Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario

Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India 721302
{sandipan,sudeshna,anupam.basu}@cse.iitkgp.ernet.in

Abstract

This paper describes our work on building Part-of-Speech (POS) tagger for Bengali. We have use Hidden Markov Model (HMM) and Maximum Entropy (ME) based stochastic taggers. Bengali is a morphologically rich language and our taggers make use of morphological and contextual information of the words. Since only a small labeled training set is available (45,000 words), simple stochastic approach does not yield very good results. In this work, we have studied the effect of using a morphological analyzer to improve the performance of the tagger. We find that the use of morphology helps improve the accuracy of the tagger especially when less amount of tagged corpora are available.

1 Introduction

Part-of-Speech (POS) taggers for natural language texts have been developed using linguistic rules, stochastic models as well as a combination of both (hybrid taggers). Stochastic models (Cutting et al., 1992; Dermatas et al., 1995; Brants, 2000) have been widely used in POS tagging for simplicity and language independence of the models. Among stochastic models, bi-gram and tri-gram Hidden Markov Model (HMM) are quite popular. Development of a high accuracy stochastic tagger requires a large amount of annotated text. Stochastic taggers with more than 95% word-level accuracy have been developed for English, German and other European Languages, for which large labeled data is available. Our aim here is to develop a stochastic POS tagger for Bengali but we are limited by lack of a large annotated corpus for Bengali. Simple HMM models do not achieve high accuracy when the training set is small. In such cases, additional information may be coded into the HMM model to achieve higher accuracy (Cutting et al., 1992). The semi-supervised model described in Cutting et al. (1992), makes use of both labeled training text and some amount of unlabeled text. Incorporating a diverse set of overlapping features in a HMM-based tagger is difficult and complicates the smoothing typically used for such taggers. In contrast, methods based on Maximum Entropy (Ratnaparkhi, 1996), Conditional Random Field (Shrivastav, 2006) etc. can deal with diverse, overlapping features.

1.1 Previous Work on Indian Language POS Tagging

Although some work has been done on POS tagging of different Indian languages, the systems are still in their infancy due to resource poverty. Very little work has been done previously on POS tagging of Bengali. Bengali is the main language spoken in Bangladesh, the second most commonly spoken language in India, and the fourth most commonly spoken language in the world. Ray et al. (2003) describes a morphology-based disambiguation for Hindi POS tagging. System using a decision tree based learning algorithm (CN2) has been developed for statistical Hindi POS tagging (Singh et al., 2006). A reasonably good accuracy POS tagger for Hindi has been developed using Maximum Entropy Markov Model (Dalal et al., 2007). The system uses linguistic suffix and POS categories of a word along with other contextual features.

2 Our Approach

The problem of POS tagging can be formally stated as follows. Given a sequence of words $w_1 \dots w_n$, we want to find the corresponding sequence of tags $t_1 \dots t_n$, drawn from a set of tags T. We use a tagset of 40 tags[1]. In this work, we explore supervised and semi-supervised bi-gram

[1] http://www.mla.iitkgp.ernet.in/Tag.html

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 221–224,
Prague, June 2007.

HMM and a ME based model. The bi-gram assumption states that the POS-tag of a word depends on the current word and the POS tag of the previous word. An ME model estimates the probabilities based on the imposed constraints. Such constraints are derived from the training data, maintaining some relationship between features and outcomes. The most probable tag sequence for a given word sequence satisfies equation (1) and (2) respectively for HMM and ME model:

$$S = \arg\max_{t_1 \ldots t_n} \prod_{i=1,n} P(w_i \mid t_i) P(t_i \mid t_{i-1}) \quad (1)$$

$$p(t_1 \ldots t_n \mid w_1 \ldots w_n) = \prod_{i=1,n} p(t_i \mid h_i) \quad (2)$$

Here, h_i is the context for word w_i. Since the basic bigram model of HMM as well as the equivalent ME models do not yield satisfactory accuracy, we wish to explore whether other available resources like a morphological analyzer can be used appropriately for better accuracy.

2.1 HMM and ME based Taggers

Three taggers have been implemented based on bigram HMM and ME model. The first tagger (we shall call it **HMM-S**) makes use of the supervised HMM model parameters, whereas the second tagger (we shall call it **HMM-SS**) uses the semi supervised model parameters. The third tagger uses **ME** based model to find the most probable tag sequence for a given sequence of words.

In order to further improve the tagging accuracy, we use a Morphological Analyzer (MA) and integrate morphological information with the models. We assume that the POS-tag of a word *w* can take values from the set $T_{MA}(w)$, where $T_{MA}(w)$ is computed by the Morphological Analyzer. Note that the size of $T_{MA}(w)$ is much smaller than T. Thus, we have a restricted choice of tags as well as tag sequences for a given sentence. Since the correct tag *t* for *w* is always in $T_{MA}(w)$ (assuming that the morphological analyzer is complete), it is always possible to find out the correct tag sequence for a sentence even after applying the morphological restriction. Due to a much reduced set of possibilities, this model is expected to perform better for both the HMM (HMM-S and HMM-SS) and ME models even when only a small amount of labeled training text is available. We shall call these new models **HMM-S+MA**, **HMM-SS+ MA** and **ME+MA**.

Our MA has high accuracy and coverage but it still has some missing words and a few errors. For the purpose of these experiments we have made sure that all words of the test set are present in the root dictionary that an MA uses.

While MA helps us to restrict the possible choice of tags for a given word, one can also use suffix information (i.e., the sequence of last few characters of a word) to further improve the models. For HMM models, suffix information has been used during smoothing of emission probabilities, whereas for ME models, suffix information is used as another type of feature. We shall denote the models with suffix information with a **'+suf'** marker. Thus, we have – **HMM-S+suf**, **HMM-S+suf+MA**, **HMM-SS+suf** etc.

2.1.1 Unknown Word Hypothesis in HMM

The transition probabilities are estimated by linear interpolation of unigrams and bigrams. For the estimation of emission probabilities add-one smoothing or suffix information is used for the unknown words. If the word is unknown to the morphological analyzer, we assume that the POS-tag of that word belongs to any of the open class grammatical categories (all classes of Noun, Verb, Adjective, Adverb and Interjection).

2.1.2 Features of the ME Model

Experiments were carried out to find out the most suitable binary valued features for the POS tagging in the ME model. The main features for the POS tagging task have been identified based on the different possible combination of the available word and tag context. The features also include prefix and suffix up to length four. We considered different combinations from the following set for obtaining the best feature set for the POS tagging task with the data we have.

$$F = \{ w_i, w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2}, t_{i-1}, t_{i-2}, |pre| \le 4, |suf| \le 4 \}$$

Forty different experiments were conducted taking several combinations from set '*F*' to identify the best suited feature set for the POS tagging task. From our empirical analysis we found that the combination of contextual features (current word and previous tag), prefixes and suffixes of length ≤ 4 gives the best performance for the ME model. It is interesting to note that the inclusion of prefix and suffix for all words gives better result instead of using only for rare words as is described in Ratnaparkhi (1996). This can be explained by the fact that due to small amount of annotated data, a significant number of instances are not found for most of the word of the language vocabulary.

3 Experiments

We have a total of 12 models as described in subsection 2.1 under different stochastic tagging schemes. The same training text has been used to estimate the parameters for all the models. The model parameters for supervised HMM and ME models are estimated from the annotated text corpus. For semi-supervised learning, the HMM learned through supervised training is considered as the initial model. Further, a larger unlabelled training data has been used to re-estimate the model parameters of the semi-supervised HMM. The experiments were conducted with three different sizes (10K, 20K and 40K words) of the training data to understand the relative performance of the models as we keep on increasing the size of the annotated data.

3.1 Training Data

The training data includes manually annotated 3625 sentences (approximately 40,000 words) for both supervised HMM and ME model. A fixed set of 11,000 unlabeled sentences (approximately 100,000 words) taken from CIIL corpus[2] are used to re-estimate the model parameter during semi-supervised learning. It has been observed that the corpus ambiguity (mean number of possible tags for each word) in the training text is 1.77 which is much larger compared to the European languages (Dermatas et al., 1995).

3.2 Test Data

All the models have been tested on a set of randomly drawn 400 sentences (5000 words) disjoint from the training corpus. It has been noted that 14% words in the open testing text are unknown with respect to the training set, which is also a little higher compared to the European languages (Dermatas et al., 1995)

3.3 Results

We define the tagging accuracy as the ratio of the correctly tagged words to the total number of words. Table 1 summarizes the final accuracies achieved by different learning methods with the varying size of the training data. Note that the baseline model (i.e., the tag probabilities depends only on the current word) has an accuracy of 76.8%.

Method	Accuracy		
	10K	20K	40K
HMM-S	57.53	70.61	77.29
HMM-S+suf	75.12	79.76	83.85
HMM-S+MA	82.39	84.06	86.64
HMM-S+suf+MA	84.73	87.35	88.75
HMM-SS	63.40	70.67	77.16
HMM-SS+suf	75.08	79.31	83.76
HMM-SS+MA	83.04	84.47	86.41
HMM-SS+suf+MA	84.41	87.16	87.95
ME	74.37	79.50	84.56
ME+suf	77.38	82.63	86.78
ME+MA	82.34	84.97	87.38
ME+suf+MA	84.13	87.07	88.41

Table 1: Tagging accuracies (in %) of different models with 10K, 20K and 40K training data.

3.4 Observations

We find that in both the HMM based models (**HMM-S** and **HMM-SS**), the use of suffix information as well as the use of a morphological analyzer improves the accuracy of POS tagging with respect to the base models. The use of MA gives better results than the use of suffix information. When we use both suffix information as well as MA, the results is even better.

HMM-SS does better than **HMM-S** when very little tagged data is available, for example, when we use 10K training corpus. However, the accuracy of the semi-supervised HMM models are slightly poorer than that of the supervised HMM models for moderate size training data and use of suffix information. This discrepancy arises due to the over-fitting of the supervised models in the case of small training data; the problem is alleviated with the increase in the annotated data.

As we have noted already the use of MA and/or suffix information improves the accuracy of the POS tagger. But what is significant to note is that the percentage of improvement is higher when the amount of training data is less. The **HMM-S+suf** model gives an improvement of around 18%, 9% and 6% over the **HMM-S** model for 10K, 20K and 40K training data respectively. Similar trends are observed in the case of the semi-supervised HMM and the ME models. The use of morphological restriction (**HMM-S+MA**) gives an improvement of 25%, 14% and 9% respectively over the **HMM-S** in case of 10K, 20K

[2] A part of the EMILE/CIIL corpus developed at Central Institute of Indian Languages (CIIL), Mysore.

and 40K training data. As the improvement due to MA decreases with increasing data, it might be concluded that the use of morphological restriction may not improve the accuracy when a large amount of training data is available. From our empirical observations we found that both suffix and morphological restriction (**HMM-S+suf+MA**) gives an improvement of 27%, 17% and 12% over the HMM-S model respectively for the three different sizes of training data.

The Maximum Entropy model does better than the HMM models for smaller training data. But with higher amount of training data the performance of the HMM and ME model are comparable. Here also we observe that suffix information and MA have positive effect, and the effect is higher with poor resources.

Furthermore, in order to estimate the relative performance of the models, experiments were carried out with two existing taggers: TnT (Brants, 2000) and ACOPOST[3]. The accuracy achieved using TnT are 87.44% and 87.36% respectively with bigram and trigram model for 40K training data. The accuracy with ACOPOST is 86.3%. This reflects that the higher order Markov models do not work well under the current experimental setup.

3.5 Assessment of Error Types

Table 2 shows the top five confusion classes for HMM-S+MA model. The most common types of errors are the confusion between proper noun and common noun and the confusion between adjective and common noun. This results from the fact that most of the proper nouns can be used as common nouns and most of the adjectives can be used as common nouns in Bengali.

Actual Class (frequency)	Predicted Class	% of total errors	% of class errors
NP(251)	NN	21.03	43.82
JJ(311)	NN	5.16	8.68
NN(1483)	JJ	4.78	1.68
DTA(100)	PP	2.87	15.0
NN(1483)	VN	2.29	0.81

Table 2: Five most common types of errors

Almost all the confusions are wrong assignment due to less number of instances in the training corpora, including errors due to long distance phenomena.

[3] http://maxent.sourceforge.net

4 Conclusion

In this paper we have described an approach for automatic stochastic tagging of natural language text for Bengali. The models described here are very simple and efficient for automatic tagging even when the amount of available annotated text is small. The models have a much higher accuracy than the naïve baseline model. However, the performance of the current system is not as good as that of the contemporary POS-taggers available for English and other European languages. The best performance is achieved for the supervised learning model along with suffix information and morphological restriction on the possible grammatical categories of a word. In fact, the use of MA in any of the models discussed above enhances the performance of the POS tagger significantly. We conclude that the use of morphological features is especially helpful to develop a reasonable POS tagger when tagged resources are limited.

References

A. Dalal, K. Nagaraj, U. Swant, S. Shelke and P. Bhattacharyya. 2007. *Building Feature Rich POS Tagger for Morphologically Rich Languages: Experience in Hindi*. ICON, 2007.

A. Ratnaparkhi, 1996. *A maximum entropy part-of-speech tagger*. EMNLP 1996. pp. 133-142.

D. Cutting, J. Kupiec, J. Pederson and P. Sibun. 1992. *A practical part-of-speech tagger*. In Proc. of the 3rd Conference on Applied NLP, pp. 133-140.

E. Dermatas and K. George. 1995. *Automatic stochastic tagging of natural language texts*. Computational Linguistics, 21(2): 137-163.

M. Shrivastav, R. Melz, S. Singh, K. Gupta and P. Bhattacharyya, 2006. *Conditional Random Field Based POS Tagger for Hindi*. In Proceedings of the MSPIL, pp. 63-68.

P. R. Ray, V. Harish, A. Basu and S. Sarkar, 2003. *Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Processing*. ICON 2003.

S. Singh, K. Gupta, M. Shrivastav and P. Bhattacharyya, 2006. *Morphological Richness Offset Resource Demand – Experience in constructing a POS Tagger for Hindi*. COLING/ACL 2006, pp. 779-786.

T. Brants. 2000. TnT – *A statistical part-of-sppech tagger*. In Proc. of the 6th Applied NLP Conference, pp. 224-231.