SSN_NLP@INLI-FIRE-2017: A Neural Network Approach to Indian Native Language Identification

D. Thenmozhi
SSN College of Engineering
Chennai, Tamilnadu
theni_d@ssn.edu.in

Kawshik Kannan
SSN College of Engineering
Chennai, Tamilnadu
kawshik98@gmail.com

Chandrabose Aravindan
SSN College of Engineering
Chennai, Tamilnadu
aravindanc@ssn.edu.in

ABSTRACT

Native Language Identification (NLI) is the process of identifying the native language of non-native speakers based on their speech or writing. It has several applications namely authorship profiling and identification, forensic analysis, second language identification, and educational applications. English is one of the prominent language used by most of the non-English people in the world. The native language of the non-English speakers may be easily identified based on their English accents. However, identification of native language based on the users posts and comments written in English is a challenging task. In this paper, we present a neural network approach to identify the native language of an Indian speaker based on the English comments that are posted in microblogs. The lexical features are extracted from the text posted by the user and are used to build a neural network classifier to identify the native language of the user. We have evaluated our approach using the data set given by INLI@FIRE2017 shared task.

KEYWORDS

Neural Network, Machine Learning, Language Recognition, Indian Native Language Identification

1 INTRODUCTION

Native Language Identification (NLI) is the process of automatically identifying the native language of a person based on her/his speech or writing in another language. It has several applications namely authorship profiling and identification [2], forensic analysis [3], second language identification [8] and educational applications [11]. Several research work have been reported on NLI based on the speakers text [13], [5], [1], [4], [7], [9] and their speech [10], [12]. English is one of the commonly used languages by many people in the world and several shared tasks on NLI have been conducted since 2013 to identify the native language based on English text and speech. In this work, we have focused on the shared task of INLI@FIRE2017 (co-located with the Forum for Information Retrieval Evaluation (FIRE), 2017) which aims to identify the native language of Indians based on their comments posted in social media in English [6]. The focus of the task is to develop techniques for identifying the native languages namely Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu from a set of Facebook comments.

2 PROPOSED APPROACH

We have implemented a supervised approach for this INLI task. The steps used in our approach are given below.

- Preprocess the given text
- Extract linguistics features for training data
- Build a neural network model from the features of training data
- Predict class label for the instance as any of the six languages namely Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu using the model

We have implemented our methodology in Python for the INLI task. The data set used to evaluate the task consists of a set of training data for six Indian languages and test data. The number of training instances are 207, 211, 203, 200, 202 and 210 for the languages Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu respectively and number of test instances are 783. The steps used in our approach are explained in detail in the following subsections.

2.1 Feature Extraction

As a preprocessing step, all the 'xml' tags are removed from the given text and only the body part of the given input is considered for further processing. The punctuations like “, ”, - , _, ‘, and, ’ are removed from the text and the terms such as n't, &, 'm, 'll are replaced as 'not', 'and', 'am', and 'will' respectively before extracting the features. Each term of the text is annotated with parts of speech (POS) information such as noun, verb, adjective, adverb, and determiner. In general, nouns present in the text can be used as features. However, adjectives may also be helpful to identify the native language. For example, from the post 'I attended my kutty brother Rams birthday party', the adjective 'kutty' may used to identify the language as Tamil. So, in our approach, we have considered nouns and adjectives as features. All forms of nouns (NN*) namely NN, NNS and NNP, and all forms of adjectives (JJ*) JJ, JJR and JJS are extracted from the text. The feature set is constructed by lemmatizing each extracted term and by eliminating all the duplicate terms. We have obtained the bag of words (BOW) by processing all the text of given training data.

We have used the NLTK tool kit[1] to preprocess the given data and to annotate the text with POS information. The Wordnet Lemmatizer was used to lemmatize the terms that are extracted from POS information. We have obtained a total of 12067 features from training data. We have used the boolean model to construct the feature vectors for the instances of training data.

2.2 Language Identification

We have applied a neural network approach to identify the native language of the user. The set of BOW features along with the class labels namely Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu from training data are used to build a model using a simple neural network with two hidden layers. The features are

[1] http://www.nltk.org/

extracted for each instance of test data with unknown class label '?', similar to training data using the features identified from training data. One of the label from the given labels namely Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu is identified for the test data instances using the built model.

We have used the Keras framework[2] with Tensorflow backend to implement a neural network classifier for this problem. The number of BOW features (12067) constitutes the number of neurons for the input layer of the network. We have used a sequential model of Keras to construct our neural network. We have added two hidden layers with number of neurons as 64 and 32 respectively with 'RELU' activation function. The output layer was added by specifying the number of neurons as 6 (to classify the instance to one of the 6 languages) with an activation function 'SOFTMAX'. We used 'sparse_categorical_crossentropy' loss function with 'SGD' optimizer to compile the model. We trained the model with a batch_size of 10 for 100 epochs and obtained a training accuracy of 98.1%.

3 RESULTS AND DISCUSSIONS

Our approach for native language identification has been evaluated based on the metrics namely precision, recall, and F1 measure for each language and also overall accuracy. The results obtained by our approach are presented in Table 1. A comparative study of results of all the participants of INLI@FIRE2017 is available in [6].

Table 1: Performance on Test Data

Class	Precision	Recall	F1-measure
BE	46.20	76.20	57.60
HI	49.40	16.30	24.60
KA	39.60	48.60	43.60
MA	31.70	21.70	25.80
TA	27.50	49.00	35.30
TE	27.00	21.00	23.60

We have obtained an overall accuracy of 38.80% using our neural network approach for Indian native language identification task. This is very poor compared to the training accuracy of 98.1% and is an indication of over-fitting. We need to explore using regularization techniques such as dropout during training to avoid this.

4 CONCLUSION

We have presented a system that uses a neural network model for identifying the native language, namely Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu, of Indians from the English comments posted by them in social media. We have extracted the linguistics features from training data to build a neural network model with two hidden layers. The data set given by INLI@FIRE2017 shared task has been used to evaluate our methodology. We have obtained an overall accuracy of 38.80%. This is very poor compared to the training accuracy and indicates over-fitting. Regularization techniques such as dropout may be used to improve generalization. A lexical database may be used to correct terms such as pls, sry, fyi, etc., present in social media text for improving the performance of the system. The performance may improve if we select only the significant features using χ^2 feature selection [14].

[2]https://keras.io/

ACKNOWLEDGMENTS

We would like to thank the management of SSN Institutions for funding the High Performance Computing (HPC) lab where this work is being carried out.

REFERENCES

[1] Serhiy Bykh and Detmar Meurers. 2014. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization.. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Ireland, 1962–1973.

[2] Dominique Estival, Tanja Gaustad, Ben Hutchinson, Son Bao Pham, and Will Radford. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. ACL, Australia, 263–272.

[3] John Gibbons. 2003. *Forensic linguistics: An introduction to language in the justice system*. Wiley-Blackwell.

[4] Radu-Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? A language-independent approach to native language identification.. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 1363–1373.

[5] Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, Atlanta, Georgia, 111–118.

[6] Anand Kumar M, Barathi Ganesh HB, Shivkaran S, Soman K P, and Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification. In *Notebook Papers of FIRE 2017*. CEUR Workshop Proceedings, Bangalore, India.

[7] Shervin Malmasi and Mark Dras. 2017. Native Language Identification using Stacked Generalization. *arXiv preprint arXiv:1703.06541* (2017).

[8] Shervin Malmasi, Mark Dras, et al. 2014. Language Transfer Hypotheses with Linear SVM Weights.. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Qatar, 1385–1390.

[9] Elham Mohammadi, Hadi Veisi, and Hessam Amini. 2017. Native Language Identification Using a Mixture of Character and Word N-grams. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, Copenhagen, Denmark, 210–216.

[10] Taraka Rama and Çağrı Çöltekin. 2017. Fewer features perform well at Native Language Identification task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 255–260.

[11] Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. ACL, Portland, Oregon, USA, 924–933.

[12] Charese Smiley and Sandra Kübler. 2017. Native Language Identification using Phonetic Algorithms. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 405–412.

[13] Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. *Proceedings of COLING 2012* (2012), 2585–2602.

[14] D Thenmozhi, P Mirunalini, and Chandrabose Aravindan. 2016. Decision Tree Approach for Consumer Health Information Search.. In *FIRE (Working Notes)*. CEUR, Kolkata, India, 221–225.