The 3rd International Symposium on Engineering, Energy and Environment
Pullman King Power Hotel, Bangkok: 17-20 November 2013

Clustering English Writing Errors Based on Error Category Prediction

Brendan Flanagan[a*], Chengjiu Yin[b], Kiyota Hashimoto[c], Sachio Hirokawa[b]

[a] *Graduate School of Information Science and Electrical Engineering, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan*
[b] *Research Institute for Information Technology, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan*
[c] *School of Humanities and Social Sciences, Osaka Prefecture University, 1-1 Gekuen-cho, Naka-ku, Sakai 599-8231, Japan*

Abstract

It is important for language learners to determine and reflect on their writing errors in order to overcome weaknesses. Each language learner has their own unique writing error characteristics and therefore has different learning needs. In this paper, SVM machine learning is applied to the writings of English language learners on the language learning SNS website Lang-8. These writings have been manually classified into 15 error categories to determine the errors in a sentence. Feature selection was used to improve the performance of the resulting classifier model. This is then utilized to analyze 142,465 sentences from journals written by online language learners to determine their errors and user characteristics.

Keywords: Language learning; user characteristics; writing error categories; machine learning.

1. Introduction

Language learning outside traditional classroom settings has increased in popularity along with increased usage of the Internet. There are many websites that utilize different methods ranging from more traditional techniques to social network services (SNS) based learning platforms, of which the later is of particular interest. These SNS website bring together language learners from across the global and are based around the idea of language exchange. Native speakers correct the writings of a learner studying their native language. In principal, these learners then correct the writings of a learner studying their own native language. For example, person A is a native Japanese speaker who is learning English as a foreign language and posts an English sentence on the website. Person B who is a native English speaker corrects the sentence. Person B is also learning Japanese as a foreign language and posts a

* Corresponding author. *E-mail address*: b.flanagan.885@s.kyushu-u.ac.jp.

sentence on the website in Japanese which is then corrected by person A. This mutually beneficial environment helps learners to achieve their respective goes of learning a foreign language, which in turn is another foreign language learner's mother tongue.

Mutual correction websites contain a large amount of foreign language writing correction data. Taking advantage of this data can help to further enhance the effectiveness of language learning. Using data from Lang-8 (http://www.lang-8.com), the authors of this paper have in past research categorized the errors in sentences manually by hand and built a quiz system [1]. They have a long-term research goal of creating tools for autonomous web based language learners that will enable them to analyze their errors and provide feedback according to their weak points. To this point it is important to identify the error characteristics of learners. These errors may occur either systematically as the result of erroneous learning in the past and as a result continue to make the same errors over and over again.

While there have been a number of researches into the errors of academic level foreign language writing, research into the foreign language writings of the general populous, in particular on language learning SNS websites, are of interest because the learners that are using these sites are likely to be autonomous learners without access to language specialists teachers to provide analysis of their writing and corrective feedback.

In this paper, the writings, in particular the diaries, of language learners on the website Lang-8 are analyzed using machine learning techniques. The method of feature selection [2] is used to improve the prediction performance of error category given in Flanagan [3]. The constructed models are used to cluster 142,465 sentences of Lang-8 data.

2. Related work

2.1. English writing error categories and corpora

Previous empirical studies on the writings of foreign language students been undertaken in academic settings to enabled the control of influencing factors, such as: subject and environment. Kroll [4] compared the difference of writings that were conducted in classroom where learners had a fixed amount of time, and the home environment, where they would have more time and less pressure. English teachers categorized errors manually and the frequency of occurrence was used to compare the writings in the two different environments. Weltig [5] looked at the effect of different categorizes of errors on the scoring given by English teachers for the writings of foreign language learners. Using similar error categories as Kroll [6], it was found that the frequency of certain error categories had more of an influence on the overall score than others. The sample data in this paper was prepared for machine learning by using similar categories to Kroll and Weltig for manually identifying errors in sample pair sentences from Lang-8. Language learners in academic settings have access to language specialists such as teachers that can provide analysis and corrective feedback, however this is not as readily available to autonomous learners. To fill this gap, we have a goal of creating tools for these learners to enable them to a certain extent to be given some feedback and analysis similar to that provided by language specialists.

Sugiura et al. [6], discuss corpus design and reviewed the International Corpus of Learner English (ICLE). Based on the corpus weaknesses identified, they set about compiling a new English learner corpus and a parallel corpus or native English speakers, called NICE (Nagoya Interlanguage Corpus of English). Using this corpus they performed analysis using mechanical text features, such as: type, token, number of sentences, and average word length to compare the language learners performance with native speakers.

Miki [7] looks at the use a parallel corpus that is constructed using the essay writings of foreign language learners and exact forms of the sentences that are provided by native English language speakers. NICE was used as a dataset to exam how Japanese English language learners use "I think" in comparison with native speakers. Unlike other studies on the over usage of expression which focus on quantifying the errors, by using a parallel corpus they were able to determine how the expression was being inappropriately used to augment the language learners writing. Miyake et al. [8] also used the same method and NICE parallel corpus to examined the use of "there" with the long-term intention of identifying the "Japaneseness" and "nativeness" relating to the use of constructions.

2.2. SVM error categorization

Previous studies have estimated errors in English text by using SVM and other types of machine learning algorithms. Hirano et al. [9] investigated the use of search engine results to detect article errors in English technical papers. The sentences were syntactically parsed to produce a parts of speech tagged sentence, and then a search query was created based on the structure of the sentence. The number of hits from the resulting search query was then counted and used to determine if the input sentence contained an error. Tanimoto et al [10] examined using the number of search results as a indicator in an attempt to identify erroneous words in English sentences. NICE (Nagoya Interlanguage Corpus of English) was used in tri-grams and 4-grams as training data for SVM machine learning to create a model that can determine if an English sentence contains an error.

3. Prediction of Error Category by SVM

3.1. Automatic Prediction of Error Category Using SVM

The error categories of each sentence were manually determined in Flanagan [3] and machine learning method SVM was used to generate models to predict the error categories of writing errors. The same training data used and optimal feature selection is used in the present paper. SVM is often used due to its ability to handle large numbers of attributes. In fact, there are more than 400,000 words in the target writings in Lang-8 data and so SVM is appropriate choice. However, the prediction performance in Flanagan [3] was not satisfactory as the highest score was less than 0.4. The present paper searches for an optimal feature selection as reported in Sakai [2], where the target data was also short sentences.

3.2. Optimal Feature Selection

We applied SVM-light using all of 399 data in Flanagan [3] as training and test data to construct 15 models with respect to each error category. Then each model was applied to an imaginary sentence that consists of a single word. The score is used as the predicted score of the word with respect to the error category. The feature words with respect a category c is determined as the union of $POS(c,N)$ and $NEG(c,N)$, where $POS(c,N)$ is the set of words whose c-score is positive and in the top N, and $NEG(c,N)$ is the set of words whose c-score is negative and in the bottom N. The sentences are vectorized with those words, and F-measure is evaluated with 10 fold cross validation. Then the optimal F-measure is evaluated among N=1,2,3,...,10,20,30,...,100, 200,300,...,900,1000. In Fig.1, (a) and (b) are plots of F-measures for error categories 2 and 42, where the optimal choice for (a) is N=900, and (b) is N=800. We also investigated the existence of an optimal N for the other error categories.

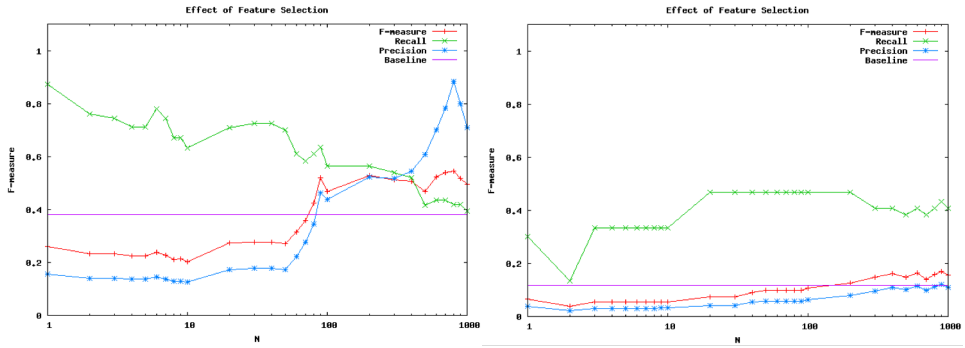

Fig 1. Effect of Feature Selection for (a) Error Category 2 and (b) Error Category 42.

Table 1. The words and tags from the model created using SVM.

Err		Feature words
42	Spelling	shopping e went e:e i:e phrase china day friend what
28	Lexical/phrase choice	which m it am would student in d:in here girl
38	Article errors	e:the i:the e:a the i:a a man e:A university e:This
36	Preposition	i:in e:in d:at at e:for e:at e:on on i:on two
19	Verb formation	i:ing e:ing ing didn e:to entrance d e:eat d:eating collage

An example of feature words that can be extracted for use by learners and teachers alike were reported in Flanagan [3] as seen in Table 1. Extracting feature words could enable learners and teachers to understand the words that are catalysts in their particular writing errors.

3.3. Improvement of Prediction Performance by Optimal Feature Selection

Table 2 and Fig. 2 shows the F-measures by baseline [3] and the proposed method with respect to the 15 error categories. The F-measures is greater than 0.4 in the five categories (19 Verb formation, 28 Lexical/phrase choice, 36 Singular for plural, 38 Article errors, 42 Spelling). In all cases, the prediction performance is improved.

Table 2. Prediction Performance (F-measure) Compared

Category	Optimal N	Description	Feature Selection	Baseline
2	900	Subject formation	0.1695	0.1169
3	200	Verb missing	0.1490	0.1077
6	20	Dangling/misplaced modifier	0.0403	0.0105
11	500	Word order	0.2843	0.2248
13	300	Extraneous words	0.1552	0.0718
17	500	Tense	0.1040	0.0917
19	800	Verb formation	0.4508	0.2828
25	700	Ambiguous/unlocatable refer	0.1746	0.1087
28	200	Lexical/phrase choice	0.5001	0.3672
30	500	Word form	0.1172	0.0750
33	300	Singular for plural	0.3100	0.1910
36	700	Preposition	0.4688	0.2948
37	700	Genitive	0.2115	0.0957
38	500	Article errors	0.5264	0.3652
42	800	Spelling	0.5452	0.3807

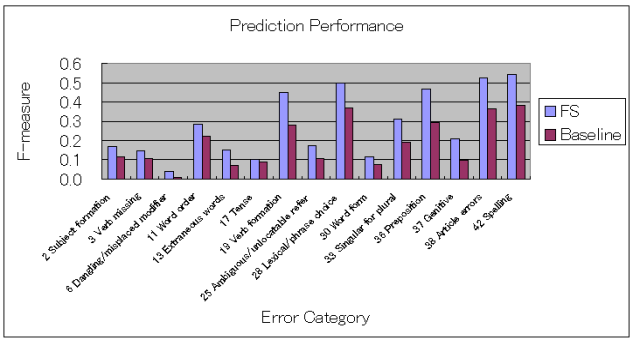

Fig.2 Comparison of Prediction Performance

4. Clustering based on Error Category Prediction

We calculated the score of each sentence if it contains an error among 15 error categories by applying the constructed model. The target is 142,465 sentences, posted from October 9, 2011 to January 6, 2012, which are written in English and are corrected some way.

As explained in the previous section, there were 10 models constructed for each error category. The score of a sentence with respect an error category is calculated as the average of the score of the result obtained by applying the 10 models. 15 scores correspond to 15 error categories form a vector representation of a sentence. We then applied the clustering tool CLUTO to cluster sentences into 20 clusters. Fig.3 shows the results.

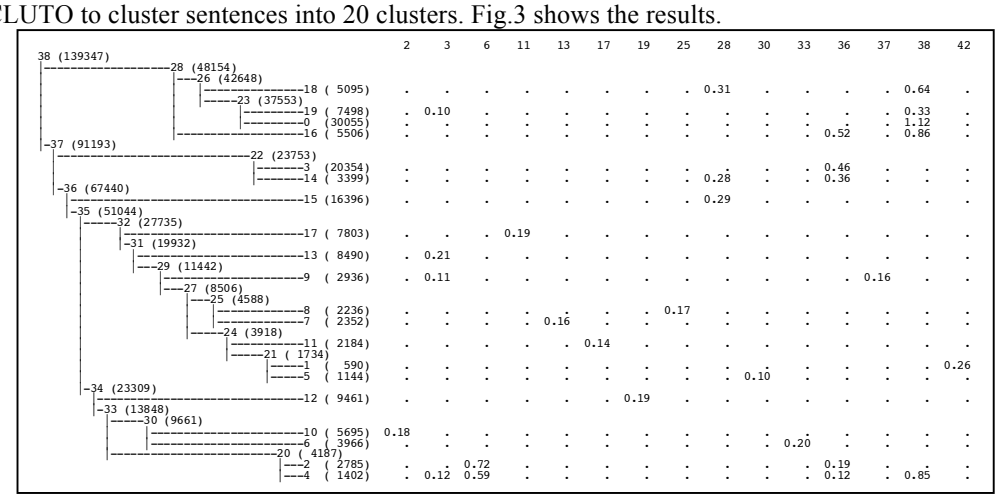

Fig. 3 Clustering of Error Sentences based on Predicted Error Category Score Vectors

The number of sentences is represented in the brackets following the cluster ID, and only the average scores of over 0.1 of error category prediction are shown. This visualization is very helpful to understand the huge amount of target data. We can see that cluster 28 contains 1/3 of all the data and corresponds to the error category 38 (article error). The dendrogram (clustering tree) closely represents the structure of the cluster. Cluster 0 is the core part of cluster 28 whose sentences contain only article errors. Cluster 19 contains article errors as well as verb-missing errors (category 3). The cluster 18 contains lexical or phrase choice errors (category 28). The cluster 16 contains preposition errors (category 36). Thus, the tree represents not only the clustering of sentences but also the clustering of error categories.

Indeed, we can interpret that article errors (category 38) are the largest errors and occur with preposition error (category 36), lexical/phrase choice error (category 28) and verb missing (category 3).

The cluster 22 is the next tightly connected large cluster with 23,753 sentences. This cluster corresponds to preposition error (category 36) and lexical/phrase choice error (category 28). The next largest pure cluster is the cluster 15 that contains only lexical/phrase choice error (category 28). The above observation gives top-down view of error causes of writings and how error categories co-occur together. The error category 3, 6, 23, 36 and 33 occurs in multiple lines. This indicates that the learning and teaching of those kinds of writing errors do not have to be considered independently. On the other hand, if we see each category in vertical line, we find isolated error categories. For example, the category 2, 11, 13, 17, 19, 25, 30, 33, 37 and 42 appears only in one cluster. The learning and teaching of those errors can be handled independently.

5. Conclusion and Further Work

The present paper applied SVM and feature selection to improve the error category prediction of writing errors of English learners' writing on the language learning SNS web site Lang-8. A total of 142,465 writings were classified and clustered based on the score vectors. In future work we will undertake detailed analysis and evaluate the validity of the results.

6. Acknowledgments

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

The authors would also like to thank the reviewers for their valuable constructive criticism that was provided during the review process.

References

1. C. Yin, S. Hirokawa, B. Flanagan, T. Suzuki, Y. Tabata, Mistake Discovery and Generation of Exercises Automaticity in Context, Proc. of LTLE2012 (2012) 163.
2. T. Sakai, S. Hirokawa, Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, (2012) 360.
3. B. Flanagan, C. Yin, T. Suzuki, S. Hirokawa, Intellegent Computer Classification of English Writing Errors, Proc. KES-IIMS2013, (2013) 174.
4. B. Kroll, What does time buy? ESL student performance on home versus class compositions, In B. Kroll (Ed.), Second language writing: Re- search insights for the classroom, Cambridge University Press, 1990, 140.
5. M. S. Weltig, Effects of language errors and importance attributed to language on language and rhetorical-level essay scoring, Spaan Fellow Working Papers in Second or Foreign Language Assessment Vol 2, 1001 (2004) 53.
6. M. Sugiura, M. Narita, T. Ishida, T. Sakaue, R. Murao, K. Muraki, A Discriminant Analysis of Non-native Speakers and Native Speakers of English. Proceedings of the Corpus Linguistics Conference CL2007, (2008) 27.
7. N. Miki,A new parallel corpus approach to Japanese learners' English, using their corrected essays. Themes in Science and Technology Education, 3(1-2), (2011) 159.
8. H. Miyake, T. Tsushima, On the features of there constructions used by Japanese speakers of English, The Journal of Humanities & Natural Sciences, 132, (2012) 55.
9. T. Hirano, Y. Hirate, H. Yamana, Detecting Article Errors in English using Search Engines, DBSJ Letters 6.3, (2007) 13. (in Japanese)
10. T. Tanimoto, M. Ohta, Examination of English Error Detection Using the Number of Search Results, DEIM Forum 2012, 9.1 (2012). (in Japanese)