Are BERTs Sensitive to Native Interference in L2 Production?

Zixin Tang
The Pennsylvania State University
zqt5035@psu.edu

Prasenjit Mitra
The Pennsylvania State University
pmitra@psu.edu

David Reitter
Google Research
reitter@google.com

Abstract

With the essays part from The International Corpus Network of Asian Learners of English (ICNALE) and the TOEFL11 corpus, we fine-tuned neural language models based on BERT to predict English learners' native languages. Results showed neural models can learn to represent and detect such native language impacts, but multilingually trained models have no advantage in doing so.

1 Introduction

With globalization progress, more and more people start acquiring more than one language. Therefore, computational models and theories to better understand multilingualism among machines and humans start receiving more and more attention. Even though language models (LMs) succeeded in a wide range of tasks in natural language understanding and processing (i.e. Enguehard et al., 2017; Linzen and Leonard, 2018; Mueller et al., 2020), how to do such models process and represent languages and knowledge remain unclear. We also have few ideas about these models' abilities when facing various types of language input, such as input from non-native speakers or speakers with lower proficiency. In this study, we implemented state-of-the-art neural LMs to predict non-native (L2) speakers with different language backgrounds through native language identification tasks. Furthermore, we investigated whether cross-lingual training can help identify a language learner's writing patterns, which his native languages (L1s) can cause interference. This study wants to expand current knowledge on representations and cross-lingual components in transformer-based LMs.

2 Related Work

With developments in computer science, data sciences, and cognitive and language sciences, artificial language models (LMs) based on language representations and embeddings have received attention in multiple areas. Current deep neural LMs based on various architectures can capture syntactic information through natural language input and utilize language knowledge to complete different tasks, such as identifying specific sentence structures (Enguehard et al., 2017; Kelly et al., 2020; Marvin and Linzen, 2018) and predicting grammar acceptability (Warstadt et al., 2019). Beyond training on a single language, multilingual neural language models could process cross-lingual tasks such as language identification and translation. Studies found that language representations in these types of models are transferable or overlapped across languages to create universal representations of grammatical structures (Chi et al., 2020). Models can also develop language-neutral components for better word alignment and improve performance on several cross-linguistic tasks (Libovickỳ et al., 2019). However, studies indicated that these LMs have lower performance in cross-linguistic tasks such as machine translations (Libovickỳ et al., 2019). Comparing to monolingual models, multilingual LMs have relatively lower performance and unbalanced performance in syntactic-related tasks between English and other languages (Mueller et al., 2020).

Even though transformer-based LMs showed impressive abilities in multiple tasks, few studies tried to analyze whether and how multilingual LMs identify and distinguish writers' native interference in their L2 production caused by their L1s. One task for detecting such interference across speakers is the native language identification task, which predicts a speaker's L1 using his language production in L2. By sharing representations and transferring knowledge across languages (Putnam et al., 2018; Hartsuiker et al., 2004), multilingual speakers can have different structures, preferences, and variations for language comprehension and production than monolingual speakers, result in cross-linguistic behaviors such as cross-lingual priming effects (Bernolet et al., 2007; Hartsuiker et al., 2016, 2004; Gries and Kootstra, 2017) and syntactic preferences in language processing (Hsiao and Gibson, 2003). Previously, statistical-based language models (N-gram models) dominated the tasks (Malmasi et al., 2017), as statistical models are more sensitive in detecting syntactic and morpho-syntactic patterns. Embedding models, on the other hand, are more sensitive in capturing semantic and lexical information and less sensitive in classifying syntactic patterns and differences (Vajjala and Banerjee, 2017).

Given the high performance in syntactic-related tasks (Devlin et al., 2018; Mueller et al., 2020), comparing to previous embedding-based models, transformer-based models may perform better in detecting native language interference. Such advantages can help models better identify specific patterns in L2 learners. Furthermore, LMs with multilingual pre-training may be benefited in identifying different non-native speakers using additional knowledge across languages. However, existed results contradicted such assumptions, indicating that multilingual LMs generally performed worse than monolingual models in the same tasks (Mueller et al., 2020). To better understand transformer-based LMs, multilingual pre-training experiences, and language representations in LMs, our study aims to use native language identification task to answer the following research questions:

1. Can transformer-based LMs detect native interference among L2 learners with different L1s?
2. Can multilingual-LMs transfer knowledge from other languages to gain advantages in such native language prediction tasks?

3 Methods

3.1 Corpora and data pre-processing

We used two different corpora in this study: the TOEFL11 corpus (Blanchard et al., 2013) used in previous native language identification shared task (Malmasi et al., 2017), and the written essays from The International Corpus Network of Asian Learners of English (ICNALE)[1] (Ishikawa, 2013). The TOEFL11 corpus contains TOEFL essays written by English learners with 11 native languages, while ICNALE corpus contains short essays written by English learners from 10 regions and English native speakers. We used document-level input to capture general writing patterns and styles among writers. We excluded essays with less than 50 words for future processes. Both corpora have held-out testing data sets: for TOEFL11, the development set serves as the held-out testing set; for ICNALE, we randomly selected 20% essays to serve as the testing set. Table 1 is the detailed description for each corpus.

Two potential factors could interfere with the final result: speakers' proficiency in the target language (English) and the semantic/lexical information inside essays. To minimize effects from the proficiency, we further developed two types of subsets to examine whether and how proficiency influences transformer-based LMs in native language identification tasks during the fine-tuning process:

- Proficiency subsets (**B1 and Mid set**): serve as proficiency control during the fine-tuning process. These subsets included essays written by intermediate level (B1) English learners in ICNALE and speakers with median TOEFL scores (Mid) from TOEFL11.
- Balanced subset (**BA set**): serve as data input control during the fine-tuning process to provide better comparisons to proficiency subsets. These subsets had the same number of essays as the proficiency subsets, formed by randomly selected essays from the original fine-tuning set.

We also developed another version of testing sets by randomly shuffling the original texts in the essays. Through the shuffling process, we minimized the impact of syntactic information, leaving most of the lexical and semantic information available inside the testing sets. This shuffling test set can examine whether LMs rely on lexical and semantic information to process native language identification tasks.

3.2 Baseline and models for comparison

For models fine-tuned on ICNALE data, essays were converted into measures of syntax complexity using syntax complexity analyzer developed by Lu (2010). The baseline model is a multiple logistic regression model developed with outcome measures. Since there is no previous identification task involves ICNALE, we did not include any other models for comparison.

For models fine-tuned on TOEFL11 data, we

[1] The ICNALE corpus: http://language.sakura.ne.jp/icnale/

Corpus	# of documents	# of prompts	Length	# of L1 background
ICNALE	5600 (**1200**)	2	173 - 479	11
TOEFL11	9892 (**1100**)	6	100 - 795	11

Table 1: Corpus Description. Bold numbers are the size of testing sets.

used the baseline unigram model mentioned in Malmasi et al. (2017). We also included some models mentioned in Malmasi et al. (2017) with the best performance in statistic-based models and embedding-based models. The best statistic-based model (stacked SVM, Cimino and Dell'Orletta, 2017) is an SVM-based model stacked with lexical, morpho-syntactic, and syntactic features. Examples of features include word-level and character-level N-grams, part-of-speech N-grams, and dependency N-grams. The embedding-based model (Doc2Vec, Vajjala and Banerjee, 2017) is a document-level embedding model with concatenated representations from 11 individual trained models representing 11 categories in the original data set.

3.3 Model fine-tuning

For this study, we fine-tuned pre-trained BERT models to predict a speaker's L1 background through different sets of corpora mentioned above. Introduced in 2018, the BERT model is still considered a top pre-trained model in current natural language processing studies. One advantage of using BERT comparing to other non-pre-trained models is its rich knowledge in lexical information, syntactic representations, and semantic networks, which comes from the large scale of training data (Devlin et al., 2018). For this study, two different BERT models will be fine-tuned with multiple datasets: the 12-layer English BERT_base model (BERT-base) and the 12-layer multilingual BERT_base model (mBERT-base).

We fine-tuned all models with an additional linear classification layer through the Transformer package. The classification layer is directly fine-tuned on the final [CLS] embedding with the default Adams optimizer, a learning rate of 2e-5, and an early-stop of 8 epoch. The performance of each model was reported with F-scores using the held-out testing set.

4 Results

Overall, all models with different architectures and fine-tuning data sets captured some features among speakers with different L1 backgrounds, as accuracy is better than random guessing and baseline models. Table 2 and Table 3 showed the results of models trained on ICNALE and TOEFL11 corpus.

	Full	B1	BA
BERT-base	0.86	**0.84**	**0.85**
mBERT-base	**0.87**	0.83	0.83
Random	0.15	0.18	0.15
Baseline	0.43	0.41	0.43

Table 2: Models weighted F-scores for Native Language Identification: ICNALE

	Full	Mid
BERT-base	0.72	0.64
mBERT-base	0.68	0.62
Doc2Vec	0.71	-
Stacked SVM	**0.88**	-
Random	0.09	0.09
Baseline	0.71	-

Table 3: Models weighted F-scores for Native Language Identification: TOEFL11. Results of Doc2Vec (Vajjala and Banerjee, 2017) and Stacked SVM (Cimino and Dell'Orletta, 2017) come from Malmasi et al. (2017).

Performance. The models fine-tuned with TOEFL11 data did not beat the state-of-the-art native language identifier using statistical models. For models fine-tuned with ICNALE data, controlling learners' proficiency did not critically impact models in identification. This showed that speaker proficiency did not strongly interfere with language background identification, indicating the robustness of native language interference across learners' L2 written production. Shuffling word orders did impact all models severely in prediction (Table 4), which showed that BERT models cannot identify and detect native language interference solely with lexical information in the essays. In summary, transformer-based LMs can capture and involve representations of word orders, general writing patterns, and sentence structures to detect English learners' native languages, and rely less on lexical or semantic information.

Multilingual pre-training. When using the

Corpus	Full	Shuffled
ICNALE	0.86 (**0.87**)	0.24 (**0.29**)
TOEFL11	0.72 (**0.68**)	0.46 (**0.26**)

Table 4: Performance of BERT (**mBERT**) models on original and shuffle text. All models were fine-tuned without proficiency control.

same data sets, models fine-tuned on mBERT had slightly lower or similar results comparing to models fine-tuned on BERT. Additional knowledge across multiple languages did not help in predicting speakers' syntax background in their L1s.

5 General Discussion

While transformer-based LMs reached impressive results in language processing and generation, we still know little about how LMs acquire, process, and store language representations. Our study used native language identification task to investigate 1) whether transformer-based models (BERT & mBERT) can detect native interference in written production; 2) whether multilingual pre-training help in identifying native interference. Overall, the results were negative, indicating 1) compared to statistical models, transformer-based models are less capable of capturing native impacts and interference, and 2) additional knowledge across multiple languages in mBERT did not help predict such impacts across multilingual speakers.

5.1 Statistical vs. embedding models

Compared to state-of-the-art native language identifiers relying on statistical methods (such as N-grams) with F-scores above 0.85, the performance of embedding models is still worse. Previous studies (Vajjala and Banerjee, 2017; Jing et al., 2020) argued that such worse predictions from embedding models indicate embedding models might have a stronger ability in capturing semantic information; while statistical-based language models might have better performance in capturing morphosyntactic and syntactic information.

Another explanation for the current lower performance for our embedding models relates to the knowledge structures in BERTs. Previously, studies indicated that for BERTs, the mid-layers have the best performance in syntactic-related tasks (Kelly et al., 2020). Such results may indicate that, unlike previous statistical models that reach the highest performance in output layers, embedding-based models like BERTs may store different types of knowledge across different layers: the deeper the layers are, the more abstract concepts they will store. Therefore, mid-layer embeddings may contain and utilize more syntactic and morpho-syntactic information than final-layer embeddings, leading to a better performance in language interference identification. The simplicity of our classification layer may also lead to lower performance across our BERT-based models. Future studies can further investigate how embeddings from different layers of BERTs respond to these similar tasks, and how different classification algorithms change the overall performance.

5.2 Multilingualism

Another key result from this study, as mentioned previously, is that cross-lingual pre-training did not help in predicting English learner's L1 background. Based on previous findings of language-neutral components within the models (Libovickỳ et al., 2019) and existences of universal linguistics (Chi et al., 2020), we expected that models trained with multiple languages should be more sensitive to syntactic preferences or specific patterns in English interfered with by L2 syntactic knowledge. Unfortunately, similar to the study by Mueller et al. (2020), we did not find advantages of cross-lingual pre-training for native language identification tasks. We had two potential explanations for such results: the training data size and the representation structures in mBERT.

Through the pre-training process, mBERT uses a smaller training English corpus for the mBERT compared to the English BERT: English BERT has an addition Book corpus as training data (Devlin et al., 2018)[2]. The smaller English training set may lead to less knowledge and representations for English in mBERT, making it a less proficient "speaker" than English BERT. Previous research also found similar deficits related to language resources and training data size. Mueller et al. (2020) found that mBERT models perform best in English-related tasks; Wu and Dredze (2020) found that lower-resource languages had significantly lower performance than higher-resource languages when handling the same tasks.

The second explanation for mBERT's lower performance belongs to the structures of knowledge within the models. Even though existing evidence

[2] https://github.com/google-research/bert/blob/master/multilingual.md

showed the possibilities of forming components and representations across languages within embedding models (Chi et al., 2020; Libovickỳ et al., 2019; Artetxe et al., 2019). However, Libovickỳ et al. (2019) pointed out, these cross-lingual components are still not sufficient enough for higher-level tasks such as automatic machine translation. Our study provided further evidence on such insufficiency, indicating that while identifying native languages across English learners, knowledge from other languages and the language-general component may not involve, even when interference might appear in writers' original writing.

5.3 Future studies

To build a better multilingual model, we need further investigations on the multilingual pre-training processes, the cross-linguistic components, and how they interact across languages within multilingual LMs. Even though our models did not reach our expectations, future studies can help better understand why it happened. Specifically, we can focus on three areas: layer-wise embeddings within embedding models, size for pre-training data and models, and cross-lingual representations. For embeddings within models, future studies can use probing tasks for layer-wise embeddings to help us better understand how contextual embedding models process, represent, and utilize syntactic information and knowledge. Second, future studies can compare mBERT with smaller monolingual BERT models to balance out the impact on BERT's richer knowledge in high-resource languages. Lastly, we need more studies and methods to measure the cross-lingual components in multilingual models, which help us better understand multilingualism and cross-lingual knowledge in state-of-the-art LMs. In short, more studies, models, and theories are needed to develop LMs with better abilities in multilingual and cross-linguistic tasks to serve multilingual communities better in practical applications.

6 Summary

While transformer-based language models and multilingual language models start dominating the natural language process area, little is known about the knowledge structures within such models. In this study, we explored Bert-like models' performance in identifying a speaker's native language. Our models did not outperform the state-of-the-art native language identifier based on statistical-based models. Multilingual pre-training did not improve the performance in predicting native language, indicating the insufficiencies for detecting writing styles and structural patterns among different multilingual speakers in current LMs. Further explorations are needed to investigate how multilingual LMs store, process, communicate, and transfer representations and knowledge across languages.

Acknowledgements

This study was supported by the National Science Foundation (PAC 1734304). We thank Dr. Michael T. Putnam, Dr. M. Alex Kelly, and Saranya Venkatraman for their comments and suggestions on the original manuscript. We also thank the anonymous reviewers for providing helpful comments on the earlier draft.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Sarah Bernolet, Robert J Hartsuiker, and Martin J Pickering. 2007. Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):931.

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.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. *arXiv preprint arXiv:2005.04511*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of rnns with multi-task learning. *arXiv preprint arXiv:1706.03542*.

Stefan Th Gries and Gerrit Jan Kootstra. 2017. Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, 20(2):235–250.

Robert J Hartsuiker, Saskia Beerts, Maaike Loncke, Timothy Desmet, and Sarah Bernolet. 2016. Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. *Journal of Memory and Language*, 90:14–30.

Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological science*, 15(6):409–414.

Franny Hsiao and Edward Gibson. 2003. Processing relative clauses in chinese. *Cognition*, 90(1):3–27.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, 1:91–118.

Wang Jing, Matthew A Kelly, and David Reitter. 2020. Do we need neural models to explain human judgments of acceptability? In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 1289–1295.

M Alex Kelly, Yang Xu, Jesús Calvillo, and David Reitter. 2020. Which sentence embeddings and which layers encode syntactic structure? In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 2375–2381.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. *arXiv preprint arXiv:1807.06882*.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. *arXiv preprint arXiv:2005.00187*.

Michael T Putnam, Matthew Carlson, and David Reitter. 2018. Integrated, not isolated: Defining typological proximity in an integrated multilingual architecture. *Frontiers in psychology*, 8:2212.

Sowmya Vajjala and Sagnik Banerjee. 2017. A study of n-gram and embedding representations for native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–248.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. *arXiv preprint arXiv:1909.02597*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.