# A Shallow Neural Network for Native Language Identification with Character *N*-grams

**Yunita Sari**
Department of Computer Science
University of Sheffield, UK
y.sari@sheffield.ac.uk

**Muhammad Rifqi Fatchurrahman** and **Meisyarah Dwiastuti**
Department of Computer Sciences and Electronics
Universitas Gadjah Mada, Indonesia
{muh.rifqi.fatchurrahman, meisyarah.dwiastuti}@gmail.com

## Abstract

This paper describes the systems submitted by GadjahMada team to the Native Language Identification (NLI) Shared Task 2017. Our models used a continuous representation of character *n*-grams which are learned jointly with feed-forward neural network classifier. Character *n*-grams have been proved to be effective for style-based identification tasks including NLI. Results on the test set demonstrate that the proposed model performs very well on essay and fusion tracks by obtaining more than 0.8 on both F-macro score and accuracy.

## 1 Introduction

Native Language Identification (NLI) is the task of identifying the native language (L1) of the speakers in which English is usually their second language (L2). Given $F = \{f_1, f_2, ..., f_3\}$ be a set of written or speech responses and $K = \{k_1, k_2, k_m\}$ a pre-defined set of native languages (L1), the NLI task is to assign L1 to each of the responses in $F$. This task is often considered as a subset of author profiling task which currently focuses more on age and gender identification (Lopez-Monroy et al., 2014; Johannsen et al., 2015; Rangel Pardo et al., 2016).

The growing interest in this field is due to the applicability of this task to support language learners by providing a tailored feed-back about their errors. Swan and Smith (2001) argued that speakers of different native languages tend to make different mistakes. Thus, targeted feed-back is expected to improve the process of language learning (Tetreault et al., 2013).

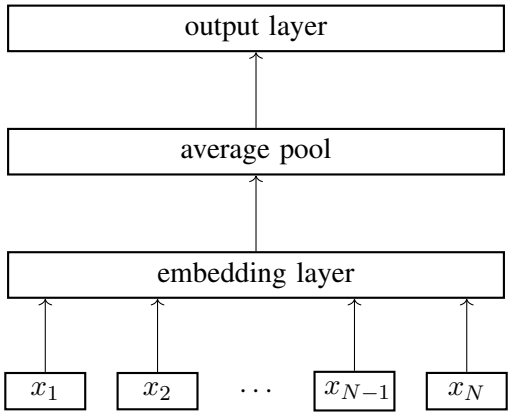
The NLI Shared Task 2017 (Malmasi et al., 2017) is the continuation of the first task that has been held in 2013 (Tetreault et al., 2013). This year's task aims to combine written responses (essay) and spoken responses (speech transcript and i-vector acoustic features) for identifying 11 native language classes.

To address the NLI Shared Task 2017 problem, we adopted an approach that has been applied for authorship attribution task (Sari et al., 2017). In this approach, continuous representations of character *n*-grams are used jointly with feed-forward neural network classifier. The methods performed very well on essay and fusion tracks by obtaining more than 0.8 on both F-macro score and accuracy. However, due to the poor hyper-parameter setting and the limitation of training data, we only managed to get around 0.5 on speech track for both evaluation scores, using only the speech transcripts.

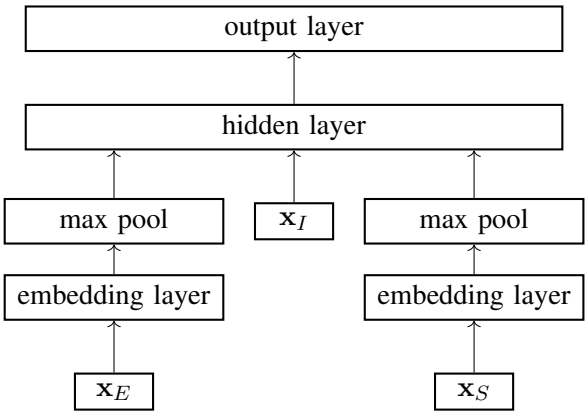
The paper is organised as follows: Section 2 provides a review of relevant work in NLI. We then explain our methodology in Section 3. The next section describes our experiments including the description of the dataset and the details of training and hyper-parameter tuning. Result and discussion are presented in Section 5. Finally, conclusion and future work are drawn in Section 6.

## 2 Related Work

The first NLI shared task (Tetreault et al., 2013) was introduced in 2013 with a total of 29 teams participated across three different subtasks. The dataset for the task was TOEFL11 corpus (Blan-




(a) Model for essay and speech tracks




(b) Model for fusion track

chard et al., 2013) consists of 11,000 essays written by a high-stakes college-entrance test taker. Same as this year's task, there are 11 native languages covered. Tetreault, et. al reported that majority of the participant addressed the problem by utilising powerful machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression. In term of features, word, character and POS *n*-grams were the most common used features.

One of the interesting findings from the first NLI task is simple features such as words, word forms, sequential word combinations, and sequential POS combinations turn out to be effective indicators for identifying L1. Jarvis et al. (2013) who implemented those features successfully secured the best systems in the first NLI task by obtaining 10-fold cross-validated accuracy of 84.5% and overall accuracy of 83.6% on the test set. In addition, they reported that a model with character *n*-grams achieved similar accuracy to the best model involving lexical and POS *n*-grams.

Following the first NLI task, Ionescu et al. (2014) extended their submission system by implementing character *n*-grams with two kernel classifiers namely Kernel Ridge Regression (KRR) and Kernel Discriminant Analysis (KDA). Their result outperformed Jarvis, et. al by 1.7% on the overall accuracy. Character *n*-grams have been known for its impressive performance in style-based text analysis task such as authorship attribution (Peng et al., 2003; Stamatatos, 2013; Schwartz et al., 2013). It has advantages of capturing stylistic and morphological information (Koppel et al., 2011; Sapkota et al., 2015) regardless of the language. This has motivated us to utilised character *n*-grams in our system.

In addition to written responses, recent trend starts to consider spoken responses (speech transcripts and audio features) for NLI task. Incorporating spoken responses has produced good result for dialect identification (Malmasi et al., 2016; Zampieri et al., 2017). However compared to audio features, speech transcripts are less useful since ambiguity is more pronounced in written transcripts.

## 3 Methodology

In this section, we describe our models and features used in our NLI system. First, we present the details of the features. Then we explain our model architectures which use shallow feed-forward neural network.

### 3.1 Features

There are two types of features used in our system: character *n*-grams and i-vectors. We used only character *n*-grams features in essay and speech tracks and combined them with i-vectors for fusion track. The details of the features are explained as follows:

- **Character n-grams:** This substring takes $n$ characters constructing the text along the whole text as features. We set the vocabulary to 70 most common characters including letters, digits, and some punctuation marks as conducted by Zhang et al. (2015). We followed Sari et al. (2017) who represented the features as continuous vectors. The idea of representing *n*-grams in continuous space was introduced by Joulin et al. (2017) who proposed an efficient model for text classification called fastText. Instead of using a single value of $n$, we applied a range of $n$ values

from three to six grams.

- **i-vectors:** i-vector or identity vector is one of feature representation that commonly used in speech processing. It is a low-dimensional vector derived from mapping sequence of speech frames (Dehak et al., 2011). The i-vectors correspond to the speech transcriptions and have a length of 800. We used the provided i-vectors without any additional pre-processing.

### 3.2 Model Architecture

Our model adopted fastText architecture which was proposed by Joulin et al. (2017). FastText represents a document with an average of feature embeddings for the features present. The probability distribution over the labels then is simply predicted using *softmax* function. However, instead of working on word level, we chose to work on character level, since it is found to be more suitable for the task. Figure 1a shows the model that we used for both essay and speech tracks. In that figure, $x_n$ represents a single character *n*-gram, while $N$ is the maximum sequence which the value is fixed. In our experiment, feature embeddings are learned during training.

For fusion track, we extended the first model with an auxiliary input to accommodate i-vectors as presented in Figure 1b. We also added one hidden layer with the size of 128 right before the output layer. Slightly different with the first model, in the fusion track we used *max pooling* as it produced higher performances. Both of the models were implemented using Keras (Chollet et al., 2015) with Tensorflow backend.

### 3.3 Baseline Systems

As a benchmark, the organiser developed baseline systems which use SVM as the classifier. Essay and speech transcript are represented as bag-of-words (BoW). The baseline results on the test set are presented in the Table 1.

## 4 Experiments

### 4.1 Dataset

The dataset provided by Educational Testing Service (ETS) contains test responses from a standardised assessment of English proficiency for academic purposes. It consists of 13,200 English essays (written responses) and 13,200 of 45 seconds English speech transcriptions (spoken responses). In addition to that, i-vectors of the speech audios are generated in lieu of the audio files. The essays typically range in length from 300 to 400 words and the transcriptions typically contain approximately 100 words.

The test responses are from 13,200 different test takers. Each test taker contributed one essay and one speech transcription. There are 11 native languages (L1) covered, including Arabic (ARA), Chinese (CHI), French (FR), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). The organiser set the 11,000 samples from the dataset for training purpose, 1,100 for development and the rest as the test set.

### 4.2 Hyper-parameter Tuning and Training Details

During training, we tried different combinations of hyper-parameter configurations. However, only the configurations of the best run are reported in this paper.

**Feature hyper-parameters**. For essay and speech tracks, we only used character $n$-grams features. We set the range of $n$-gram values from 2 to 5. The sequence lengths were set to 6,000 for essay and 4,000 for speech. Meanwhile, for fusion track in addition to the character *n*-grams, i-vectors were used. The *n*-grams range was set to 3 to 5. In order to reduce the input dimensions, we decreased the length of the sequence to 1,500 for essays and 300 for speech transcriptions.

**Model hyper-parameters.** The best run for essay and speech tracks used embedding size of 25 with dropout rate of 0.75. For fusion model, embedding size for both essay and speech representation was set to 128. Between the layers, we put dropout with the probability of 0.5.

**Training.** For all sub-tracks, the models were trained using Adam Optimizer (Kingma and Ba, 2015) with cross-entropy loss. We also implemented early stopping procedure in order to avoid over-fitting. We set batch size of 64 for essay and fusion tracks; and 32 for speech track. Number of epochs for essay, speech, and fusion tracks were set to 80, 50, and 100 respectively. For both essay and speech tracks, learning rate of 0.005 was used.

| Track | Baseline System | | GadjahMada System | |
|---|---|---|---|---|
| | F1-score | Accuracy | F1-score (macro) | Accuracy |
| Essay | 0.7104 | 0.7109 | **0.8107** | **0.8109** |
| Speech (transcription only) | **0.5435** | **0.5464** | 0.5084 | 0.5073 |
| Fusion (essay, speech transcripts, i-vectors) | 0.7901 | 0.7909 | **0.8414** | **0.8409** |

Table 1: Submission Results

While fusion track used learning rate of 0.0002, higher rates did not make any improvement.

## 5 Results and Discussion

Table 1 shows our submission results for all the sub-tracks. We participated in closed- training subtask in which we only used provided training data to train our models. Results on the table present that our systems performed very well on the essay and fusion tracks. Our systems outperformed the baseline systems with accuracy of 0.8109 and 0.8409 on the essay and fusion tracks respectively. However, the system failed to produce similar performances on speech track. Our system produced accuracy of 0.5073 which is lower than the baseline. This might happen due to the poor hyper-parameter tuning. Note that on speech track, we only utilised speech transcripts.

Similar to the previous NLI shared task results, character *n*-grams demonstrate their effectiveness for capturing style in written responses. We believe that speakers of each native language have their own learning experiences which are reflected in their responses. The speaker's characteristics are better captured in written responses than speech transcripts. Written responses are significantly longer compared to speech transcripts which make it better on providing information about the speaker. In addition to that, speech transcripts are less useful since ambiguity is more pronounced (Malmasi et al., 2016). Audio features in the form of i-vectors help to improve the performance. Our results on the fusion track are higher than the results on other tracks.

In order to get more insight into the classification results, confusion matrices for the best run in each sub-track are presented in Figure 2. In the essay and fusion tracks, it can be seen that German (GER) speaker are the easiest class to identify with more than 90% on the accuracy. It is also interesting to highlight that the system is mistakenly identified several native language classes that have morphological and lexical similarities, for example: Chinese (CHI), Japanese (JPN) and Korean (KOR); Hindi (HIN) and Telugu (TEL); French (FRE), Italian (ITA) and Spanish (SPA). However in the speech track as shown in Figure 2b, in most classes the system made correct predictions no more than 50% of the total samples. It demonstrates that spoken response in the form of speech transcripts is not good enough to be used as feature.

## 6 Conclusion and Future Work

This paper presents our submission approaches for NLI Shared Task 2017. Results on the test set show our model that utilises shallow feed-forward neural network with character *n*-grams features could effectively identify the native language (L1) of the speaker. Our proposed model performed very well on the essay and fusion tracks but failed to achieve similar scores on the speech track. It is interesting to note that character *n*-grams mostly works for any style-based classification tasks including NLI. More details analysis on the languages with similar lexical and morphological forms can be an interesting work to explore. Indicative features for those languages are essential since most incorrect predictions were made on those groups.

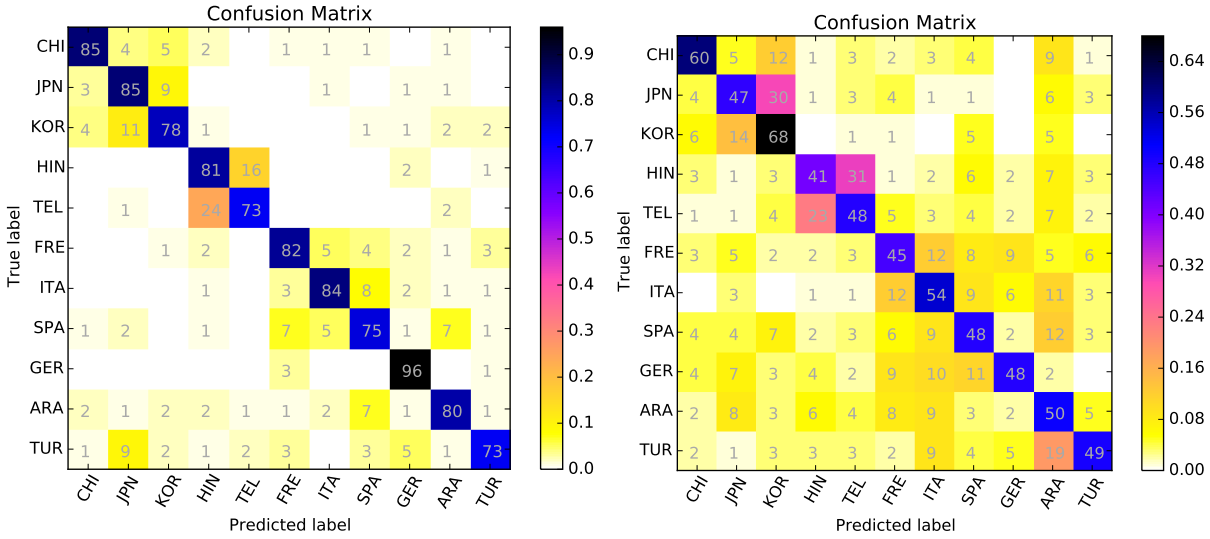
## References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. Technical report, Educational Testing Service.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

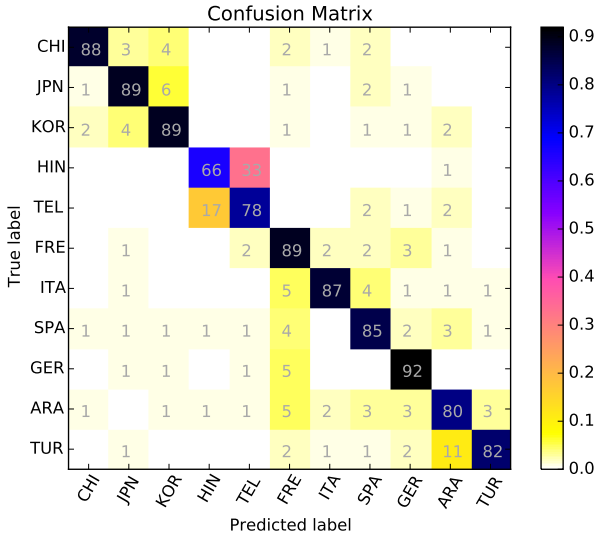
Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. 2011. Language recognition via i-vectors and dimensionality reduction. In *Twelfth Annual Conference of the International Speech Communication Association*.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? A language-independent approach to native




(a) Essay track

(b) Speech track




(c) Fusion track

Figure 2: Confusion matrices for the best run in each sub-track

language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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*. Association for Computational Linguistics, Atlanta, Georgia, pages 111–118.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Beijing, China, pages 103–112.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 427–431.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceeding of the 3rd International Conference for Learning Representations, ICLR 2015*. San Diego, CA.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45(1):83–94.

A. Pastor Lopez-Monroy, Manuel Montes Gomez, and Hugo Jair-Escalante. 2014. Using Intra-Profile Information for Author Profiling. In *Notebook for PAN at CLEF 2014*.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Copenhagen, Denmark.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the VarDial Workshop*. Osaka, Japan.

Fuchun Peng, Dale Schuurmanst, Vlado Kesel, and Shaojun Wan. 2003. Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics, EACL 2003*. Budapest, Hungary.

Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*. CLEF and CEUR-WS.org, CEUR Workshop Proceedings.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL HLT 2015*. Association for Computational Linguistics, Denver, Colorado, pages 93–102.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 267–273.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micromessages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1880–1891.

Efstathios Stamatatos. 2013. On the Robustness of Authorship Attribution Based on Character n-gram Features. *Journal of Law and Policy* 21(2):421–439.

Michael Swan and Bernard 1937 Smith. 2001. *Learner English : a teacher's guide to interference and other problems*. Cambridge (U.K.) Cambridge University Press, 2nd ed edition. :A Teacher's Guide to Interfernce and Other Problems.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, GA, USA.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain, pages 1–15.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS 2015*. MIT Press, Cambridge, MA, USA, pages 649–657.