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 of short English essays (Blanchard et al., 2013), composed as part of the *Test of English as a Foreign Language (TOEFL)* 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 §4. We report our results in §5 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*... [$^{\text{NS}}_{\text{S}}$ Here we can talk about learning and regularization and give a high-level overview of features.]

All essays were tagged with the Stanford part-of-speech tagger (Toutanova et al., 2003). We did not parse the dataset.

4 Model Overview

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.

4.1 Motivation

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 begin by motivating our choice of features.

POS *n*-grams Part-of-speech n-grams were used in various text-classification tasks.

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. We therefore use the prompt ID in conjunction with other features.

Document length The number of tokens in the text is highly correlated with the author's level of fluency, which in turn is correlated with the author's L1.

Pronouns The use of pronoun varies greatly among different authors. We use the same list of 25 English pronouns that Volansky et al. (Forthcoming) use for identifying translationese.

Punctuation Similarly, different languages use punctuation differently, and we expect this to taint the use of punctuation in non-native texts.

Passives English uses passive voice more frequently than other languages. Again, the use of passives in L2 can be correlated with the author's L1.

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 author's L1.

Cohesive markers These are 40 function words (and short phrases) that have a strong discourse function in texts ('however', 'because', 'in fact', etc.) Translators tend to spell out implicit utterances and render them explicitly in the target text (Blum-Kulka, 1986). We use the list of Volansky et al. (Forthcoming).

Cohesive verbs This is a list of manually compiled verbs that are used, like cohesive markers, to spell out implicit utterances ('indicate', 'imply', 'contain', etc.)

Function words Frequent tokens, which are mostly function words, have been used successfully for various text classification tasks. Koppel and Ordan (2011) define a list of 400 such words, of which we only use 100 (using the entire list was not significantly different).

Contextual function words To further capitalize on the ability of function words to discriminate, we define 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. We also define 3-grams consisting of one or two function words and the POS tag of the third word in the 3-gram.

Lemmas The content of the text is not considered a good indication of the author's L1, but many text categorization tasks use lemmas (more precisely, the stems produced by the tagger) as features approximating the content.

Misspelling features Clearly, the spelling errors that learners make in English depend on the phonological properties of their L1. [$^{\text{S}}_{\text{W}}$???]

Restored tags We focus on three important token classes defined above: punctuation marks, function words and cohesive verbs. We first remove words in these classes from the texts, and then recover the most likely hidden tokens [$^{\text{Y}}_{\text{T}}$ To Shuly: I removed 'POS'. We recover words, not POS tags, I used the word 'tags' to distinguish between words to recover and all other words. Probably the word 'tag' is misleading, I changed it to 'token'] in a sequence of words, according to an n-gram language model trained on all essays in the training corpus corrected with a spell checker and containing both words and hidden tokens. This feature should capture specific words or punctuation marks that are consistently omitted (deletions), or misused (insertions, substitutions). To restore hidden tokens we use the `hidden-ngram` utility provided in SRILM (Stolcke, 2002).

Brown clusters [$^{\text{S}}_{\text{W}}$???]

4.2 Main Features

First, we use the following four feature types as the core of our model. Whenever counts are mentioned, we use the log of the count as the feature. We report the accuracy of using each feature type, in isolation, on the training set.

POS Part-of-speech n-grams. Features were extracted to count every POS 1-, 2-, and 3-gram in each document. 53.92%. [$^{\text{S}}_{\text{W}}$ But the table says 55.18]

FreqChar Frequent character n-grams. We experimented with character n-grams: The number of character 1-, 2-, and 3-grams. This yielded 69.94% accuracy. We then refined the feature to include only those character n-grams that are observed more than m times in the corpus are considered. [$^{\text{S}}_{\text{W}}$ n ranges from 1 to 4, and m is set to ???. 74.12%]

CharPrompt Conjunction of the character n-gram features defined above with the prompt ID. 65.09%.

Brown Brown clusters. [$^{\text{S}}_{\text{W}}$???]

The accuracy of the classifier on the development set using these four feature types is reported in Table 1.

Feature Group	# Params	Accuracy (%)	ℓ_2
POS	540,947	55.18	1.0
+ FreqChar	1,036,871	79.55	1.0
+ CharPrompt	2,111,175	79.82	1.0
+ Brown	5,664,461	81.09	1.0

Table 1: Dev set accuracy with MAIN feature groups, added cumulatively. The number of parameters is always a multiple of 11 (the number of classes). Only ℓ_2 regularization was used for these experiments; the penalty was tuned on the dev set as well.

4.3 Additional Features

To the basic set of features we now add more specific, linguistically-motivated features, each adding a small number of parameters to the model. As above, we indicate the accuracy of each feature type in isolation.

DocLen Document length in tokens. 11.81%.

Punct Counts of each punctuation mark. 27.41%.

Pron Counts of each pronoun. 22.81%.

Position Positional token frequency. We use the counts for the first two and last three words before the period in each sentence as features. 53.03%.

PsvRatio The proportion of passive verbs out of all verbs. 12.26%.

CxtFxn Contextual function words. Bi-grams yield 62.79%, tri-gram 62.32%.

Misspell Spelling correction edits. [$^{S}_{W}$???]. 37.29%.

Restore Counts of substitutions, deletions and insertions of predefined tokens that we restored in the texts. 47.67%

Table 2 reports the empirical improvement that each of these brings independently when added to the main features (§4.2).

Feature Group	# Params	Accuracy (%)	ℓ_2
MAIN + Position	6,153,015	81.00	1.0
MAIN + PsvRatio	5,664,472	81.00	1.0
MAIN	5,664,461	81.09	1.0
MAIN + DocLen	5,664,472	81.09	1.0
MAIN + Pron	5,664,736	81.09	1.0
MAIN + Punct	5,664,604	81.09	1.0
MAIN + Misspell	5,799,860	81.27	5.0
MAIN + Restore	5,682,589	81.36	5.0
MAIN + CxtFxn	7,669,684	81.73	1.0

Table 2: Dev set accuracy with MAIN features plus additional feature groups, added independently. ℓ_2 regularization was tuned as in Table 1 (two values, 1.0 and 5.0, were tried for each configuration; more careful tuning might produce slightly better accuracy). Results are sorted by accuracy; only three groups exhibited independent improvements over the MAIN feature set.

4.4 Discarded Features

We also tried several other feature types that did not improve the accuracy of the classifier on the development set.

Cohesive markers Counts of each cohesive marker. 25.71%.

Cohesive verbs Counts of each cohesive verb. 22.85%.

Function words Counts of function words. 42.47%. This feature is subsumed by the highly discriminative CxtFxn feature.

5 Results

The full model that we used to classify the test set combines all features listed in Table 2. Using all these features, the accuracy on the development set is [$^{S}_{W}$???], and on the test set it is 81.5%. Table 3 lists the confusion matrix on the test set, as well as precision, recall and F_1-score for each L1.

[$^{S}_{W}$ Analysis? Error analysis? Observations?]

6 Conclusion

Acknowledgments

This research was supported by a grant from the Israeli Ministry of Science and Technology.[$^{NS}_{S}$ anything from the CMU side?]

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.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies*, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

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`.

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR	Precision (%)	Recall (%)	F_1 (%)
ARA	80	0	2	1	3	4	1	0	4	2	3	80.8	80.0	80.4
CHI	3	80	0	1	1	0	6	7	1	0	1	88.9	80.0	84.2
FRE	2	2	81	5	1	2	1	0	3	0	3	86.2	81.0	83.5
GER	1	1	1	93	0	0	0	1	1	0	2	87.7	93.0	90.3
HIN	2	0	0	1	77	1	0	1	5	9	4	74.8	77.0	75.9
ITA	2	0	3	1	1	87	1	0	3	0	2	82.1	87.0	84.5
JPN	2	1	1	2	0	1	87	5	0	0	1	78.4	87.0	82.5
KOR	1	5	2	0	1	0	9	81	1	0	0	80.2	81.0	80.6
SPA	2	0	2	0	1	8	2	1	78	1	5	77.2	78.0	77.6
TEL	0	1	0	0	18	1	2	1	1	73	3	85.9	73.0	78.9
TUR	4	0	2	2	0	2	2	4	4	0	80	76.9	80.0	78.4

Table 3: Official test set confusion matrix with the full model. [$^{\text{NS}}_{\text{S}}$ which direction is predicted vs. gold?] Accuracy is 81.5%.

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 and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1132`.

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`.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Procedings of International Conference on Spoken Language Processing*, pages 901–904, 2002. URL `citeseer.ist.psu.edu/stolcke02srilm.html`.

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.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*

Technology (HLT-NAACL 2003), pages 173–180, Edmonton, Canada, June 2003. 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`.