

Part of Speech Tagging for Bengali with Hidden Markov Model

Sandipan Dandapat, Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India 721302

{sandipan,sudeshna}@cse.iitkgp.ernet.in

Abstract

This report describes our work on Bengali Part-of-speech tagging (POS) for the NLP AI Machine Learning contest 2006. We use a Hidden Markov Model (HMM) based stochastic tagger. The tagger makes use of morphological and contextual information of words. Since only a small labeled training set is provided (41,000 words), a HMM based approach does not yield very good results. In this work, we have used a morphological analyzer to improve the performance of the tagger. Further, we have made use of semi-supervised learning by augmenting the small labeled training set provided with a larger unlabeled training set (100,000 words). The tagger has an accuracy of about 89% on the test data provided.

1 Introduction

Part-of-Speech (POS) tagging for natural language texts are developed using linguistic rules, stochastic models and a combination of both (hybrid taggers). Stochastic models (DeRose, 1988; Cutting, 1992; Dermatas, 1995; Mcteer 1991; Merialdo, 1994) have been widely used POS tagging for simplicity and language independence of the models. Among stochastic models, bi-gram and tri-gram Hidden Markov Models (HMM) are quite popular. TNT (Brants, 2000) is another widely used stochastic trigram HMM tagger which uses a suffix analysis technique to estimate lexical probabilities for unknown tokens based on properties of words in the training corpus which share the same suffix. Development of a stochastic tagger requires 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. The problem is quite difficult here due to lack of such large annotated corpus.

Simple HMM models do not work well when small amount of labeled data are used to estimate the model parameters. Additional information are coded into HMM model to achieve high accuracy for POS tagging (Cutting, 1992). For example, Cutting et al (1992) propose an HMM model that uses a lexicon and an untagged corpus for accurate and robust tagging. The semi-supervised (Cutting, 1992; Kupiec, 1992; Merialdo, 1994) model makes use of both labeled training text and some amount of unlabeled text. Small amount of labeled training text are used to estimate a model. Then the unlabeled text are used to find a model which best describe the observed data. The well known Baum-Welch algorithm is used to estimate the model parameters iteratively until convergence.

2 Our Approach

Our approach is a combination of both supervised and unsupervised learning for training a bi-gram HMM. 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. We are using a Morphological Analyzer (R. Maitra, 2004) for Bengali in our POS tagging technique. The Morphological Analyzer takes a word as input and gives all possible POS tags for the word. This in turn restricts the set of possible tags for a given word.

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 , which satisfies:

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

For each of the model, the model parameters are estimated based on the provided training data. Unlabelled text is used to re-estimate the model parameters during semi-supervised learning. The model parameters are re-estimated using Baum-Welch algorithm which is a special case of the Expectation Maximization algorithm.

2.1 Taggers

We define the baseline model as the one where the tag probabilities depend only on the current word:

$$P(t_1 \dots t_n | w_1 \dots w_n) = \prod_{i=1, n} P(t_i | w_i) \quad (2)$$

In this model each word in the test data will be assigned the tag which occurred most frequently for that word in the training data.

Two taggers have been implemented based on bigram HMM 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 to find the most probable tag sequence for a given sequence of words. Both the models use a set of 27 tags (T) for every word in a sentence and the most probable tag sequence is determined using the Viterbi algorithm.

In order to further improve the tagging accuracy, we integrate morphological information with the HMM-S and HMM-SS 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 performs better for both the HMM models even when only small labeled training text is available. We shall call these new models **HMM-S + MA** and **HMM-SS + MA**.

2.2 Unknown Word Hypothesis

While processing unknown text, the tagger often encounters unknown words. In our model, words are unknown when they are not included in the training text, and neither the morphological restriction of tags for that particular word is available (i.e. unknown to the morphological analyzer). When the words are not in the training text, the transition and symbol emission probabilities are estimated by add-one smoothing. All the experiments conducted in section 2.3 are based on the above smoothing criteria. 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).

2.3 Experiments

All parts of the training text are used to estimate the model parameters of the supervised bi-gram HMM. Some unlabeled texts along with the training text are used to estimate the model parameters of the bigram semi-supervised HMM. The most probable tag sequence is estimated based on equation (1) and the models described in section 2.1. The models HMM-S+MA and HMM-SS+MA use a morphological analyzer to restrict the possible tags of a word, but in reality it is possible that the morphological analyzer is neither complete nor sound. Two different experiments are carried out on both supervised and semi-supervised model to understand the relative performance of the tagger under two situations: (1) when the morphological restriction is neither sound nor complete (we shall call these HMM-S + IMA and HMM-SS + IMA) and (2) when the morphological restriction is sound and complete (we shall call these HMM-S + CMA and HMM-SS + CMA). All the experiments are tested on the same test set.

One experiment is conducted based on unknown word hypothesis for further improvement of HMM-S + IMA model as described in the previous section. We also carry out some experiments with the freely available ACOPOST¹ tagger, which is based on a supervised trigram HMM with suffix tree information for unknown words.

2.4 Data Used for the Experiments

The supervised model parameters are estimated from the annotated text corpus (provided during the competition²). The training data includes 3085 sentences (approximately 41,000 words). For unsupervised learning, the HMM learned through supervised training is considered as the initial model. The model parameters are then re-estimated using the Baum-Welch algorithm by training it over a fixed set of 11,000 unlabeled sentences (approximately 100,000 words) taken from CIIL corpus.

¹ <http://acopost.sourceforge.net/>

² The training data includes both the data provided in non-privileged and privileged mode.

2.5 Tag Set and Corpus Ambiguity

The training data has been manually annotated using a tag set consisting of 27 grammatical tags. The *corpus ambiguity* is defined as the mean number of possible tags for each word of the corpus. It has been observed that the corpus ambiguity in the training text is 2.09 which is much larger compare to the European languages (Dermatas, 1995). This may be one of the reasons of relatively lesser accuracy of the tagging task.

3 System Performance

In order to measure the performance of the system we use the annotated test data (developmental data set) that has been provided by NLP AI. The development data set consists of 365 sentences (approximately 6000 words). It has been noticed that some of the data in the development data set are also present in the training set. This may lead to a slight amplification in the accuracy measured on the test set compared to the actual accuracy of the systems.

The performance of the system has also been measured on the test data. The test data set consist of 458 sentences (5127 words) which is totally different from training data.

3.1 Tagging Accuracy for the Development Data Set

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 when the complete annotated data is used for training. Figure 1 shows the improvement in accuracy of each of the models along with the increase in the size of annotated training data. Note that the baseline model has an accuracy of 69.11%.

Method	Precision
Baseline	69.11
ACOPOST	83.45
HMM-S	74.53
HMM-S + IMA	78.65
HMM-S + CMA	88.83
HMM-SS	73.77
HMM-SS + IMA	77.98
HMM-SS + CMA	89.65

Table 1. Tagging accuracies of different models

It is interesting to note that the use of morphological restriction (HMM-S + CMA and HMM-SS+CMA) gives an improvement of around 15% over the HMM-S model. Even the incomplete

morphological analyzer boosts up the performance of both the models by around 4%.

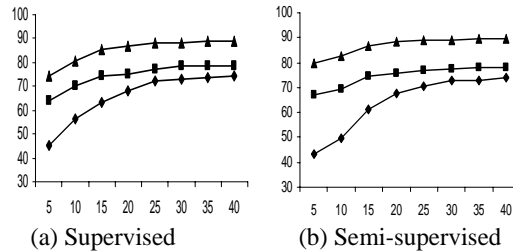


Figure 1. The accuracy growth of two different tagging models are shown in the figure; horizontal axis represent the *size of the training corpus (in 1000x words)* and the vertical axis represents the *Tagging Accuracy(in %)*. The lines with diamond, square and triangle respectively denotes the performance of a model without morphological restriction, with a incomplete morphological restriction and with complete and sound morphological restriction on the tags.

Another significant observation is that the accuracy of the HMM-SS and HMM-SS + IMA models are slightly less than the HMM-S and HMM-S + IMA model. Usually, the semi-supervised model (HMM-SS) has a higher accuracy than the supervised model (HMM-S), because the former method takes care of the problem of over-fitting, which otherwise a serious problem with supervised learning over a small training set. However, as mentioned previously, due to the presence of the training data in the test set the accuracy of the supervised learning algorithm is slightly better than the unsupervised one. Nevertheless, the use of a sound and complete morphological restriction (HMM-SS + CMA) gives an improvement of 1% in semi-supervised model compare to the supervised HMM model.

It is also noted that assumption for unknown words in section 2.2 leads to a 4% rise in the tagging accuracy from 77.98% for the HMM-SS+IMA model to 81.43%.

In order to estimate the effect of over-fitting in the HMM-S model, we split the test data into three sets: (1) known – the texts that are also present in the annotated training set, (2) unknown – the texts that are not present in any of the training sets, and (3) seen – the text that is present in the unannotated training set but not in the annotated training set (note that this is meaningful only for the HMM-SS models). The results of the experiments on these three sets of data are shown in Figure 2. The experiments have been con-

ducted only on the HMM-S+CMA and HMM-SS+CMA models. The know, unknown and seen test sets consist of 165, 100 and 100 sentences respectively.

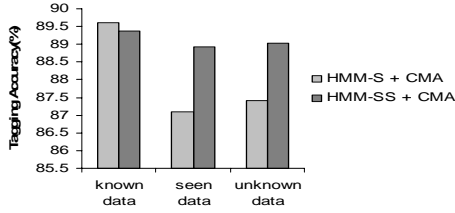


Figure 2. Tagging accuracy of known, seen and unknown data by both supervised and semi-supervised bigram HMM model

Though the supervised learning method performs better than semi-supervised learning method for the known data set, a higher accuracy (nearly 2%) has been achieved by semi-supervised learning for both seen data and unknown data over the supervised model. In contrast to our expectation, semi-supervised learning method performs slightly better for the unknown data set compared to the seen data set. This can be explained by the fact that the mean corpus ambiguity of the seen data is 1.23, whereas the mean corpus ambiguity of the known data set is 1.13.

3.2 Tagging Accuracy for the Test Data Set

The test data set has been tagged using **HMM-SS + IMA** as described in 2.1. We also use unknown word hypothesis to assume the POS tags for the unknown words as described in section 2.2. The overall tagging accuracy on the test data set is 84.32%. The precision and recall values for each of the grammatical classes have been listed in Table 2 for the test data set.

Type	Precision (%)	Recall (%)	Type	Precision (%)	Recall (%)
CC	83.93	91.56	RP	100	63.46
INF	100	12.50	SYM	100	99.78
JJ	75.20	76.42	UH	100	75.00
JVB	0	0	VFM	81.08	85.37
NEG	95.45	100	VJJ	0	0
NN	82.66	91.65	VNN	90.91	40.82
NNC	50.00	19.05	VRB	60.98	67.57
NNP	76.25	77.87	INTF	85.19	67.65
NVB	0	0	NLOC	85.16	77.30
PRP	95.72	93.18	NNPC	57.45	49.09
QF	73.33	37.29	PREP	84.16	80.19
QT	100	25.00	QFNUM	94.70	91.24
QW	88.89	53.33	VAUX	92.44	74.30
RB	71.96	69.37			

Table 2: Result obtained for the test set by semi-supervised learning along with morphological restriction (HMM-SS + IMA) and unknown word hypothesis.

4 Conclusion

In this report we have described an approach for automatic stochastic tagging of natural language text. The models described here are very simple and efficient for automatic tagging even when the amount of available labeled text is small. The models have a much higher accuracy than the naive baseline model. However, the performance of the current system is not as good as that of the best POS-taggers available for English and other European languages. The best performance is achieved for the semi-supervised learning model along with morphological restriction on the possible grammatical categories of a word.

References

- L. E. Baum. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, 3:1-8, 1972.
- T. Brants. 2000. TnT – A statistical part-of-speech tagger. In *Proc. of the 6th Applied NLP Conference*, pp. 224-231.
- 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.
- S. J. DeRose. 1988. Grammatical Category disambiguation by statistical optimization. *Computational Linguistics*, 14:31-39.
- J. Kupiec. 1992. Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language*, 6.
- R. Maitra. 2000. Inflectional Morphological Analyzers for Hindi and Bengali Languages. M. Tech. Thesis. Dept. of CSE. Indian Institute of Technology Kharagpur.
- M. Mcteer, R. Schwartz and R. Weischedel. 1991. Empirical studies in part-of-speech labeling. *Proceedings Of the 4th DARPA Workshop on Speech and Natural Language*, pp. 331-336.
- B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155-171.
- A. J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transaction on Information Theory*, 13:260-269.