Mangalore University INLI@FIRE2018: Artificial Neural Network and Ensemble Based Models for INLI

Hamada A. Nayel[1][1234−5678−9012] and H. L. Shashirekha[2]

[1] Department of Computer Science,
Faculty of Computers and Informatics,
Benha University, Benha, Egypt
hamada.ali@fci.bu.edu.eg
[2] Department of Computer Science,
Mangalore University, Mangalore, India
hlsrekha@gmail.com

Abstract. In this paper, the systems submitted by Mangalore University team for Indian Native Language Identification (INLI) task have been described. Native Language Identification (NLI) has different applications such as social media analysis, authorship identification, second-language acquisition and forensic investigation. We submitted three systems using Artificial Neural Network (ANN) model and Ensemble approach. All the three submitted systems achieved the same accuracy of 35.30% and secured second rank over all submissions for the task.

Keywords: Artificial Neural Network · Ensemble Learning · Native Language Identification.

1 Introduction

Native Language Identification (NLI) aims at identifying the native language (L1) of users written in another or later learned language or speech (L2). NLI is an important task that has many applications in different areas such as social-media analysis, authorship identification, second language acquisition and forensic investigation. In forensic analysis [5], NLI helps to glean information about the discriminant L1 cues in an anonymous text. Second Language Acquisition (SLA) [14] studies the transfer effects from the native languages on later learned language. In academics, automatic correction of grammatical errors is an important application of NLI [15]. NLI can be used as a feature in authorship identification task [4] which aims at assigning a text to one of the predefined list of authors. Authorship identification is used in terrorists communications investigation [1] and digital crime investigation [3].

2 Task Description

In this era, social media is overwhelming our lives. Majority of people are communicating and discussing different topics using different platforms of social media such as Google+, Facebook and Twitter. While communicating with each other Indians prefer to use English because their native languages are different.In addition, most software and keyboards does not support input using Indian languages characters. So, people are using standard English keyboard to write their own words as transliterated words.

The task [8] aims at identifying the native language of the writer from the given Facebook comment written in English language. Six Indian languages - Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu are considered for this shared task.

2.1 Related Work

Many researchers have explored the task of NLI for various applications. Jarvis et al. [6], used SVM to create a model for NLI and reported an accuracy of 83.6%. N-grams, PoS tags and lemmas have been used to create feature space model for training the classifier. They tested the performance of their system using TOEFL11 dataset [2]. The TOEFL11 is a collection of essays written by learners from the following native languages backgrounds: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. In this work, the feature set was not sufficient to cover the characteristics of different languages. Tetreault et al. [16] used ensemble approach to build a classifier to improve the performance of base classifiers. A wide range of features were used to build an ensemble of logistic regression learners. Such features include word and character *n*-gram, PoS, function words, writing quality markers and spelling errors. In addition, a set of syntactic features such as Tree Substitution Grammars and dependency features extracted using the Stanford parser[3] have been used. The system evaluated using TOEFL11 and International Corpus of Learner English (ICLE) datasets have resulted in state of the art accuracies of 90.1% and 80.9% respectively.

Nayel and Shashirekha [9, 12] used SVM and ensemble approach for the first version of INLI and achieved accuracies of 47.60% and 47.30% respectively.

3 Approaches

3.1 Artificial Neural Networks

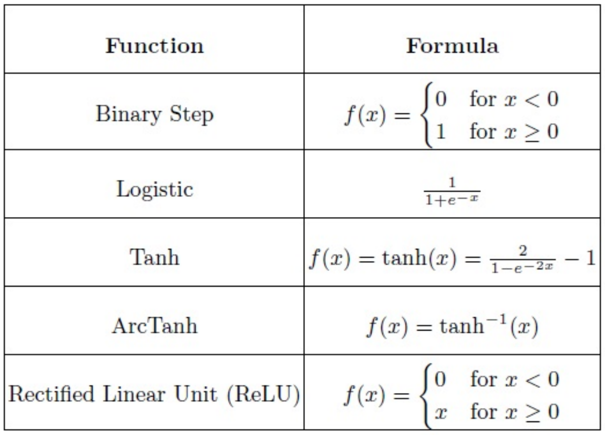
Artificial Neural Networks (ANNs) are inspired by the mechanism of brain computation, which consists of computational units called neurons. The connections

[3] http://nlp.stanford.edu:8080/parser/

between ANNs and the brain are in fact rather slim. In the metaphor, a neuron has scalar inputs with associated weights and outputs. The neuron multiplies each input by its weight, sums them and transforms to a working output through applying a non linear function called activation function. Table 1 shows examples of activation functions. The structure of the biological neuron and an example of an artificial neuron model with n inputs and one output is shown in Figures 1(a), 1(b) respectively. In this example, a neuron receives simultaneous inputs $X = (x_1, x_2, \ldots, x_n)$ associated with weights $W = (w_1, w_2, \ldots, w_n)$, a bias b and calculate the output as:

$$y = f(W \cdot X + b) \tag{1}$$

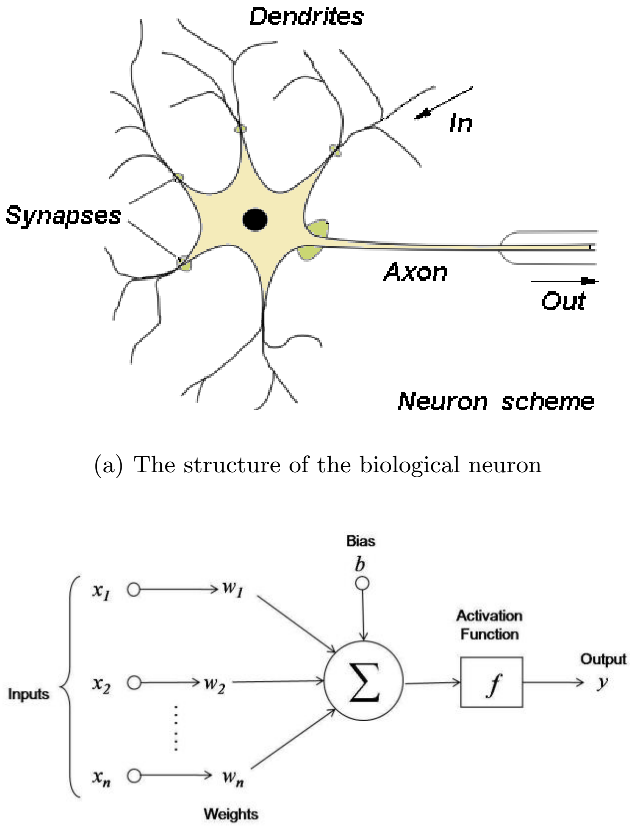
where f is the activation function.



Function	Formula
Binary Step	$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Logistic	$\frac{1}{1+e^{-x}}$
Tanh	$f(x) = \tanh(x) = \frac{2}{1-e^{-2x}} - 1$
ArcTanh	$f(x) = \tanh^{-1}(x)$
Rectified Linear Unit (ReLU)	$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$

Table 1. Examples of activation functions

ANN comprises of a large number of neurons within different layers. An ANN model basically consists of three layers: an input layer, a number of hidden layers and an output layer. Input layer contains a set of neurons called input nodes, which receive raw inputs directly. The hidden layers receive the data from the input nodes and are responsible for processing these data by calculating the weights of neurons at each layer. These weights are called connection weights and are passed from one node to another. Number of nodes in hidden layers influences the number of connections. During training phase connection weights are adjusted to be able to predict the correct class label of the input. Output layer receives the processed data and uses its activation function to generate final output. This kind of ANN where information flows in one direction is called feed-forward ANN. Figure 2 shows an example of a feed-forward ANN with two hidden layers. An ANN is called fully connected if each node in a layer is connected to all nodes in the subsequent layer.



(a) The structure of the biological neuron

(b) A simple neuron example

Fig. 1. A typical BioNER system with an example

3.2 Ensemble Approach

Most of the classification tasks use a single classifier. However, for some data some classifier may give good results while other classifier may not perform well. Further, there is no generic rule which helps to choose a classifier for a particular application and data. So, instead of experimenting the single classifiers one by one in search of good results it will be beneficial to pool several such classifiers and then take the collective decision similar to the decision taken by a committee rather than an individual. This technique which overcomes the weakness of some classifiers using the strength of other classifiers is termed as "ensemble". Ensemble approach has been applied for different tasks such as BioNER [11, 13], word segmentation [10] and word sense disambiguation [7].

INLI considers set of Indian languages, namely Kannada (KA), Tamil (TA),

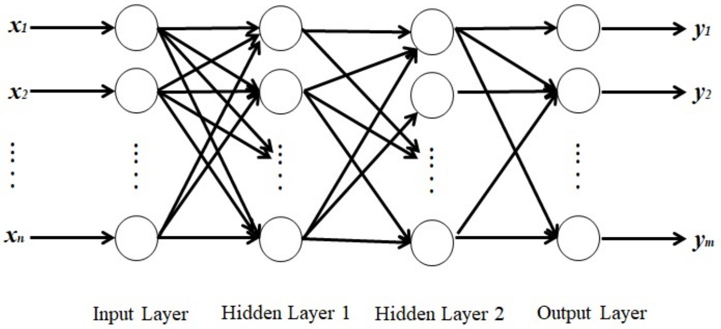


Fig. 2. A Simple Feed-Forward ANN Structure

Hindi (HI), Telugu (TE), Bengali (BE) and Malayalam (MA). Considering the languages as a set of classes $L = \{KA, TA, HI, TE, BE, MA\}$ and comments as individual instances, the task of identifying the native language can be considered as a classification problem that assigns one of the predefined languages of L to a new unlabelled comment.

The general framework of our system is as described in [12]. Vector space model using Term Frequency/Inverse Document Frequency (TF/IDF) has been used to represent comments. ANN based classifier is designed for the first and second submissions. The hidden layer of first submission contains 70 neurons and the activation function is logistic function. The hidden layer of second submission contains 80 neurons and the activation function is the identity function. Ensemble approach using majority voting technique has been used for designing the third submission. Four ANN based models with different parameters (shown in Table 2)) have been used as base classifiers to build the ensemble classifier.

Table 2. Parameters of different base models

Model Number	Number of neurons in hidden layer	Activation function
1	70	Logistic
2	80	Logistic
3	80	Tanh
4	80	Identity

4 Results and Discussion

Accuracy and class-wise Precision (P), Recall (R) and F-measure have been used for evaluating the submitted systems [9]. Cross-Validation (CV) technique has been used while building the systems. Table 3 shows the 10-fold CV accuracy for the three systems.

Table 3. Cross validation accuracies for the three submitted systems

	Accuracies in %		
	Submission 1	Submission 2	Submission 3
	89.68	90.48	89.68
	85.60	84.80	85.60
	87.10	87.90	87.90
	91.87	90.24	91.06
	91.87	92.68	92.68
	84.55	82.93	82.93
	88.62	89.43	89.43
	90.16	90.16	90.98
	86.88	85.25	86.07
	88.52	86.88	86.88
Mean=	88.49	88.08	88.32

Performance evaluation of the first, second and third submissions are shown in Tables 4, 5 and 6 respectively. The accuracy of each of the submitted system is 35.30% and all of them rank second among all the submissions.

In all the three submissions, the lowest and the best performance was reported for Hindi and Bengali language respectively among all submissions. Most of native speakers of Indian languages have knowledge of Hindi which affects while writing their comments in English.

5 Conclusion

In this work, ANN and Ensemble based classifiers have been used to design systems for INLI 2018. All designed classifiers reported the same accuracy and

Table 4. Performance evaluation of first system

Class Label	**Confusion matrix**						**Class-wise results**		
	BE	HI	KA	MA	TA	TE	P	R	F-measure
BE	79	24	17	28	43	16	47.00	38.20	42.10
HI	19	14	12	42	24	27	13.90	10.10	11.70
KA	16	20	106	26	47	35	37.20	42.40	39.60
MA	19	19	36	87	26	13	36.6	43.50	39.70
TA	10	12	31	24	61	2	26.6	43.60	33.10
TE	25	12	83	31	28	71	43.3	28.40	34.30
Overall Accuracy							**35.30%**		

Table 5. Performance evaluation of second system

Class Label	**Confusion matrix**						**Class-wise results**		
	BE	HI	KA	MA	TA	TE	P	R	F-measure
BE	80	20	18	29	43	17	47.60	38.60	42.70
HI	19	12	11	44	24	28	12.60	8.70	10.30
KA	13	21	112	28	41	35	38.10	44.80	41.20
MA	23	18	36	86	23	14	36.6	43.00	39.50
TA	8	15	30	23	57	7	25.80	40.70	31.60
TE	25	9	87	25	33	71	41.30	28.40	33.60
Overall Accuracy							**35.30%**		

achieved the second rank over all submissions for the task. This work can be improved using different structures of ANN and using deep learning model. In addition, improving input representation will improve the performance of systems.

References

1. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems **20**(5), 67–75 (Sep 2005)
2. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., Chodorow, M.: Toefl11: ”A corpus of non-native english”. ETS Research Report Series **2013**(2) (2013)
3. Chaski, C.E.: Whos at the keyboard? ”Authorship attribution in digital evidence investigations”. International Journal of Digital Evidence **4**(1), 1–13 (2005)

Table 6. Performance evaluation of third system

Class Label	**Confusion matrix**						**Class-wise results**		
	BE	HI	KA	MA	TA	TE	P	R	F-measure
BE	79	24	17	28	43	16	47.00	38.20	42.10
HI	19	14	12	42	24	27	13.90	10.10	11.70
KA	16	20	106	26	47	35	37.20	42.40	39.60
MA	19	19	36	87	26	13	36.60	43.50	39.70
TA	10	12	31	24	61	2	26.60	43.60	33.10
TE	25	12	83	31	28	71	43.30	28.40	34.30
Overall Accuracy							**35.30%**		

4. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for english emails. In: ”Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics”. pp. 263–272 (2007)
5. Gibbons, J.: Forensic linguistics: ”An introduction to language in the justice system”. Wiley-Blackwell (2003)
6. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification pp. 111–118 (2013)
7. Klein, D., Toutanova, K., Ilhan, H.T., Kamvar, S.D., Manning, C.D.: Combining heterogeneous classifiers for word-sense disambiguation. In: Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8. pp. 74–80. WSD ’02, Stroudsburg, PA, USA (2002)
8. Kumar, A., Ganesh, B., P, S.K.: Overview of the INLI@FIRE-2018 Track on Indian Native Language Identification. In: ”workshop proceedings of FIRE 2018, FIRE-2018”. Gandhinagar, India, December 6-9, CEUR Workshop Proceedings (2018)
9. Kumar, A., Ganesh, B., Shivkaran, P, S.K., Rosso, P.: Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification. In: ”Notebook Papers of FIRE 2017, FIRE-2017”. Bangalore, India, December 8-10, CEUR Workshop Proceedings (2017)
10. Min, K., Ma, C., Zhao, T., Li, H.: BosonNLP: ”An Ensemble Approach for Word Segmentation and POS Tagging”. In: Natural Language Processing and Chinese Computing. pp. 520–526. Springer International Publishing (2015)
11. Nayel, H.A., Shashirekha, H.L.: Improving NER for Clinical Texts by Ensemble Approach using Segment Representations. In: ”Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)”. pp. 197–204. NLP Association of India, Kolkata, India (December 2017)
12. Nayel, H.A., Shashirekha, H.L.: Mangalore-University@INLI-FIRE-2017: ”Indian Native Language Identification using Support Vector Machines and Ensemble approach”. In: Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017. pp. 106–109 (2017)
13. Nayel, H.A., Shashirekha, H.L., Shindo, H., Matsumoto, Y.: Improving Multi-Word Entity Recognition for Biomedical Texts. International Journal of Pure and Applied Mathematics **118**(16), 301–3019 (2017)

14. Ortega, L.: Understanding Second Language Acquisition. Hodder Education, Oxford (2009)
15. Rozovskaya, A., Roth, D.: 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. pp. 924–933. Portland, Oregon, USA (June 2011)
16. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native Tongues, Lost and Found: " Resources and Empirical Evaluations in Native Language Identificatio". In: "Proceedings of COLING 2012". pp. 2585–2602 (2012)