Cross-corpus Native Language Identification via Statistical Embedding

Francisco Rangel
Universitat Politècnica de València
Valencia - Spain
francisco.rangel@autoritas.es

Paolo Rosso
Universitat Politècnica de València
Valencia - Spain
prosso@dsic.upv.es

Alexandra L. Uitdenbogerd
RMIT University
Melbourne - Australia
sandra.uitdenbogerd@rmit.edu.au

Julian Brooke
Thomson Reuters - Canada
julian.brooke@gmail.com

Abstract

In this paper, we approach the task of native language identification in a realistic cross-corpus scenario where a model is trained with available data and has to predict the native language from data of a different corpus. We have proposed a statistical embedding representation reporting a significant improvement over common single-layer approaches of the state of the art, identifying Chinese, Arabic, and Indonesian in a cross-corpus scenario. The proposed approach was shown to be competitive even when the data is scarce and imbalanced.

1 Introduction

Native Language Identification (NLI) is the task of identifying the native language (L1) of a writer based solely on a textual sample of their writing in a second language (L2), for example, essays in English by students from China, Indonesia or Arabic-speaking countries. NLI is very important for education, since it can lead to the provision of more targeted feedback to language learners about their most common errors. It is also of interest for forensics, security and marketing. For example, knowing the possible native language of the user who wrote a potentially threatening message may help to better profile that user and the potential scope of the threat.

The first Native Language Identification shared task was organised in 2013 (Tetreault et al., 2013). The twenty-nine teams had to classify essays written in English (L2) in one of the eleven possible native languages (L1). The most common features were word, character and POS n-grams, and the reported accuracies rose to 83.6%. The Support Vector Machine (SVM) has been the most prevalent classification approach. Furthermore, participants were allowed to train their models with external data, specifically *i)* any kind of external data, excluding TOEFL[1] (Blanchard et al., 2013); or *ii)* any kind of external data, including TOEFL. Participants such as Brooke and Hirst (Brooke and Hirst, 2013) combined data from sources such as Lang8,[2] ICLE[3] (Granger, 2003), FCE[4] (Yannakoudakis et al., 2011), and ICNALE[5] (Ishikawa, 2011). The reported accuracies show that, when training only with external data, the results fall to 56.5%. Recently, the 2017 Native Language Identification Shared Task (Malmasi et al., 2017) has been organised with the aim of identifying the native language of written texts, alongside a second task on spoken transcripts and low dimensional audio file representations as data (although original audio files were not shared). The organisers included the macro-averaged F1-score (Yang and Liu, 1999) since it evaluates the performance across classes more consistently. Although deep learning approaches were widely used, the best results (up to 88.18%) were achieved with classical methods such as SVM and n-grams. Despite participants being allowed to use external data, there were no such submissions, possibly also due to the poor results obtained in the previous edition (56.5% of accuracy).

We are interested in the following cross-corpus scenario: a model trained with data from external sources (e.g. social media). The authors in (Malmasi and Dras, 2015) used the EF Cambridge Open Language Database (EFCamDat)[6] (Geertzen et al., 2013) for training and

[1] https://www.ets.org/research/policy_research_reports/publications/report/2013/jrkv

[2] http://www.lang8.com

[3] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

[4] https://ilexir.co.uk/datasets/index.html

[5] http://language.sakura.ne.jp/icnale/

[6] https://corpus.mml.cam.ac.uk/efcamdat1/

Proceedings of the 2nd Workshop on Stylistic Variation, pages 39–43
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

TOEFL for evaluation, and vice versa. They trained a linear-SVM with several features such as function word unigrams and bigrams, production rules and part-of-speech unigrams, bigrams and trigrams, and the combination of all of them. The authors reported an accuracy of 33.45% when training with EFCamDat and evaluated on TOEFL, and an accuracy of 28.42% when training on TOEFL, and evaluated on EFCamDat, in contrast to the accuracy of 64.95% obtained when evaluating intra-corpus. The authors in (Ionescu et al., 2016) evaluated String Kernels in a cross-corpus scenario (TOEFL11 for training and TOEFL11-Big (Tetreault et al., 2012) for evaluation). They reported significant improvements over the state of the art with accuracies up to 67.7%. The authors explain these results by arguing ”that string kernels are language independent, and for the same reasons they can also be topic independent”.

In this work, we propose to follow the methodology represented in Figure 1. Given a set of corpora C, we learn a model with all the corpora together except c, which is used to evaluate the model. To evaluate the task, we have proposed a statistical embedding representation that we have compared with common single-layer approaches based on n-grams, obtaining encouraging results.

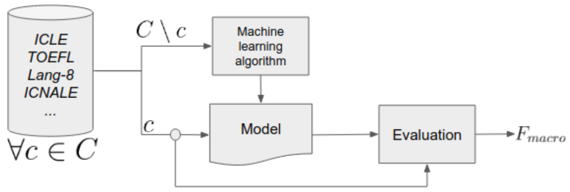

Figure 1: Evaluation methodology.

2 Corpora

Due to the typical geographical origins of students registered in Australian universities, our interest is in Arabic (AR), Chinese (CH) and Indonesian (ID). Arabic is incorporated in TOEFL and Lang8, as well as Chinese. Indonesian is included in Lang8 and ICNALE. The number of documents per corpus and language is shown in Table 1. As can be seen, classes are very imbalanced in most cases. Furthermore, in the case of Indonesian, figures for the ICNALE corpus are very small[7].

NL	Corpus	L1	Others
AR	Lang8	1,139	23,931
	TOEFL	1,103	10,997
CH	Lang8	22,549	16,102
	TOEFL	1,102	10,998
ID	Lang8	1,143	23,923
	ICNALE	8	74

Table 1: Number of documents in each corpus. *L1* corresponds to the documents written by authors of the native language to be identified. *Others* comprise all the documents written by authors of the other native languages in the corpus.

3 Low Dimensionality Statistical Embedding

As shown in (Brooke and Hirst, 2012; Ionescu et al., 2014), single-layer representations such as n-grams are able to obtain competitive results in a cross-corpus scenario. However, n-grams use to be filtered in order to reduce dimensionality, and generally the most frequent ones are selected. Nevertheless, omitting some of the rarest terms is fairly common and necessary for top performance. We propose a Low Dimensionality Statistical Embedding (LDSE) to represent the documents on the basis of the probability distribution of the occurrence of all their terms in the different languages, i.e. L1. Furthermore, LDSE represents texts without the need of using external resources or linguistic tools, nor preprocessing or feature engineering. The intuition is that the distribution of weights for a given document should be closer to the weights of its corresponding native language. The proposed representation relies on descriptive statistics to carry out the comparison among distributions. Formally, we represent the documents as follows.

We calculate the *tf-idf* weights for the terms in the training set *D* and build the matrix Δ. Each row represents a document d_i, each column a vocabulary term t_j, and each cell represents the *tf-idf* weight w_{ij} for each term in each document. Finally, $\delta(d_i)$ represents the assigned native language *c* to the document *i*.

$$\Delta = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1m} & \delta(d_1) \\ w_{21} & w_{22} & \dots & w_{2m} & \delta(d_2) \\ \dots & \dots & \dots & \dots & \\ w_{n1} & w_{n2} & \dots & w_{nm} & \delta(d_n) \end{bmatrix}, \qquad (1)$$

Eq. 2 shows how we obtain the term weights $W(t, c)$ as the ratio between the weights of the documents belonging to a given native language c

[7] We have used the merged text set from the ICNALE Written Essays 2.0

and the total distribution of weights for that term.

$$W(t,c) = \frac{\sum_{d \in D/c=\delta(d)} w_{dt}}{\sum_{d \in D} w_{dt}}, \forall d \in D, c \in C \quad (2)$$

As shown in Eq. 3, these term weights are used to obtain the representation of the documents.

$$d = \{F(c_1), F(c_2), ..., F(c_n)\} \sim \forall c \in C, \quad (3)$$

Each $F(c_i)$ contains the set of features showed in Eq. 4, with the following meaning: *i)* average and median values of the document term weights; *ii)* minimum and maximum values of the weights in the document; *iii)* first and third quartiles of the weights distribution; *iv)* Gini (Gini, 1912) indexes (to measure the distribution skewness and kurtosis); and *vi)* the nine first moments (Bowman and Shenton, 1985) (the more moments two distributions share, the more similar they are).

$$F(c_i) = \{avg, median, min, max, \\ Q_1, Q_3, G_1, G_2, M_2, .., M_{10}\} \quad (4)$$

Finally, these weights are learned with a machine learning algorithm. We have tested several machine learning algorithms and we report the ones with the best results: *i)* Naive Bayes in Lang8 for Arabic and Indonesian, as well as in ICNALE for Indonesian; *ii)* Simple Logistic in TOEFL for Arabic; *ii)* SVM in TOEFL and Lang8 for Chinese; and *iii)* Neural Networks in the Student Writing Task (SWT) for the three languages.

As can be seen, this representation reduces the dimensionality to only 17 features per class by statistically embedding the distribution of weights of the document terms, but unlike methods such as PCA or LSA, it takes into account all the terms in the corpus instead of removing those ones that contribute less. We have evaluated several machine learning algorithms and reported the best results obtained.

We have used both character and word n-grams with SVM to compare our proposal since they are the most common features used in the state of the art. We have iterated n from 1 to 5 with the top 100, 500, 1,000 and 5,000 most frequent terms.

4 Experimental scenario

In this section we report and discuss the obtained results. Firstly, we focus on the described corpora for the languages of interest. Then, we analyse as a case study the Australian academic scenario. Due to the imbalance of the data, we use a macro-averaged F1-score which gives the same importance to the different classes no matter their size.

4.1 Results on NLI corpora

Although NLI has most commonly approached as multi-class, the difficulty lining up multiple languages across multiple corpora means that we instead focus here on the one versus all (1va) formulation; we note that in practice multi-class NLI using SVMs is of realized using 1va SVM classification (Brooke and Hirst, 2012), so our results here should extend directly to the multi-class case. Results are presented in Table 2. The second and third columns show respectively the corpus used for training and test: Lang8 and TOEFL include Arabic and Chinese, whereas Indonesian is included in Lang8 and ICNALE. The fourth column shows the best result obtained by the baseline:[8], whereas the fifth column shows the result obtained with LDSE.

NL	Train	Test	Base	LDSE	%
AR	Lang8	TOEFL	54.75	65.30	19.27
	TOEFL	Lang8	51.10	59.60	16.63
CH	Lang8	TOEFL	53.25	56.95	6.95
	TOEFL	Lang8	50.10	52.30	4.39
ID	Lang8	ICNALE	73.05	86.15	17.93
	ICNALE	Lang8	53.75	61.35	14.14

Table 2: Results in macro-averaged F1-score. The baseline corresponds to the best result obtained with character or word n-grams. The last column shows the improvement percentage achieved by LDSE over the baseline.

As can be seen, LDSE significantly[9] outperforms the best results obtained with n-grams for all languages and setups, with improvements from 4.39% up to to 19.27%. The highest improvement has been obtained for Arabic, although the best results were achieved for Indonesian. It is worth mentioning that, as shown in Table 1, the ICNALE corpus is very small: 8 documents for Indonesian and 74 documents for the other 9 native languages. Due to that, especially in the case of evaluating on ICNALE, a small variation in the identification can cause a high variability in the results.

[8] The best results have been obtained with the following setups *i)* character 5-grams; *ii)* word 1-grams; *iii)* character 4-grams; *iv)* character 4-grams; *v)* word 1-grams; *vi)* character 2-grams. In all the cases the 1,000 most frequent n-grams were selected.

[9] T-Student at 95% of significance was used.

Despite Chinese being larger and less imbalanced in Lang8 (as can be seen in Table 1), the overall results are lower and closer to the baseline. No matter the language, the best results have been obtained when training with Lang8. This may be due to the larger size of this dataset, and especially to the freedom of choice of their authors to write about different topics.

4.2 Case study

With the aim of investigating the performance of our approach in the Australian academic scenario, we have tested LDSE on the Student Writing Task (SWT) corpus for the three native languages and compared the results obtained for the previous corpora. SWT contains 32 essays of 200–300 words, written by Computer Science PhD students studying in Australia. The students had 16 different native languages, including: Arabic (4), Chinese (5), and Indonesian (4). The essays all discuss the same topic, being the relative merit of three algorithms.

To train our model, we have used Lang8 together with TOEFL in the case of Arabic and Chinese, and Lang8 with ICNALE in the case of Indonesian. No data from SWT was used for training.

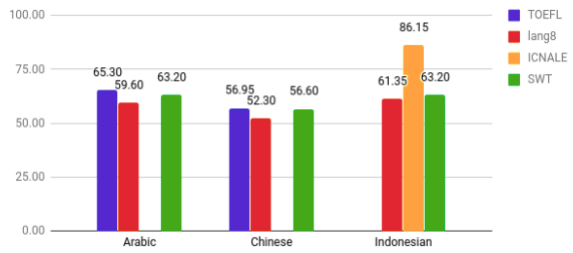

Figure 2: Comparative results of the LDSE model on the different corpora.

As shown in Figure 2, results on SWT are similar to those obtained on Lang8 and TOEFL on the three languages. Specifically, results obtained on TOEFL are slightly better, whereas they are slightly worse in the case of Lang8, without statistical significance in any case. However, results obtained on ICNALE are significantly higher.

5 Conclusions and future work

In this work, we have approached the task of identifying the native language of authors based on their written text in English, focussing on the languages of the main geographical origins of students in the Australian academic environment: Arabic, Chinese, and Indonesian.

We have proposed the LDSE statistical embedding approach that considers descriptive statistics such as the distribution skewness and kurtosis (Gini indexes) as well as the moment information to represent the documents of the three different classes (native languages). We have evaluated LDSE on the available corpora, showing a higher performance than SVM approaches based on n-grams that obtained the bests results in the NLI previous shared tasks. Finally, we have evaluated LDSE also on the written essays of the SWT case study, showing its competitiveness from a cross-corpus perspective despite the small size and imbalance degree of the corpus.

Although it is typical to treat NLI as a multi-class problem instead of 1-vs-all, the main difficulty would be to line up multiple languages across multiple corpora. Furthermore, our interest is in identifying whether the L1 is Arabic, Chinese or Indonesian. We aim to address NLI from multi-class perspective as a future work.

Acknowledgments

The research of the second author was partially funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project. The work on the data of authors whose native language is Arabic as well as this publication were made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the first two authors.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series* 2013(2).

Kamiko O Bowman and Leonard R Shenton. 1985. Method of moments. *Encyclopedia of statistical sciences* 5:467–473.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. *Proceedings of COLING 2012* pages 391–408.

Julian Brooke and Graeme Hirst. 2013. Using other learner corpora in the 2013 nli shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications.* pages 188–196.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*.

CW Gini. 1912. Variability and mutability, contribution to the study of statistical distribution and relaitons. *Studi Economico-Giuricici della R* .

Sylviane Granger. 2003. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* 37(3):538–546.

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)*. pages 1363–1373.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2016. String kernels for native language identification: Insights from behind the curtains. *Computational Linguistics* 42(3):491–525.

Shinichiro Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the icnale project. *Corpora and language technologies in teaching, learning and research* pages 3–11.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1403–1409.

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 Innovative Use of NLP for Building Educational Applications*. pages 62–75.

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*. pages 48–57.

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* pages 2585–2602.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 42–49.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 180–189.