Fusion of Simple Models for Native Language Identification

Fabio N. Kepler*
University of Pampa, Alegrete, Brazil
INESC-ID, Lisbon, Portugal
fabio@kepler.pro.br

Ramon F. Astudillo*
INESC-ID, Lisbon, Portugal
ramon@astudillo.com

Alberto Abad*
INESC-ID, Lisbon, Portugal
IST, University of Lisbon, Portugal
alberto.abad@inesc-id.pt

Abstract

In this paper we describe the approaches we explored for the 2017 Native Language Identification shared task. We focused on simple word and sub-word units avoiding heavy use of hand-crafted features. Following recent trends, we explored linear and neural networks models to attempt to compensate for the lack of rich feature use. Initial efforts yielded f1-scores of 82.39% and 83.77% in the development and test sets of the fusion track, and were officially submitted to the task as team L2F. After the task was closed, we carried on further experiments and relied on a late fusion strategy for combining our simple proposed approaches with modifications of the baselines provided by the task. As expected, the i-vectors based subsystem dominates the performance of the system combinations, and results in the major contributor to our achieved scores. Our best combined system achieves 90.1% and 90.2% f1-score in the development and test sets of the fusion track, respectively.

1 Introduction

Native Language Identification (NLI) is the task of identifying a person's native language (L1) based on that person's written or spoken content in a learned language (L2). The task has gained increased interest from various research communities, which led to the first shared task in 2013 (Tetreault et al., 2013). In 2016, a sub-challenge was held at Interspeech (Schuller et al., 2016) on identifying the native language based on spoken responses in English, in contrast to the NLI shared task, which was based on written responses.

The NLI Shared Task 2017 is the next instance in this series of shared tasks (Malmasi et al., 2017), with the distinction of featuring both written and spoken based responses as available data. Spoken responses were available in the form of speech transcriptions and i-vectors, not actual audio files. Systems could compete in three tracks: ESSAYS, where only the provided written essays data could be used; SPEECH, where only the speech transcriptions and possibly i-vectors could be used; and FUSION, where all three datasets could be combined. The task provided a single development labeled dataset and two different unlabeled test sets: one for the ESSAYS and SPEECH tracks, and another for the FUSION track. Additionally, each system was allowed to participate in an open or closed sub-track depending on whether any external data was used or not, respectively.

In this paper we describe the approaches we took in the NLI Shared Task 2017, specifically in the FUSION closed track, where we participated as team L2F. After having officially submitted a system to the track, we performed further experiments and developed additional systems, including a late fusion one that performs 7 absolute points above the system we submitted.

The best performing systems on a variety of Natural Language Processing (NLP) and Information Retrieval problems, including NLI, are ensembles of complex models that employ a myriad of high-level features (Malmasi and Dras, 2017). There are, however, some systems with simple features that are able to surpass complex ensembles, like the previous state of the art in NLI by Ionescu et al. (2014).

One way of not relying on specially engineered features is to follow the current trend on using Neural Networks (NN) and Deep Learning (DL)

*All authors contributed equally.

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 423–429
Copenhagen, Denmark, September 8, 2017.

techniques (and doing parameter tuning instead). Although DL approaches have achieved several state of the art results in NLP, this is not the case yet for NLI.

Our line of approach for this task was to benefit from the power of fusion systems while avoiding complex feature engineering and exploring the usefulness of DL techniques.

2 Related Work

There are several works on NLI based on essays, most of which are analyzed by Malmasi (2016). The current state of the art is the recent work of Malmasi and Dras (2017), which uses ensembles of several classifiers over a large set of features. The previous state of the art was the work of Ionescu et al. (2014), which used only character p-grams as features.

A recent trend has been the use of speech transcripts and audio features for tasks like dialect identification (Malmasi et al., 2016; Zampieri et al., 2017) or of only spoken responses for NLI, like in the 2016 Computational Paralinguistics Challenge (ComParE, Schuller et al. (2016)). The best performing system in ComParE 2016 was the work of Abad et al. (2016), which also employs a fusion of systems and highlights the importance of i-vectors acoustic features.

3 Methodology and Data

The NLI Shared Task 2017 combines the basic written essay approach with the spoken response approach by providing a written essay, a speech transcript, and an i-vector for each subject. For a thorough description of the datasets, including the number of samples for training, development and test, and the 11 L1 classes, see Malmasi et al. (2017).

Given that the task allowed for the fusion of all these data, we experimented with several approaches targeting a final fusion system, all of which we describe below.

3.1 Language Identification Techniques

Following the success in applying language identification techniques to L1 identification in speech (Abad et al., 2016), we explored language identification techniques in an initial stage. We trained the well known *langid* tool (Lui and Baldwin, 2011) using the data-sets provided in the shared task. The technique implemented in *langid* combines a Naive Bayes classifier with byte n-grams and no assumption over word boundaries. Unfortunately, no results outperforming the baseline could be attained with *langid*.

Character n-grams are a common feature in NLI systems and have been shown to provide strong results (Koppel et al., 2005; Ionescu et al., 2014). The low performance we attained with *langid* might therefore be related with particularities of the tool. It is also possible that specific tuning of algorithms for language identification might not be suitable for L1 identification.

3.2 Sub-word Features

Together with part-of-speech, character-level features are a commonly used feature for NLI (Malmasi, 2016). Upon manual inspection of the essays and speech transcriptions corpora of the shared task, it became clear that spelling or transcription errors were present with high frequency. This is a scenario in which sub-word units can play an important role for two main reasons. On the one hand, sub-word units help alleviate the effect of rare words that do not appear in the training corpus, also known as Out of Vocabulary Words (OOVs). On the other hand, they can capture systematic sub-word patterns, such as typographical or transcription errors, that can be specific to a particular L1 profile.

As an alternative to sub-word units based on character n-grams, we explored the use of the Byte Pair Encoding (BPE) approach (Sennrich et al., 2015). This simple approach, that has recently help to achieve state-of-the-art results in machine translation (Sennrich et al., 2015), provides a middle ground between character and word models. BPE is a well known compression technique that is here employed to iteratively merge the characters or sequences of characters that are most common into new tokens. The resulting vocabulary contains many or the original word tokens as well as fragments of frequent character sequences and individual characters.

Initial experiments explored the use of BPE tokens as a replacement for word tokens in the baseline system. This yielded however no notable improvements over the provided features. One possible limitation on the use of BPE features compared to Sennrich et al. (2015) is the lack of Recurrent Neural Networks to capture context. With-

out them, the use of sub-word units might destroy some useful information at the word token level. For this reason, further experiments included n-grams of BPE units as features with $n = 1, 2, 3$. It has to be taken into account that n-grams of BPE features might not only capture whole words but also sub-word patterns within and across word boundaries. The use of n-grams together produced however no improvements compared to the baseline system.

To provide some additional complementarity in the final ensemble, a Naive Bayes model was trained on the same features. Despite its simplicity, the model became competitive after introducing the n-gram features. Minor improvements over the baseline on the ESSAYS dataset were then attained by using BPE sub-word units (as we will see in Section 5, Table 1) and were kept for the final ensemble due to its complementarity. After determining the optimal features, the system parameters were tuned using the development set. A value of 10000 new BPE symbols was determined as optimal. The Naive Bayes classifier smoothing, equivalent to an uniform Dirichlet prior for the likelihood estimation, was set to $1e^{-4}$.

3.3 Neural Networks

Neural Networks are being successfully applied to a varying set of NLP problems. Following the current trend, we developed several architectures and tested them over the ESSAYS and FUSION tracks.

A common choice for treating sequence data like text are Recurrent Neural Networks (RNN), usually in their Long-Short Term Memory (LSTM) or Gated Recurrent Units (GRU) flavors, which are better able to capture long dependencies than plain RNNs. We decided to use GRUs (Chung et al., 2015) since they are faster to train and provide similar results to LSTMs.

We ended up building two networks: one for the ESSAYS tracks (NN-ESSAYS) and another for the FUSION track (NN-FUSION). The network for the FUSION track uses all available data as input: essays, transcripts, and i-vectors. The network for the ESSAYS track only uses the tokens in the essays.

Our final architecture for the NN-ESSAYS network is composed by the following layers:

- An embedding layer mapping input identifiers to a 300-dimensional space;
- A feed-forward layer with 300 units and ReLU (Nair and Hinton, 2010) activations;
- A bidirectional GRU layer with 300 units;
- A max-pooling layer applied across the time dimension;
- A feed-forward layer with 11 units (one for each language) and softmax activation.

The architecture for the NN-FUSION network is essentially similar but has to deal with the multiple inputs:

- The essay and transcript inputs each pass through the first four layers as in NN-ESSAYS before being concatenated;
- Each sample i-vector goes through a 400 units, ReLU activated feed-forward layer before being concatenated with the resulting concatenation above;
- A final softmax layer is then applied.

Several different architectures were tested, but none yielded results outperforming the baseline. As we will see in Section 5, the i-vectors dominate over the other features.

3.4 I-vector system

The success of the i-vector (Dehak et al., 2011a) framework in speaker recognition tasks has motivated the investigation of its application to other related fields, including language recognition (Martınez et al., 2011; Dehak et al., 2011b), where it has become the current *de facto* standard for acoustic Spoken Language Recognition (SLR), and more recently L1 recognition (Abad et al., 2016).

In the Total-variability modeling approach – so-called i-vector approach – the variability present in the high-dimensional GMM super-vector is jointly modeled as a single low-rank total-variability space. The low-dimensionality total variability factors extracted from a given speech segment form a vector, named i-vector, which represents the speech segment in a very compact and efficient way. Thus, the total-variability modeling is used as a factor analysis based front-end extractor.

In this work, the 800 dimensionality i-vectors provided in the task were used to build a new

acoustic L1 classifier. First, we apply i-vector centering and whitening (Garcia-Romero and Espy-Wilson, 2011) that is known to contribute to a reduction of the channel variability. Moreover, the resulting centered and whitened i-vectors are normalized to be of unit length.

Second, we explored different classifiers on the top of the processed i-vectors. Like in Abad et al. (2016), in which log-linear and non-linear classifiers based on feed-forward networks were investigated, we could observe that the i-vector front-end already provides a very good separation of the classes which leads to similar results for the different modeling techniques.

In particular, we tried to model the distribution of i-vectors for each language with a single mixture Gaussian distribution with full covariance matrix shared across different target languages since it has proven very effective (Martınez et al., 2011; Abad et al., 2016). However, in this case, this approach showed very similar performance to the baseline classifier: a multi-class one-vs-rest logistic regression classifier. Consequently, we opted for the baseline logistic regression approach.

4 Calibration and Fusion Back-End

In this work, we carried out calibration and fusion of the systems at the output score level using the FoCal Multi-class Toolkit[1]. For that purpose, every single sub-system is forced to produce an 11-element score vector $\mathbf{s}_i$ corresponding to each of the target languages. Then, a Linear Logistic Regression (LLR) is trained to fuse the score outputs generated by the selected sub-systems in order to produce fused well-calibrated log-likelihoods $\mathbf{l}$ as follows:

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b}, \quad (1)$$

where α_i is the weight for sub-system i and $\mathbf{b}$ is the language-dependent shift. For this challenge, the language with the highest fused log-likelihood is the hypothesized L1 language.

Notice that, in contrast to Abad et al. (2016), the use of a Gaussian Back-End to transform the score-vector of each individual sub-system before the LLR stage has not been applied, since it did not reveal to contribute for improved language identification in the validation experiments.

[1] https://sites.google.com/site/nikobrummer/focalmulticlass

During the development of our systems, the LLR fusion parameters were trained and evaluated on the development set using a kind of 2-fold cross-validation: development data was randomly split in two halves, one for parameter estimation and the other for assessment. This process was repeated using 10 different random partitions so that the mean and variance of the systems' performance could be computed. This method allowed for a comparison and ranking of the different subsystems under study. Then, for the trial submissions, no partition was made and all the development data was used to train the LLR fusion.

The final combined system for the FUSION track, which we call FINAL-FUSION, consists in the LLR fusion of the following 5 systems: i) ESSAY baseline; ii) speech transcriptions baseline; iii) the i-vector system described in Section 3.4; iv) the NN-ESSAYS system described in Section 3.3; and v) the BPE system described in Section 3.2. We also evaluated a LLR-FUSION system consisting in i), iv) and v) on the ESSAYS track.

5 Results

We first show the results over the development set in order to justify our approach choices, beginning with the ESSAYS track. The official evaluation metric is the macro averaged F1 score.

The organizers provided an already strong baseline at 72% F1 over essays. As we can see in Table 1, our BPE based systems and NN system were only able to be on par with the baseline, with the Naive Bayes using BPE n-grams only slightly surpassing it. However, as shown in Table 2, the Naive Bayes approach is indeed very complementary to the baseline. It performs well above the baseline for German, Italian, and Spanish, while performing much worse for Arabic and Telugu. The fusion system results confirms this hypothesis, showing the best result for all languages. The NN model also shows complementarity in a smaller scale that still provides a positive impact in the final ensemble.

Considering the FUSION track, both NN-FUSION and FINAL-FUSION systems significantly surpassed the baselines, as can be seen in Table 3. This is due mainly because of the use of the i-vectors in both systems. The NN-FUSION system without i-vectors, for example, performed 5 points worse than the baseline with no i-vector (not shown in the tables).

System	F1 (macro)	Accuracy
Baseline (1)	0.7230	0.7236
langid (2)	0.5469	
Naive Bayes 1-gram (3)	0.5912	0.5918
NN-ESSAYS (4)	0.7127	0.7145
Naive Bayes 1,2,3-grams (5)	0.7210	0.7227
Naive Bayes BPE 1,2,3-grams (6)	**0.7294**	**0.7309**
LLR-FUSION (1)+(4)+(6)	0.7949	0.7945

Table 1: Results for the ESSAYS track over the development dataset for the baseline, the *langid*, the Naive Bayes with and without BPE, the essay Neural Network (NN-ESSAYS), and the LLR-FUSION systems. The best result, excluding the LLR-FUSION, is highlighted in bold.

L1	Baseline	NB+BPE	NN-ESSAYS	FINAL-FUSION
ARA	**0.74**	0.67	0.65	0.76
CHI	**0.75**	0.74	0.74	0.84
FRE	0.74	**0.77**	0.72	0.81
GER	0.79	**0.85**	0.81	0.93
HIN	**0.69**	0.66	**0.69**	0.71
ITA	0.76	**0.83**	0.80	0.86
JPN	0.74	**0.76**	0.69	0.82
KOR	0.69	**0.70**	0.68	0.75
SPA	0.61	**0.68**	0.66	0.73
TEL	**0.73**	0.65	0.71	0.77
TUR	**0.72**	0.70	0.67	0.77
avg	0.72	**0.73**	0.71	0.79

Table 2: F1 (macro) scores on the ESSAYS track over the development dataset for the baseline, the Naive Bayes with BPE (NB+BPE), the essay Neural Network (NN-ESSAYS), and the FINAL-FUSION systems. The best result for each language, excluding the FINAL-FUSION, is highlighted in bold.

System	F1 (macro)	Accuracy
Baseline fusion	0.7500	0.7500
Baseline fusion+i-vectors	0.7809	0.7827
NN-FUSION	0.8238	0.8245
FINAL-FUSION	0.9011	0.9009

Table 3: Results for the FUSION track over the development dataset for the baselines, the fusion Neural Network (NN-FUSION), and the FINAL-FUSION systems.

5.1 Test set

As previously mentioned, the NLI Shared Task 2017 provided two test datasets, one for the ESSAYS and SPEECH tracks, and one for the FUSION track. We focused on the FUSION track test set for comparing the systems described above.

Table 4 shows the results of our two best single systems trained only on the ESSAYS dataset. The difference in performance is proportional to that on the development set shown in Table 1.

Table 5 shows the results on the FUSION track for our two fusion systems trained on all available data: essays, speech transcriptions, and i-vectors. The NN-FUSION was the only system we officially submitted to the task. After advancing with the other complementary systems and the FINAL-FUSION one, the results we achieved make it clear Neural Networks are not the best alterna-

tive for combining multiple sources of information, at least not in the simple way we approached it.

6 Discussion

Concerning the text component of the problem, we focused on simple word and sub-word features avoiding excessive use of hand engineered features. We also tested linear and recurrent neural networks based classifiers to attain complementary models. We then relied on fusion methods for combining our simple approaches and the task provided i-vectors.

Compared with the existing results, performance on the text component of the tasks was limited, with small improvements over the baseline. The obtained models were however complementary to each other and the baseline system, providing additional gains when ensembled. The use of BPE has shown as well to be a possible alternative to other sub-word units usually employed in NLI systems.

The outstanding performance of the i-vectors, consistent with findings of previous works, is the main driver in the final system's performance. The main improvements shown reside therefore in the alternative fusion strategy followed for the different systems.

Following the current trend, a possible line of work is to explore sub-word information combined with recurrent or convolutional neural architectures. In addition, more complex neural architectures can also be explored, like hierarchical classification models with attention, which are recently obtaining good results in other document classification problems.

Acknowledgements

Fabio Kepler gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for several experiments reported in this paper. This work was partially supported by Portuguese national funds through – Fundação para a Ciência e a Tecnologia (FCT) under Project UID/CEC/50021/2013.

References

Alberto Abad, Eugénio Ribeiro, Fábio Kepler, Ramón Fernández Astudillo, and Isabel Trancoso. 2016. Exploiting phone log-likelihood ratio features for the detection of the native language of non-native english speakers. In *Interspeech 2016*. pages 2413–2417.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Lille, France, volume 37 of *Proceedings of Machine Learning Research*, pages 2067–2075. http://proceedings.mlr.press/v37/chung15.html.

Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011a. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(4):788–798.

Najim Dehak, Pedro A Torres-Carrasquillo, Douglas A Reynolds, and Reda Dehak. 2011b. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*. pages 857–860.

Daniel Garcia-Romero and Carol Y. Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*.

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)*. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. *Intelligence and Security Informatics* pages 41–76.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*. Citeseer.

Shervin Malmasi. 2016. *Native Language Identification: Explorations and Applications*. Ph.D. thesis. http://hdl.handle.net/1959.14/1110919.

Shervin Malmasi and Mark Dras. 2017. Native Language Identification using Stacked Generalization. *arXiv preprint arXiv:1703.06541* .

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

System	F1 (macro)	Accuracy
Baseline	0.7109	0.7112
NN-ESSAYS	0.7273	0.7269
Naive Bayes BPE 1,2,3-grams	0.7445	0.7404

Table 4: Results over the FUSION test dataset trained only on the ESSAYS for the baseline and our two best single systems.

System	F1 (macro)	Accuracy
Baseline fusion	0.7790	0.7790
Baseline fusion+i-vectors	0.7900	0.7900
NN-FUSION	*0.8377*	*0.8391*
FINAL-FUSION	0.9018	0.9018

Table 5: Results for the FUSION track over the test dataset for the baseline and our two fusion systems. The emphasized system, NN-FUSION, was the only officially submitted to the task.

Third DSL Shared Task. In *Proceedings of the VarDial Workshop*. Osaka, Japan.

David Martınez, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka. 2011. Language recognition in ivectors space. In *Interspeech*. pages 861–864.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the International Conference on Machine Learning*. pages 807–814.

Bjrn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Interspeech 2016*. pages 2001–2005. https://doi.org/10.21437/Interspeech.2016-129.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

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 Building Educational Applications Using NLP*. 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.