Improving Native Language Identification by Using Spelling Errors

Lingzhen Chen
DISI
University of Trento
Trento, Italy
lzchen.cs@gmail.com

Carlo Strapparava
HLT
Fondazione Bruno Kessler
Trento, Italy
strappa@fbk.eu

Vivi Nastase
Inst. for Computational Linguistics
University of Heidelberg
Heidelberg, Germany
nastase@cl.uni-heidelberg.de

Abstract

In this paper, we explore spelling errors as a source of information for detecting the native language of a writer, a previously under-explored area. We note that character n-grams from misspelled words are very indicative of the native language of the author. In combination with other lexical features, spelling error features lead to 1.2% improvement in accuracy on classifying texts in the TOEFL11 corpus by the author's native language, compared to systems participating in the NLI shared task[1].

1 Introduction

Native Language Identification (NLI) aims to determine the native language (L1) of a writer based on his or her writings in a second language (L2). Though initially motivated by the study of cross-linguistics influence, the value of NLI is not limited to education. Potentially, it is also very valuable in academic, marketing, security and law enforcement fields. Identifying the native language is based on the assumption that the L1 of an individual impacts his or her writing in L2 due to the language transfer effect.

We focus here on the influences from L1 that surface as spelling errors in L2. Crossley and McNamara (2011) showed that syntactic patterns and lexical preferences from L1 appear in L2 systematically, and are very informative for identifying the writer's native language. Texts written by authors with the same L1 also exhibit similarities with respect to the errors within.

In terms of spelling, in particular, both the sound of the words in different languages and the mapping from sounds to letters in L2 vs. L1, as well as the particular conventions of writing can have a visible impact. In Italian, for example, each vowel has only one pronunciation, so it is very common for Italian writers to confuse the use of vowels in English: the English e can correspond to the sounds written either as i or e in Italian. In Arabic, on the other hand, vowels are rarely written, and this could cause writers to miss vowels when writing in English. In Chinese, since it uses a completely different writing system compared to English, there might be a higher probability for authors to make spelling errors when it comes to complicated words, because the mapping from English sounds to letters of the Roman alphabet is not one-to-one. We test whether we are able to capture some of these phenomena by going below the word level to character level and testing their usefulness as features for identifying the native language of the author.

Spelling errors have been used as features for NLI since Koppel et al. (2005). They considered syntax errors and eight types of spelling errors such as repeated letters, missing letters, and inversion of letters. The relative frequency of each error type with regard to the length of the document is used as feature values. By combining these with common features such as function words, they obtained a classification accuracy of 80.2% on a subcorpora of ICLE*v1* that consists of five languages. More recently, Nicolai et al. (2013) focused on the misspelled part of a word rather than the type of spelling errors. They used pairs of correct and misspelled parts in a word as features. Lavergne et al. (2013) adopted a similar approach to represent the spelling errors by the inner-most misspelled substring compared to the correct word. Combined with other features, they obtained a test accuracy of 75.29% on the TOEFL11 dataset.

Character n-grams have been explored, but not particularly for representing spelling errors. Brooke and Hirst (2012) showed that using char-

[1] https://sites.google.com/site/nlisharedtask2013/home

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 542–546
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2086

acter unigrams, bigrams, and trigrams, the test accuracy can reach 37.4% for 7-class NLI on a subset of ICLE*v2* corpus. For the NLI 2013 shared task, Lahiri and Mihalcea (2013) used as features character trigrams represented by their raw frequencies. It leads to a test accuracy of 57.77%, which shows that how often character combinations are used is indicative of the L1 of an author.

Using complete words to represent spelling errors would not capture regularities that go beyond a single misspelled instance – like the preference of using i instead of e by Italian writers. We investigate the representation of spelling errors through character n-grams with size up to 3. We assess the effectiveness of using such feature representation for NLI and its contribution when combined with word and lemma n-grams, whose effectiveness has already been established (Gyawali et al., 2013; Jarvis et al., 2013). We report high classification results when using only spelling errors, and an improvement of 1.2 percentage points in accuracy, compared to the best results obtained in NLI shared task, when using spelling errors in combination with word and lemma features.

2 Data

The experiments are performed on the TOEFL11 corpus (Blanchard et al., 2013) of English essays written by non-native English learners as part of the *Test of English as a Foreign Language* (TOEFL). We also use the ICLE*v2* corpus (Granger et al., 2009) for extracting additional spelling errors. The TOEFL11 corpus is not the most recent corpus for the NLI task, but is by far one of the largest learner corpora that is balanced in terms of both topics and L1 languages.

The TOEFL11 corpus contains 4 million tokens in 12,100 essays written by authors whose native language (L1) is one of: Arabic (ARA), Chinese (ZHO), French (FRA), German (DEU), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL) or Turkish (TUR). For each target L1, the number of essays is equal (900 in the train set, 100 in the development set and 100 in the test set). The distribution of the number of essays per topic is not perfectly balanced across different L1s, but they are rather close: the average number of essays per topic is 1,513 and the standard deviation is 229.

3 Methods

We aim to analyze the impact of spelling errors on identifying the native language of the author. We will analyze them separately and in combination with commonly used features.

3.1 Features

Word n-grams Previous work on NLI avoided lexical features due to the topic bias in the dataset. This is not an issue with the TOEFL11 corpus, since the topic distribution over the different L1s is quite balanced, as described in the previous section. Hence we can use as features the word unigrams, bigrams and trigrams. We filter out those that do not appear in at least two texts in the corpus. In the extraction of word n-grams, we ignore the use of all punctuations. Word n-gram features have as value their weighted frequencies using the Log-Entropy weighting technique. (Dumais, 1991)

Lemma n-grams This feature represents the lexical choice of a writer, regardless of the inflected form of a word. We use the n-grams of lemma produced by the TreeTagger tool (Schmid, 1994) with size up to 3. The feature extraction rule and the feature weighting scheme is the same as for word n-grams.

Character n-grams The frequency of character or character sequence usage captures information about the preference of an author in using certain sounds, combinations of sounds, prefixes or suffixes. We represent texts using character n-grams and assign them the value v_i – their relative frequency with respect to the set of n-grams (unigram, bigram, trigram) that they belong to:

$$v_i = \frac{count(i)}{\sum_{j \in S_i} count(j)}$$

where S_i is the collection of n-grams that have the same gram size as n-gram i.

Spelling errors The errors made by a writer while writing in their L2 may result from sound-to-character mappings in L1, writing preferences or other biases. These could be strong indicators for an author's L1. We extract spelling errors using the *spell* shell command. There are 34,233 unique spelling errors in the TOEFL11 corpus. We looked for additional sources of spelling errors, and extracted a list of 12,488 unique misspelled words from the ICLE*v2* corpus that are produced by writers with the common L1 as in the TOEFL11 corpus (namely Spanish, French,

Italian, German, Turkish, Chinese and Japanese). They are referred to here as *Spelling_error_ICLE* and *Spelling_error_TOEFL* in the later part of this paper. Spelling errors are binary features.

Spelling errors as character n-grams Every misspelled word in a text will be represented as character n-grams, where $n = 1..3$. Special characters marking the start and end of a word will be part of the n-grams. The value of these features is their relative frequencies, as for character n-grams.

3.2 Classifiers

Following the proven effectiveness of Support Vector Machine (SVM) by numerous experiments on text classification tasks, we adopt the use of linear SVM for NLI. In particular, we use linear SVM implemented by `scikit-learn` package (Pedregosa et al., 2011) to perform the multi-class classification.

3.3 Experiment Setup

The data is pre-processed by lower casing the tokenized version of the corpus. Each text is represented through the sets of features described above. The feature size of word n-grams up to size 3 are over 500,000 and that of lemma n-grams up to size 3 are over 400,000. The combination of the two is over 600,000. The hyper-parameter C of the linear SVM is set to 100, an optimal setting obtained by cross-validation on the train set. The performance is evaluated by classification accuracy, as was done in the NLI shared task. We test the performance of the used feature sets through a 10-fold cross-validation on the train+development set before the final run on the test set.

4 Results

The classification accuracies obtained by using different features and feature combinations are presented in Table 1. The feature sets include the word, lemma, character n-grams up to size 3 (denoted as *word_ngrams*, *lemma_ngrams*, *char_ngrams* respectively), the misspelled words in *Spelling_error_ICLE* or in *Spelling_error_TOEFL* (denoted as *word_error_icle* or *word_error_toefl*), and the character n-grams up to size 3 extracted from *Spelling_error_ICLE* or *Spelling_error_TOEFL* (denoted as *char_error_icle* or *char_error_toefl*).

In terms of classifying by a single type of feature, word n-grams are the most indicative one,

Type of Feature	10 Fold	Test
(1) *word_ngrams*	83.63 (±1.38)	84.16
(2) *lemma_ngrams*	83.18 (±1.46)	84.00
(3) *word_error_toefl*	35.05 (±1.42)	32.55
(4) *word_error_icle*	24.42 (±1.12)	26.45
(5) *char_ngrams*	66.27 (±0.93)	67.27
(6) *char_error_icle*	65.03 (±1.14)	65.73
(7) *char_error_toefl*	65.27 (±1.21)	66.45
(1) + (2)	83.91 (±1.50)	84.32
(1) + (2) + (3)	83.82 (±1.53)	84.27
(1) + (2) + (4)	83.90 (±1.49)	84.73
(1) + (2) + (5)	83.92 (±1.31)	84.64
(1) + (2) + (6)	83.85 (±1.26)	**84.82**
(1) + (2) + (7)	83.81 (±1.26)	**84.82**
Jarvis et al. (2013)	**84.50**	83.60
Nicolai et al. (2013)	58.50	81.70

Table 1: Classification accuracy of using different features by 10-fold cross-validation on the train+development set and test on the test set, the accuracy scores are in %. The values in bracket are the standard deviation of accuracy scores in 10-fold cross-validation.

which is consistent with the results reported by other researchers (Jarvis et al., 2013; Nicolai et al., 2013). Using the combination of word n-grams and lemma n-grams improves the performance of using only word n-grams by less than 0.2%. It is not a significant improvement, mainly because there is a big overlap in the features in these two categories. Word-level spelling errors when used on their own do not perform well, with one of the causes being sparseness. When combining them with other features (word n-grams and lemma n-grams), *word_error_toefl* does not seem to provide any additional information for improving classification accuracy. By combining the *word_error_icle* with lemma n-grams and word n-grams, however, we observe a small increase in classification accuracy of 0.4%. Main reason for this is that by using *word_error_icle*, we incorporate errors that are more common (since they occurred in two different corpora – ICLE*v2* and TOEFL11 corpus).

From the classification results obtained by using feature (5), (6) and (7), it is also worth noting that by using only the character n-grams extracted from the spelling errors in *Spelling_error_TOEFL* (feature size: 5797), or even filtered by those that also appear in *Spelling_error_ICLE* (feature size:

3907), the accuracy is almost as the same as the one obtained by using all character n-grams (feature size: 12601). It shows that the character n-grams in the spelling errors are a most indicative part in all the character n-grams when it comes to identifying the L1 of an author.

Using character n-grams extracted from spelling errors works better than directly using misspelled words. It further implies that while the misspelled words might differ from one another in the text, the misspelled parts share similarities. The classifier trained by combining word n-grams, lemma n-grams and character n-grams extracted from *Spelling_error_ICLE* or from *Spelling_error_TOEFL* both reach the best test accuracy in our experiments - 84.82%, which is 1.2% better than the best result reported by Jarvis et al. (2013) in the NLI 2013 shared task. We note that including only the character n-grams extracted from the spelling errors ((1)+(2)+(6) or (1)+(2)+(7)) leads to better results than including all the character n-grams ((1)+(2)+(5)). It supports the hypothesis that spelling errors capture relevant information about the writer's native language at a character level.

Table 2 includes some of the most informative spelling errors made by writers with different L1s. They were selected based on their weights after training the SVM with word n-grams, lemma n-grams and spelling errors (as word). They seem to confirm our starting hypothesis regarding the various phenomena of language – and script – transfer that can influence spelling errors.

As shown in the table, the informative errors for each target language are quite different. French writers tend to double the character in the word, for example, they misspell *"personally"* as *"personnaly"* and *"developed"* as *"developped"*. It is also apparent that Japanese writers and Italian writers tend to misspell the vowels in a word. This may be the result of rules of word pronunciations in their own languages, and sound to letter mappings in their L1. Arabic writers tend to omit vowels. It could be due to the fact that vowels are rarely written in Arabic language and the writers carry this habit in their writing in English.

The results confirm the usefulness of features representing spelling errors, particularly at a subword level. On their own, they perform on a par with character level representation of the document. They also bring improvement in performance when combined with word and lemma n-grams, indicating that they provide at least partly complementary information to the frequently used word n-grams or lemma n-grams, which on their own have high performance.

L1	Word
ARA	*evry, experince, diffrent, advertisment, statment*
DEU	*knowlegde, advertisment, successfull, freetime, neccessary*
FRA	*generaly, personnaly, litterature, independant, developped*
HIN	*theoritical, sucess, enviornment, sucessful, gandhi*
ITA	*indipendent, specialistic, tecnology, studing, istance*
JPN	*actualy, youg, shoud, peple, convinient*
KOR	*poors, newpaper, eventhough, becaus, thesedays*
SPA	*conclution, consecuences, succesful, responsabilities, enviroment*
TEL	*oppurtunities, hardwork, intrested, atleast, donot*
TUR	*altough, spesific, easly, succesful, turkish*
ZHO	*sociaty, knowlege, easiler, sucessful, improtant*

Table 2: Most informative spelling errors made by writers with different L1

5 Conclusion

In this work, we investigate the usefulness of spelling errors for the native language identification task. The experiments show that representing spelling errors through character n-grams captures interesting phenomena of language transfer. Both on their own and combined with customarily used word n-grams, they have high performance in terms of accuracy, when tested on the TOEFL11 corpus and compared to participating systems in the NLI shared task. In future work, it would be interesting to characterize the spelling errors with respect to similarity to specific L1s, and further explore the hints that they provide with respect to the author's native language.

Acknowledgments

We would like to thank EST for making the TOEFL11 dataset public and Fondazione Bruno Kessler for providing the facilities to conduct the related experiments. We also thank the anonymous reviewers for their detailed and insightful comments.

References

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. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*. pages 391–408. http://aclweb.org/anthology/C/C12/C12-1025.pdf.

Scott A. Crossley and Danielle S. McNamara. 2011. Shared features of l2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing* 20(4):271–285. https://doi.org/10.1016/j.jslw.2011.05.007.

Susan T. Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers* 23(2):229–236. https://doi.org/10.3758/BF03203370.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *The International Corpus of Learner English: Handbook and CD-ROM, version 2*. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium. https://doi.org/10.1017/S0272263110000641.

Binod Gyawali, Gabriela Ramírez-de-la-Rosa, and Thamar Solorio. 2013. Native language identification: a simple n-gram based approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA*. pages 224–231. http://aclweb.org/anthology/W/W13/W13-1729.pdf.

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, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA*. pages 111–118. http://aclweb.org/anthology/W/W13/W13-1714.pdf.

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*. ACM, New York, NY, USA, KDD '05, pages 624–628. https://doi.org/10.1145/1081870.1081947.

Shibamouli Lahiri and Rada Mihalcea. 2013. Using n-gram and word network features for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA*. pages 251–259. http://aclweb.org/anthology/W/W13/W13-1732.pdf.

Thomas Lavergne, Gabriel Illouz, Aurélien Max, and Ryo Nagata. 2013. Limsi's participation to the 2013 shared task on native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA*. pages 260–265. http://aclweb.org/anthology/W/W13/W13-1733.pdf.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2013, June 13, 2013, Atlanta, Georgia, USA*. pages 140–145. http://aclweb.org/anthology/W/W13/W13-1718.pdf.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/tree-tagger1.pdf.