AMRITA_CEN@ICON-2015: Part-of-Speech Tagging on Indian Language Mixed Scripts in Social Media

Anand Kumar M
Center for Excellence in Computational
Engg and Networking,
Amrita Vishwa Vidyapeetham,
Coimbatore, India.
m_anandkumar@cb.amrita.edu

Soman K P
Center for Excellence in Computational
Engg and Networking,
Amrita Vishwa Vidyapeetham,
Coimbatore, India.
kp_soman@amrita.edu

Abstract

Large volumes of unstructured text are generated in Social media platforms like blogs, Facebook and Twitter. Stylistic and linguistic variations are the major challenges in handling these texts. In multilingual nation like India, Code mixing is a usual style observed in social media conversations. Multilingual users often use the Roman script, which is a popular mode of expression, instead of native script for generating content in Social media platform. Plenty of Roman transliterated data available on the Web for Indian languages. This paper explains our approach on POS tagging on mixed scripts in the ICON-2015 tools contest. The utterances are written in Roman script and the word level language is given as additional information. SVM based machine learning system with relevant features is developed for tagging the words with its corresponding part-of-speech tags. We also explore few experiments with mixed script word embeddings as features to train the SVM based classifier.

1 Introduction

Currently, the extensive use of internet in multilingual population provides plenty of opportunities for major and exhaustive analyses of mixed language use in online media platform [1]. Compared with standard text corpora, the text used in social media illustrates lot of differences in style and variations. The primary challenges in handling the mixed scripts are spell variations, phonetic typing, creative spelling and abbreviations [2] [3]. In this paper we addressed the POS Tagging problem in mixed social media scripts. We Observe that Indian social media mixed script often contains English as the mixing language. Processing and analyzing mixed-language data requires identification of languages at the word level.

Part-of-Speech tagging is considered as a key task in most of the language processing applications. POS taggers for Indian languages are well studied discipline in language processing research. Existing Indian language POS taggers for normal text is not directly suitable for Social media text because of its informal style and mixed nature. Most of the social media content in Indian languages is generated in Roman mixed form instead of native script. POS tagging of these mixed scripts are challenging and interesting area of research in social media text analytics especially from the multilingual nation like India. The ICON-2015 tools contest addresses the POS tagging for mixed script task in three Indian languages viz, Hindi, Bengali and Telugu. Here we used Support Vector Machines based classifier for training the developed system. SVMs are successfully applied to Indian language processing [4] [5] [6] [12]. Mixed script POS tagging is a less studied area of research in language processing. Very few related works [8] [9] [10] [11] are only exists in Code mixed POS tagging.

2 Dataset description

The ICON-2015 NLP tools contest, POS tagging for Code mixing text is designed for evaluating team’s ability to identify the POS tags for code three (Hindi, Bengali and Telugu) mixed Indian languages. Organizers released the code mixed train and test set for each languages.

Table 1. Dataset details

	Training			Testing		
	Hindi	**Bengali**	**Telugu**	**Hindi**	**Bengali**	**Telugu**
Utterances	728	2837	638	376	1458	279
Tokens	15839	24638	4316	11212	13561	2254
Average	21.76	8.68	6.76	29.82	9.3	8.08

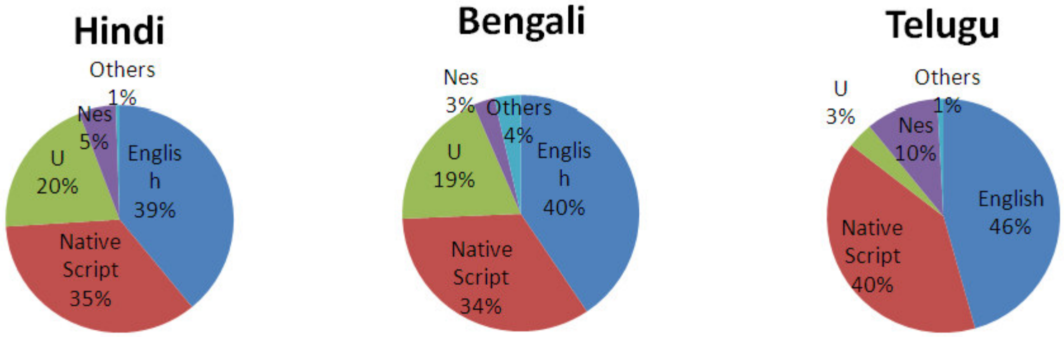

Figure 1. Percentage of different word level information exists in all the three languages.

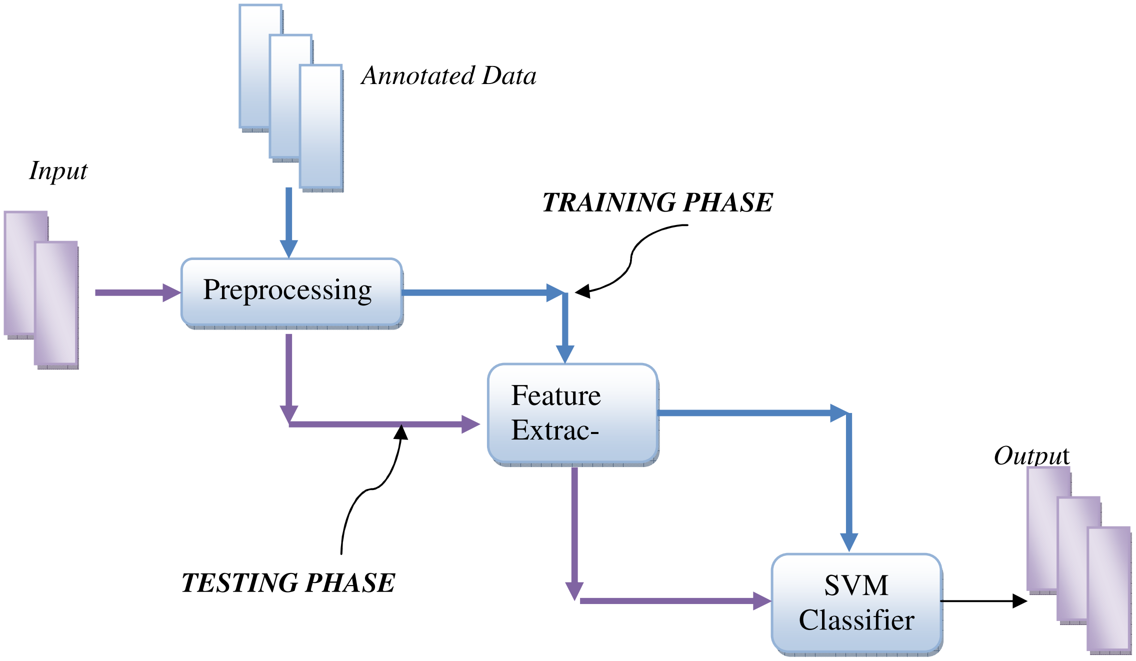

Figure 2. Methodology

Training dataset contains tokens in Roman form and their word level information along with its Pos tag. Word level information is language of the token or named entity or universal tags.etc. Test data contains only token and its word level information. Number of utterance and tokens size (excluding utterance break) of train and test dataset for all the three languages are illustrated in Table.1. Average tokens in each utterance is also calculated and shown in Table.1. Among the three languages, Hindi contains the highest average number of tokens per utterance. Approximately around fifty present (utterance size) of the training data is given for testing in all the three languages. Figure 1. illustrates the percentage of different word level information exist in all the three languages. The interesting fact from the figure.1 is that, in code mixed text, most of the tokens are in English compared with their native script. This clearly shows the significant influence of English in Indian language's Code-mixed Social media text. Hindi and Telugu contain *universal* tag, where Telugu fails to get. Named entities are higher in Telugu compared with Hindi and Bengali languages.

Table 2. Primary POS tag counts

	Hindi	Bengali	Telugu
N_NN	2349	5161	1112
RD_PUNC	1091	3984	1
V_VM	1997	3055	684
N_NNP	1134	2437	505
PSP	1354	1414	81

Table.2. explores the primary POS tag counts in the training dataset. We have taken 5 major POS categories, which are, Noun, Punctuation, Verb, Proper noun and post position.

3 Methodology

The system is developed with the training dataset provided by organizers. We have submitted constrained and unconstrained runs for all the three languages. The overall methodology is illustrated in Figure.2. In pre-processing, utterance break (<*ub*>) is included in between the utterances.

Constrained System*:* The preprocessed data of each language is given to the feature extraction module which extracts the defined features (explained in subsection 3.1). Then, these feature vectors are trained with the SVM based classifier SVMLight [7].

Table 3. Feature set Details

Features	Symbols
Token features	
Word Features (unigram, Bigram, Trigram)	w_{-1}, w_{0}, w_{+1}
Language tags(unigram, Bigram, Trigram)	l_{-1}, l_{0}, l_{+1}
POS tags	p_{-1}, p_{0}, p_{+1}
Prefixes and Suffixes	$P_3\, S_3,$
Binary Features	
Starts with #	*SH*
Starts with @	*SA*
Starts Capital	*SC*
Contains Punctuation	*CP*
Full Capital	*FC*
Contain "http"	*CH*
Punctuation Features	
Contain Apostrophe	*CA*
QM, Hyphen, Comma, Parenthesis, square bracket, Colon	*CQ,CHy,CC ,CP, CS,CCo*
Other Features	
Length	*l*
Position	*pos*

Unconstrained System*:* In this unconstrained method, we have developed a single system with merging all the training data of Hindi, Bengali and Telugu. The training data consists of 44793 tokens from 4203 utterances, with the average 10.66 tokens per utterance. The testing data consists of 27027 tokens from 2113 utterances, with the average of 12.79 per utterance. The main reason behind the integration of training dataset is that, we observed near 40% of English words in the given code-mixed training data of each language. We believe that the English tokens in one language will be useful for improving the performance of other language systems. More importantly the word level tagset and POS tagset used in Hindi and Bengali are same. But in the case of Telugu, number of word level tags and the Pos tags are different. *Universal* tag and *Punctuation* tags are not exist in Telugu language. We have used the same features of constrained system to train the classifier. Finally, single model is used to predict the POS tags of test dataset in each language.

3.1 Feature Extraction

In this system development, we have provided more importance to feature extraction as this decides the performance of the machine learning classifier. The common features like, words, prefixes and suffixes of the word, binary features, punctuation feature are used to train the classifier. For prefix and suffix feature, first and last three characters of the current token is considered. The punctuation mark such as question mark, comma, and parenthesis are also taken as feature. For training the system the current token, the token which are above and below are also taken as feature for deciding the current token's part-of-speech. Detailed feature set is illustrated in Table.3.

4 Experimentations and Results

We have developed constrained and unconstrained system for Indian code mixed data. Percentage of unknown words in the test dataset is given in the Table.4. Telugu test data contains 45% of unseen words which is higher compared with other two languages. Before testing the system, we have done 10 fold cross validation in the train dataset. Constrained cross validation accuracies are listed in the Table.5. Due to the large number of unique words in Telugu dataset, cross validation fails and shows very less accuracy. Compared with Hindi language, Bengali system performs better in known and unknown words.

In unconstrained, we have tried two different systems, one is word embedding based and another is trained with the features explained in subsection 3.1. In word embedding based method, we feed all the training utterances to word2vec [13] tool, with the dimension *d = 10.* In order to capture the context, we have integrated previous and next vector with the current vector (so the final feature size is 30). We failed to include the word level information in word2vec and we have not tried for other dimensions, these are main drawbacks in our word embedding based system. Even though we have taken small *d* and fail to add word level information, the cross validated accuracy of unknown words are higher compared with the general rich feature based system. Unconstrained cross validated accuracies are explained in Table.6.

Table 4. Unknown word percentage in Test dataset

	Hindi	Bengali	Telugu
Known	79.28	78.45	54.45
Unknown	20.72	21.55	45.55

Table 5. Constrained Cross Validation Accuracies

	Hindi	Bengali	Telugu
Known	80.1	85.37	-
Unknown	23.98	39.89	-
Overall	66.18	75.37	-

Table 6. Unconstrained Cross Validation Accuracies

	Word2Vec	Rich Features
Known	78.50	82.30
Unknown	46.68	44.38
Overall	71.97	74.58

Table 7. Accuracies of constrained and unconstrained System

	Hindi	Bengali	Telugu
Constrained	75.58%	78.50%	73.30%
Unconstrained	73.66%	76.73%	68.16%

Table 8. Accuracies of major POS categories

	Constrained			Unconstrained		
	Hindi	Bengali	Telugu	Hindi	Bengali	Telugu
N_NN	79.83%	79.80%	77.64%	81.57%	76.18%	68.85%
RD_PUNC	98.30%	99.11%	0.00%	99.11%	98.93%	56.25%
V_VM	83.32%	81.87%	84.54%	88.78%	79.46%	69.81%
N_NNP	67.54%	55.47%	99.09%	59.94%	50.18%	80.91%
PSP	75.67%	89.38%	52.38%	60.62%	88.86%	53.97%

The organizers accuracy is shown in Table.7. Our performance is far better than the other team's performance submitted in this contest. Accuracies of major POS categories are also shown in Table.8.

5 Conclusion and Future Scope

The ICON-2015 tools contest addresses the POS tagging for mixed script task in three Indian languages viz, Hindi, Bengali and Telugu. Here we used Support Vector Machines based classifier for training the developed system. We have submitted constrained as well as unconstrained runs. Compared with all the unconstrained system, the Telugu system obtains 68% accuracy. The main reason for this is, Telugu tagset are conflict with other two languages. Even though we obtain 68% for Telugu, we are in the first place compared with all the teams. As a future work, we will be focusing in deep learning based features from unlabeled utterances. Deeper result analysis, like word level information based accuracies, will also help us to better understanding the existing system. Because of high influence in English terms, we would like to investigate how English Social media text POS tagger reacts in the performance of the Indian mixed scripts.

6 Acknowledgments

First and foremost, we would like to thank ICON-2015 organizers for organizing a wonderful NLP event in India. We also like to thank Dr. Amitav Das for initiating this research along with the tools contest.

Reference

[1] Dong Nguyen and A. Seza Dogruoz. 2013. Word level language identification in online multilingual communication. In Proceedings of the 2013 Con- ference on Empirical Methods in Natural Language Processing , pages 857–862

[2] Gambäck, Björn, and Amitava Das. "On Measuring the Complexity of Code-Mixing." *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*. 2014.

[3] Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query expansion for mixed-script information retrieval. In Proc. of SIGIR , pages 677–686. ACM Association for Computing Machinery.

[4] Anand Kumar, M., Rajendran, S., Soman, K.P. Tamil word sense disambiguation using support vector machines with rich features (2014) International Journal of Applied Engineering Research, 9 (20), pp. 7609-7620.

[5] Anand Kumar, M., Dhanalakshmi, V., Soman, K.P., Rajendran, S. Factored statistical machine translation system for English to Tamil language (2014) Pertanika Journal of Social Science and Humanities, 22 (4), pp. 1045-1061

[6] Kumar, M. Anand, S. Rajendran, and K. P. Soman. "Cross-Lingual Preposition Disambiguation for Machine Translation." Procedia Computer Science 54 (2015): 291-300.

[7] T. Joachims. Svmlight: Support vector machine. 1999, http://svmlight. joachims. org/, University of Dortmund, 19(4).

[8] 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*.

[9] Kalika Bali, Yogarshi Vyas, Jatin Sharma, and Monojit Choudhury. 2014. ''i am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching, EMNLP.

[10] Jamatia, A. and Das, A., 2014. Part-of-Speech Tagging System for Indian Social Media Text on Twitter. *सocial-ई ndia 2014*, *2014*, p.21.

[11] Chakma, K. (2014). Revisiting Automatic Transliteration Problem for Code-Mixed Romanized Indian Social Media Text. *सocial-ई ndia 2014*, *2014*, 42.

[12] Anand Kumar, M , Soman, K. P. "AMRITA_CEN@ FIRE-2014: Morpheme Extraction and Lemmatization for Tamil using Machine Learning." Proceedings of the Forum for Information Retrieval Evaluation. ACM, 2014.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.