Language Identification in Mixed Script Social Media Text

S. Nagesh Bhattu
CoE on Analytics, IDRBT
Castle Hills Road#1,Masab Tank
Hyderabad-500057, India
nageshbs@idrbt.ac.in

Vadlamani Ravi
CoE on Analytics, IDRBT
Castle Hills Road#1,Masab Tank
Hyderabad-500057, India
vravi@idrbt.ac.in

ABSTRACT

With the spurt in usage of smart devices, large amounts of unstructured text is generated by numerous social media tools. This text is often filled with stylistic or linguistic variations making the text analytics using traditional machine learning tools to be less effective. One of the specific problem in Indian context is to deal with large number of languages used by social media users in their roman form. As part of FIRE-2015 shared task on mixed script information retrieval, we address the problem of word level language identification. Our approach consists of a two stage algorithm for language identification. First level classification is done using sentence level character n-grams and second level consists of word level character n-grams based classifier. This approach effectively captures the linguistic mode of author in social texting enviroment. The overall weighted F-Score for the run submitted to FIRE Shared task is 0.7692. The sentence level classification algorithm which is used in achiving this result has an accuracy of 0.6887. We could further improve the accuracy of sentence level classifier further by 1.6% using additional social media text crawled from other sources. Naive Bayes classifier showed largest improvement (5.5%) in accuracy level by the addition of supplementary tuples. We also observed that using semi-supervised learning algorithm such as Expectation Maximization with Naive Bayes, the accuracy could be improved to 0.7977.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Information filtering.

Keywords

Classification

1. INTRODUCTION

With the proliferation of social tools like twitter, facebook, etc.. large volumes of text is being generated on a daily basis. Traditional machine learning tools used for text analysis such as Named Entity Recognization(NER) or Parts of Speech Tagging or parsing, are dependent on the premise that the text provided for them are in purer form. They achieve their objective using cooccurrence patterns of features. It has been observed by many studies that social media text when fed to such machine learning algorithm, is often plagued by the excessive out of vocabulary words(sparsity of features). The FIRE-2015 shared task1 addresses the language identification task as well as Named Entity recognition in the context of Indian social fora, where the number of languages used is more than 10 and all of them share vocabulary excessively.

Table 1: Number of Words in each languge

Lang	No of words	Lang	No of words
Bengali	2207	Hindi	2457
English	5115	Kannada	1302
Gujarati	1741	Malayalam	1862
Marathi	1265	Tamil	1886
Telugu	3462		

To understand the complexity of the task we have posed the primary problem of word-level language identification as a multi-class classification using word lists gathered for each language. These word-lists are obtained using the method suggested in Gupta et al. [2012]. We have converted these words into n-gram representation and built a classifier based on multi-class logistic regression McCallum [2002] and multi-class SVM (support vector machine) Crammer and Singer [2002]. We conducted this experiment by taking n-gram representation of each word of the training data as an instance (leaving the NE words). This experiment yielded an accuracy of 57%-54% respectively. The number of words used in these word lists are given in table 1. The multi-class Logistic functions likelihood is defined as below

$$p(y|x) = \frac{exp(\lambda_y.\mathbf{F}(x,y))}{\sum_{y'} exp(\lambda_{y'}.\mathbf{F}(x,y'))} \quad (1)$$

Here y is the label associated with instance x. The instance x is expressed in some feature representation $\mathbf{F}(x,y)$. In the current work, feature representation is n-gram represenation of words. λ_y are class specific parameters learnt during maximum likelihood based training process.

As the text segments are typically social media posts, the number of languages with in a text segment can be safely assumed to be 2. Using this corpus level prior knowledge, we built a two-stage classification algorithm. The first stage consists of identification of sentence level language. We used character-level n-grams of each of the sentences as training data for building sentence level classifier. We have used 1,2,3,4,5 grams of all the words in the sentence as the features. We divided the input training data into 80-20 splits using 5-fold cross validation. We built a multi-class classifier using softmax, Naive Bayes and Naive Bayes EM and SVM algorithms using the training data. Among these Naive Bayes EM is a semi-supervised learning algorithm, which

Table 2: Class-Wise Disribution in the training data

bn-en	215	ml-en	131
en	679	mr-en	200
gu-en	149	ta-en	318
hi-en	383	te-en	525
kn-en	272		

Table 3: Accuracy Results of Cross-Validation on Training Data

Method	Accuracy
Naive Bayes	0.7419
MaxEnt	0.8436
Multi-Class SVM	0.8123
Naive Bayes EM	0.7454

uses EM algorithm for improving the test data accuracy. In preparing such training data, we have removed the URLs, X, NE tagged words. The 5-fold cross-validation classification accuracy are reported in Table 3. We tried varying the number of n-grams to 3 and 4, which has the effect of depreciating the accuracy 3-6%. The class-wise distribution of documents in the training data is given in Table 2.

We get 82% accuracy when we applied the multi-class Logistic regression based classifier trained on the above data. We have experimented with latent Dirichlet based topic model for this with 100 as the number of topics which was not providing accuracy levels beyond 60%. The results of classification using training data are as given in table 3.

1.1 Word-Level Classification

After identifying the language pair used for writing a particular posting, we further build binary classifiers for each of the language pairs namely, *bn-en,gu-en,kn-en,hi-en, ml-en,mr-en,ta-en,te-en*. We use the words in the table 1 to build the binary classifiers. We used Logistic regression based binary classifier which are giving 92-94% accuracy on training data. The character n-grams (where n is set to 5) are used as features for the binary classifier. The approach suggested in Täckström and McDonald [2011] uses latent variable models for using document level sentiment ratings to infer sentence level classifier.

This approach of using word-level binary classifier works well as long as the length of the words is sufficiently long to capture the n-gram characteristics of the language of our interest. But, as we see, tweets often contain stylistic variations which reduce the length of words signinficantly. When the length of words is below 3, the words do not carry n-grams representative of target language. To address this problem we use words with in a window of two length on either side to make for the sparsity of features of shorter words. This is also a heuristic approach effectively used in the other works such as Han and Baldwin [2011].

Named Entity detection for short text is much harder task, as word-colocations and POS tagging do not work well with mixed script. Ritter et al. [2011] have proposed a solution based on word-clusters from a large collection of twitter corpus. We use the tool provided by authors of Ritter et al. [2011] for english tweets and a small lexicon of named entities for all the other languages for dealing with Named-Entity detection.

Table 4: Results as submitted for the test run

Tag	F1 Score
MIX	0.57
MIX-en-bn	0
MIX-en-kn	0
MIX-en-ml	0
MIX-en-te	0
NE	0.387409201
NE-ml	0
NE-L	0.2791
NE-O	0
NE-OA	0
NE-P	0.2187
NE-PA	0
Others	0
X	0.9555
bn	0.7749
en	0.831
gu	0
hi	0.6125
kn	0.8215
ml	0.8132
mr	0.745
ta	0.8582
te	0.6148
tokensAccuracy	77.5231
tokensCorrect	9302
utterances	792
utterancesAccuracy	18.0556
UtterancesCorrect	143
Average F-measure	**0.6845007667**
Weighted F-Measure	**0.769245171**

2. EXPERIMENTS

We have used McCallum [2002] for multi-class classification. The table 4 contains the F1 scores of language identification and Named entities. These are the results of test run submitted for the FIRE workshop. We reported F1 score which is a representative measure capturing both precision and recall. As we have adopted 2-stage algorithm for word level language identification, the classification accuracy of the first-level(sentence level) classificaion is most imprtant for the further processing. As we can see in the results there are languages like Gujarati which go misclassified by the classifier, having zero F1 score. Named entity detection is typically addressed using sequence level features which are quite unreliable in short-message context. Our test run results are limited to the presence in the training data.

2.1 Errors and Analysis

The error analysis is not complete without making the classifier further accuracte. In this regard, we have manually tagged the test data sentences to be of one of the languages

Table 5: Sentence Classification Accuracy on Test Data

Method	Accuracy Training-Data	Accuracy Training-Data-Expanded
Naive Bayes	0.7204	0.7751
MaxEnt	0.6887	0.7052
Naive Bayes EM	0.7684	0.7977

of our interest mixing with english. The authors are confident in tagging 6 of these languages, we depended on other resources for distinguishing malayalam and tamil. As this shared task largely focuses on english being the mixing language we had to classify any of the training data sentences into 9 classes. In order to improve this sentence level language identification task, we collected tweets of the 8 languages of our interest using seed words of each of these languages. The seed words are chosen in such a way that the resultant tweets retrieved belonged to the mixed script category. We added atleast 500 tweets for each class to make up a total number of labeled sentences (sentence level tags) to be 7656. We compare the accuracies of various learning algorithms using this expanded training dataset. As we can see accuracy has increased signinficantly by 6% for Naive Bayes and 3% for Logistic regression. We report the summary of experiments conducted in Table 5. The second column shows the classification accuracy of sentence level language identification with the training data provided in FIRE shared task. The third column shows the accuracy results using the expanded training dataset. We can observe that the semi-supervised version of Naive Bayes (Naive Bayes EM) is superior among all the classifiers. We can also observe that Naive Bayes classifier is benefitted the most (5.5% increase) by the supplementary tuples added to the training data. Naive Bayes EM is improved by approximately 3% and MaxEnt is improved by 1.6%.

3. CONCLUSION

In the current study we addressed the language identification problem in mixed script socail media text, at the word level, involving multiple indian languages namely *bengali*, *gujarati*, *hindi*, *kannada*, *malayalam*, *marathi*, *tamil*, *telugu*. Observing that the social media mixed script posts often involve english as the mixing language and can involve atmost one more other language as the length of the messages are quite short, we used a two-stage classification approach for sentence level langauge mode of the author and then a binary classifier for distinguishing english and each of the specific languages listed above. The test run submitted has given overall weighted F-measure of 0.7692. The sentence level classification accuracy was 68.87%. We could further improve this accuracy to 79.77% using abundantly available social media tweets crawled using seed words of specific langauge.

Acknowledgement

We sincerely thank the active participation of members of CoE Analytics, IDRBT, especially K. Sai Kiran, B. Shiva Krishna for helping us in sentence level labeling, and word level labeling. We thank IDRBT for providing the research environment for executing this task.

References

Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292.

Gupta, K., Choudhury, M., and Bali, K. (2012). Mining hindi-english transliteration pairs from online hindi lyrics. In Chair), N. C. C., Choukri, K., Declerck, T., DoÄ§an, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Täckström, O. and McDonald, R. (2011). Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 569–574. Association for Computational Linguistics.