A Memory Based POS Tagger for Bengali

Kamal Sarkar JUKAMAL2001@YAHOO.COM
Arup Ratan Ghosh GHOSH.MAILME@GMAIL.COM
Computer Science & Engineering Department,
Jadavpur University,
Kolkata – 700 032, India

Abstract

We present in this paper a memory based learning technique for Bengali POS tagging. Our developed memory based POS tagger has been compared with a trigram HMM based POS tagger and a baseline tagger and the results show that the performance of our proposed POS tagger for Bengali performs better than a HMM based tagger and a baseline tagger which it is compared to.

1. Introduction

Part-of-Speech (POS) tagging is the task of assigning grammatical categories (noun, verb, adjective etc.) to words in a natural language sentence. POS tagging can be used in parsing, word sense disambiguation, information extraction, machine translation, question answering, chunking etc. Assigning a tag to each word in a sentence is not a trivial problem, because many of the most common words of a language are ambiguous (for example, *can* can be an auxiliary ('to be able'), a noun ('a metal container'), or a verb ('to put something in such a metal container')).

Most POS tagging algorithms fall into one of two classes: rule-based tagger and stochastic taggers. For Bengali language, the POS taggers which have been reported in the literature are basically stochastic POS taggers.

The previous works on Bengali POS tagging use Hidden Markov Model(HMM) (Brants, 2000; Sarkar et al.,2012), Maximum entropy model (Dandapat et al., 2007), conditional random fields (Ekbal et al., 2007) to find out the solutions for Bengali POS tagging problem. In our own work, we have used Memory-Based Learning (MBL) techniques for Bengali POS tagging. Unlike the earlier HMM based POS tagging approaches (Brants, 2000; Sarkar et al.,2012), our proposed memory based POS tagger is less affected by the sparse data problem since MBL provides a solution to the sparse data problem via an implicit similarity-based smoothing scheme. Moreover memory based learner can directly handle string features that facilitates defining the context of the word.

2. Proposed Memory Based POS Tagging Approach

Memory based learning (MBL) is a type of supervised inductive learning technique based on similarity-based classification (Zavrel and Daelemans, 1999). We use *K*-nearest neighbor (KNN) algorithm as a memory based learner for implementation of our proposed Bengali POS tagger. The construction of the POS tagger for a specific corpus is achieved in the following way. Given an annotated corpus, the training set is prepared by creating the feature vectors for known words (words occurring in the annotated corpus). A feature vector for a word is created with the values of the defined features shown in the table 1. The feature vectors for the known words are labeled with the tag (considered as class label) associated with the corresponding occurrence of the word in the annotated corpus. During the training phase, the labeled pattern vectors are created and added to the memory by the learner. During tagging, each word in the text to be tagged is represented as a feature vector with values of the defined features shown in table 1 and the unlabeled test pattern is compared with the labeled patterns stored in the memory and the distances are computed using Euclidean measure. Finally, the input is labeled as the class that is the mode of the classes of *K nearest* pattern vectors selected from the memory. For our work, *K* is set to 3 for which we obtain the best average results on our dataset. The set of features used for forming a feature vector corresponding to a word occurred in the annotated corpus is shown in table 1.

Appearing in Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013. Copyright 2013 by the author(s)

Table 1. Feature set for vector representation of an instance (or example)

FEATURE NAMES	DESCRIPTION	TYPE
HYPHENATED	whether the word is hyphenated.	Binary
IS_FOUR_DIGIT	whether the word is a four digit number	Binary
IS_TWO_DIGIT	whether the word is a two digit number	Binary
NUMERIC_OTHER THAN_TWO_OR_ FOUR_DIGIT	whether the word is not two digit or four digit number	Binary
CONTAINING SPECIAL SYMBOL	Whether the word contains any non alpha-numeric symbol	Binary
WORD_LENGTH_ CUT-OFF	Whether the word length is more than three characters	Binary
INFLECTION-LIST	Whether the longest possible suffix of a word is found in the pre-computed inflection list for a particular lexical category	Nominal (category name)
RARE_WORD	Whether the word is rare (frequency <= 2) in the training set or not	Binary
TAG-FREQUENCY DISTRIBUTION	Number of times a word occurred with each possible tag in the training corpus	Continuo us
SUFFIX	Suffix of the word(suffix length = 5)	String
PREFIX	Prefix of a word (prefix length = 3)	String
LEFT CONTEXT	All the previous words preceding the current word in the sentence.	String
RIGHT CONTEXT	All the words following the current word in the sentence.	String
PREV_N_TAGS	The pervious-tag feature will hold the tag label of the previously tagged word	String

3. Evaluation and Results

Accuracy of tagging is computed as the ratio of number of matched tags to the total number of tags with duplicates. We test our proposed tagger on NLTK dataset[1]. The NLTK dataset contains a total of 895 Bengali sentences tagged using 26 POS tags. The performance of our developed tagger is compared with a trigram HMM based Bengali tagger presented in (Sarkar et al., 2012) and the unigram baseline tagger that assigns each token to the class it occurred in most often in the training set. The HMM based tagger has been developed according to the settings as described in (Sarkar et al., 2012). For evaluation of each system, 10-fold cross validation is done. The performance scores for our proposed memory-based tagger, the HMM tagger and the baseline tagger are 80.77, 78.68 and 67.37 respectively.

4. Conclusion

A memory based POS tagger for Bengali has been presented in this paper. Since the memory based learner is slow in nature, we can apply an efficient *k* nearest neighbor search algorithm (Zhang and Srihari, 2004) for speeding up the proposed POS tagger.

References

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

Dandapat, S., Sarkar, S., Basu, A. Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor scenario, *Proceedings of the Association for Computational Linguistic*s, pp. 221-224, 2007.

Ekbal A., et al. Bengali part of speech tagging using conditional random field. *In Proceedings of the 7th International Symposium of Natural Language Processing(SNLP-2007),* Thailand, pp. 131-136, 2007.

Sarkar, K., Gayen, V. A Practical Part-of-Speech Tagger for Bengali, *In the proceedings of third International Conference on Emerging Applications of Information Technology (EAIT 2012)*

Zavrel, J. and W. Daelemans.. Recent advances in memory -based part-of-speech tagging. *In VI Simposio Internacional de Comunicacion Social*, pages 590–597, 1999

Zhang, B., and Srihari. S. N. Fast k-nearest neighbor classification using cluster-based trees. *Pattern Analysis and Machine Intelligence,* IEEE Transactions on 26.4 (2004): 525-528.

[1] http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml