Chinese Grammatical Error Diagnosis Using Ensemble Learning

Yang Xiang, Xiaolong Wang†, Wenying Han, and Qinghua Hong
Intelligent Computing Research Center,
Harbin Institute of Technology Shenzhen Graduate School, China
{xiangyang.hitsz, hanwenying09, hongqh65}@gmail.com,
†wangxl@insun.hit.edu.cn

Abstract

Automatic grammatical error detection for Chinese has been a big challenge for NLP researchers for a long time, mostly due to the flexible and irregular ways in the expressing of this language. Strictly speaking, there is no evidence of a series of formal and strict grammar rules for Chinese, especially for the spoken Chinese, making it hard for foreigners to master this language. The CFL shared task provides a platform for the researchers to develop automatic engines to detect grammatical errors based on a number of manually annotated Chinese spoken sentences. This paper introduces HITSZ's system for this year's Chinese grammatical error diagnosis (CGED) task. Similar to the last year's task, we put our emphasis mostly on the error detection level and error type identification level but did little for the position level. For all our models, we simply use supervised machine learning methods constrained to the given training corpus, with neither any heuristic rules nor any other referenced materials (except for the last years' data). Among the three runs of results we submitted, the one using the ensemble classifier Random Feature Subspace (HITSZ_Run1) gained the best performance, with an optimal F1 of 0.6648 for the detection level and 0.2675 for the identification level.

1 Introduction

Automatic grammatical error detection for Chinese has been a big challenge for NLP researchers for a long time, mostly due to the flexible and irregular ways in the expressing of this language. Different from English which follows grammatical rules strictly (i.e. subject-verb agreement, or strict tenses and modals), the Chinese language has no verb tenses or numbers and endures heavily for the incompleteness of grammatical elements in a sentence (i.e. the zero subject or verb or object). Some examples are shown below in Table 1.

	Examples
1.	四月/最/熱。 April is the hottest.
2.	我/一/看到/你/就/覺得/非常/開心。 I feel very happy as soon as I see you.
3.	他們很高興。 They are very happy

Table 1. Some typical examples for special grammatical usage in Chinese.

In the above table, the first sentence contains no verb elements in the Chinese version. In the Chinese language, the adjectives will not co-occur with copulas in many cases. So if we add a *be* (*是*) into the sentence (四月/是/最/熱), it will be grammatically incorrect. In the second sentence, the conjunction *就* has nothing to do with the meaning of the whole sentence, but it is a necessary grammatical component when collocate with the word *一* to express the meaning of *as soon as*. The adverb *很* is an essential element for the third sentence which corresponds to the word *very* in the English version. However, we can simply remove *very* but cannot remove *很* due to some implicit grammatical rules. Overall, the expression of the Chinese language is flexible and the grammar of Chinese is complicated and sometimes hard to summarize, so that it is very difficult for foreign language learners to learn Chinese as the second language.

The CFL14 and 15 shared tasks provide a platform for learners and researchers to observe various cases of grammatical errors and think deep-

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 99–104,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

er about the intrinsic of these errors. The goal of the shared task is to develop computer-assisted tools to help detect four types of grammatical errors in the written Chinese. The error types include *Missing, Redundant, Disorder* and *Selection*. And in last years shared task, several groups submitted their report, employing different supervised learning methods in which some groups obtained good results in detection and classification (Yu et al., 2014). Similar to the last year's task, we put our emphasis mostly on the error detection level and error type identification level but did little for the position level although this year's task includes the evaluation on this level.

In this paper, we use supervised learning methods to solve the error detection and identification sub tasks. Different from most of previous work, we didn't use any external language materials except for the dataset for the year 2014's shared task. What we adopt include feature extraction, data construction and ensemble learning. We also report some of our observations towards the errors and summarize some conceivable rules, which might be useful for future developers. At last, we analyze the limitation of our work and propose several directions for improvement.

The following of this paper is organized as: Section 2 briefly introduces the literature in this community. Section 3 shows some observations towards the data provided. Section 4 introduces the feature extraction and learning methods we used for the shared task. Section 5 includes experiments and result analysis. And future work and conclusion are arranged at last.

2 Related Work

In the community of grammatical error correction, more work focused on the language of English such as those researches during the CoNLL2013 and 2014 shared tasks (Ng et al., 2013; Ng et al., 2014). A number of English language materials and annotated corpus can be used such that the research on this language went deeper. However, the resource for Chinese is far from enough, and very few previous works are related to Chinese grammatical error correction. Typical ones are the CFL 2014 shared task (Yu et al., 2014) ant the task held in this year. Following, we briefly introduce some previous work related to Chinese grammatical error diagnosis.

Wu et al. proposed two types of language models to detect the error types of word order, omission and redundant, corresponding to three of the types in the shared task. Chang et al. (2012) proposed a probabilistic first-order inductive learning algorithm for error classification and outperformed some basic classifiers. Lee et al. (2014) introduced a sentence level judgment system which integrated several predefined rules and N-gram based statistical features. Cheng et al. (2014) shown several methods including CRF and SVM, together with frequency learning from a large N-gram corpus, to detect and correct word ordering errors.

In the last year's shared task, there are also some novel ideas and results for the error diagnosis. Chang et al. (2014)'s work included manually constructed rules and rules that automatically generated, the latter of which are something like frequent patterns from the training corpus. Zhao et al. (2014)'s employed a parallel corpus from the web, which is a language exchange website called Lang-8, and used this corpus to training a statistical machine translator. Zampieri and Tan (2014) used a journalistic corpus as the reference corpus and took advantage of the frequent N-grams to detect the errors in the data provided by the shared task. NTOU's submission for the shared task was a traditional supervised one, which extracted word N-grams and POS N-grams as features and trained using SVM (Lin et al., 2014). In their work, they also employed a reference corpus as the source of N-gram frequencies.

Our submission was similar to NTOU's work whereas we didn't use any large scale textural corpus as references. Our target was to see to what extent can the supervised learner learn only from the limited resource and what types of classifiers perform better in this task.

3 Data Analysis

We show some of our observations towards the training data in this section. What we observed are some frequent cases among the error types *Missing* and *Redundant*.

For the error type *Missing*, we noticed that errors often occur in some certain cases. For example, the auxiliary word 的 (of/'s) accounts for 11.35% in all the *Missing* sentences (and 7.93% sentences contain 的 in the training data are incorrect). One of the most frequent missing cases is the missing between an adjective (~est for short) and a noun. For instance, 最好(的) 電影院 (the best cinema), 附近(的) 飯店(a near restaurant), and 我(的)日常生活(my daily life). From the English translation we see that there is no 's or of in the phrase such as *the girl's dress* (女孩

的衣服) or *a friend of mine* (我的一個朋友), but in the grammar of Chinese, a 的 is inserted due to the incompleteness of the expressions.

For the error type *Redundant*, the word 了(an auxiliary word related to a perfect tense) accounts for 10.88% in all the *Redundant* sentences (and 21.78% sentences contain 了 are incorrect). The word is redundant when the sentence contains nothing related to a perfect tense. For instance, 我第一次去(了) 英國留學。 (I studied abroad in Britain for the first time.) and 當時他不老(了)。(He wasn't old at that time.). So we can judge whether the word is redundant according to the tense of the sentence.

Words that are grammatical incorrect are almost function words, which behave differently in the grammars for Chinese and English (or other languages). Typical examples are 是(is), 都 (auxiliary), 有 (be), 會(will), 在 (in/at), 要(will), etc. However, we didn't do much towards specific words in our research but only recognize there should be some frequent rules that we can follow. And we will further discuss some proposals later.

4 Supervised Learning

In this work, neither did we use any external corpora except for the dataset for the year 2014's shared task, nor are any language specific heuristic rules or frequent patterns included. We were going to see what kind of features and what type of supervised learners can benefit this problem most. As declared previously, we did little for the position level extraction, so we introduce mostly on feature extraction, model selection and the construction of the training data.

4.1 Feature Extraction

For this task, we tried several kinds of features such words, POS (part-of-speech), as well as syntactic parse trees and dependency trees. Finally, we find that POS Tri-gram features perform stably and generate the best results. Therefore, we define the POS Tri-gram for sentential classification at first.

For each word in a sentence, we extract the following triple as the Tri-gram for this word: <POS-1, POS, POS+1>. And for the beginning and the ending of a sentence, we add two indicators to make up the column vectors. For example, in the sentence這/一天/很/有意思。(This day is very interesting.), the sentence-level POS features are (r, m, zg, l) the features for the word這 (This) are <start, r, m>[1].

In addition, we extract the relative frequency (probability) for each triple based on the CLP 14 and 15 dataset as *P*(<*POS-1, POS, POS+1*>). In the experiment, we noticed that the frequency features are also good indicators to detect candidates for grammatical errors.

To summarize, we extract two types of POS Tri-gram features: the binary Tri-gram and the probabilistic Tri-gram. The binary Tri-gram demands that if the sentence contains this Tri-gram (i.e. <start, r, m>), the corresponding position in the gram vector (the union set of all possible Tri-grams after removing those with very low frequencies) is set to be 1. For probabilistic Tri-gram, the position is set to be the relative frequency (the proportion for the Tri-gram).

4.2 Supervised Learning

After feature extraction, we put the features into several supervised learners. We use a series of single classifiers such as Naïve Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM) and Maximum Entropy (ME), and ensemble learners Adaboost (AB), Random Forest (RF) and Random Feature Subspace (RFS). RF is an ensemble of several DTs, each of which samples training instances with replacement and samples features without replacement. RFS is an ensemble classifier based on feature sampling which takes results trained on different feature subspace as majority voters. The classifiers are from Weka (Hall et al., 2009).

We take those training sentences with annotated errors as positive instances and subsample the correct sentences as negative ones. Through tuning towards the proportion of negative instances, we discovered that the number of negative instances also affected the final results.

5 Experiment and Analysis

In the experiment, we use the training data from this year's and last year's shared tasks. Table 2 lists the number of sentences for each type in the training data. Since the scale of this year's data is really small, we add last year's corpus into the training data and do cross validations in the training steps. Table 2 lists the number of sentences for each error type in these two years' dataset.

Our experiments cover training data construction, feature selection and supervised learning.

[1] The POS tags are generated by LTP (Liu et al., 2011)

Method	Detection Level				Identification Level			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ME	0.4985	0.3235	0.0028	0.0055	0.4975	0.1154	0.00075	0.0015
SVM	0.5144	0.6091	0.0803	0.1418	0.4969	0.4677	0.0453	0.0825
NB	0.5146	0.5562	0.1448	0.2297	0.4771	0.3765	0.0698	0.1177
DT	0.6255	0.6285	0.6140	**0.6211**	0.5249	0.5321	0.4128	0.4649
RFS	0.6284	**0.7479**	0.3873	0.5103	0.6064	**0.7245**	0.3433	0.4658
RF	0.6510	0.7173	0.4985	0.5882	**0.6121**	0.6817	0.4208	0.5203
AB	**0.6654**	0.7177	**0.5453**	0.6197	0.6105	0.6700	**0.4355**	**0.5279**

Table 3. CV results based on POS Tri-gram features

Method	Detection Level				Identification Level			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ME	0.4985	0.3235	0.0028	0.0055	0.4975	0.1154	0.00075	0.0015
SVM	0.5144	0.6091	0.0803	0.1418	0.4969	0.4677	0.0453	0.0825
NB	0.5145	0.5443	0.1783	0.2686	0.4661	0.3532	0.0815	0.1324
DT	0.6306	0.6354	**0.6130**	0.6230	0.5285	0.5375	0.4087	0.4644
RFS	0.6574	0.7338	0.4940	0.5905	0.6200	0.7005	0.4193	0.5246
RF	0.6588	**0.7554**	0.4695	0.5791	**0.6300**	**0.7305**	0.4120	0.5269
AB	**0.6618**	0.6899	0.5878	**0.6347**	0.5951	0.6323	**0.4545**	**0.5289**

Table 4. CV results based on POS Tri-gram and probability features

Error type	No. in 15	No. in 14
Correct	2205	5541
Disorder	306	710
Redundant	430	1803
Missing	620	2201
Selection	849	827

Table 2. Error type distribution for the two years' shared tasks.

We tried several groups of training data, different combinations of features and a variety of classifiers in the training phase.

5.1 Training Data Construction

As mentioned previously, the sentences that contain no grammatical errors behave as the negative instances for training. To avoid imbalance between the positive and negative instances, negative ones were randomly selected to construct the training set. At last, we divided the training data into 8 parts and used 8-fold cross validation (CV) for the classifiers. We found that, when we selected 4000 negative instances, the system achieved the best results.

5.2 Feature Selection

As mentioned in §4.1, we investigate the features POS Tri-gram and POS Tri-gram + POS Tri-gram probability. We report the CV results generated by four single classifiers and three ensemble classifiers in Table 3 and Table 4 for the two set of features, respectively. The results have been optimized through tuning the parameter settings for each classifier.

From the results, we find that the ensemble classifiers generally perform better than the single ones, and AB achieves the best results for detection and identification.

5.3 Final Results

Among the three runs of results we submitted, the first run is the best. We show the results in Table 5 and compare them with the CV results.

Accuracy	Precision	Recall	F1
Detection Level			
0.509	0.5047	0.974	0.6648
Identification Level			
0.173	0.2401	0.302	0.2675

Table 5. The final results

This submission is generated by the ensemble classifier RFS by using POS Tri-gram and probability features. We see that the performance of the identification level greatly falls behind that in the cross validation. One of the possible reasons for this gap, we consider is the setting of instances, which may be quite distinct between the training and the testing data. And another possible reason is the reasonability of the probability features.

5.4 Analysis

Compare the results generated by the two feature sets (Table 1), it can be seen that the second fea-

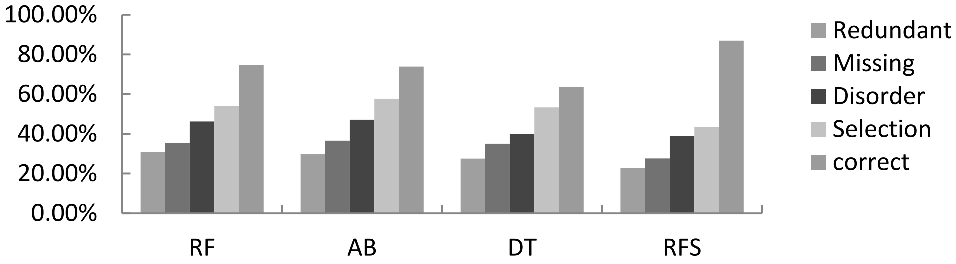

Figure 1. Accuracy of the four error types and the correct type on four classifiers that perform best.

ture set outperforms the first, on both the detection level and the identification level. To some extent, it indicates that the patterns for the grammatical phrases may frequently occur in the datasets.

Further, we pick up the last four classifiers which perform relatively better on the task data, including DT and three ensemble classifiers, and do statistical analysis on the true positive rates during cross validation (Figure 1). The results reveal that the difficulty on judging decreases from Redundant, Missing to Disorder and Selection. In addition, the accuracy for the correct label is not quite high, leading to a number of false negative sentences.

Through observation, we found several cases might affect the predicting results. A typical case is that a grammatically wrong sentence can be corrected through several ways, corresponding to more than one error types. For example, the sentence 他馬上準備上學 (He is preparing for school.) can be classified to any of the four types:

	Correct Sentence	Type
1.	他(馬上)準備上學	Redundant
2.	他(準備)(馬上)上學	Disorder
3.	他(很快地)準備上學	Selection
4.	他馬上要準備上學(了)	Missing

Table 6. Example on multiple ways for correction.

All the four directions are reasonable but the dataset only provide the third one. Therefore, these data may create confusion for classification and should be considered in the future work. In addition, some annotation maybe not so cleat, for instance in the sentence 但是這幾天我發現(到)你有一些生活上不好的習慣 (But these days I noticed some bad habits on you in your daily life). The given annotation is *selection*, but we think *redundant* is much more reasonable.

6 Future Work

According to the observations towards the training data, we think the following direct proposal is learning from the position level, just as the shared task demands. On this level, we can extract more pointed features, integrating both syntactic and semantic ones. Besides, for the sentential level classification, the deep neural network based methods (i.e. Convolutional Neural Networks) are expected, with traditional features or embeddings, to detect more structured rules. In addition, we deem that dependency tree features may be useful and should be further developed. And improvement may also be achieved by mining the confusion in annotation (i.e. the difference between *selection* and *redundant*).

7 Conclusion

In this paper, we introduce the ensemble learning based method used in the CFL shared task for Chinese grammatical error diagnosis. We report some of our observations towards the training data, features and learners we used in our experiments. Different from most previous work, we didn't use any other external language corpus for reference and we didn't use any rules either. The results show that the ensemble methods perform better than the single classifiers based on our simple features. From the results, we see space for further development.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (61272383, 61173075 and 61203378), and the Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20120613151940045).

Reference

Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo. 2012. Error Diagnosis of Chinese Sentences using Inductive Learning Algorithm and Decomposition-based Testing Mechanism. ACM Trans. Asian Language Information Processing.

Tao-Hsing Chang, Yao-Ting Sung, Jia-Fei Hong and Jen-I Chang. 2014.KNGED:a Tool for Grammatical Error Diagnosis of Chinese Sentences. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 48-55.

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In Proceedings of COLING 2014, pp. 279-289.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, et al. 2014. A Sentence Judgment System for Grammatical Error Detecgtion. In Proceedings of COLING 2014: System Demonstrations, 67–70.

Chuan-Jie Lin and Shao-Heng Chan. 2014. Description of NTOU Chinese Grammar Checker in CFL 2014. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 75-78.

Ting Liu, Wanxiang Che, Zhenghua Li. 2011. Language Technology Platform. Journal of Chinese Information Processing. 25(6), pp. 53-62.

Hwee Tou Ng, Siew Mei, Yuanbin Wu, Christian Hadiwinoto and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2013 Shared Task). Sofia, Bulgaria.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2014 Shared Task). Baltimore, Maryland.

Chung-Hsien Wu, Chao-Hong Liu, Harris Matthew and Liang-Chih Yu. 2010. Sentence Correction Incorporating Relative Position and Parse Template Language Models, IEEE Trans. on Audio, Speech and Language Processing, vol. 18, no. 6, pp. 1170-1181.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang (2014). Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 42-47.

Marcos Zampieri and Liling Tan. 2014. Grammatical Error Detection with Limited Training Data: The Case of Chinese. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 69-74.

Yinchen Zhao, Mamoru Komachi and Hiroshi Ishikawa. 2014. Extracting a Chinese Learner Corpus from the Web: Grammatical Error Correction for Learning Chinese as a Foreign Language with Statistical Machine Translation. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 56-61.