Lexicon Development and POS Tagging using a Tagged Bengali News Corpus

Asif Ekbal [1] and **Sivaji Bandyopadhyay** [2]
Department of Computer Science and Engineering, Jadavpur University, Kolkata, India 700032
Email: asif.ekbal@gmail.com [1], sivaji_cse_ju@yahoo.com[2]

Abstract

Lexicon development and Part of Speech (POS) tagging are very important for almost all Natural Language Processing(NLP) application areas. The rapid development of these resources and tools using machine learning techniques for less computerized languages requires appropriately tagged corpus. A tagged Bengali news corpus has been developed from the web archive of a widely read Bengali newspaper. This corpus is then used for lexicon development and POS tagging.

Tagged Bengali News Corpus Development

Newspaper is a huge source of readily available documents. A tagged corpus has been developed from the web archive of a very well known and widely read Bengali News Paper. The development of the tagged Bengali news corpus includes language resource acquisition using a web crawler, language resource creation which includes HTML file cleaning and code conversion, as well as language resource annotation that involves defining a tag set and subsequent tagging of the news corpus. Code conversion is necessary to convert the dynamic fonts used in the newspaper into the standard Indian Standard Code for Information Interchange (ISCII) form, which can be processed for various text processing tasks. At present, the corpus contains 34 million wordforms and it is available in both ISCII and UTF-8 formats.

A news corpus, whether in Bengali or in any other language has different parts like title, date, reporter, location, body etc. To identify these parts in a news corpus, the following tagset has been defined: header (Header of the news document), title (Headline of the news document), t1 (1st headline of the title), t2 (2nd headline of the title), date (Date of the news document), bd (Bengali date), day (Day), ed (English date), reporter (Reporter-name), agency (Agency providing news), location (the news location), body (Body of the news document), p (Paragraph), table (information in tabular form), tc (Table Column), and tr (Table row).

Lexicon Development from the Corpus

The tagged Bengali news corpus has been used to develop a Bengali lexicon that is a list of Bengali root words derived from the corpus along with its basic part of speech information. An unsupervised learning method has been used for the lexicon development. No extensive knowledge about the language is required except the knowledge of the different inflections that can appear with the different words in Bengali. In Bengali, there are five different parts of speech namely: noun, pronoun, verb, adjective and indeclinable (postpositions, conjunctions, and interjections). Noun, verb and adjective are the open class of part of speech for Bengali. Initially, all the words (inflected and uninflected) are extracted from the tagged corpus and added to the database. A list of inflections that may appear with the noun words is kept and at present the list has 27 entries. In Bengali, the verbs can be organized into 20 different groups according to their spelling patterns and the different inflections that can be attached to them. Original word-form of a verb word often changes when any suffix is attached to the verb. At present, there are 214 different entries in the verb inflection list. Noun and verb words are tagged by looking at their inflections. Some inflections may be common to both nouns and verbs. In these cases, more than one root words will be generated for a wordform. The POS ambiguity is resolved by checking the number of occurrences of these possible root words along with the POS tags as derived from the other wordforms. Pronoun and indeclinable are basically closed class of part of speech in Bengali and these are added to the lexicon manually. It has been observed that adjectives in Bengali generally occur in four different forms based on the suffixes attached. The first type of adjectives can form comparative and superlative degree by attaching the suffixes *-tara* and *-tamo* to the adjective word. These adjective stems are stored in the lexicon with adjective POS. The second set of suffixes (e.g., *-gato*, *-karo* etc.) identifies the POS of the wordform as adjective if only there is a noun entry of the desuffixed word in the lexicon. The third group of suffixes (e.g., *-janok*, *-sulav* etc.) identifies the POS of the wordform as adjective and the desuffixed word is included in the lexicon with noun POS. The last set of suffixes identifies the POS of the wordform as adjective.

Hidden Markov Model Based POS Tagging

A POS tagger based on the modified Hidden Markov Model (HMM) has been developed using a portion of the tagged Bengali news corpus. We have used a tagset having 27 different tags, developed by the International Institute of Information Technology, Hyderabad [1] for Indian languages and which is still in the process of being standardized. The task of Part of Speech (POS) tagging 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 $W = w_1, w_2, w_3 \ldots w_n$. The tagging problem becomes equivalent to searching for $argmax_T P(T) * P(W|T)$, by the application of Bayes' law. Generally, the most probable tag sequence is assigned to each sentence following the Viterbi algorithm (Viterbi 1967). We have used tri-gram model, i.e., the probability of a tag depends on two previous tags, and then we have, $P(T) = P(t_1|\$) \times P(t_2|\$, t_1) \times P(t_3|t_1, t_2) \times P(t_4|t_2, t_3) \times \ldots \times P(t_n|t_{n-2}, t_{n-1})$, where an additional tag '\$' (dummy tag) has been introduced to represent the beginning of a sentence. Due to sparse data problem, the linear interpolation method has been used to smooth the tri-gram probabilities as follows: $P'(t_n|t_{n-2}, t_{n-1}) = \lambda_1 P(t_n) + \lambda_2 P(t_n|t_{n-1}) + \lambda_3 P(t_n|t_{n-2}, t_{n-1})$ such that the λs sum to 1. The values of λs have been calculated by the method given in (Brants 2000).

To make the Markov model more powerful, ***additional context dependent features*** have been introduced to the emission probability in this work that specifies the probability of the current word depends on the tag of the previous word and the tag to be assigned to the current word. Now, we calculate $P(W|T)$ by the following equation: $P(W|T) \approx P(w_1|\$, t_1) \times P(w_2|t_1, t_2) \times \ldots \times P(w_n|t_{n-1}, t_n)$. So, the emission probability can be calculated as:
$P(w_i|t_{i-1}, t_i) = \frac{freq(t_{i-1}, t_i, w_i)}{freq(t_{i-1}, t_i)}$.

Here also the smoothing technique is applied rather than using the emission probability directly. The emission probability is calculated as: $P'(w_i|t_{i-1}, t_i) = \theta_1 P(w_i|t_i) + \theta_2 P(w_i|t_{i-1}, t_i)$, where θ_1, θ_2 are two constants such that all θs sum to 1. In general, the values of θs can be calculated by the same method that was adopted in calculating λs.

Handling of unknown words is an important issue in POS tagging. Viterbi algorithm attempts to assign a POS tag to the unknown words. Specifically, three different approaches have been adopted to take care of the unknown words in Bengali. For words which have not been seen in the training set, $P(w_i|t_i)$ is estimated based on features of the unknown words, such as whether the word contains a particular suffix. At present, there are 435 suffixes; many of them usually appear at the end of verb, noun and adjective words. A null suffix is also kept to take care of those words that have none of the suffixes in the list. The probability distribution of a particular suffix with respect to specific POS tags is generated from all words in the training set that share the same suffix. Apart from suffix analysis, two other features have been included that tackle tokens of digits and symbols. A named entity recognition (NER) system (Ekbal & Bandyopadhyay 2007), based on pattern directed shallow parsing, has been used to take care of the unknown words in Bengali. Lexicon is used to further handle the unknown words.

[1] http://ltrc.iiit.net/nlpai_contest06/iiit_tagset_guidelines.pdf

Table 1: Lexicon Statistics

Iteration	1	2	3	4	5
News Doc.	9737	19929	39924	69951	99651
Sentences	0.22	0.49	1.02	1.79	2.55
Wordforms	2.77	5.98	12.53	21.53	30.61
Distinct Wordforms	0.10	0.15	0.23	0.37	0.526
Root-words	0.03	0.04	0.065	0.09	0.128

Table 2: Performance of the POS Tagger for Bengali

Type	Accuracy (in %)
HMM	83.04
HMM + unknown word features	86.38
HMM + unknown word features + NER	88.45
HMM + unknown word features +NER +Lexicon	91.6

Experimental Results

Table 1 shows the results for lexicon development. Except news documents, the number of sentences, wordforms, distinct wordforms and root words have been mentioned in millions. The lexicon has been checked manually for correctness and it has been observed that the accuracy is approximately 79.6%. The list of rootwords can be automatically corrected to a large degree by using the named entity recognizer for Bengali (Ekbal & Bandyopadhyay 2007) to identify the named entities in the corpus in order to exclude them from the lexicon. The number of root words increases as more and more news documents are considered in the lexicon development.

The POS tagger has been trained with a training set of approximately 31,190 wordforms. The POS tagger has been tested on the manually tagged corpus of 5967 wordforms that is used as the Gold standard test set to evaluate the POS tagger.The POS tagger initially demonstrated 83.04% accuracy for the Bengali test set. The accuracy increases upto 91.6% with the inclusion of the different techniques, adopted for handling the unknown words. The results have been presented in Table 2. The popular language independent TnT tagger has been trained on the same training set and then run on the same test set. It has demonstrated an accuracy of 86.2%.

References

Brants, T. 2000. TnT a statistical parts-of-speech tagger. In *Proceedings of the sixth International Conference on Applied Natural Language Processing ANLP-2000*. 224–231.

Ekbal, A., and Bandyopadhyay, S. 2007. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In *Proceedings of the fifth International Conference on Natural Language Processing, ICON-2007*. 123–128.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transaction on Information Theory* 13(2):260–269.