Classification of English language learner writing errors using a parallel corpus with SVM

Flanagan, Brendan
Graduate School of Information Science and Electrical Engineering, Kyushu University

Yin, Chengjiu
Research Institute for Information Technology, Kyushu University

Suzuki, Takahiko
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：International Journal of Knowledge and Web Intelligence. 5 (1), pp.21-35, 2014
バージョン：

Classification of English language learner writing errors using a parallel corpus with SVM

Brendan Flanagan*

Graduate School of Information Science and Electrical Engineering,
Kyushu University,
6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, Japan
E-mail: b.flanagan.885@s.kyushu-u.ac.jp
*Corresponding author

Chengjiu Yin, Takahiko Suzuki and Sachio Hirokawa

Research Institute for Information Technology,
Kyushu University,
6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, Japan
E-mail: yin@cc.kyushu-u.ac.jp
E-mail: suzuki@cc.kyushu-u.ac.jp
E-mail: hirokawa@cc.kyushu-u.ac.jp

Abstract: In order to overcome mistakes, learners need feedback to prompt reflection on their errors. This is a particularly important issue in education systems as the system effectiveness in finding errors or mistakes could have an impact on learning. Finding errors is essential to providing appropriate guidance in order for learners to overcome their flaws. Traditionally the task of finding errors in writing takes time and effort. The authors of this paper have a long-term research goal of creating tools for learners, especially autonomous learners, to enable them to be more aware of their errors and provide a way to reflect on the errors. As a part of this research, we propose the use of a classifier to automatically analyse and determine the errors in foreign language writing. For the experiment in this paper, we collected random sentences from the Lang-8 website that had been written by foreign language learners. Using predefined error categories, we manually classified the sentences to use as machine learning training data. This was then used to train a classifier by applying SVM machine learning to the training data. As the manual classification of training data takes time, it is intended that the classifier would be used to accelerate the process used for generating further training data.

Keywords: error classification; language learning; SVM; machine learning; writing errors.

Reference to this paper should be made as follows: Flanagan, B., Yin, C., Suzuki, T. and Hirokawa, S. (2014) 'Classification of English language learner writing errors using a parallel corpus with SVM', *Int. J. Knowledge and Web Intelligence*, Vol. 5, No. 1, pp.21–35.

Biographical notes: Brendan Flanagan received his BS in Information Technology (Computing Studies) from RMIT University in 2010. Since 2013, he has been a graduate student studying advanced information technology at the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include: text mining, search engines, and CSCL (Computer Supported Collaborative Learning).

Chengjiu Yin received his PhD degree from the Department of Information Science and Intelligent Systems, Tokushima University, Japan, in 2008. He is an Assistant Professor in the Research Institute for Information Technology, Kyushu University. Currently he is committing himself in mobile learning, ubiquitous computing, language learning, text mining and social learning. He is a member of JSiSE, JSET, and APSCE.

Takahiko Suzuki received his BS in Physics, MS in Engineering, and Dr. of Engineering in Faculty of Engineering of Kyushu University in 1981, 1989, and 1991. Since 1997, he has been an Associate Professor in the Research Institute for Information Technology of Kyushu University. His research interests include: text mining, natural language processing, and location information.

Sachio Hirokawa received his BS and MS degrees in Mathematics and PhD in Interdisciplinary Graduate School of Engineering Sciences from Kyushu University in 1977, 1979 and 1992. Since 1997, he has been a Professor in the Research Institute for Information Technology of Kyushu University. His research interest includes search engine, text mining, and computational logic.

This paper is a revised and expanded version of a paper entitled ‘Intelligent computer classification of English writing errors’ presented at KES/IIMSS2013 Conference, Sesimbra, Portugal, 26–28 June 2013.

1 Introduction

In the last decade or so with the global spread of the internet, the number of people studying languages on the web has increased. Of particular interest are sites that offer a social or collaborative approach to study languages, and are often based on a social networking service (SNS) platform. Traditional classroom-based language study offered interaction with other learners and feedback from teachers and peers. To some extent these SNS-based websites offer some feedback and interaction that might otherwise be absent in autonomous learners studies. These language learning SNS sites work on the language exchange function, where native speakers of the target language offer corrections and feedback to the language learners. An example of this would be: Learner A is a native Japanese speaker who is studying English. Learner A writes a diary on their page of the SNS site in English and then offers it to all the native English speakers on the site to read and correct their mistakes. Learner B is a native Australian English speaker who is studying Japanese. Learner B reads Learner A’s diary in English, corrects the errors and provides feedback. Conversely, Learner B might also write a diary in the language they are learning and then Learner A could reciprocate by correcting Learner B’s writing errors and providing feedback. We refer to this process of language exchange as mutual correction. Mutual correction websites contain numerous foreign language writings that have been created by learners. These have then been corrected by speakers

of the target language and could be seen essentially as a crude crowd-sourced foreign language writing parallel corpus. In this paper we use machine learning to analyse the writings collected from a leading SNS-based mutual correction website, Lang-8 (http://www.lang-8.com).

Previously the authors of this paper have used data from mutual correction sites to build a system that offers automatically generated fill-in-the-blanks quizzes. The intention of the system is to be used by learners to practice their writing errors in an effort to prevent future mistakes (Yin et al., 2012). By determining the particular characteristics of the learner, and then practicing the quizzes that focus on their weaknesses has been shown to increase the effectiveness of learning. In the creation of the system, the errors in the candidate sentences were classified manually by hand, which is a time consuming process that takes effort and skill. In order to make the system fully independent, a method of automatically detecting and classifying errors in candidate sentences is required.

As there have been remarkable advances in machine learning research recently, we propose that machine learning techniques could be used to automatically detect and classify the errors in foreign language writing sentences. A machine learning classifier model for error detection could be created and used to determine the characteristics of the learners' errors. Remedial quizzes could be provided based on these characteristics to augment their usual studies.

This research is part of a long-term goal to provide language learners, and particularly autonomous learners, with tools to determine the error characteristics of their learning. 500 corrected sentence pairs by learners of English foreign writing on the Lang-8 website were chosen at random. The corrected sentence pairs were then manually classified into error categories. These error categories were based on the previous research investigated by Kroll (1990), Polio and Fleck (1998) and Weltig (2004) to examine the characteristics of foreign language writings. Using these error categories, we manually detected and classified the sample sentence pairs into error categories.

The raw sentence pairs from the Lang-8 website were marked up with tags that are supposed to represent the changes that have been made by English speakers providing feedback. These tags are applied by users and are not methodically implemented to indicate the inserted, deleted, or edited text. On further investigation we found that the tags did not accurately indicate the changes and therefore could not be used for the purpose of our intended research. To overcome this problem we processed the sentence pairs using an alignment algorithm to extract the actual edits provided in the feedback by the English speaker. The results of this process were then used to retag the edits accurately. This data in conjunction with the words of the sentence pairs was analysed for machine learning. The purpose of this paper is to evaluate the prediction performance of using an SVM classifier to detect errors in English foreign language writing.

2 Related work

2.1 Errors in language learner writings

Foreign language writing experiments are often conducted in controlled environments as outside influences can have an impact on the performance of works produced by learners. Most of the previous research in this field has aimed to control these factors by

undertaking experiments in academic settings. This has enabled researchers to control the subject of the works produced by the learners, and other factors such as time limits and environment.

Kroll (1990) investigated the difference between writings that were produced in the highly controlled environment of a classroom and those that were produced at home where time was not limited and the learner could have more time to think about their composition. Kroll hypothesised that students might be able to produce better writing in an environment in which they have less pressure and more time to think about the task at hand. An experiment was conducted and the essays of foreign language writings were graded by the frequencies of errors categories that occurred. These frequencies were then compared and it was found that there was not a statistically significant difference between the writings produced in the different environments.

Polio and Fleck (1998) examined whether additional revisions of essays influenced the linguistic accuracy of the content as it is theoretically interesting to researchers in the areas of second language acquisition and second language writing pedagogy. Polio and Fleck (1998) built on the error categories used in previous research reported by Kroll (1990). However Polio and Fleck (1998) concluded that the practical implications in the context of writing assessment might be too small.

Weltig (2004) investigated the influence of writing error categories on the scores of essays by foreign language writing learners. This research built on the error categories that were used in the two previously introduced works by Kroll (1990) and Polio and Fleck (1998). A combination of their defined error categories was used. Weltig (2004) introduced additional error categories as it was thought they could have an influence on the scoring of writings, such as spelling errors and punctuation errors. The results of the investigation revealed that certain errors do have a greater influence on the overall score attributed to the writings. The error categories defined in these works were used as a basis for the error categories in this paper.

2.2 Writing error detection using data mining

Various methods can be used for the automated detection of writing errors, as machine/statistical learning algorithms. Spell checkers commonly available in word processors have been used in the previous research combined with other techniques for writing error detection. Koppel et al. (2005) used the MS Word spell checker with a sentence tagger and an n-gram corpus to detect errors. The native language of ESL (English as a second language) learners was determined by stylistic text feature (function word selection, errors and syntax) analysis of their writings. The use of n-grams for the classification of texts has featured numerous times in the previous research for both the general classification of texts and also detection of errors. Schwarm and and Ostendorf (2005) and Petersen and Ostendorf (2009) used n-grams combined with support vector machine classifiers to find appropriate reading material for students according to their reading level. Brockett et al. (2006) approached the problem by using techniques that are usually synonymous with phrasal statistical machine translation. They used a parallel corpus of texts that were made up of ESL learner writings with both pre- and post-editing correction similar to that found on Lang-8. Bailey and Meurers (2008) examined the use of machine learning methods to augment feedback from computer-aided language learning systems by using the shallow matching features to detect meaning errors. They focused on the analysis of short answers to reading comprehension problems. They

achieved an accuracy of almost 90% for learner response content error detection on a learner corpus collected from real-life ESL learners completing assigned exercises.

Hirano et al. (2007) used the frequency of results from a web search engine to check if a sentence from a technical paper contains an article error. It was stipulated that as the language used in technical papers is more complex than simple phrases, it is difficult to use a search engine to determine if there is an error or not as the number of search results is often too small to have any significance. It was proposed that using queries built based on the results of POS (parts-of-speech) tagging would better serve as a determiner if the sentence contains an error. Tanimoto and Ohta (2012) examined using the number of search results as an 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.

Some previous researches (Han et al., 2004, 2006; Chodorow et al., 2007) have used maximum entropy classifiers to detect article errors (incorrect use of: A, an, the…). Parts of speech tags and local context words were used to determine the probability of noun phrases. This technique was found to be superior than past techniques, however it was noted that the classifier lacked the ability to determine the context of previously mentioned entities. Tetreault and Chodorow (2008) also used a maximum entropy classifier augmented with combination features and a series of thresholds to detect preposition errors (incorrect use of a word expressing the relation between a noun/pronoun and another word or element in the phrase). It was found that the system could detect up to 84% of preposition errors. A disadvantage of using this approach is that it cannot automatically model the interactions among features.

Gamon et al. (2008) used decision trees to perform error detection and correction for prepositions and definite/indefinite determiners on a reduced feature set using an n-gram corpus. Overall evaluation of the system was positive in providing error detection and also suggesting a correction. It was noted that the biggest challenge was solving false positives as it can confuse non-native speakers. Tsur and Rappoport (2007) applied machine learning techniques to study the effect of language transfer, which is a major topic in second language acquisition (SLA). Language transfer studies the effect that a learner's native language has on foreign language study. They hypothesised that language transfer affects the level of basic sounds and short sound sequences, manifested by the words that people choose when writing in a second language. Thus, foreign language words are strongly influenced by native language sounds and sound patterns. They applied SVM machine learning to train a classifier using the International Corpus of Learner English (ICLE) in an effort to realise the hypothesis.

Others have focused on the classification of questions and evaluation of the quality of English in formal scientific papers. Suzuki et al. (2003) classified question sentences by n-gram and SVM analysis. The characteristic features of question types were identified by n-gram word attributes. SVM learning was applied to the data to create a question classifier model. It was found to be superior to conventional methods when tested using 10,000 sample questions. Zhang and Lee (2003) looked at what types of machine learning are effective for classifying questions. They analysed the TREC English corpus in the form of words, n-grams and sentence trees as training data for machine learning. When only using surface text features for sentence classification, it was found that the prediction performance of SVM is superior to four other machine learning algorithms: Nearest Neighbors, Naive Bayes, Decision Tree, and Sparse Network of Winnows.

Kobayashi et al. (2012) investigated analysis by random forests and the frequencies of words and parts of speech tagged n-grams as features to determine the quality of formal English scientific papers. Using this method they were able to attain an accuracy of 77.75% when classifying a corpus as either poor or good papers.

3 Vectorisation of error sentences for categorisation

In order to evaluate the classification of errors in English sentences, the following process was undertaken to construct basic data. Firstly, 500 corrected sentences written in English were chosen at random from diaries written by language learners on the Lang-8 website. However, in some cases large portions of the sample sentences were rewritten or contained comments that would reduce the effectiveness of machine learning and were removed, leaving 399 candidate sentences.

Analysis was performed not on just the sentences, but on pairs of sentences: the original sentence that contains errors, and the corrected sentence that contains tagged edited words. These sentence pairs are a result of mutual corrections that have occurred on the Lang-8 website. In this paper, the GETA search engine (http://geta.ex.nii.ac.jp/geta.html) was used to index the original and corrected sentence pairs. Word is usually stemmed when building an index, however it was decided that the indexed words should not be stemmed as analysis was performed at the word level.

In Lang-8, the edits made by English speakers on the sentences are marked up using span tags, such as . The class attribute of these span tags changes depending on action of the English speaker. If a word is removed then the sline class is applied. Classes that describe the font colour and weight are also used, such as f_bold, f_red, and f_blue. However the intention with which these classes are assigned is unregulated and not uniformly applied across the all sentences. In this paper, it was decided that because of the inconsistency of tag use that better results would be achieved by using an alignment algorithm to programmatically detect and tag changes in sentence pairs. Table 1 shows an example untagged sentence.

Table 1 An example of an original and corrected sentence pair

Original sentence	I woke up alone, with lose memory, lying on the white beach, not knowing where I was.
Corrected sentence	I woke up alone, with no memory, lying on a white beach, not knowing where I was.

As seen in this example, “lose” and “the” are corrected with “no” and “a”. These corrections are identified using the alignment algorithm and the results are tagged as: delete:lose, delete:the, insert:no, and insert:a. In the search engine that was used in this paper the corrections are expressed as d:lose, d:the, i:no, and i:a along with the other words in the sentence. The corrections were also added without distinguishing whether the edit is an insertion or deletion, and were indexed as: e:lose, e:the, e:no, and e:a.

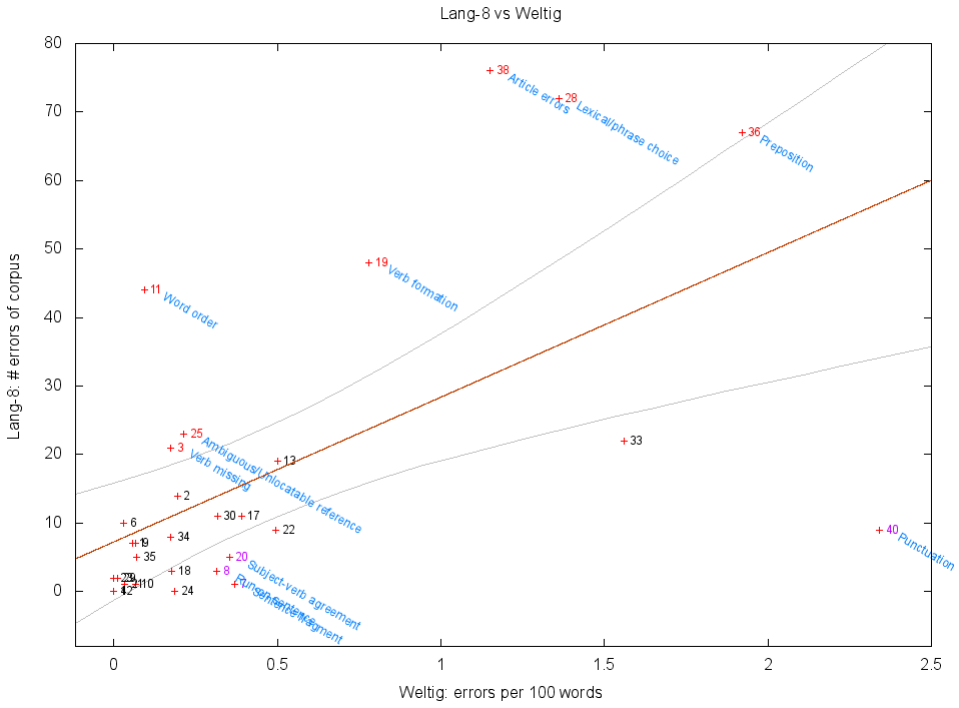
These sentences were classified into 42 error categories by the first author of this paper whose language is English. It was determined that the above example contains

errors of two categories: Error number 38, which is an article error, and error number 41, which is a negation error. These errors are indexed in the search engine as c:38 and c:41 respectively. The three indexes for error category, edited words and non-edited word are then vectorised. Using this it is then possible to determine if a sentence has an article error by examining if it contains "i:a, d:the, e:a, and e:the". It also makes it possible to determine if the sentence contains a negation error by checking if it contains "i:no, and e:no". Simple classification would analyse just the words of the sentence. However we analyse the information about the corrections along with the words of the sentence to determine the error categories with the sentence.

Table 2 Indexed example sentence

c:38/ c:41/
d:lose/ d:the/ i:no/ i:a/
e:the/ e:lose/ e:a/ e:no/
the/ a/ woke/ no/ not/ on/ white/ memory/ with/ lying/
beach/ up/ i/ knowing/ where/ alone/ was/ lose/

Figure 1 Error correlation of Lang-8 vs Weltig (see online version for colours)



A special use search engine was built using indexes as shown in Table 2. The information about the error categories, c:38, c:41, was not used in the classification of error categories.

4 Error categories of English compositions

A subset of 500 pairs of sentences was selected for error pattern categorisation. After removing invalid pairs, 399 pairs of sentences were manually categorised into 42 error types that were defined based on previous research by Kroll (1990) and Weltig (2004). As both utilise a different set of error number lists for their analysis, a merged error number list was created.

Linear regression analysis was used to establish whether a correlation exists between the frequency of errors in the common categories of previous studies (Kroll, 1990; Weltig, 2004) and that of the Lang-8 error analysis. As shown in Figure 1 and Table 3, the results of the analysis show that there is a significant correlation, with a critical alpha level of $p < 0.05$, and $t = 4.3509$, 4.4179, and 3.8011 for Kroll Class, Kroll Home, and Weltig, respectively.

Table 3 Linear regression analysis results

	Kroll (Class)	*Kroll (Home)*	*Weltig*
r^2	0.6351	0.6409	0.5834
t	4.3509	4.4179	3.8011
p	0.0002	0.0001	0.0007
y	2.9376 + 4.2918x	4.9722 + 3.6384x	7.2613 + 21.1171x

The feedback provided by English speakers often contained several different types of error pattern corrections within a single response. Taking this into consideration, the sentences that contain more than one error type were categorised as having multiple errors accordingly. Some feedback contained comments about the correction and/or multiple suggestions for a single word or phrase. A majority was to do with lexical or phrase choices and categorised as lexical or phrase choice errors accordingly.

Table 4 Outlier error categories and relation to Lang-8 error frequency

More freq. in Lang-8		*Less freq. in Lang-8*	
#	Error Cat.	#	Error Cat.
3	verb missing	7	sentence fragment
11	word order	8	run-on sentence
19	verb formation	20	subject-verb agreement
25	ambiguous/unlocatable reference	40	punctuation
28	lexical/phrase choice		
36	preposition		
38	article errors		

These correlations were then used to identify possible outlier errors, not residing within the 95% confidence interval. A total of 22 different error categories were found outside the 95% confidence interval, with 11 of these errors being common across all three regressions analyses. These common outlier errors suggest a characteristic difference in the frequency of errors on Lang-8 compared to those from an academic setting, such as Kroll (1990) and Weltig (2004). This may be a result of the differences in influencing

factors, such as motivation, the subject of the writing, and personal factors (age, socioeconomic background, etc.).

As seen in Table 4, seven error categories occur more frequently on Lang-8 when compared to results from Kroll and Weltig. Of these, the error categories 'word order', 'verb formation', 'preposition' and 'article errors' are considerably outside the 95% coincidence interval and therefore occur more frequently in the writings on Lang-8 when compare to the previous research results. This therefore could be seen as a characteristic of the types of errors occurring in writings on Lang-8.

5 Evaluation of error categorisation using SVM

An evaluation of error categorisation using SVM to classify the errors of 399 sentences with all the data as training data is shown below in Table 5. It should be noted that the columns in this table are sorted by F-measure in descending order. The prediction performance of the classification of errors 36 (preposition), 42 (spelling), 2 (subject formation) and 28 (lexical/phrase choice) is more than 90%. However, as this evaluation analyses all the data as training data it cannot be used as a general evaluation of the prediction performance.

Table 5 Evaluation of the classification of error categories

Error category	*Precision*	*Recall*	*F*	*Accuracy*
36	0.9310	0.9643	0.9474	0.9850
42	0.9773	0.8958	0.9348	0.9850
2	1.0000	0.8571	0.9231	0.9950
28	0.8696	0.9677	0.9160	0.9724
38	0.2698	1.0000	0.4250	0.5388
19	0.1845	1.0000	0.3116	0.5238
11	0.1201	1.0000	0.2145	0.3208
33	0.0955	1.0000	0.1743	0.5013
25	0.0806	1.0000	0.1493	0.4286
3	0.0599	1.0000	0.1131	0.2531
17	0.0521	1.0000	0.0990	0.5439
13	0.0492	1.0000	0.0939	0.3709
6	0.0488	1.0000	0.0930	0.5113
37	0.0478	1.0000	0.0913	0.5013
30	0.0461	1.0000	0.0881	0.4812

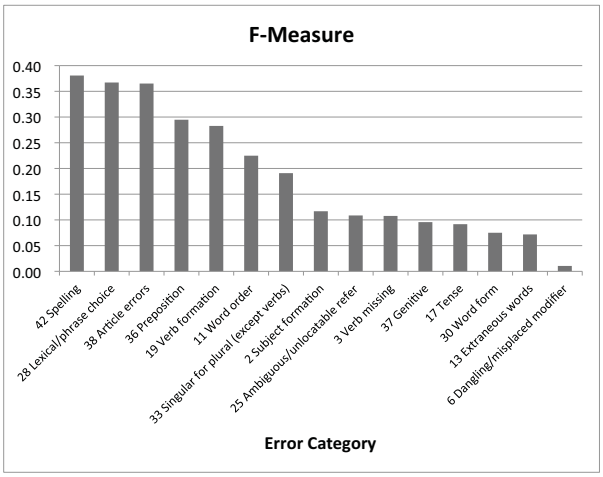
We then used ten-fold cross-validation to evaluate the prediction performance of the classifier. All 399 sentences were then randomly divided into 10 even groups. In each group 90% of the data was used for SVM training to generate a model. The prediction performance of the classifier was then tested using the remaining 10% of the data from the same group. The average of ten test results for each error category is used as a measure of the prediction performance of each classifier respectively. These results are displayed in Table 6, Figure 2, and Figure 3.

Table 6 Evaluation of the classification of errors by ten-fold cross-validation

	Error type	*Number of samples*	*Precession*	*Recall*	*F*	*Accuracy*
42	spelling	48	0.4153	0.3906	0.3807	0.7780
28	lexical/phrase choice	62	0.3109	0.5206	0.3672	0.7218
38	article errors	68	0.2265	0.9857	0.3652	0.4023
36	preposition	56	0.2049	0.5742	0.2948	0.6288
19	verb formation	43	0.1865	0.6881	0.2828	0.6547
11	word order	37	0.1472	0.6514	0.2248	0.5999
33	singular for plural	21	0.1129	0.8000	0.1910	0.5796
2	subject formation	14	0.0758	0.3333	0.1169	0.5217
25	ambiguous/unlocatable refer	20	0.0687	0.2833	0.1087	0.4843
3	verb missing	19	0.0585	0.8250	0.1077	0.2647
37	genitive	10	0.0539	0.4667	0.0957	0.4941
17	tense	10	0.0588	0.4167	0.0917	0.3633
30	word form	10	0.0418	0.3833	0.0750	0.4491
13	extraneous words	12	0.0385	0.6500	0.0718	0.4516
6	dangling/misplaced modifier	10	0.0063	0.0333	0.0105	0.5078

Table 6 shows the overall results of the tests along with the number of sentence samples for each error category. The table is sorted by the F-measure of each of the models in descending order.

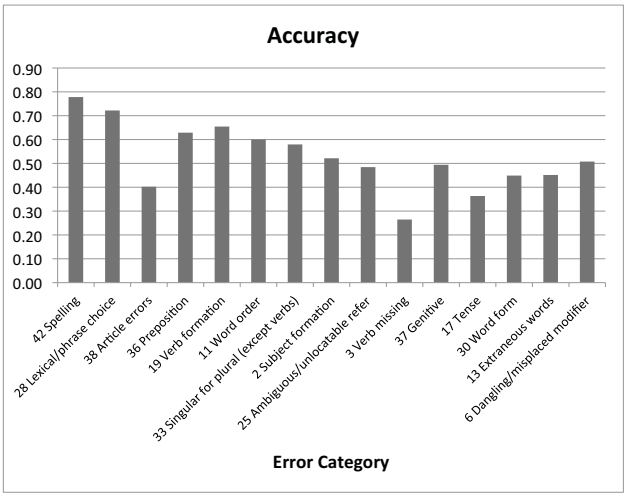
Figure 2 Error classification evaluation for each category (see online version for colours)



Note: F-measure, ten-fold cross-validation

The F-measure performance of each model is displayed in Figure 2. As you see, the F-measure of all the models is less than 40%, with error category 42 (spelling), 28 (lexical/phrase choice) and 38 (article errors) being the more effective models with an F-measure of only 38.07%, 36.72%, and 36.52%, respectively. On the lower end of the scale the model for error category 6 (dangling/misplaced modifier) has an F-measure of 1.05%.

Figure 3 Error classification evaluation for each category (see online version for colours)



Note: Ten-fold cross-validation, accuracy

The accuracy of the generated models also varies for each error category. As shown in Figure 3, the model for error category 42 (spelling) has the greatest in all the models with an accuracy of 77.80%. Error category 3 (verb missing) has the lowest accuracy in all the models at 26.47%.

Overall, the prediction performance of the classifier as seen above cannot be considered effective enough for practical use. Figures 4 and 5 are plots of correlations between the number of samples, F-measure, and accuracy for each of the error category models. A positive correlation can be seen in both plots, indicating that as the number of samples increases so does the F-measure and accuracy of the evaluation. This suggests that if the samples for each error category were increased to an adequate number then the prediction performance of the classifier would also increase accordingly.

Looking at the results in Figure 4, the error category models that were trained using a small number of samples generally have a smaller F-measure than those with a greater number of samples. Therefore one can expect if 100 manually categorised samples were used to train each error category it would result in an F-measure of around 80%.

Figure 4 Correlation between the number of data samples and the f-measure of the evaluation (see online version for colours)

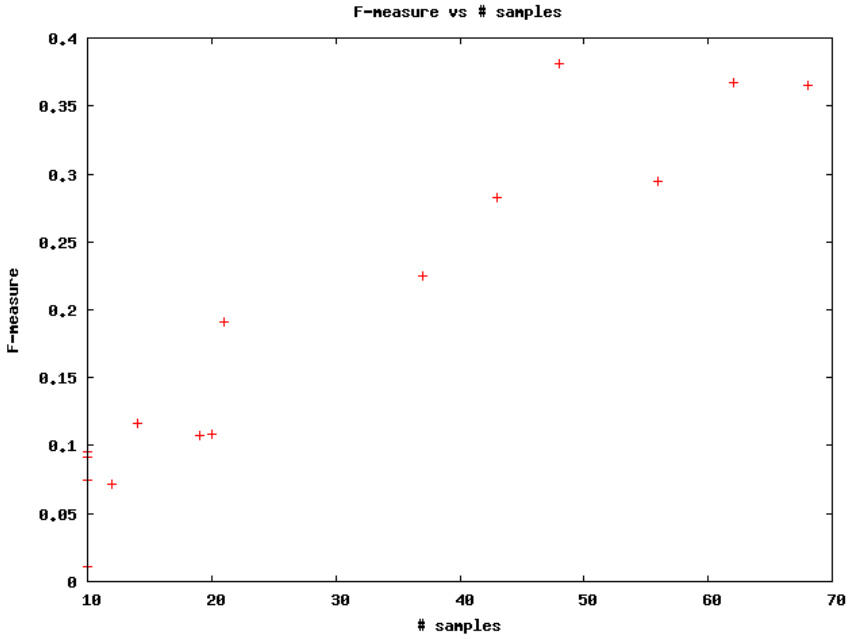
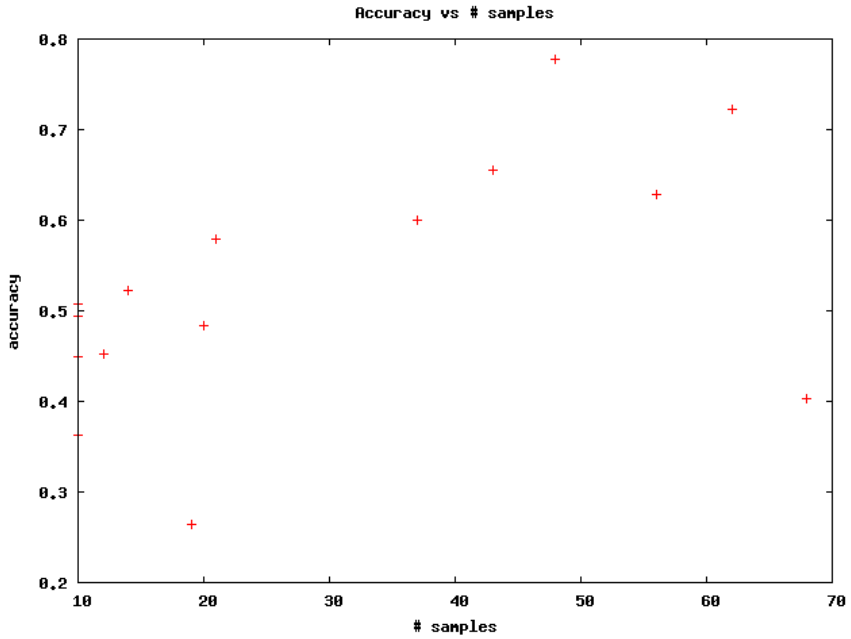


Figure 5 Correlation between the number of data samples and the accuracy of the evaluation (see online version for colours)



A similar correlation can also be seen in Figure 5 with the accuracy of models increasing along with the number of samples.

6 Detailed analysis

A score for each word or tag can be extracted from the model created by applying SVM to the training data. As shown in Table 7, error category 38 (article) has the features that consist of tags, such as "e:the, i:the, e:a, and i:a". Error category 36 (preposition) has the following tags as the features of the error "i:in, e:in, d:at, e:for, e:at, e:on, and i:on". The ability to extract such information from the model enables the confirmation of the features associated with the error types in the corrections. The feature "ing" can be expected for error category 19 (verb formation). The error features associated with error category 42 (spelling) are "e", "e:e", and "i:e" can be seen as common spelling errors in words such as conv-a-rsation, and ev[e]ryone.

Table 7 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

7 Conclusions

In this paper, we manually classified the errors contained in sample sentences from diaries written in the mutual correction language-learning site Lang-8. The errors were classified into categories based on previous research (Kroll, 1990; Weltig, 2004). The sample sentence pairs had tags indicating the edits in the corrections, however it was determined that these did not always correctly reflect the true corrections, and were removed. An alignment algorithm was then used to programmatically identify the corrections that had been made, and the edited words were then tagged as 'inserted' or 'deleted' accordingly. These tags, along with the manually classified error categories and the other words in the original sentence, were then indexed to build a special use search engine. This search engine index was then used as training data for SVM machine learning to create a model for error category classification.

This model was then evaluated using ten-fold cross-validation. 399 sentences used as sample data were divided randomly into ten even groups, with 90% of the sample data used for training and the remaining 10% used for model verification. The F-measure for each error category was less than 40%. However, the results did show a significant positive correlation between the number of data samples, F-measure and accuracy of the model. Thus it can be expected that if the number of samples is increased to 100 manually identified samples, then it is expected that the model will produce an F-measure of roughly 80%. Therefore by increasing the training data it is expected to produce a

reasonable level of performance for error category classification. As manual classification of error takes a significant amount of time and labour, the current model will be used to classify error categories that will then be checked manually to verify the error category. This is expected to accelerate the process of generating training data samples that then can be used to further improve the model.

In the future we plan to increase the amount of manually classified training data to investigate if an efficient SVM classification model can be attained for determining languages learner's error characteristics.

References

Bailey, S. and Meurers, D. (2008) 'Diagnosing meaning errors in short answers to reading comprehension questions.' *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pp.107–115.

Brockett, C., Dolan, W.B. and Gamon, M. (2006) 'Correcting ESL errors using phrasal SMT techniques', *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp.249–256.

Chodorow, M., Tetreault, J.R. and Na-Rae, H. (2007) 'Detection of grammatical errors involving prepositions', *SigSem '07 Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pp.25–30.

Gamon, M., Gao, J., Brockett, C. Klementiev, A., Dolan, W.B., Belenko, D. and Vanderwende, L. (2008) 'Using contextual speller techniques and language modeling for ESL error correction', *Proceedings of the Third International Joint Conference on Natural Language Processing*, Vol. 1, pp.449–456.

Han, N.R., Chodorow, M. and Leacock, C. (2004) 'Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus', *LREC 2004 Fourth International Conference on Language Resources and Evaluation*, pp.1625–1628.

Han, N.R., Chodorow, M. and Leacock, C. (2006) 'Detecting errors in English article usage by non-native speakers', *Natural Language Engineering*, Vol. 12, No. 2, pp.115–129.

Hirano, T., Hirate, Y. and Yamana, H. (2007) 'Detecting article errors in english using search engines', *DBSJ Letters*, in Japanese, Vol. 6, No. 3, pp.13–16.

Kobayashi, Y., Tanaka, S. and Tomiura, Y. (2012) 'Pattern recognition of english scientific papers using N-grams' *Information Fundamentals and Access Technologies (IFAT)*, Vol. 12, No. 1, pp.1–6.

Koppel, M., Schler, J. and Zigdon, K. (2005) 'Determining an author's native language by mining a text for errors', *KDD '05 Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp.624–628.

Kroll, B. (1990) '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.

Petersen, S.E. and Ostendorf, M. (2009) 'A machine learning approach to reading level assessment', *Computer Speech and Language*, Vol. 23, No. 1, p.89–106.

Polio, C. and Fleck, C. (1998) '"If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions', *Journal of Second Language Writing*, Vol. 7, No. 1, pp.43–68.

Schwarm, S.E. and Ostendorf, M. (2005) 'Reading level assessment using support vector machines and statistical language models', *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, pp.523–530.

Suzuki, J., Sasaki, Y. and Maeda, E. (2003) 'Question type classification using word attribute n-gram and statistical machine learning', *Transactions of Information Processing Society of Japan*, (in Japanese), Vol. 44, No. 11, pp.2839–2853

Tanimoto, T. and Ohta, M. (2012) 'Examination of English error detection using the number of search results' *DEIM Forum 2012*, (in Japanese), Vol. 9, No. 1.

Tetreault, J.R. and Chodorow, M. (2008) 'The ups and downs of preposition error detection in ESL writing', *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1, pp.865–872.

Tsur, O. and Rappoport, A. (2007) 'Using classifier features for studying the effect of native language on the choice of written second language words', *CACLA '07 Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pp.9–16.

Weltig, M.S. (2004) '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 Volume 2 2004*, Vol. 2, pp.53–82.

Yin, C., Hirokawa, S., Flanagan, B., Suzuki, T. and Tabata, Y. (2012) 'Mistake discovery and generation of exercises automaticity in context', *2012 IIAI International Conference on Advanced Applied Informatics (IIAIAAI)*, pp.163–167.

Zhang, D. and Lee, W.S. (2003) 'Question classification using support vector machines', *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.26–32.