

Intelligent computer classification of english writing errors

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

<https://hdl.handle.net/2324/1444096>

出版情報 : Frontiers in Artificial Intelligence and Applications. 254, pp.174-183, 2013-12-01

バージョン :

権利関係 :

Intellegent Computer Classification of English Writing Errors

Brendan FLANAGAN ^{a,1} and Chengjiu YIN ^b and Takahiko SUZUKI ^b
and Sachio HIROKAWA ^b

^a*Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan*

^b*Research Institute for Information Technology, Kyushu University, Fukuoka, Japan*

Abstract. An important issue in education systems is the ability to determine the characteristics of learners and then provide intelligent and informed guidance in response. The authors of this paper have a long-term research goal to provide language learners with the ability to determine and improving their weaknesses. However, to achieve this goal a sizable amount of manually classified data is required. The task is both time consuming and labor intensive. In this paper a system was built to help intelligently classify the errors in an English learner's writings into categories (Kroll 1990, Weltig 2004). Using a randomly selected manually classified sample as training data, it was determined that there is a positive correlation between the number of samples for each error category and the effectiveness of the model created by applying SVM machine learning to the writings of language learners on the Lang-8 website. It is intended that the classification results will be used to accelerate the manually process classification and increase the amount of training data available for use.

Keywords. Error classification, SVM, machine learning, writing errors.

Introduction

In recent years, language learning on the Web outside the traditional classroom setting has been increasing in popularity. In particular, sites that serve as a language exchange, bring together native speakers from different language backgrounds are prevalent. For example, person A is a native Japanese speaker who is learning English as a foreign language. A posts an English sentence on the website, and then the sentence is corrected by person B who is a native English speaker. B is also learning Japanese as a foreign language and posts a sentence on the website in Japanese. This is then corrected by person A who is native Japanese speaker. This mutually beneficial environment helps learners to achieve their respective goals of learning a foreign language, which in turn is another foreign language learner's mother tongue. The data used in this paper is from Lang-8 (<http://www.lang-8.com>), which is a leading mutually beneficial foreign language learning writing correction website.

Mutual correction websites contain a large amount of foreign language writing correction data. Taking advantage of this data can help to further enhance the

¹ Corresponding Author: Graduate School of Information Science and Electrical Engineering, Kyushu University, Hakozaki 6-10-1 8285 Fukuoka, Japan, E-mail: bflanagan.kyudai@gmail.com.

effectiveness of language learning. Using data from Lang-8, the authors of this paper have in past research categorized the errors in sentences manually by hand and built a quiz system [8].

By determining the particular weakness of learners (student-specific error patterns), and then repeatedly having the learner practice quizzes that focus on these weaknesses, it has been shown to increase the effectiveness of learning. However, the manually categorization of error patterns requires significant time and effort.

In recent years there have been remarkable advances in machine learning research. In particular, Support Vector Machine (SVM) has become a standard technique for efficient classification of fixed dimensional vectors. SVM's high performance classification techniques have been used in many fields [5]. Therefore, error classification was undertaken using SVM in this paper.

A long-term goal of the author's research is to extracting the error characteristics of learners. In this paper, 500 corrected sentence pairs written in English were randomly selected from diaries written by language learners on the Lang-8 website. These were then manually classified into error categories that were derived from previous research that was conducted by Kroll [6] and Weltig [7]. The first author of this paper, who is a native English speaker, manual classified the sample sentences into error categories.

On Lang-8 the original and corrected sentences are available in pairs, and the corrected sentence is marked up with tags to show where a native speaker has inserted, deleted and edited text. However after investigation it was found that these tags do not always reflect the actual changes that have been made. In light of this, the sample sentences pairs were processed using an alignment algorithm to extract the actual edits that had been made by the native speaker. The edited words were then tagged as insert or delete, and used along with the other words in the sentences for the classification of the error category contained within the sentence. Using this as training data for machine learning, SVM machine learning was used to classify and evaluate of the performance the classification.

1. Related Work

Previous empirical studies on the writings of foreign language students have predominantly been undertaken in academic settings. This has enabled the control of influencing factors, such as: subject matter, conditions, and environment in which the writing was conducted. Kroll [6] compared the difference of writings that were conducted in classroom where learners had a fixed amount of time, and the home environment, where it was postulated that students would have more time and less pressure to write. English teachers categorized errors manually and the frequency of occurrence was used to compare the writings in the two different environments. Weltig [7] 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.

Previous studies have estimated errors in English text by using SVM and other types of machine learning algorithms. Hirano et al. [2] 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 [4] 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.

Others have focused on the classification of questions and evaluation of the quality of English in formal scientific papers. Suzuki et al [1] examined using n-grams and SVM in the classification of question sentences. They proposed a method for finding n-gram word attributes for identifying effective characteristic features of question types for classification. These features were then used as training data for SVM machine learning to create a question classifier model. It was found to be superior compared to conventional methods when tested using 10,000 sample questions. Zhang et al. [9] looked at what types of machine learning are effective for classifying questions. They used the TREC English corpus in the form of words, n-grams and sentence trees as training data for machine learning. They determined that by only using surface text features that SVM was superior in classifying sentences when compare to four other machine learning algorithms: Nearest Neighbors, Naive Bayes, Decision Tree, and Sparse Network of Winnows. Kobayashi et al. [3] used 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 of good papers.

2. Vectorization of Error Sentences for Categorization

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 as such was 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. 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. As the analysis was performed at the word level, it was decided that the indexed words should not be stemmed, which is a practice usually used when building an index.

In Lang-8, the edits made by native 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 native speaker. If a word is removed then the `sline` class is applied. Classes that describe the font color 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 as such not uniformly applied across 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. An example untagged sentence is shown below in Table 1.

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 sentence, "lose" and "the" are corrected with "no" and "a". These corrections are identified using the alignment algorithm and as a 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 as such 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 native 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. These three indexes: error category, edited words and non-edited word, are then vectorized. 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. It is thought that by using the information contained in the corrections and not just the general words of the sentence that it is possible to determine the category of errors contained within a 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/

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

3. Error Categories of English Compositions

A subset of 500 pairs of sentences was selected for error pattern categorization. After removing invalid pairs, 399 pairs of sentences were manually categorized into 42 error types that were defined based on previous research by Kroll [5] and Weltig [12]. As

both utilize a different set of error number lists for their analysis, a merged error number list was created.

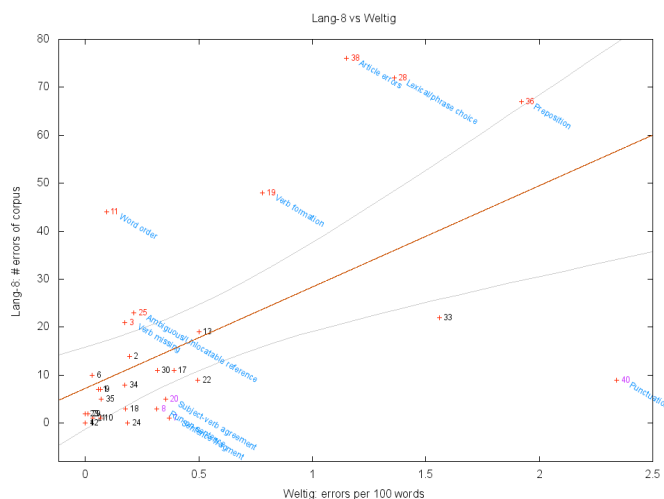


Figure 3. Error correlation of Lang-8 vs Weltig.

Linear regression analysis was used to establish whether a correlation exists between the frequency of errors in the common categories of previous studies (Kroll [5] and Weltig [12]) and that of the Lang-8 error analysis. 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 native 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 categorized as having multiple error patterns accordingly. Some feedback contained comments about the correction and/or multiple suggestions for a single word or phrase that were mostly to do with lexical or phrase choices and categorized accordingly.

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 errors frequency of writings and corrections on Lang-8 when compared to those from an academic setting, such as: Kroll and Weltig. 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).

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		

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 previous research results. This therefore could be seen as a characteristic of the types of errors that occurring in writings on Lang-8.

4. Evaluation of Error Categorization using SVM

An evaluation of using SVM to classify the errors of 399 sentences into categories when using all the data as training data is shown below in Table 5. It should be noted that the columns of this table have been sorted by F-measure in descending order. The effectiveness of classification of errors 36 (Preposition), 42 (Spelling), 2 (Subject formation) and 28 (Lexical/phrase choice) is more than 90%. However, as this evaluation uses all the data as training data it can't be used as general evaluation of the effectiveness.

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

Table 6, Figure 2, and Figure 3 display the results of a 10-fold cross-validation that were determined by taking the average of 10 tests. These tests were conducted on all 399 sentences randomly divided into 10 even groups, of which 90% was used as training data and the remaining 10% as test data. The number of sentences for each error category are displayed along with the evaluation results in Table 6. The F-measure of all the errors is less than 40%.

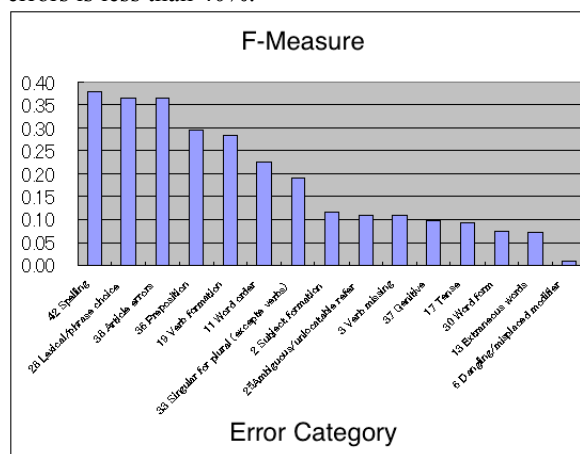


Figure 2. Error classification evaluation for each category (F-measure, 10-fold cross-validation).

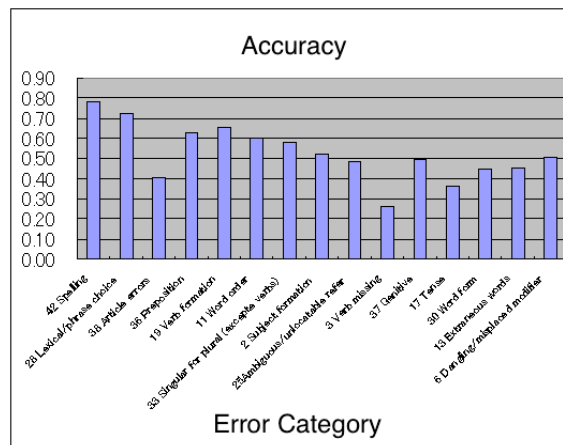


Figure 3. Error classification evaluation for each category (10-fold cross-validation, Accuracy)

Figures 4 and 5 are plots of the correlation between the number of samples, F-measure, and Accuracy for each of the error categories. 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 which intern implies that the effectiveness of classification increases. Looking at the results in Figure 4, one can expect a F-measure of 80% if there are 100 manually categories samples for each error.

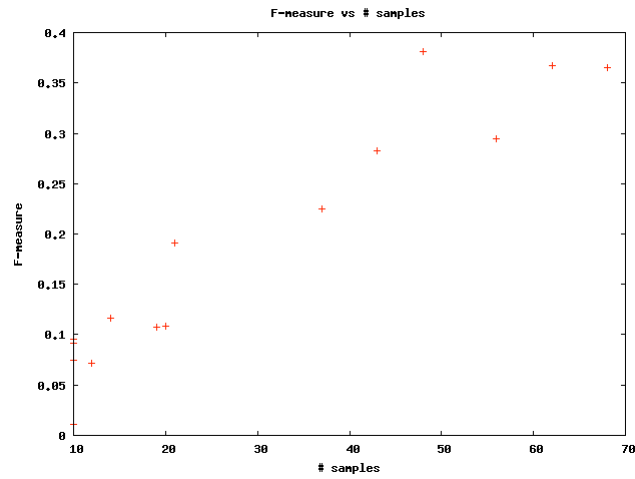


Figure 4. Correlation between the number of data samples and the F-measure of the evaluation.

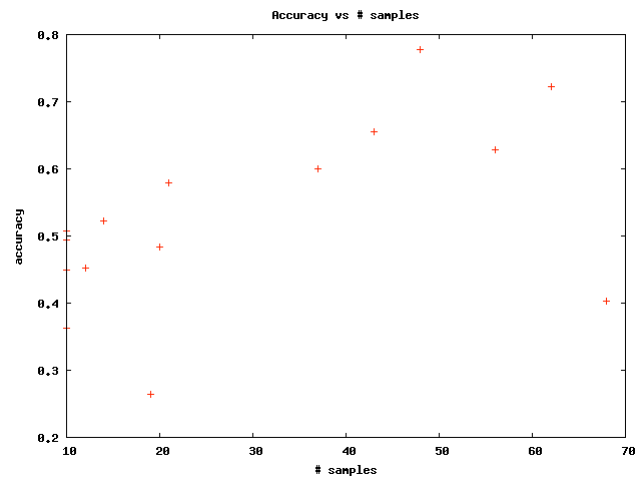


Figure 5. Correlation between the number of data samples and the Accuracy of the evaluation.

Table 10. Evaluation of the classification of errors into categories by 10-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

5. Detailed Analysis

A score for each word or tag can be extracted from the model created by applying SVM to the training data. 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

6. Conclusion

In this paper, the first author who is a native English speaker manually classified the errors contained in sample sentences from diaries written on the mutual correction language-learning site, Lang-8. The errors were classified into categories based on previous research (Kroll [6], Weltig [7]). The sample sentence pairs collected from the site contained tags that marked up the edits made in the corrections, however it was determined that these did not always correctly reflect the true corrections, and as such 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 that for error category classification.

This model was then evaluated using 10-fold cross-validation. 399 sentences used as sample data were divided randomly into 10 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 labor, 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

- [1] J. Suzuki, Y. Sasaki, E. Maeda, Question Type Classification Using Word Attribute N-gram and Statistical Machine Learning, *Transactions of Information Processing Society of Japan*, **44.11** (2003), 2839-2853. (in Japanese)
- [2] T. Hirano, Y. Hirate, H. Yamana, Detecting Article Errors in English using Search Engines, *DBSJ Letters* **6.3** (2007), 13-16. (in Japanese)
- [3] Y. Kobayashi, S. Tanaka, Y. Tomiura, Pattern Recognition of English Scientific Papers Using N-Grams, *Information Fundamentals and Access Technologies (IFAT)*, **12.1** (2012).
- [4] T. Tanimoto, M. Ohta, Examination of English Error Detection Using the Number of Search Results, *DEIM Forum 2012*, **9.1** (2012). (in Japanese)
- [5] A. Karatzoglou, D. Meyer, K. Hornik, Support vector machines in R, *Journal of Statistical Software*, **15.9** (2006), 1-28. (in Japanese)
- [6] 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: Cambridge University Press (1990). 140-154.
- [7] M. S. Weltig, Effects of language errors and importance attributed to language on language and rhetorical-level essay scoring, *Spain Fellow Working Papers in Second or Foreign Language Assessment Volume 2 2004*, **1001** (2004), 53.
- [8] C. Yin, S. Hirokawa, B. Flanagan, T. Suzuki, Y. Tabata, Mistake Discovery and Generation of Exercises Automaticity in Context, *Proc. of LTLE2012*, (2012).
- [9] D. Zhang, W. S. Lee, Question classification using support vector machines, In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, (2003), 26-32.