Experiments with POS Tagging Code-mixed Indian Social Media Text

Prakash B. Pimpale
KBCS, CDAC Mumbai
prakash@cdac.in

Raj Nath Patel
KBCS, CDAC Mumbai
rajnathp@cdac.in

Abstract

This paper presents Centre for Development of Advanced Computing Mumbai's (CDACM) submission to NLP Tools Contest on POS Tagging For Code-mixed Indian Social Media Text (POSCMISMT) 2015 (collocated with ICON 2015).

We submitted results for Hindi, Bengali, and Telugu languages mixed with English. In this paper we have described our approaches to the POS tagging techniques we exploited for this task. Machine learning techniques have been used to POS tag the mixed language text. Distributed representations of words in vector space (word2vec) for feature extraction and Log-linear part-of-speech tagging for POS tagging have been tried. We report our work on all three languages Hindi, Bengali and Telugu mixed with English.

1 Introduction

In this paper we present our experiments for POS tagging code-mixed Indian social media text. The evolution of social media platforms – such as blogs, micro-blogs (e.g., Twitter), and chats (e.g., Facebook messages) – has created many new sources for information access and language technology. But the same has presented many new challenges, making it one of the prime present-day research areas.

Most of the Indians and many other non English speakers across the world don't always use Unicode to write something in social media, they make use of transliteration and frequently insert English elements through code-mixing and anglicisms, and often mix multiple languages to express their thoughts.

English still is the principal language for social media communications, but this kind of multilingual content is growing and calls for the development of language technologies for languages other than English. If we observe twitter and facebook feeds of Indians, it's full of frequent code-mixing. It's not a surprise given the diverse linguistic culture across India. But this poses additional difficulties for automatic Indian social media text processing.

Part-of-speech (POS) is an essential prerequisite for most of the NLP applications. POS tagging for English text is now a mature tool and a lot of work is in progress for English social media text. The work on POS tagging for code-mixed languages is a recent topic and not much has been done for it in the Indian context.

Vyas et. al. (2014) created a multi-level annotated corpus of Hindi-English code-mixed text from Facebook forums, and explored language identification, back-transliteration, normalization and POS tagging of this data. They used tools like CRF++ based tagger and Stanford POS tagger for experimentation. Jamatia and Das (2014) created good amount of labeled corpus using amazon mechanical turk and bootstrapping. They experimented with various machine learning techniques for POS tagging and reported Random Forest to be the best one among what they tried.

We have used Stanford log-linear Part-Of-Speech tagger (Toutanova et. al., 2003) for tagging, word2vec (Mikolov et. al., 2013) for feature extraction and WEKA (Hall et. al., 2009) for machine learning.

The rest of the paper is organized as follows. In section 2, we discuss data-sets followed by experiments and results in section 3. The submission to shared task has been discussed in section 4 and the conclusion and future work in section 5.

2 Data-sets

We have used 80% of the training data shared by POSCMISMT detailed in Table 1 for the experiments.

Testing for the experiments was done using remaining 20% data. But the system for final submission was trained using the complete data shared. The submitted systems were evaluated against test corpus, by the organizers.

	Hi+En				Bn+En				Ta+En			
	Total	**Hi**	**En**	**O**	**Total**	**Bn**	**En**	**O**	**Total**	**Ta**	**En**	**O**
Train	15955	5546	6178	4231	24638	8330	9973	6335	4315	1716	1969	630
Test	11212	411	8553	2248	13561	46715	5459	3431	2255	1155	819	281

Table 1. Train and Test data – Number of tokens. Hi: Hindi, Bn: Bengali, Ta: Telugu, En: English, O: Others i.e punctuations, acronyms, named entities, mixed language words and other universal symbols.

3 Experiments and Results

We have used Stanford POS tagger (Toutanova et. al., 2003) available on Stanford Natural Language Processing group's website for constrained training and result submission. And unconstrained training and result submission has be done using word2vec (Mikolov et. al., 2013) and WEKA (Hall et. al., 2009).

3.1 POS Tagging using Stanford POS tagger: Constrained

The constrained result submission needed to be done using system trained on data provided by POSCMISMT only. We trained Stanford POS tagger using train data provided. Basically this POS tagger learns a log-linear conditional probability model from tagged text, using a maximum entropy method. The POS tag of input word is then decided by the model based on context and surrounding tags of the word. The architecture (arch property) we used for training was: words(-2, 2), order(1), prefix(6), suffix(6), unicodeshapes(1).

3.2 POS tagging using Machine Learning: Unconstrained

We used WEKA to experiment with application of various machine learning techniques to the POS tagging problem. Various combinations of following word features were used for training and testing the system.
F1) Language of the word
F2) Language of the previous word
F3) Language of the next word
F4 & F5) POS tags of the previous 2 word
F6 & F7) POS tags of the next 2 word's similar words
F8) Position of the word in sentence

In sentence with length L, words located at positions 1, 2, L and L-1 were assigned required number of default feature values for previous and next languages and POS tags.

For POS tag of the next word's similar word we used distributed representations of words in vector space. We trained a word2vec model using sentences from train and test set detailed in Table 1. And whenever we needed POS tag of the next word, we followed one of the following steps.

1) The word was looked up in the list from training data, if it was found, the most frequent POS of that word was used. If it was not in the list, we followed next step.
2) The nearest word list was fetched using word2vec model trained on train and test set. And the most frequent available POS tag of the nearest word was used instead. If this failed i.e no nearest word was found in the training set, we followed next step.
3) The most frequent POS tag from the training set was used instead.

We reserved 20% of the train data for the purpose of evaluation. Table 2 details some of significant results we obtained during the experiments on this test set.

	Hi+En	**Ta+En**
Decision Tree J48	44.60	50.30
Decision Tree Random Forest	43.00	47.00
Naive Bayes	40.40	46.30
Multilayer Perceptron	39.30	41.10

Table 2. Experimental Results: F1 Measures in %.

4 Submission to the Shared Task

From the Table 2 we can see that J48 decision tree gave better results and so that was used to train the final system for submission. The submitted results were evaluated by organizers. These results by organizers have been detailed in the table 3 below. The results for Hi+En unconstrained are very low and on prima facie it looks like an error. We plan to investigate that.

5 Conclusion and Future Work

In this paper, we presented two techniques for POS tagging of code-mixed Indian social media text. The method used for constrained submission is performing well, but lack of the quality training data doesn't allow to do much with it. On other hand, use of the distributed vector representation of words in feature engineering may allow us to use unlabeled data for training. The results are encouraging and future work can be focused on obtaining more social media corpus and using that for the better feature representation.

	Constrained	**Unconstrained**
Hi+En	71.11	6.84
Bn+En	75.46	Not Submitted
Ta+en	71.04	48.03

Table 3. Consolidated Results. Accuracy in %.

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Jamatia, A., & Das, A. (2014). Part-of-Speech Tagging System for Indian Social Media Text on Twitter. स ocial-ई ndia 2014, 2014, 21.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In*Advances in neural information processing systems* (pp. 3111-3119).

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1(pp. 173-180). Association for Computational Linguistics.

Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014, October). Pos tagging of english-hindi code-mixed social media content. In Proceedings of the First Workshop on Codeswitching, EMNLP.