Identifying the L1 of non-native writers: the CMU-Haifa system

Chris Dyer[*] Manaal Faruqui[*] Noam Ordan[†] Nathan Schneider[*]
Yulia Tsvetkov[*] Naama Twitto[†] Shuly Wintner[†]

[*]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
`cdyer@cs.cmu.edu`

[†]Department of Computer Science
University of Haifa
Haifa, Israel
`shuly@cs.haifa.ac.il`

Abstract

Given a dataset of English essays composed by non-native speakers, as part of the TOEFL exam, we identify with high accuracy the native language of the authors. We use standard text classification techniques, but define sophisticated classifiers that are sensitive to the specific patterns observed in the English of authors whose first language is structurally different. We describe the various features used for classification, as well as the settings of the classifier that yielded the highest accuracy.

1 Introduction

The task we address in this work is identifying the native language (*L1*) of non-native English authors. More specifically, given a dataset (Blanchard et al., 2013) of short English essays, composed as part of the TOEFL exam (of English as a foreign language) by authors whose native language is one out of 11 possible languages (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish), our task is to identify that language.

This task has a clear empirical motivation. Non-native speakers make different errors when they write English, depending on their native language (Swan and Smith, 2001); understanding the different types of errors is a prerequisite for correcting them (Leacock et al., 2010), and systems such as the one we describe here can shed interesting light on such errors. Tutoring applications can use our system to identify the native language of students and offer better-targeted advice. Forensic linguistic applications are sometimes required to determine the L1 of authors (Estival et al., 2007a,b). Additionally, we believe that the task is interesting in and of itself, providing a better understanding of non-native language. We are thus equally interested in defining *meaningful* features whose contribution to the task can be linguistically interpreted.

We address the task as a multiway text-classification task; we specify our methodology in §3. As in other author attribution tasks (Juola, 2006), the choice of features for the classifier is crucial; we discuss the features we define in §**??**. We report our results in §7 and conclude with suggestions for future research.

2 Related work

The task of L1 identification was introduced by Koppel et al. (2005a,b), who work on the International Corpus of Learner English (Granger et al., 2009), which includes texts written by students from Russia, the Czech Republic, Bulgaria, France, and Spain. The texts lengths range from 500 to 850 words. The classification method is a linear SVM, and features include 400 standard function words, 200 letter *n*-grams, 185 error types and 250 rare part-of-speech (POS) bi-grams. Ten-fold cross-validation results on this dataset are 80% accuracy.

The same experimental setup is assumed by Tsur and Rappoport (2007), who are mostly interested in testing the hypothesis that an author's choice of words in a second language is influenced by the phonology of his or her L1. They confirm this hypothesis by carefully analyzing the features used by Koppel et al., controlling for potential biases.

Wong and Dras (2009, 2011) are also motivated by a linguistic hypothesis, namely that syntactic errors in a text are influenced by the author's L1. Wong and Dras (2009) analyze three error types statistically, and then add them as features in the same experimental setup as above (using LIBSVM with a radial kernel for classification). The error types are subject-verb disagreement, noun-number disagreement and misuse of determiners. Addition of these features does not improve on the results of Koppel et al.. Wong and Dras (2011) further extend this work by adding as features horizontal slices of parse trees, thereby capturing more syntactic structure. This improves the results significantly, yielding 78% accuracy compared with less than 65% using only lexical features.

Kochmar (2011) uses a different corpus, the Cambridge Learner Corpus, in which texts are 200-400 word long, and are authored by native speakers of five Germanic languages (German, Swiss German, Dutch, Swedish and Danish) and five Romance languages (French, Italian, Catalan, Spanish and Portuguese). Again, SVM is the classification device. Features include POS *n*-grams, character *n*-grams, phrase-structure rules (extracted from parse trees), and two measures of error rate. The classifier is evaluated on its ability to distinguish between pairs of closely-related L1s, and the results are usually excellent.

A completely different approach is offered by Brooke and Hirst (2011). Since training corpora for this task are rare, they use mainly L1 (blog) corpora. Given English word bi-grams $\langle e_1, e_2 \rangle$, they try to assess, for each L1, how likely it is that an L1 bi-gram was translated literally by the author, resulting in $\langle e_1, e_2 \rangle$. Working with four L1s (French, Spanish, Chinese, and Japanese), and evaluating on the International Corpus of Learner English, the results are lower than 50%.

Our dataset in this work is different, and consists of TOEFL essays written by speakers of eleven different L1s (Blanchard et al., 2013), distributed as part of the First Native Language Identification Shared Task (Tetreault et al., 2013). We use a plethora of features; some of them are inspired by previous work outlined above, but many are motivated by other author attribution tasks, in particular identification of *translationese*, the language of translated texts (Volansky et al., Forthcoming).

3 Methodology

Characteristics of the dataset. Development, train, test sets.

For classification we use *creg*...

Pre-processing: POS-tagging, etc.

4 Model Overview

[$^{\text{NS}}_{\text{S}}$ I think we should break this up into sections by feature group (below). Here we can talk about learning and regularization and give a high-level overview of features.]

We define a large arsenal of features, our motivation being both to improve the accuracy of classification and to be able to interpret the characteristics of the language produced by speakers of different L1s. In this section we define the features and motivate their use.[1] While some of the features were used in the works surveyed in §2, many are novel, and are inspired by the features used to identify translationese by Volansky et al. (Forthcoming). We also report the accuracy of using each feature type, in isolation, on the training set.

Character *n*-grams The number of character 1-, 2-, and 3-grams. 69.94%.

Frequent character *n*-grams Only those character *n*-grams that are observed more than *m* times in the corpus are considered. ???

POS *n*-grams The number of POS 1-, 2-, and 3-grams. 53.92%.

Document length The number of tokens in the text. 11.81%.

Pronouns The number of each pronoun. 22.81%.

Punctuation The number of each punctuation mark. 27.41%.

Passives The ratio of verbs to passive verbs. 12.26%.

Positional token frequency The choice of the first and last few words in a sentence is highly constrained, and may be significantly influenced by the authors L1. We use the counts (???) of the first and last three words in each sentence as features. 53.03%.

Cohesive markers These are function words (and short phrases) that have a strong discourse func-

[1]Whenever counts are mentioned, we use the log of the count as the feature.

tion in texts, contributing to its cohesiveness. 25.71%.

Cohesive verbs ???. 22.85%.

Function words The number of occurrences of each word from a pre-defined list of 100 most frequent words in English (excluding punctuation). 42.47%.

Contextual function words, bigrams Pairs consisting of a function word from the list mentioned above, along with the POS tag of its adjacent word. This feature captures patterns such as verbs and the preposition or particle immediately to their right, or nouns and the determiner that precedes them. 62.79%

Contextual function words, trigrams Same as above, but counting 3-grams consisting of two function words and the POS tag of the third word in the 3-gram. 62.32%.

Lemmas The number of each of the most frequent lemmas in the text. ???. 58.95%.

Prompt Conjunction of the character n-gram features defined above with the prompt; since the prompt contributes information on the domain, it is likely that some words (and, hence, character sequences) will occur more frequently with some prompts than with others. 65.09%.

Misspelling features ???. 37.29%.

Brown ???

Restored ???

5 Main Features

These four feature groups form the core of our model.

5.1 POS: part-of-speech sequences

5.2 FreqChar: frequent character n-grams

5.3 CharPrompt: character n-grams paired with the prompt ID

5.4 Brown: Brown clusters

6 Additional Features

Each of these adds a small number of parameters to the model. We report the empirical improvement that each of these brings independently when added to the main features (§5). The full model combines all features.

6.1 CxtFxn: Contextual function words[$^{NS}_{S}$ does this subsume pronouns?]

6.2 DocLen: Document length in tokens

6.3 Misspell: Spelling correction edits

6.4 Position: [$^{NS}_{S}$?]

6.5 Pron: pronouns[$^{NS}_{S}$?]

6.6 PsvRatio: Ratio of passive to active voice verbs[$^{NS}_{S}$ or is it the proportion of passive verbs?]

6.7 Punct: Count of each punctuation mark

6.8 Restore: LM-restored function words

6.9 Discarded Features

[$^{NS}_{S}$ things that didn't help in our preliminary experiments]

7 Results

8 Conclusion

Acknowledgments

This research was supported by a grant from the Israeli Ministry of Science and Technology.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. TOEFL11: A corpus of non-native English. Technical report, Educational Testing Service, 2013.

Julian Brooke and Graeme Hirst. Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium, 2011. Presses universitaires de Louvain.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. TAT: An author profiling tool with application to Arabic emails. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 21–30, Melbourne, Australia, December 2007a. URL `http://www.aclweb.org/anthology/U07-1006`.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, Melbourne, Australia, 2007b.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. *International Corpus of Learner English*. Presses universitaires de Louvain, Louvain-la-Neuve, 2009. ISBN 978-2-87463-143-6.

Patrick Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3): 233–334, 2006. URL `http:/dx.doi.org/10.1561/1500000005`.

Ekaterina Kochmar. Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge, 2011.

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

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 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, Chicago, IL, 2005b. ACM.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool, 2010. URL `http://www.morganclaypool.com/doi/abs/10.2200/S00275ED1V01Y201006HLT009`.

Michael Swan and Bernard Smith. *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge Handbooks for Language Teachers. Cambridge University Press, 2001. ISBN 9780521779395.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 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*, Atlanta, GA, USA, June 2013. Association for Computational Linguistics.

Oren Tsur and Ari Rappoport. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0602`.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Literary and Linguistic Computing*, Forthcoming.

Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December 2009. URL `http://www.aclweb.org/anthology/U09-1008`.

Sze-Meng Jojo Wong and Mark Dras. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1148`.