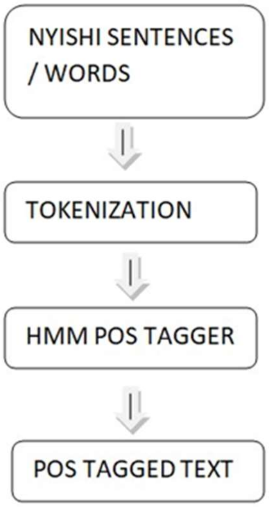


Fig 3 Architecture of Nyishi POS tagger (Image by author)

5. Methodology : Hidden Markov Model

HMM taggers select the best tag for a particular word sequence before selecting the best tag combination that optimizes the calculation by

P (word | tag) * P (tag | previous n tags).

An "emission probability" in the HMM model is given by $P(W|T)$, which represents the likelihood of witnessing the input phrase or word sequence W given the state sequence T. Additionally, we may determine the probability $P(T)$ of producing the state sequence T using the state transition probabilities. P(ni|ni-1ni-2) represents the "transition probability" between two states. Formally, the challenge is to identify the tag sequence T that maximizes the probability $P(T|W)$ is argmax $P(T|W)$ for a set of words W.

We can determine $P(T|W)$ using the Bayes' rule for conditional probabilities.

P(T|W) = [P(W|T) * P(T)/P(W)].

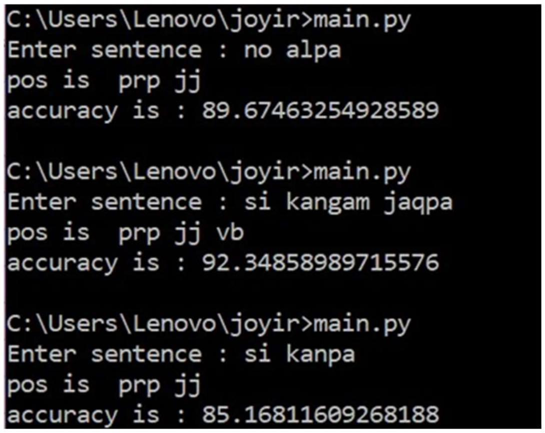
Since the word sequence $P(W)$ has the same probability for each sequence, we may ignore it, resulting in

$P(N|W) = P(W|N) P(N)$, thus $P(N|W)$ wants to be maximized as **argmax P(W|T) * P(T).**

Our POS tagger is created using the Viterbi algorithm and the HMM. Regardless of word order, the HMM model allocated tags to words based on data from the corpus and training set. Sentences from a dataset are first selected as a sample for POS tagging. To calculate the likelihood that the sentences are correct, these datasets are then projected to an HMM. To determine the probability of transition, the phrases were broken up into sections. In a sentence, a transition probability is the likelihood that a noun will come after a model or a model will come after a noun. To determine which word is a noun and which word is a model, emission probabilities of these datasets must be determined after receiving the transition probabilities for the relevant datasets. The estimated probability is obtained by multiplying the number of outcomes that these two probabilities can produce by one another. The POS correctness is tested as the final stage in this technique. The results of the HMM are then put to the test to determine their accuracy as POS for NLP.

6. Result:

This section goes into detail about how the suggested methodology was put into practice. This section discusses the results of utilizing HMM for POS tagging in the Nyishi language. HMM is used to train a phrase once it is randomly input. After training, the likelihood of the phrase transition and emission is established, and an estimated probability is computed based on this probability calculation. NLP computes the sentence's accuracy based on the POS test results. The most accurate POS for NLP, with an average accuracy of 89.90%, is found after analysis of the HMM results. The precision of POS Tag is displayed in Figure 4.



```
C:\Users\Lenovo\joyir>main.py
Enter sentence : no alpa
pos is  prp jj
accuracy is : 89.67463254928589

C:\Users\Lenovo\joyir>main.py
Enter sentence : si kangam jaqpa
pos is  prp jj vb
accuracy is : 92.34858989715576

C:\Users\Lenovo\joyir>main.py
Enter sentence : si kanpa
pos is  prp jj
accuracy is : 85.16811609268188
```

Fig 4 Accuracy of POS tagging (Image By author)

7. Conclusion and future work

For Nyishi, a language used frequently in Arunachal Pradesh but with little computational linguistic research prior to our work, we have produced good POS tagging results. With just a 30000 sentences training corpus, we were able to achieve an average tagging accuracy of 89%.The Nyishi tag set which was not present prior to the start of this project, is our greatest accomplishment. We have developed an existing POS tagging technique, but the language for which we are working lacks both annotated corpora and a predefined tag set. The accuracy results of the system are achieved by training using Nyishi corpus and then testing the system. And for future work, more new data can be added in the corpus and accomplish the main motive of this work by reducing the ambiguity of the language. This work will definitely prove as an important resource for any future work on Nyishi language, NLP perspective and Arunachal as a whole.

References

1. Reshamwala, Alpa, Dhirendra Mishra, and Prajakta Pawar. "Review on natural language processing." IRACST Engineering Science and Technology: An International Journal (ESTIJ) 3, no. 1 (2013): 113-116.
2. Martinez, Angel R. "Part-of-speech tagging." Wiley Interdisciplinary Reviews: Computational Statistics 4, no. 1 (2012): 107-113.
3. Siram Joyir, Koj Sambyo, and Achyuth Sarkar. "Partof-Speech (POS) for the Nyishi Language." In Advances in Information Communication Technology and Computing, pp. 191-199.
4. Navanath S, Dhrubajyoti D, Utpal S, Jugal K “part of speech tagger for Assamese text”

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec,
Singapore, 4 August 2009, pages 33–36

5. Joshi, N., Darbari, H., & Mathur, I. "HMM based POS tagger for Hindi" Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013), pp. 341–349 (2013).

6. Braja G. P, Khumbar D, Dipankar D, S.
Bandyopadhyay "part of speech tagger (POS) tagger for korborok "Proceedings of COLING 2012: Posters, pages 923–932

7. Singh, T. D., Ekbal, A., and Bandyopadhyay, S. (2008). Manipuri POS tagging using CRF and SVM: A language-independent approach. In proceeding of 6th International Conference on Natural Language Processing (ICON-2008), pages 240-245.

8. Sunita wajra, ParthaPakray, Saralin Lyngdoh, Arnab
Kumar Maji "Khasi language as dominant Part of speech (POS) ascendant in NLP". Proceeding of international conference on computational intelligence & IOT (ICCIIoT)2018, page no 109-115. (2018).

9. Jayaweera, A.J.P.M.P. and Dias "Part of Speech (POS) tagger for Sinhala language". proceedings of the Annual Research Symposium 2011

10. Kumawat, Deepika, and Vinesh Jain. "POS tagging approaches: A comparison." International Journal of Computer Applications 118, no. 6 (2015).

11. Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18, no. 5 (2011): 544-551.

12. Hirschberg, Julia, and Christopher D. Manning. "Advances in natural language processing." Science 349, no. 6245 (2015): 261-266.

13. Xue, Xiaorui, and Jiansong Zhang. "Part-of-speech tagging of building codes empowered by deep learning and transformational rules." Advanced Engineering Informatics 47 (2021): 101235.

14. Chotirat, Saranlita, and Phayung Meesad. "Part-ofSpeech tagging enhancement to natural language processing for Thai wh-question classification with deep learning." Heliyon 7, no. 10 (2021): e08216.

15. AlKhwiter, Wasan, and Nora Al-Twairesh. "Part-ofspeech tagging for Arabic tweets using CRF and BiLSTM." Computer Speech & Language 65 (2021): 101138.

16. Afini, Umriya, and Catur Supriyanto. "Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger." In 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), pp. 237-240. IEEE, 2017.

17. Drovo, Mah Dian, Moithri Chowdhury, Saiful Islam Uday, and Amit Kumar Das. "Named entity recognition in Bengali text using merged hidden markov model and rule based approach." In 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1-5. IEEE, 2019.