Prediction of Learner Native Language by Writing Error Pattern

Brendan Flanagan[1(✉)], Chengjiu Yin[2], Takahiko Suzuki[3], and Sachio Hirokawa[3]

[1] Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
b.flanagan.885@s.kyushu-u.ac.jp

[2] Faculty of Arts and Science, Kyushu University, Fukuoka, Japan
yin.academic@gmail.com

[3] Research Institute for Information Technology, Kyushu University, Fukuoka, Japan
{suzuki,hirokawa}@cc.kyushu-u.ac.jp

Abstract. The native language of a foreign language learner can have an effect on the errors they make because of similarities or differences between the two languages. In order to provide effective error prediction and correction for non-native English language learners it is important to identify their specific characteristic error patterns that are influenced by their native language. In this paper, we examine analyzing error detection scores to predict the native language of an English language learner. 15 categories of error detection scores are combined to create an error prediction score vector representation of each sentence. The native language is predicted by training an SVM classifier with the error vectors. The results are compared to an SVM classifier trained with just word representations of the learner writing sentences.

Keywords: Native language prediction · Writing errors · SVM classifier

1 Introduction

As a result of increased globalization facilitated by the Internet, the number of foreign language learners has increased. In particular, the numbers of people who speak English as a second or foreign language are increasing. Graddol [1] suggests that 80 % of communication in English is among non-native speakers. It has been estimated that there are over a billion second or foreign language speakers of English, which is the native language of only approximately 400 million people [2]. As many automated correction methods are targeted at native speakers, there is an increasing need for second or foreign language targeted tools to correct their characteristic errors.

The native language of a foreign language learner can have an effect on the errors they make because of similarities or differences between the two languages. In order to provide effective error prediction and correction for non-native English language learners it is important to identify their specific characteristic error patterns that are influenced by their native language. In our previous research into the prediction of

P. Zaphiris and A. Ioannou (Eds.): LCT 2015, LNCS 9192, pp. 87–96, 2015.
DOI: 10.1007/978-3-319-20609-7_9

writing errors in foreign language writing [3], we identified the differences and similarities of error co-occurrence characteristics of learners who's native language are: Chinese, Japanese, Korean, Spanish, and Taiwanese. In particular, by error clustering analysis we found that some languages were quite similar in their characteristics, such as: Japanese and Korean, while others only shared a few common characteristics.

The writing error categories used in our research are based on previous empirical studies on the writings of foreign language students in academic settings [4, 5]. A parallel corpus of original and corrected sentences was collected from the writings of learners on the popular online language learning SNS lang-8.com. A randomly selected subset of sentences from the corpus was manually categorized by hand into 15 writing error categories. This subset was then analyzed to train and evaluate SVM classifiers for each of the writing error categories [6–8]. The error category classifiers output a score for 15 error categories is a vector representation of the analyzed sentence.

In this paper, we will compare the prediction performance of two SVM models: one created by analyzing error category prediction vectors, and another created by analyzing word vectors. This can be thought of as comparing two different viewpoints: the error category prediction vector viewpoint, and the word vector viewpoint.

2 Related Work

2.1 Native Language Prediction

Wong [9] analyzed learner writing with an extension of adaptor grammars for detecting colocations not only at the word level, but also for parts-of-speech and functional words. Classification was performed at the document level by parsing individual sentences of the learner's writing to detect the native language with the final prediction based on a majority score of the sentences. Some notable characteristic features of languages extracted by this method were also discussed.

Brooke et al. [10] suggested that the International Corpus of Learner English (ICLE) corpus, which is commonly used in research into native language prediction of learner writing, has problems that can lead to misleading performance evaluation. It was argued that the problem stems from the way the corpus was built, and proposed other methods and sources to collect data that might be useful in the task of native language prediction. An evaluation was undertaken on data collected from a language learning SNS, Lang-8.com, and it was shown to be useful for the task. In this paper, we analyze data collected from Lang-8.com for the purpose of native language prediction by writing error prediction vector.

In 2013, Tetreault et al. [11] organized a shared task on native language identification of learners through analysis of their writing. A new corpus named TOEFL11, which contains essays in English by learners from 11 different native languages and was provided as the shared data set on which the participants conducted analysis. Jarvis et al. [12] was a participating group with a high identification performance. A variety of features were analyzed in the identification task, such as: word n-grams, parts-of-speech n-grams, character n-grams, and lemma n-grams. An SVM classifier was trained and the prediction performance was evaluated of several different models with varying combinations of features.

In this paper, we investigate the difference in prediction performance of an SVM classifier trained with writing error prediction vectors and an SVM classifier trained with basic word features.

2.2 Native Language Prediction by Error Analysis

Koppel et al. [13], investigated predicting a learner's native language by analyzing writing errors detected with MS Word and a Brill based parts-of-speech tagger in addition to other features, such as: function words, letter n-grams, and rare part-of-speech bigrams. They analyzed a sub-corpus of ICLE containing learner writings by learners with the following native languages: Russia, Czech Republic, Bulgaria, France and Spain. It was found that most classification errors occurred between writings from Slavic languages. An overall accuracy of 80 % was achieved using all features.

Kochmar [14], predicted the native languages of Indo-European learners through binary classification tasks preformed with linear kernel SVM. Compare to previous studies a larger set of learner native languages were examined. These native languages were divided into two main groups: Germanic and Romance, with intergroup prediction performance accuracy ranging from 68.4 % to 100 %. The features analyzed for prediction ranged from general words and n-grams, to different error types that had been manually tagged within the corpus.

Bestgen et al. [15], investigated the used of error patterns in the identification of the native languages of learners. They analyzed the manually tagged errors within the ICLE. The 46 error types that have been tagged in the corpus were used to predict the native language of 223 learner writings. Three groups of native languages were chosen: French, German, and Spanish. They identified that using just errors as a predictor of native language an accuracy of 65 % could be achieved. Discriminative error types for the three native languages were identified by comparing the mean relative frequency significance difference of each error category. They impact of proficiency on the results was also examined and resulted in improved predictive discrimination between French and German learners. In conclusion it is mentioned that it still remains to be seen if the same prediction performance can be achieved through the automatic detection of writing errors, instead of relying on manual classification by hand.

In this paper, we endeavor to investigate the prediction performance of automatic error detection as a predictor of the native language of learners.

3 Data

The data for analysis in this paper was collected from lang-8.com, a language learning SNS site. The target data was learner journals that were written in English and posted on Lang-8 during the period from Oct 9 2011 to Jan 6 2012. A total of 57,776 journals written in English were collected. Within these journals, there were 142,465 sentences that had been corrected by native English speakers who are members on the lang-8.com site. As the corrections are made at the sentence level, analysis undertaken in this paper is by sentence units. The native and target languages of the learner were also collected and each sentence was annotated with this information accordingly. An alignment algorithm was used to identify the corrected words within the sentence and were tagged as either insert or delete and also generally as an edit. Figure 1 shows the sentence distribution of the five main learner native languages who's English journals were corrected by a native English speaker. A large majority are Japanese natives who have written 100,432 corrected sentences. The other main learner native languages are in descending order: Chinese, Korean, Taiwanese, and Spanish, which each makeup more than 2 % of the total sentences.

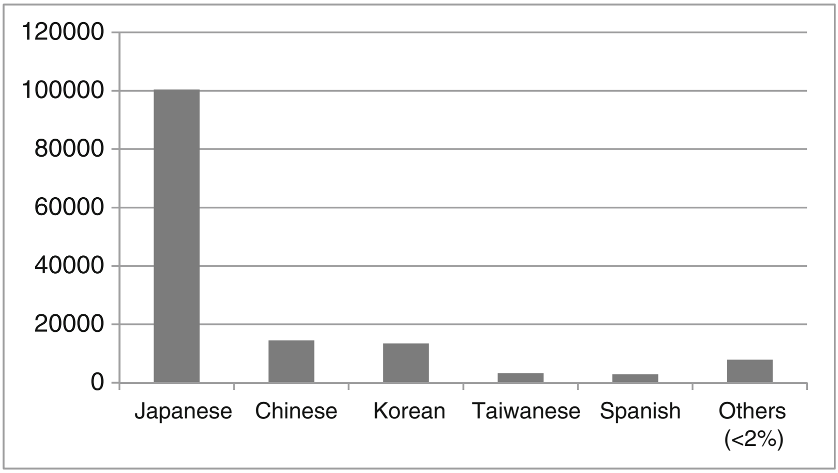

Fig. 1. Distribution of learner native languages

4 Error Prediction and Error Vector

In previous research [6–8], the authors have predicted 15 different writing error types by SVM classifier. These errors were selected from a larger list of 42 error types in [16] because of the frequency in annotated data. Table 1 lists the 15 writing error type descriptions along with the original error category number.

In this paper, we predict the errors of sentences by SVM models that were trained and evaluated using 10-fold cross validation. As a result of this evaluation there are 10 models for each writing error type. The prediction for each error type is made up of the average of the 10 scores from the models. The predictions are then combined to form an error vector representation for each sentence as seen in Fig. 2.

The distribution of predicted errors for each of the five main learner native languages is displayed in Fig. 3.

5 Trivial Biased Words

Initially an SVM model was trained to predict the native language of learners just by analyzing the words in their writings, however the prediction performance was higher than expected, so we investigated the characteristic feature words for each language.

Table 1. Predicted error categories

Category	Description
2	Subject formation
3	Verb missing
6	Dangling/misplaced modifier
11	Word order
13	Extraneous words
17	Tense
19	Verb formation
25	Ambiguous/unlocatable referent
28	Lexical/phrase choice
30	Word form
33	Singular for plural
36	Preposition
37	Genitive
38	Article
42	Spelling

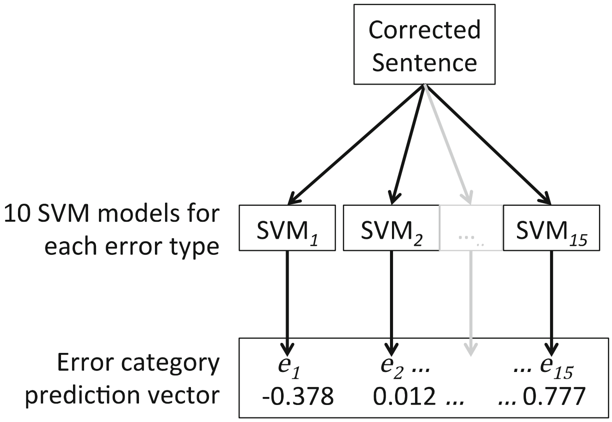

Fig. 2. The process of creating error vector representations of each sentence

An SVM model was trained for each learner native language by analyzing all of the data. These models were analyzed to calculate and rank all of the feature words by weight.

Feature words with a high positive weight are characteristic of that particular learner group. In Table 2, the top 10 positive and negative weight feature words for native Japanese learners of English are shown. Many high positive words are directly related to Japan, were as low negative words are related to other countries. Therefore, these words are trivial biased words that have been influenced by the nation or culture

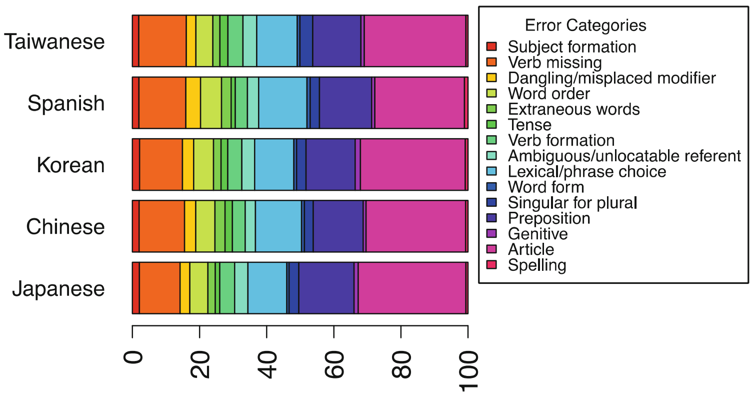

Fig. 3. Distribution of predicted errors for each language

Table 2. Top 10 positive and negative feature words by weight for native Japanese learners of English.

Top Positive 10		Top Negative 10	
Word	Weight	Word	Weight
north	1.0305	taiwan	-1.2025
japan	1.0073	campus	-1.251
tokyo	0.6735	soju	-1.26
japanese	0.572	beijing	-1.3393
peninsula	0.5502	pepero	-1.3534
jong	0.5223	korean	-1.522
kara	0.5032	kimchi	-1.5315
kyoto	0.4653	1	-1.7565
thailand	0.4447	korea	-1.7737
algerian	0.4447	seoul	-1.8214

of the learner. The characteristic feature words for each learner native language group also contained similar influences.

Other sources of trivial biased words included events that had occurred just before the collection of data from the lang-8.com website (October 2011 ~ January 2012). Table 3 contains feature words that we believe are related to the 2011 Tohoku Earthquake and Tsunami that occurred in Japan.

To reduce the influence of trivial biased words and provide a fare comparison between the proposed method of language prediction by error vector and the baseline method of prediction by words, feature words with a high frequency distribution difference between the native language groups were removed. The relative standard deviation for each word was calculated as follows:

Table 3. Biased words in the model for Japanese native language learners

Rank	Weight	word
13	0.3602	earthquake
…	…	…
24	0.3093	radiation
…	…	…
42	0.2943	nuclear

$$TDR(w,l) = \frac{TF(w,l)}{DF(l)} \tag{1}$$

$$\mathrm{s}(w) = \sqrt{\frac{\sum_{l \in L} TDR(w,l)^2}{|L|} - \left(\frac{\sum_{l \in L} TDR(w,l)^2}{|L|}\right)^2} \tag{2}$$

$$\bar{x}(w) = \frac{\sum_{l \in L} TDR(w,l)}{|L|} \tag{3}$$

$$RSD(w) = \frac{s(w)}{\bar{x}(w)} \tag{4}$$

Where Eq. 1 is the term document ratio for the word w in language set l, and TF is the term frequency and DF with the document frequency. The standard deviation and mean of the term documents ratio between languages is calculated in Eqs. 2 and 3 respectively. Then finally the relative standard deviation is shown in Eq. 4.

A list of words ranked by RSD was manually checked for words that might identify the culture or nation of the five main groups of native languages. Through these manual checks it was estimated that words with an RSD of greater than 1.25 were trivially biased towards one or more of the native languages. Figure 4 shows a plot of all words ranked by RSD in descending order, with the horizontal line at 1.25 RSD representing the maximum threshold for non-biased words used in the analysis of this paper.

6 Method and Results

To provide a fare evaluation of the two feature sets, the same method was used for training and evaluating prediction performance of error prediction vectors and word vector features. For additional comparison, we also include the prediction performance for word vectors that contain all the words of the original learner writing, including those that were identified as trivially biased in the previous section. For the word vectors, the words of each sentence were vectorized as a bag-of-words model. The error prediction vector consists of the values of 15 error prediction scores.

Separate SVM classifiers were trained for five different native languages across all three data sets. The native language prediction performance of each of these classifiers

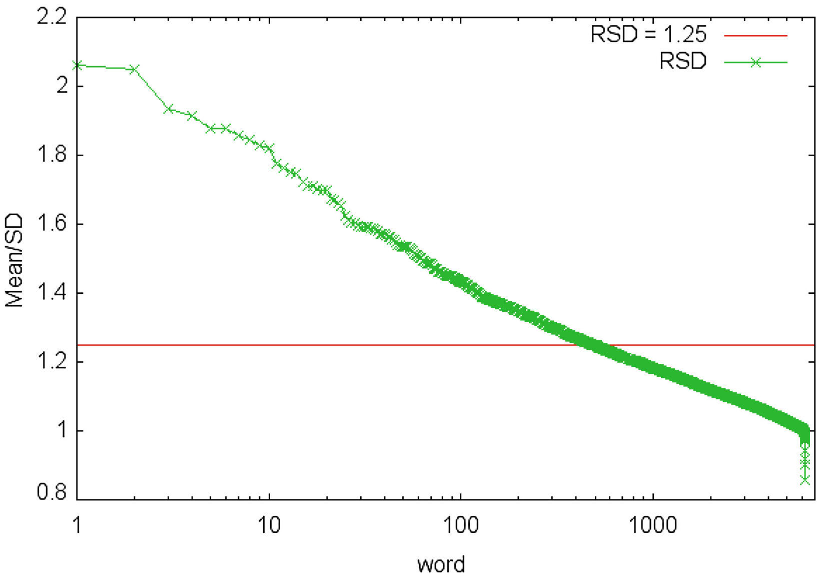

Fig. 4. The RSD distribution of word frequencies between five native languages

was evaluated by randomly sampled 10-fold cross validation, with 9:1 training to test data ratio for each of the data sets.

A comparison of the prediction performance evaluation on all three data sets for each of the five native languages is shown in Fig. 5. The prediction performance of the

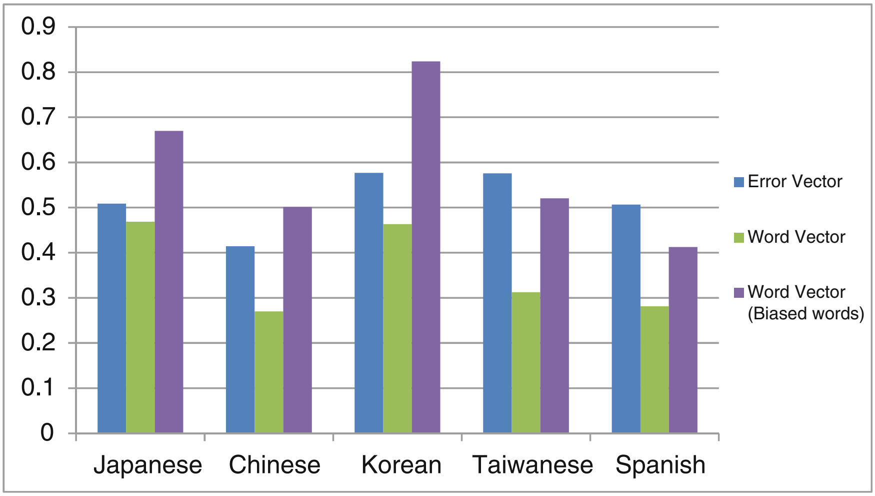

Fig. 5. Native language prediction evaluation for each vector (Accuracy, 10-fold cross validation)

word vectors that include biased words is high, especially for writings by native Korean learners. This would suggest that there are biased words that are highly characteristic of native Korean learners. The word vectors that do not contain trivial biased words have a prediction performance ranging from 36 % lower in the case of Korean, to 13 % lower for Spanish. The native language prediction performance by error prediction vector is higher than the performance of the unbiased word vector. However the prediction performance for two out of the five native languages is lower than that of the word vectors that contain all the words of the original learner writings, which we argue is influenced by biased words.

7 Conclusion

In this paper, we evaluated and compared the prediction performance of error prediction vectors and word vectors. Initial analysis indicated that the learner writing data that was collected from Lang-8.com contained trivial biases, which were in the form of differences in word use distribution due to the culture, location, and recent localized events. A method for identifying and reducing trivial biased words was proposed to alleviate the problem. SVM classifiers were then trained for three data sets: error prediction vectors, word vectors without biased words, and word vectors containing all the words from the original learner writing. The prediction performance for each data set was then evaluated with 10-fold cross validation. The prediction performance error prediction vectors were superior to the unbiased word vectors for all native languages. However, word vectors containing all words including biased words performed better in three out of five native languages.

In future work, we intend to examine in detail the results of our evaluation along with comparisons to other methods and corpora. It is also necessary to perform a search for optimal selections of error predictions to further enhance the native language prediction performance.

Acknowledgment. This work was partially supported by JSPS KAKENHI Grant Number 24500176.

References

1. Graddol, D.: English Next: Why Global English May Mean the End of English as a Foreign Language. British Council, London (2006)
2. Guo, Y., Beckett, G.H.: The hegemony of english as a global language: reclaiming local knowledge and culture in china. Convergence **40**, 117–132 (2007)
3. Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S.: Classification and clustering english writing errors based on native language. In: IIAI 3rd International Conference on Advanced Applied Informatics (IIAIAAI), pp. 318-323 (2014)
4. Kroll, B.: What does time buy? ESL student performance on home versus class compositions. In: Kroll, B. (ed.) Second Language Writing: Research Insights for the Classroom, pp. 140–154. Cambridge University Press, Cambridge (1990)

5. Weltig, M.S.: Effects of language errors and importance attributed to language on language and rhetorical-level essay scoring. In: Spaan Fellow Working Papers in Second or Foreign Language Assess. vol. 2(1001), pp. 53-81 (2004)
6. Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S.: Intelligent Computer Classification of English Writing Errors. In: Proceedings of the 6th International Conference on Intelligent Interactive Multimedia Systems and Services (IIMSS 2013) vol. 254, pp. 174-183, IOS Press (2013)
7. Flanagan, B., Yin, C., Hashimoto, K., Hirokawa, S.: Clustering English Writing Errors based on Error Category Prediction, ISEEE 2013, pp. 733-738 (2013)
8. Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S.: Classification of english language learner writing errors using a parallel corpus with svm. Int. J. Knowl. Web Intell. **5**(1), 21–35 (2014)
9. Wong, S.M.J., Dras, M., Johnson, M.: Exploring adaptor grammars for native language identification. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 699-709 (2012)
10. Brooke, J., Hirst, G.: Native language detection with 'cheap' learner corpora. In Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011), pp. 37-57, Presses universitaires de Louvain (2013)
11. Tetreault, J., Blanchard, D., Cahill, A.: 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, pp. 48-57 (2013)
12. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. NAACL/HLT **2013**, 111–118 (2013)
13. Koppel, M., Schler, J., Zigdon, K.: 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, pp. 624-628, ACM (2005)
14. Kochmar, E.: Identification of a writer's native language by error analysis. Master's thesis. University of Cambridge (2011)
15. Bestgen, Y., Granger, S., Thewissen, J.: Error patterns and automatic L1 identification. In: Approaching Language Transfer Through Text Classification, pp. 127-153 (2012)
16. Flanagan, B., Yin, C., Hirokawa, S., Hashimoto, K., Tabata, Y.: An automated method to generate e-learning quizzes from online language learner writing. Int. J Distance Educ. Technol. **11**(4), 63–80 (2013)