Part of Speech Tagging in Manipuri with Hidden Markov Model

Kh Raju Singha[1], Bipul Syam Purkayastha[2] and Kh Dhiren Singha[3]

[1] Department of Computer Science, Assam University, Silchar
Silchar, Assam 780011, India

[2] Department of Computer Science, Assam University, Silchar
Silchar, Assam 780011, India

[3] Department of Linguistics, Assam University, Silchar
Silchar, Assam 780011, India

Abstract

Part of Speech tagging in Manipuri is a very complex task as Manipuri is highly agglutinating in nature. There is no enough tagged corpus for Manipuri which can be used in any statistical analysis of the language. In this tagging model we are using tagged output of the Manipuri rule-based tagger as tagged corpus. The present paper expounds the Part of Speech Tagging in Manipuri by applying a stochastic model called Hidden Markov Model.

Keywords: *Part of Speech Tagging, Tokenization, Affix, Stochastic, Hidden Markov Model.*

1. Introduction

Part of Speech Tagging is an indispensable part of Natural Language Processing. It plays a vital role in developing of any serious applications in processing all the natural languages in the world. Part of Speech Tagging is an initial stage of linguistics text analysis like sub-category acquisition, information retrieval, machine translation, text to speech synthesis etc [19]. Assigning Part of Speech tag to each and every lexical item of an unannotated text is a tedious and time consuming task. Therefore, automatic Manipuri Part of Speech Tagger is required for the overall development of the language in the field of Natural Language Processing.

POS tagger can be classified as supervised and unsupervised. Both the supervised and unsupervised can be categorized as rule-based and stochastic models [14]. Rule-based POS tagger depends on the lexicon and handcrafted linguistics rules [19]. Hidden Markov Model is a common stochastic model widely used in Natural Language Processing. HMM based POS tagger assigns the best sequence of tags to an entire text of the test set. In general, the most probable tag sequence is assigned to each sentence following the Viterbi algorithm [14]. The task of the HMM based POS tagger is to find the sequence of POS tags $T= \{t_1, t_2, t_3 \ldots., t_n\}$ that is optimal for a word sequence $W= \{w_1, w_2, w_3 \ldots\ldots, w_n\}$.

2. Related Work

Many works related to POS tagging has been done in languages like English, Chinese, German and Arabic etc. Several POS tagger of these languages are developed by using different algorithms. For instance, English language has developed POS tagger using rule based, statistical method, neural network and transformational based method etc [14]. In the year 1992 Eric Brill has been developed a rule based POS tagger with the accuracy rate of 95-99% [2]. POS tagging of some languages like Turkish [3], Czech [5] has been attempted using a combination of hand-crafted rules and statistical learning.

Similarly, Indian languages like Hindi, Bengali, Punjabi and Dravidian languages have many POS taggers. Manish Shrivastava and Pushpak Bhattacharyya proposed POS tagger for Hindi based on HMM in the year 2008 [11]. Adopting rule based approach a POS tagger for Marathi has been developed in 2006 using a technique called SRR (suffix replacement rule) by Sachin Burange et al. [9]. A Punjabi POS tagger is also developed by Singh Mandeep, Lehal Gurpeet and Sharma Shiv in 2008 with accuracy performance of 88.86% excluding unknown words [16].

As per the literature, there is a few works related to POS tagging in Manipuri and other Tibeto-Burman languages in the Indian Sub-continent. In the year 2004, Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, P.K. Das have designed and implemented a

morphological analyzer for Manipuri language [7]. Besides, D. S. Thoudam et al. developed morphology driven Manipuri POS tagger in 2008 [13]. Furthermore, Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha designed a tagset for Manipuri POS tagging consisting of 97 tags including generic attributes and language specific attribute values based on the ILPOST framework in 2011 [17] and developed a POS tagger using rule-based approach in 2012 [19].

3. Manipuri Language

Manipuri (Meiteilon or Meiteiron) is one of the oldest languages in the South-east Asia which has its own script (Meitei Mayek) and literature. At Present, Manipuri used to write in Bengali Script from 1709 A.D. onwards i.e.; during the reign of king Pamheiba. Manipuri is widely spoken in Manipur, Assam, Tripura, Bangladesh and Myanmar, which has been included in the eight schedule of Indian Constitution since 1992. Interestingly, it is the first Tibeto-Burman language which has obtained its due place and recognition in Indian Constitution [10].

Linguistically, it belongs to the Kuki-Chin group of the Tibeto-Burman family of Languages [1] influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. The total number of people who return Manipuri as their mother tongue was 1,500,000 out of which 1,466,705 speakers reside in India (Census of India, 2001).

Manipuri is a tonal, agglutinating and verb final language. Like other OV languages adjectives may precede or follow the noun in NP constructions. As in many other Tibeto-Burman languages, adjective is not a distinct category of words in Manipuri as adjectives are derived from the intransitive verb particularly the stative verbs. Unlike other Indo-Aryan languages of the sub-continent, there is lack of relative pronoun; the relative clause is expressed by means of participle. Compounding, derivation and affixation play the major role in the word formation process of Manipuri. The occurrence of a direct relationship between case markers and case roles has made it possible for the language to show a lot of variation in valence structure of its verbs [8].

4. Manipuri Tagset

Tagset is a set of well defined grammatical classes of each lexical item and its attributes of a natural language. Therefore, tagset is a very essential part in developing a part of speech tagger of any natural language. In this paper, we are using a tagset of Manipuri consisting of 97 tags including generic attribute values and language specific attribute values [17]. It is a hierarchical tagset based on the ILPOSTS framework [12] and is customized for Manipuri to meet the morpho-syntactic requirements of the language. The framework has three layer hierarchies. The top layer has morphological categories followed by the sub-categories in the middle layer and the bottom layer has morpho-syntactic features or attributes of sub categories. A partial graphical representation of the proposed tagset "NOUN" as an example is given below:

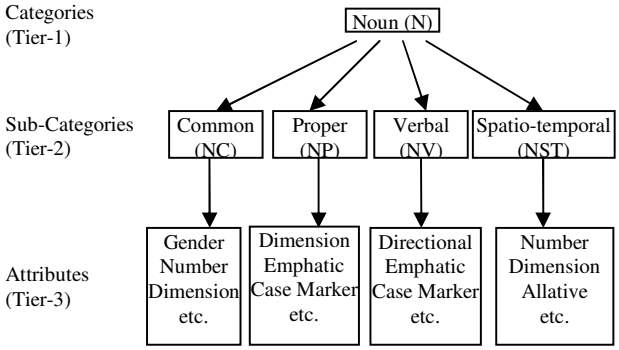

Fig. 1 A partial graphical representation of Manipuri tagset

5. Proposed System Design

The overall structural design of the proposed system including the relations between the modules is shown in the Fig.2.

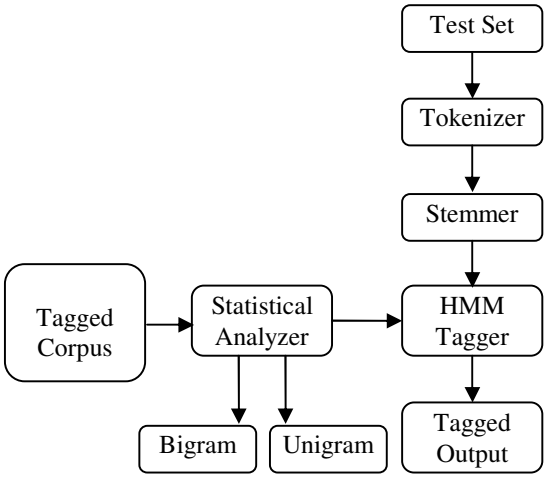

Fig. 2 Proposed System Design of HMM based Manipuri POS Tagger

The different modules associated in this design are elaborated as follows:

5.1 Tokenizer

It breaks the input text stream or test set into words, phrases, symbol, punctuation and other meaningful elements by keeping a whitespace between them. These elements are called token. Manipuri has the writing style of keeping a whitespace between each and every token except punctuation and symbol. Therefore, role of the tokenizer in this architecture is to separate the punctuation mark or symbol from the words.

5.2 Stemmer

For the processing of agglutinative language like Manipuri, stemmer plays a major role of separating affixes from the root word. In this system, the stemmer split the suffixes one at a time from right to left until the root word is found by using iterative suffix stripping technique [19].

5.3 Tagged Corpus

To perform part of speech tagging by applying stochastic technique, a tagged corpus is required. Since Manipuri has no tagged corpus, we are using the tagged output of the Rule-based Manipuri Part of Speech Tagger developed by Kh Raju Singha et.al. [19] as tagged corpus. This corpus consists of 2000 tagged lexical items.

5.4 Statistical Analyzer

Statistical Analyzer extracts the unigram and bigram probabilities from the tagged corpus [3]. In computing the n-gram probabilities, the frequency of each word and two words sequence in the corpus is determined. This computation results unigram or bigram, the probability that it occurs with a particular tag or tag sequence among all possible tags or tag sequences is determined.

Unigram Tagger: The unigram (n-gram, n=1) tagger is an uncomplicated statistical tagging algorithm. For each token, it assigns the tag that is most likely for that token's text. For example, it will assign the tag JJ to any occurrence of the word *"achouba" i.e; "big",* since "*achouba"* is used as an adjective more often than it is used as a proper noun. Before a unigram tagger can be used to tag data, it must be trained on a training corpus. It uses the corpus to determine which tags are most common for each word [15]. The unigram tagger will assign the default tag UNK (Unknown) to any token that was not encountered in the training data.

Bigram Tagger: Bigram assumption states that probability of a tag appearing depends only on the previous tag.

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i | t_{i-1})$$

Here, bigram is a set of two consecutive words; it is a special case of n-gram.

5.5 HMM Tagger

After collecting statistical data of the tagged corpus from statistical analyzer, the tagger is activated on the test set which is already tokenized and stemmed by the tokenizer and stemmer respectively. The tagger employs a sentence based approach rather than a word based approach. That is, first all the possible tags for the words and the word sequences in the sentence are determined, and then the combination of the tags with the highest probability for the whole sentence is selected.

The perception behind Hidden Markov Model tagger is a simple generalization of the "pick the most likely tag for this word" approach [15]. The unigram tagger only considers the probability of a word for a given tag t; the surrounding context of that word is not considered. On the other hand, for a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the following formula:

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

For finding the maximum probability HMM uses the Viterbi Algorithm.

Viterbi for POS tagging:

Initialization Step

For i=1 to N do
Seqscore (i, 1) = Prob (w_1 | L_i)*Prob (L_i |ø)
Backptr (i, 1)=0;

Iteration Step

For t=2 to T
For i=1 to N
Seqscore (i, t) = MAX $_{j=1, N}$ (Seqscore (i,t-1)*
Prob (L_i | L_j)) * Prob (w_t | L_i)
Backptr (i, t) = index of j that gave the MAX above
Sequence

Identification step

C (T) = I that maximizes Sequence (i, T)
For i=T-1 to 1 do
C (i) = Backptr (C (i+1), i+1)

w_1, ……..w_T : Word Sequence
L_i , ……...L_N : Lexical Categories
Prob (w_t | L_i) : Lexical Probability
Prob (L_i | L_j) : Bigram Probability

6. Results

In order to measure the performance of the system we use the manually annotated test set data using the tag set consisting of 97 morpho-syntactic categories of Manipuri as evaluation set. Accuracy percentage of the tagger is calculated using the formula given below:

$$Accuracy\ \text{percentage} = \frac{Correctly\ \text{Tagged words}}{No.\ of\ \text{words in evaluation set}} \text{ X } 100$$

It gives the accuracy of 92% and it is clear that accuracy percentage is increased with increase the size of the tagged corpus. For performing statistical tagging, we have considered only 12 tag sequences, and the result obtained from the Statistical Analyzer is very satisfactory. 80% of the sequences generated automatically for the test set were found correct when compared with the manually tagged result of those sentences.

6. Conclusion

This paper presents a model for POS tagging in Manipuri using HMM. As Manipuri has no tagged corpus, the system uses the small set of tagged sentence which is generated from Manipuri Rule-based Tagger. The system has the ability to assign tags to most of the lexical items in the test set. This tagger will be very helpful applications in processing of Manipuri language like text-based information retrieval, speech recognition and machine translation etc. The proposed system can be made more efficient by extending the bigram probability to trigram probability.

References

[1] G.A. Grierson's Linguistic Survey of India. Vol. III, Pt. III, 1976.

[2] Eric Brill. A simple rule-based part of speech tagger. In Proceedings Third Conference on Applied Natural Language Processing, ACL, Trento, Italy, 1992.

[3] K. Oflazer, I Kuruoz, "Tagging and morphological disambiguation of Turkish text". In Proceedings of 4th ACL conference on Applied Natural Language Processing Conference, 1994.

[4] James Allen, Natural Language Understanding, Benjamin/Cummings Publishing Company, 1995

[5] J. Hajic, P. Krbec, P. Kveton, K. Oliva, V.Petkevic, "A Case Study in Czech Tagging". In proceedings of the 39th Annual Meeting of the ACL, 2001.

[6] Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu "A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali", Transactions on Engineering, Computing and Technology V1 December, 2004.

[7] Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, P.K. Das, "Morphological Analyzer for Manipuri: Design and Implementation". In Proceedings of AACC, Kathmandu, Nepal, pp 123-129, 2004.

[8] P.C. Thoudam. "Problems in the Analysis of Manipuri Language."www.ciil-ebooks.net, CIIL, Mysore, 2006.

[9] Sachin Burange, Sushant Devlakar, Pushpak Bhattacharyya, "Rule Governed Marathi POS Tagging". In Proceeding of MSPIL, IIT Bombay, pp 69- 78, 2006.

[10] Kh. Dhiren Singha, "Loan Words in Manipuri", Bilingualism and North-East India, an Assam University Publication, 2008.

[11] Manish Shrivastava and Pushpak Bhattacharyya, Hindi POS Tagger Using Naïve Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge, International Conference on NLP (ICON08), Pune, India, 2008.

[12] S. Baskaran et al." Designing a Common POS-Tagset Framework for Indian Languages" The 6th Workshop on Asian Languae Resources, 2008.

[13] Thoudam Doren Singh & Sivaji Bandyopadhyay "Morphology Driven Manipuri POS Tagger", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India, January 2008.

[14] D. Jurafsky, and J. H. Martin, "Speech and Language Processing", Second edition, Published by Pearson Education, 2009.

[15] Manju K, Soumya S, Suman Mary Idicula, " Development of a POS tagger for Malayalam – An Experience" International Conference on Advances in Recent Technologies in Communication and Computing, 2009

[16] Dinesh Kumar and Gurpreet Singh Josan, "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume6–No.5, September, 2010.

[17] Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha, Arindam Roy "Developing a Tagset for Manipuri Part of Speech Tagging" Journal of Computer Science and Engineering, Volume-5-issue-1-january 2011.http://sites.google.com/site./jcseuk/volume-5-issue 1-january-2011.

[18] http://language.worldofcomputing.net/pos-tagging/rulebased-pos-tagging.html.2012.

[19] Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha "Part of Speech Tagging in Manipuri: A rule-based Approach" International Journal of Computer Applications (0975 – 8887) Volume 51– No.14, August 2012

Kh Raju Singha is a Ph.D. student in the Department of Computer Science, Assam University, Silchar.

Bipul Syam Purkayastha is working as a Professor in the Department of Computer Science, Assam University, Silchar. He is a member of IEEE and ACM journal.

Kh Dhiren Singha is working as an Associate Professor in the Department of Linguistics, Assam University, Silchar. He is a member of the Linguistic Society of India and International journal of Dravidian Linguistics.