Native Language Identification on Text and Speech

Marcos Zampieri[1], Alina Maria Ciobanu[2], Liviu P. Dinu[2]
[1]University of Wolverhampton, United Kingdom
[2]University of Bucharest, Romania
marcos.zampieri@uni-koeln.de

Abstract

This paper presents an ensemble system combining the output of multiple SVM classifiers to native language identification (NLI). The system was submitted to the NLI Shared Task 2017 fusion track which featured students essays and spoken responses in form of audio transcriptions and iVectors by non-native English speakers of eleven native languages. Our system competed in the challenge under the team name ZCD and was based on an ensemble of SVM classifiers trained on character n-grams achieving 83.58% accuracy and ranking 3rd in the shared task.

1 Introduction

Native language identification (NLI) is the task of automatically identifying non-native speakers' native language based on their foreign language production. As evidenced in Malmasi (2016) NLI is a vibrant research area in NLP and is usually modeled as single-label text classification.

NLI is based on the assumption that the mother tongue influences second language acquisition (SLA) and production. Corpora containing texts and utterances by non-native speakers are used to train systems that are able to recognize features that are prominent in the production of speakers of a particular native language. These features are subsequently used to identify texts (or utterances) that are likely to be written or spoken by speakers of the same language.

There are two important reasons to study NLI. Firstly, there is SLA. NLI methods can be applied to learner corpora to investigate the influence of native language in second language acquisition and production complementing corpus-based and corpus-driven studies. The second reason is a practical one. NLI methods can be an important part of several NLP systems including, for example, author profiling systems developed for forensic linguistics.

This paper presents the system submitted by the ZCD team to the NLI Shared Task 2017 (Malmasi et al., 2017). The organizers of the challenge provided participants with a dataset containing essays and spoken responses in form of transcriptions and acoustic features (iVectors) by non-native English speakers of eleven native languages taking a standardized assessment of English proficiency for academic purposes. Native languages included are: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. To discriminate between these eleven native languages we apply an ensemble of multiple linear SVM classifiers trained on character n-grams. The main motivation behind the choice of this approach is the success of linear SVMs and SVM ensembles in NLI and in similar text classification tasks such as dialect, language variety, and similar language identification as will be discussed in Section 2.

2 Related Work

There have been several NLI studies published in the past few years. Due to the availability of suitable language resources for English (e.g. learner corpora), the vast majority of these studies dealt with English (Brooke and Hirst, 2012; Bykh and Meurers, 2014), however, a few NLI studies have been published on other languages. Examples of NLI applied to languages other than English include Arabic (Ionescu, 2015), Chinese (Wang et al., 2016), and Finnish (Malmasi and Dras, 2014).

To the best of our knowledge, the NLI Shared Task 2013 (Tetreault et al., 2013) was the first

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 398–404
Copenhagen, Denmark, September 8, 2017.

Team	Approach	System Paper
Jarvis	SVM trained on character n-grams (1-9), word n-grams (1-4), and POS n-grams (1-4)	(Jarvis et al., 2013)
Oslo	SVM trained on character n-grams (1-7)	(Lynum, 2013)
Unibuc	String Kernels and Local Rank Distance (LRD)	(Popescu and Ionescu, 2013)
MITRE	Bayes ensemble of multiple classifiers	(Henderson et al., 2013)
Tuebingen	SVM trained on word n-grams (1-2), and POS n-grams (1-5), and syntactic features (dependencies)	(Bykh et al., 2013)
NRC	Ensemble of SVM classifiers trained on character trigrams, word n-grams (1-2), POS n-grams (2-4), and syntactic features (dependencies)	(Goutte et al., 2013)
CMU-Haifa	Maximum Entropy trained on word n-grams (1-4), POS n-grams (1-4), and spelling features	(Tsvetkov et al., 2013)
Cologne-Nijmegen	SVM classifier with TF-IDF weighting trained on character n-grams (1-6), word n-grams (1-2), and POS n-grams (1-4)	(Gebre et al., 2013)
NAIST	SVM trained on character n-grams (2-3), word n-grams (1-2), and POS n-grams (2-3), and syntactic features (dependencies and TSG)	(Mizumoto et al., 2013)
UTD	SVM trained on word n-grams (1-2)	(Wu et al., 2013)

Table 1: Top ten NLI Shared Task 2013 entries ordered by performance.

shared task to provide a benchmark for NLI focusing on written texts by non-native English speakers. A few years later, the 2016 Computational Paralinguistics Challenge (Schuller et al., 2016) included an NLI task on speech data. The NLI Shared Task 2017 combines these two modalities of non-native language production by including essays and spoken responses of test takers in form of transcriptions and iVectors.

The combination of text and speech has been previously used in similar shared tasks such as the dialect identification shared tasks organized at the VarDial workshop series (Zampieri et al., 2017) and described in more detail in Section 2.2.

In the next sections we present the most successful entries submitted for the NLI Shared Task 2013 and their overlap with methods applied to dialect, language variety, and similar language identification.

2.1 NLI Shared Task 2013

The aforementioned NLI Shared Task 2013 (Tetreault et al., 2013) established the first benchmark for NLI on written texts. Organizers of the first NLI task provided participants with the TOEFL 11 (Blanchard et al., 2013) dataset which contained essays written by students native speakers of the same eleven languages included in the NLI Shared Task 2017.

Twenty-nine teams participated in the competition, testing a wide range of computational methods for NLI. In Table 1 we list the top ten best entries ranked by performance along with their respective system description papers.

The best system by Jarvis et al. (2013) applied a linear SVM classifier trained on character, word, and POS n-grams. Seven out of the ten best entries in the shared task used SVM classifiers. This indicates that SMVs are a very good fit for NLI and motivates us to test SVM classifiers in our ensemble-based system described in this paper.

2.2 Overlap with Dialect Identification

In the last few years, we observed a significant and important overlap between NLI approaches and computational methods applied to dialect, language variety, and similar language identification. So far the overlap between the two tasks has not been substantially explored in the literature.

Members of several teams that submitted systems to the NLI Shared Task 2013, some of them presented in Table 1, also participated in the dialect identification shared tasks organized within the scope of the VarDial workshop series held from 2014 to 2017. The three related shared tasks organized at the VarDial workshop thus far are the Discriminating between Similar Languages (DSL) task organized from 2014 to 2017, Arabic Dialect Identification (ADI) organized in 2016 and 2017, and German Dialect Identification (GDI) organized in 2017.

Next we list some of the teams that adapted systems from NLI to dialect identification in the past few years.

- Variations of the string kernels method by the Unibuc team (Popescu and Ionescu, 2013) competed in the ADI task in 2016 (Ionescu and Popescu, 2016) and in 2017 (Ionescu and Butnaru, 2017) achieving the best results.

- Cologne-Nijmegen's TF-IDF-based approach (Gebre et al., 2013) competed in the DSL shared task 2015 (Zampieri et al., 2015a) as team MMS ranking among the top 3 systems.

- A variation of NRC's SVM approach (Goutte et al., 2013) competed in the DSL 2014 (Goutte et al., 2014) achieving the best results.

- Bobicev applied Prediction for Partial Matching (PPM) in the NLI shared task (Bobicev, 2013) with results that did not reach top ten performance. A similar improved approached competed in the DSL 2015 (Bobicev, 2015) ranking in the top half of the table.

- A similar approach to the one by Jarvis (Jarvis et al., 2013) that ranked 1st place in the NLI task 2013 competed in the DSL 2017 (Bestgen, 2017), achieving the best performance in the competition.

- Variations of MQ's SVM ensemble approach (Malmasi et al., 2013) have competed in the DSL 2015 (Malmasi and Dras, 2015) and the ADI 2016 (Malmasi and Zampieri, 2016), achieving the best performance in both shared tasks.

This section evidenced an important overlap between NLI methods and dialect identification methods both in terms of participation overlap in the shared tasks and in terms of successful approaches. With the exception of Bobicev (2013), most teams that were ranked among the top ten entries in the NLI shared task were also successful at the VarDial workshop shared tasks.

Detailed information about all approaches and performance obtained in these competitions can be found in the VarDial shared task reports (Zampieri et al., 2014, 2015b; Malmasi et al., 2016b; Zampieri et al., 2017) and in the evaluation paper by Goutte et al. (2016).

3 Methods

In the next sections we describe the data provided by the shared task organizers and the ensemble SVM approach applied by the ZCD team.

3.1 Data

The organizers of the NLI Shared Task 2017 provided participants with data corresponding to eleven native languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish. The training dataset consists of 11,000 essays, orthographic transcriptions of 45-second English spoken responses, and iVectors (1,000 instances for each of the eleven native languages), while the development dataset was stratified similarly, containing 100 instances for each native language.

There were individual tracks in which only the essays or only the responses could be used and a fusion track in which both the essays and the speech transcriptions (including iVectors) could be used. The test dataset, containing 1,100 instances with essays, speech transcriptions and iVectors, was released at a later date.

The use of a dataset containing text and speech is the main new aspect of the 2017 NLI task so we decide to compete in the fusion track taking both modalities into account. The approach used in our submission is described next.

3.2 Approach

We built a classification system based on SVM ensembles, following the methodology proposed by Malmasi and Dras (2015).

The idea behind classification ensembles is to improve the overall performance by combining the results of multiple classifiers. Such systems have proved successful not only in NLI and dialect identification, as evidenced in the previous sections, but also in numerous text classification tasks, among which are complex word identification (Malmasi et al., 2016a) and grammatical error diagnosis (Xiang et al., 2015). The classifiers can differ in a wide range of aspects, such as algorithms, training data, features or parameters.

In our system, the classifiers used different features. We experimented with the following features: character n-grams (with n in $\{1, ..., 10\}$) from essays and speech transcripts, word n-grams (with n in $\{1, 2\}$) from essays and speech transcripts, and iVectors. For the n-gram features we

System	F1 (macro)	Accuracy
Essays + Transcriptions + iVectors	0.8358	0.8355
Essays + Transcriptions	0.8191	0.8191
Official Baseline (with iVectors)	0.7901	0.7909
Official Baseline (without iVectors)	0.7786	0.7791
Random Baseline	0.0909	0.0909

Table 2: ZCD results and baselines for the fusion track.

used TF-IDF weighting applied on the tokenized version of the essays and speech transcripts (provided by the organizers). As a pre-processing step, we lowercased all words.

We first trained a classifier for each type of feature using the essays as input data, and performed cross-validation to determine the optimal value for the SVM hyperparameter C, searching in $\{10^{-5}, ..., 10^{5}\}$. Further, for the n-gram features we kept only those classifiers whose individual cross-validation performance was higher than 0.8. Thus, our first ensemble consisted of individual classifiers using character n-grams (with n in $\{6, 7, 8\}$) from essays and speech transcripts.

For the second ensemble, we introduced an additional classifier using the iVectors as features. To combine the classifiers, we employed a majority-based fusion method: the class label predicted by the ensemble is the one that was predicted by the majority of the classifiers. We used the SVM implementation provided by Scikit-learn (Pedregosa et al., 2011), based on the Liblinear library (Fan et al., 2008).

On the development dataset, the first ensemble (essays + speech transcripts) obtained 0.83 accuracy, and the second ensemble (essays + speech transcripts + iVectors) obtained 0.84 accuracy.

4 Results

We submitted two runs of our system. The first run included the essays and the transcriptions of responses, whereas the second run included also the iVectors. We present the results obtained by the two runs along with a random baseline and the performance of the unigram-based official baseline system in terms of F1 score and accuracy in Table 2.

The best results were achieved by the second run, reaching 83.55% accuracy and 83.58% F1 score. As can be seen in Table 2, the iVectors bring a performance improvement of about 1.6 percentage points in terms of accuracy and F1 score.

Ten teams participated in the fusion track and our best run was ranked 3rd by the shared task organizers. Ranks were calculated using McNemars test for statistical significance, a common practice in many NLI shared tasks (e.g. DSL 2016 (Malmasi et al., 2016b), and the shared tasks at WMT (Bojar et al., 2016)).

The confusion matrix of our best submission is presented in Table 3. We observed that the best performance was obtained for Japanese and the worst performance was obtained for Arabic. Not surprisingly, most confusion occurred between Hindi and Telugu. Our initial analysis indicates that this confusion occurred because of geographic proximity and not by intrinsic linguistic properties shared by these two languages, as Hindi and Telugu do not belong to the same language family - Hindi is a Hindustani language and Telugu is a Dravidian language.

5 Most Informative Features

As briefly discussed in the introduction of this paper, NLI methods can provide interesting information about patterns in non-native language that can be used to study second language acquisition and L1 interference or language transfer. For this purpose, in Table 4 we present the top ten most informative character 8-grams for each of the eleven languages in the dataset according to our classifier.

It is not surprising that named entities are very informative for our system and highly discriminative for most native languages. For example, essays and responses from China often contain place names like *China, Taipei, Taiwan*, and *Beijing*, whereas those from Turkey contain *Istanbul* and, of course, *Turkey*. These features are very frequent in essays and responses by Chinese and Turkish speakers due to topical bias and not because of any intrinsic linguistic property of Chinese or Turkish. However, in other languages, interesting linguistic patterns can be identified by looking at these features.

	CHI	JPN	KOR	HIN	TEL	FRE	ITA	SPA	GER	ARA	TUR
CHI	91	3	2	0	0	0	2	0	0	1	1
JPN	2	93	2	0	1	1	0	0	1	0	0
KOR	4	14	77	0	0	1	1	1	0	1	1
HIN	1	0	1	80	18	0	0	0	0	0	0
TEL	0	0	1	18	78	0	0	1	0	2	0
FRE	2	0	0	2	1	87	5	0	2	1	0
ITA	0	0	0	1	0	6	85	3	3	2	0
SPA	1	1	2	2	1	4	7	77	2	2	1
GER	0	1	0	3	0	3	2	1	90	0	0
ARA	2	2	2	3	2	7	1	2	1	77	1
TUR	1	2	0	3	0	2	3	1	1	3	84

Table 3: Confusion matrix on the test set.

Language	**Most Informative Features**
Arabic	|alot of | alot of|y thing | statmen|statment|e alot o|tatment |ery thin|very thi|every th|
Chinese	| i think| taiwan |i think |beijing | beijing| taipei | in chin|in china|n china | chinese|
French	| indeed |. indeed| . indee|indeed ,|ndeed , | france |developp| french |to concl|o conclu|
German	| , that | and um | germany|germany | berlin |, that t|have to | have to| um yeah|um yeah |
Hindi	| towards|towards |as compa| as comp|various | various|s compar| enjoyme| mumbai | behind |
Italian	|hink tha|nk that |ink that| in ital|n italy |in fact |in italy| in fact|i think | italian|
Japanese	|in japan|n japan | in japa|apanese | japanes|japanese| japan ,|japan , |i disagr| japan .|
Korean	| korean |in korea| in kore|n korea | however|however |korea , | korea ,| . howev|. howeve|
Spanish	| mexico |oing to |going to| going t| diferen|le that | the cit|es that |diferent|ple that|
Telugu	|ing the |hyderaba| hyderab|yderabad|derabad | subject| we can | i concl|i conclu|where as|
Turkish	| turkey |istanbul|stanbul | istanbu| uh and |n turkey| in turk|in turke|s about | . becau|

Table 4: Top ten most informative character 8-grams for each language.

In the most informative features for French, for example, we find *developp* from the French *développé* which leads to a misspelling of the English word *developed*. In Arabic we observed a number of features that indicate misspellings. The Arabic alphabet is very different from the Latin one, making spelling English words particularly challenging for native speakers of Arabic. The top ten most informative features for Arabic include word boundary errors such as *every thing* for *everything*, and *alot* for *a lot*, as well as the omission of vowels such as *statment* for *statement*.

6 Conclusion

To the best of our knowledge, the NLI Shared Task 2017 fusion track was the first shared task to provide both written and spoken data for NLI. It was an interesting opportunity to evaluate the performance of NLI methods beyond written texts.

In this paper we highlighted the overlap between NLI and dialect, language variety, and similar language identification and used an approach that achieved high results in both tasks. We applied an SVM ensemble approach trained character n-grams achieving competitive results of 83.55% accuracy ranking 3rd in the fusion track.

Even though the results obtained by our approach were not low, we believe that there is still room for improvement. In previous shared tasks (e.g. NLI 2013, DSL 2015, and ADI 2016) we observed that SVM ensembles ranked higher in the results tables than our method did in the NLI 2017. We are investigating whether the combination of features or the implementation itself can be optimized for better performance.

Acknowledgements

We would like to thank the NLI Shared Task 2017 organizers for making the dataset available and for replying promptly to all our inquiries. We further thank the anonymous reviewers for their valuable feedback.

Liviu P. Dinu is supported by UEFISCDI, project number 53BG/2016.

References

Yves Bestgen. 2017. Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain, pages 115–123.

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

Victoria Bobicev. 2013. Native language identification with ppm. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 180–187.

Victoria Bobicev. 2015. Discriminating between similar languages using ppm. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*. Hissar, Bulgaria, pages 59–65.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT*.

Julian Brooke and Graeme Hirst. 2012. Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of Language Resources and Evaluation (LREC)*. pages 779–784.

Serhiy Bykh and Detmar Meurers. 2014. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, pages 1962–1973.

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar Meurers. 2013. Combining shallow and linguistically motivated features in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 197–206.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9:1871–1874.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 216–223.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2013. Feature space selection and combination for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 96–100.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The nrc system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*. Dublin, Ireland, pages 139–145.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of Language Resources and Evaluation (LREC)*.

John Henderson, Guido Zarrella, Craig Pfeifer, and John D. Burger. 2013. Discriminating non-native english with 350 words. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 101–110.

Radu Tudor Ionescu. 2015. A Fast Algorithm for Local Rank Distance: Application to Arabic Native Language Identification. In *Proceedings of the International Conference on Neural Information Processing*. Springer, pages 390–400.

Radu Tudor Ionescu and Andrei Butnaru. 2017. Learning to identify arabic and german dialects using multiple kernels. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain, pages 200–209.

Radu Tudor Ionescu and Marius Popescu. 2016. UnibucKernel: An Approach for Arabic Dialect Identification Based on Multiple String Kernels. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. Osaka, Japan, pages 135–144.

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*. Atlanta, Georgia, pages 111–118.

André Lynum. 2013. Native language identification using large scale lexical features. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 266–269.

Shervin Malmasi. 2016. *Native Language Identification: Explorations and Applications*. Ph.D. thesis.

Shervin Malmasi and Mark Dras. 2014. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop*.

Shervin Malmasi and Mark Dras. 2015. Language identification using classifier ensembles. In *Proceedings of the VarDial Workshop*.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016a. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval*.

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*. Copenhagen, Denmark.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. Nli shared task 2013: Mq submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 124–133.

Shervin Malmasi and Marcos Zampieri. 2016. Arabic Dialect Identification in Speech Transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. Osaka, Japan, pages 106–113.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016b. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*. Osaka, Japan.

Tomoya Mizumoto, Yuta Hayashibe, Keisuke Sakaguchi, Mamoru Komachi, and Yuji Matsumoto. 2013. Naist at the nli 2013 shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 134–139.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Marius Popescu and Radu Tudor Ionescu. 2013. The story of the characters, the dna and the native language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 270–278.

Björn 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 *Proceedings of Interspeech*. pages 2001–2005.

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*. Atlanta, GA, USA.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the l1 of non-native writers: the cmu-haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 279–287.

Lan Wang, Masahiro Tanaka, and Hayato Yamana. 2016. What is your Mother Tongue?: Improving Chinese Native Language Identification by Cleaning Noisy Data and Adopting BM25. In *Proceedings of the International Conference on Big Data Analysis (ICBDA)*. IEEE, pages 1–6.

Ching-Yi Wu, Po-Hsiang Lai, Yang Liu, and Vincent Ng. 2013. Simple yet powerful native language identification on toefl11. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, pages 152–156.

Yang Xiang, Xiaolong Wang, Wenying Han, and Qinghua Hong. 2015. Chinese Grammatical Error Diagnosis Using Ensemble Learning. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*. pages 99–104.

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015a. Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*. Hissar, Bulgaria, pages 66–72.

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.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*. Dublin, Ireland, pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015b. Overview of the dsl shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*. Hissar, Bulgaria, pages 1–9.