A Deep Generative Approach to Native Language Identification

Ehsan Lotfi, Ilia Markov, Walter Daelemans
CLiPS Research Center
University of Antwerp, Belgium
{ehsan.lotfi, ilia.markov, walter.daelemans}@uantwerpen.be

Abstract

Native language identification (NLI) – identifying the native language (L1) of a person based on his/her writing in the second language (L2) – is useful for a variety of purposes, including marketing, security, and educational applications. From a traditional machine learning perspective, NLI is usually framed as a multi-class classification task, where numerous designed features are combined in order to achieve state-of-the-art results. We introduce a deep generative language modelling (LM) approach to NLI, which consists in fine-tuning a GPT-2 model separately on texts written by the authors with the same L1, and assigning a label to an unseen text based on the minimum LM loss with respect to one of these fine-tuned GPT-2 models. Our method outperforms traditional machine learning approaches and currently achieves the best results on the benchmark NLI datasets.

1 Introduction

The goal of native language identification (NLI) is to automatically identify the native language (L1) of a writer based solely on his or her texts written in the second language (L2), which can inform marketing, educational, security and forensic applications, e.g., by narrowing down the list of candidate authors of a text under investigation. NLI is an interesting example of a task which is hard to perform for humans: a study of human performance in NLI (Malmasi et al., 2015) showed that they achieve around 37% accuracy, while automated NLI systems perform in the 80%–90% accuracy range, depending on the number of languages being considered, amount of data, etc.

NLI is usually approached from a machine learning perspective as a multi-class classification problem of assigning class labels representing L1s to texts written in L2, where the main focus (of the traditional machine learning) is to design features that capture the systematic fingerprints of the first language in the second language writing (native language interference (Odlin, 1989)). Numerous feature types that capture various aspects of the interference phenomenon have been explored for NLI: spelling errors (Koppel et al., 2005; Chen et al., 2017); lexical features, e.g., word and lemma n-grams (Jarvis et al., 2013), cognates (Markov et al., 2019), etymologically-related words (Nastase and Strapparava, 2017); syntactic features, e.g., context-free grammar features (Wong and Dras, 2011), Stanford parser dependency features (Tetreault et al., 2012); stylometric features, e.g., punctuation (Markov et al., 2018a), character n-gram features (Kulmizev et al., 2017); emotion-based features (Markov et al., 2018b), etc. The combination of such features provides the best results for NLI, as shown by the two shared tasks on the NLI task organized in 2013 (Tetreault et al., 2013) and 2017 (Malmasi et al., 2017), where the two top-ranked systems (Cimino and Dell'Orletta, 2017; Markov et al., 2017) used Support Vector Machines (SVM) with a variety of engineered features.

NLI has also been approached using deep neural networks: Franco-Salvador et al. (2017) used word embeddings, reporting a slight increase in performance when they are combined with string kernels in a meta-learning set-up; Chen (2016) and Bjerva et al. (2017) used convolutional neural networks (CNN) and long short-term memory networks (LSTM), concluding that traditional methods, i.e., SVM with engineered features, appear to work better for NLI.

OpenAI's Generative Pre-trained Transformer-2 (GPT-2) (Radford et al., 2019) is a unidirectional transformer-based language model pretrained on 40 GB of text data with the objective of predicting

Proceedings of the 28th International Conference on Computational Linguistics, pages 1778–1783
Barcelona, Spain (Online), December 8-13, 2020

the next word, given all of the previous words within a text. GPT-2 is able to generate conditional synthetic texts without task-specific training, achieves the state-of-the-art zero-shot results on a variety of domain-specific language modeling tasks, and was successfully applied to various other NLP tasks such as question answering, summarization, reading comprehension, and machine translation (Radford et al., 2019). We show that these generative models, due to their ability to adapt to the style and content of the conditioning text, are able to capture L1 peculiarities and can be successfully used for identifying the first language of the author. We propose a new approach exploiting the GPT-2 model for text classification and evaluate it on the task of native language identification. Our approach consists in fine-tuning a GPT-2 model on texts written by the authors with the same L1, and assigning a label to an unseen text according to the minimum LM loss with one of the fine-tuned GPT-2 models.

The contributions of the work presented here are the following: (i) we introduce a novel approach for text classification using generative models, and (ii) we evaluate the approach on the NLI task, improving the state-of-the-art results.

2 Data

We evaluate our approach on two datasets commonly used in NLI research:

TOEFL11 (Blanchard et al., 2013): the ETS Corpus of Non-Native Written English (TOEFL11) contains 1,100 essays in English for each of the 11 L1s: Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). In total there are 12,100 essays with on average 348 tokens per essay. The essays were written in response to eight different writing prompts, all of which appear in all 11 L1 groups, by authors with low, medium, or high English proficiency. The dataset is considered a benchmark dataset for NLI and was used in two shared tasks on the NLI task (Tetreault et al., 2013; Malmasi et al., 2017).

ICLE (Granger et al., 2009): the ICLEv2 dataset consists of essays written by highly-proficient non-native college-level students of English. For comparability, we used a 7-language subset of the corpus normalized for topic and character encoding (Tetreault et al., 2012; Ionescu et al., 2014) (the so-called ICLE-NLI subset (Tetreault et al., 2012) to which we refer as ICLE). This subset contains 110 essays for each of the 7 languages: Bulgarian (BUL), Chinese (CHI), Czech (CZE), French (FRE), Japanese (JPN), Russian (RUS), and Spanish (SPA) (in total 770 essays with on average 747 tokens per essay).

3 Methodology

Instead of training a classifier, we fine-tune a pretrained generative model (GPT-2) on the subset of the training samples written by the authors with the same L1, which provides us with n GPT-2 models (n equals the number of L1s (classes)). Each of these models has learned the characteristics of a certain L1 group, which we will use to discriminate between new samples: at inference time, we feed an unseen text to all these models and calculate their LM loss. Based on the fine-tuning, we expect to get the least LM loss from the model that is trained on the same class of texts, thus the assigned label is the argmin of all models' losses. An example of assigning a label to an unseen text (Turkish L1 in this case) is shown in Figure 1[1]. Although the LM loss (as a measure of likelihood) has been used before as an evaluation metric (Mehri and Eskenazi, 2020), to the best of our knowledge, it has not been exploited in this way for text classification tasks in general and for NLI in particular.

4 Results and Discussion

To compare the performance of our approach with the previously published state-of-the-art results (Markov, 2018), we report the results in terms of classification accuracy on the TOEFL11 test set, as well as on the TOEFL11 dataset under 10-fold cross-validation (10FCV) and on the ICLE dataset under 5-fold cross-validation (5FCV). To better situate the method, several baseline models have also

[1] The LM loss scores are provided for a randomly chosen essay in the TOEFL11 dataset: 1262.txt.

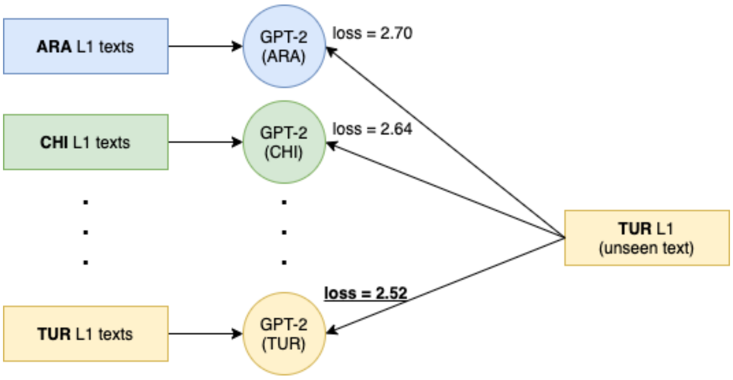

Figure 1: Example of classifying an unseen text written by a Turkish native speaker.

been selected and evaluated, including bag-of-words (BoW), BERT (Devlin et al., 2019) and an LSTM-based (Hochreiter and Schmidhuber, 1997) language model approach similar to our method but with the difference that models are trained from scratch. The implementation has been done in Pytorch (Paszke et al., 2019), using the GPT-2 medium model from the HuggingFace transformers library (Wolf et al., 2019), fine-tuned for 2-4 epochs. As for the baselines, the LSTMs are 2-layer networks with 650 hidden units per layer and a hidden dimension of 650, trained for 25 epochs with a batch size of 20. The BERT model is the standard base-uncased version trained for 12-16 epochs with a batch size of 12.

	TOEFL11	TOEFL11	ICLE
	(test set)	(10FCV)	(5FCV)
BoW (SVM)	71.1	68.7	80.6
BERT	80.8	76.3	76.8
LSTM	77.8	75.9	77.9
Best shared task 2017 (SVM)	88.2	86.4	–
Best after (previous SOTA, SVM)	88.6	86.5	93.4
Ours	**89.0**	**86.8**	**94.2**

Table 1: Results in terms of classification accuracy (%) on the TOEFL11 and ICLE datasets.

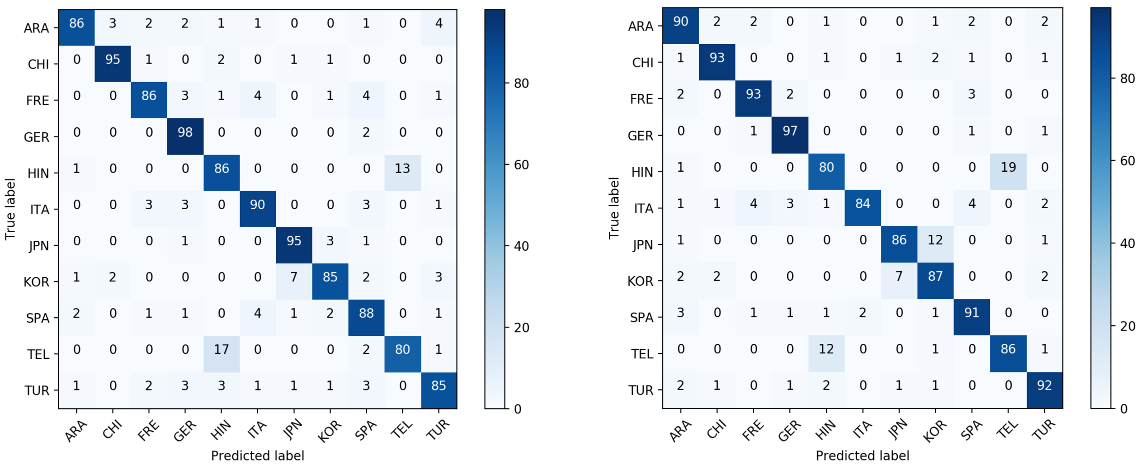

Figure 2: TOEFL11 test set: previous SOTA (left; source: (Markov, 2018)), our approach (right).

The results presented in Table 1 indicate that our method outperforms the baseline methods and the previous state-of-the-art results on both datasets for all the considered settings. In line with previous work on NLI (Chen, 2016; Bjerva et al., 2017), deep learning models (BERT and LSTM in our case) provide lower results when used for classification than the traditional machine learning approaches. This may be related to the specific nature of the NLI data, particular to the speakers of the same L1, and to the limited amount of training data. However, our approach seems to capture these peculiarities.

In a detailed, per-language view, presented through the confusion matrices for the TOEFL11 test set (Figure 2), we can note that the trends are similar to the SVM approach with a large number of engineered features (Markov, 2018): German is the easiest L1 to identify, while Hindi and Telugu, and Korean and Japanese are the most problematic L1s, with the highest degree of confusion between them. This indicates that, despite the relatively small size of the datasets (for deep learning model standards: around 70k tokens in the ICLE dataset), fine-tuning allows to focus on and learn the differentiating factors, as also evidenced by the lower LSTM results when the models are trained from scratch. Our model manages to pick up discriminative characteristics of L1 writing from this limited amount of data.

The GPT-2 model represents a language model combining many writing styles and topics. Since in our data the topics are nearly-equally distributed across the represented L1 groups (Blanchard et al., 2013), by fine-tuning, the derived models have learned to focus on the interference cues particular to different L1s. Figure 3 illustrates the way the loss of different language models varies when receiving a text token by token (L1=TUR; TOEFL11: 1262.txt). Universal surges in the loss value (e.g., first 7 tokens and also around token 45 in this case) usually correspond to typos or non-discriminative grammar errors (in this case, the aforementioned intervals contain ‘People have got different qualities from each other ...’ and ‘...take risks rathet than only...’), while singular trends can be attributed to discriminative patterns. In this case, for example, a decisive bifurcation happens around token 80, after which the TUR model maintains a winning margin against the other languages. Interestingly, the interval corresponds to ‘...they feel themselves the happiest person in the world . They think the result of this accoplishment...’, containing examples of preposition errors, which is one of the most frequent types of errors made by Turkish learners of English as L2 (Kirkgoz, 2010; Tasci and Aksu Atac, 2018).

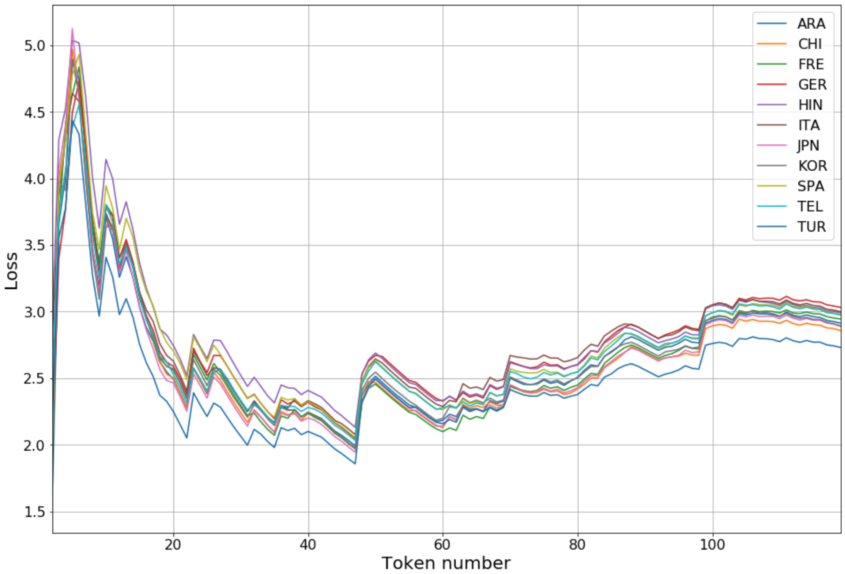

Figure 3: Variation of the LM loss value when the models receive a text token by token (L1=TUR).

5 Conclusions

We proposed a novel method to address the NLI task. The method consists in fine-tuning a GPT-2 model with the training data grouped by the L1 group into one model per L1, and selecting the fine-tuned model with the lowest loss to assign the L1 class for an unseen text. This approach, simple and not requiring any feature engineering, nevertheless achieves the best results on the datasets commonly used in NLI research. It also improves upon deep learning methods applied directly to the classification task.

The method is generic in that many text categorization tasks (both stylistic and content-based ones) can be framed this way. We therefore hypothesize that the approach can be used for any other style-based classification tasks if the (test) samples are long enough, so that the difference in the LM loss is significant. Testing this hypothesis is our main direction for future work.

6 Acknowledgement

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme and the Flemish Research Foundation through the bilateral research project FWO G070619N “The linguistic landscape of hate speech on social media”.

References

Johannes Bjerva, Gintare Grigonyte, Robert Östling, and Barbara Plank. 2017. Neural networks and spelling features for native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 235–239, Copenhagen, Denmark. ACL.

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):i–15.

Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 542–546, Vancouver, Canada. ACL.

Lingzhen Chen. 2016. Native language identification on learner corpora. Master’s thesis, University of Trento, Department of Information Engineering and Science, Trento, Italy.

Andrea Cimino and Felice Dell’Orletta. 2017. Stacked sentence-document classifier approach for improving native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 430–437, Copenhagen, Denmark. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, pages 4171–4186, Minneapolis, USA. ACL.

Marc Franco-Salvador, Greg Kondrak, and Paolo Rosso. 2017. Bridging the native language and language variety identification tasks. *Procedia Computer Science*, 112:1554–1561.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2 (ICLE).* Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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*, pages 1363–1373, Doha, Qatar. ACL.

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*, pages 111–118, Atlanta, USA. ACL.

Yasemin Kirkgoz. 2010. An analysis of written errors of Turkish adult learners of English. *Procedia - Social and Behavioral Sciences*, 2:4352–4358.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author’s native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 624–628, New York, USA. ACM.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 382–389, Copenhagen, Denmark. ACL.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and human baselines for native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, USA. ACL.

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*, pages 62–75, Copenhagen, Denmark. ACL.

Ilia Markov, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. CIC-FBK approach to native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 374–381, Copenhagen, Denmark. ACL.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2018a. Punctuation as native language interference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3456–3466, Santa Fe, USA. The COLING 2012 Organizing Committee.

Ilia Markov, Vivi Nastase, Carlo Strapparava, and Grigori Sidorov. 2018b. The role of emotions in native language identification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 123–129, Brussels, Belgium. ACL.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2019. Anglicized words and misspelled cognates in native language identification. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 275–284, Florence, Italy. ACL.

Ilia Markov. 2018. *Automatic Native Language Identification*. Ph.D. thesis, Instituto Politécnico Nacional, Mexico City, Mexico.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. ACL.

Vivi Nastase and Carlo Strapparava. 2017. Word etymology as native language interference. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2692–2697, Copenhagen, Denmark. ACL.

Terence Odlin. 1989. *Language Transfer: cross-linguistic influence in language learning*. Cambridge University Press, Cambridge, UK.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Samet Tasci and Bengü Aksu Atac. 2018. Written grammatical errors of Turkish adult learners of English: An analysis. *Journal of International Social Sciences Education*, 4(1):1–13.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2585–2602, Mumbai, India. The COLING 2012 Organizing Committee.

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, Atlanta, USA. ACL.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland. ACL.