Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

# Grammatical Error Detection with Limited Training Data: The Case of Chinese

**Marcos ZAMPIERI, Liling TAN**
*Saarland University, Germany*

**Abstract:** In this paper, we describe the UDS submission to the shared task on Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. We designed four different experiments (runs) to approach this task. All of them are variations of a frequency-based approach using a journalistic corpus as standard corpus and comparing n-gram frequency lists to both the training and the test corpus provided by the shared task organizers. The assumption behind this approach is that comparing a standard reference corpus to a non-standard study corpus using frequency-based methods levels out non-standard features present in the study corpus. These features are very likely to be, in the case of this corpus, grammatical errors. Our system obtained 60.3% f-measure at the error detection level and 25.3% f-measure at the error diagnosis level.

**Keywords:** grammatical error detection, Chinese, error diagnosis, learner corpora

## 1. Introduction

Grammatical error detection and correction is a vibrant research area in NLP. In the last couple of years much effort has been concentrated on the detection, diagnosis and correction of errors in texts written both by native speakers and by foreign language learners. For foreign language learning the practical applications of grammatical error detection are manifold, ranging from spelling and grammar checkers to essay scoring and grading.

Given this interest, a number of shared tasks have been organized in recent years. This includes the HOO 2012 preposition and determiner error correction shared task (Dale et al., 2012) held at the 2012 edition of the BEA Workshop and the Grammatical Error Correction shared tasks held at CoNLL-2013 (Ng et al., 2013) and one year later at CoNLL-2014 (Ng et al., 2014).

Similar to the previous shared tasks, this year's Grammatical Error Diagnosis for Learning Chinese as a Foreign Language provided us the opportunity to explore computational methods on diagnosis of errors committed by foreign learners of Mandarin Chinese. The shared task was designed to evaluate systems' output in two stages:

1) Error detection level: identify whether a sentence contains an error or not.
2) Error diagnosis level: classification of errors types (redundant words, missing words, word order and word selection).

Participants were required to train their systems not only to identify errors, but also to classify error types making the task more challenging. As an example, a system trained for Chinese error diagnosis was recently presented by Lee et al. (2014) obtaining 68.9% F1 score.

Apart from the difficulty of the task itself, it is important to note that the computational processing of logographic languages such as Chinese poses several difficulties to researchers used to handling character-based languages. Trivial pre-processing tasks like tokenization and segmentation are much more challenging for Chinese than for example, for English. This issue will be discussed in more detail in this paper.

In the next sections we describe the UDS submissions to the shared task commenting on the results obtained and on the strengths and weaknesses of our approach. In our submissions we used a frequency-based approach using a reference corpus to compensate the small amount of training data available.

## 2. Related Work

Grammatical error correction and detection has been the subject of a number of research papers in recent years. Shared tasks such as the aforementioned Grammatical Error Correction at CoNLL-2013 (Ng et al., 2013) and CoNLL-2014 (Ng et al., 2014) have been organized to evaluate systems' performance in correcting errors in learner corpora.

Tetreault and Chodorow (2008) presented experiments for detecting preposition errors in English texts written by non-native speakers. The authors report 84% precision and 19% recall. Heilman et al. (2012) proposed a hybrid error correction approach to the HOO 2012 shared task (Dale et al., 2012) focusing on increasing recall and F-measure scores. The authors argue that most systems take only precision into account due to the high cost of false positives (e.g. labeling grammatical sentences as ungrammatical).

More recently, Yuan and Felice (2013) proposed the use of phrase-based statistical machine translation to grammatical error correction. The application of SMT techniques to the task is not new (Brockett et al., 2006) and the performance achieved by their approach is not particularly high. However, in Yuan and Felice (2013), authors contribute in other ways, as for example, in exploring methods of generating new artificial errors to increase the size of the dataset and therefore providing more training material. The generation of artificial errors has been the subject of other research papers such as in Foster and Anderson (2009) and Felice and Yuan (2014).

As for Chinese, Yu and Chen (2012) investigated the problem of word ordering errors in Chinese texts written by Chinese foreign language learners. Authors report 71.64% accuracy using word n-grams and POS tags. Chang et al. (2012) presented a rule-based learning algorithm (first order inductive learner (FOIL)) combined with a log-likelihood function to identify error types in Chinese texts.

In this section we briefly discussed a couple of recent papers that deal with error detection, correction and diagnosis. For a comprehensive overview about the topic see Leacock et al. (2014).

## 3. Methods

Given the task description presented in section 1, we received a training corpus from the organizers containing over 12,000 labeled instances in XML format. The corpus was annotated with a unique identifier for each sentence 'sentence id', the type of mistake that each sentence contained and its respective correction. A snapshot of the corpus provided by the organizers can be seen next:

```
<ESSAY title="寫給即將初次見面的筆友的一封信">
   <TEXT>
      <SENTENCE id="B1-0112-1">我的計畫是十點早上在古亭捷運站</SENTENCE>
      <SENTENCE id="B1-0112-2">頭會戴著藍色的帽子</SENTENCE>
   </TEXT>
      <MISTAKE id="B1-0112-1">
            <TYPE>Disorder</TYPE>
            <CORRECTION>我的計畫是早上十點在古亭捷運站</CORRECTION>
      </MISTAKE>
      <MISTAKE id="B1-0112-2">
            <TYPE>Missing</TYPE>
            <CORRECTION>頭上會戴著藍色的帽子</CORRECTION>
      </MISTAKE>
 </ESSAY>
```

In our preliminary experiments we observed that the corpus provided was not sufficiently large to build robust machine learning models for grammatical error detection or diagnosis. In a similar text classification shared task using learner corpora (Tetreault et al., 2013), the amount of training data was significantly larger than the test data which allowed researchers to build more robust models based only on the given training data.

In addition to that, we had a couple of problems with the Chinese segmentation tool that we used (Chang et al. 2008) and this returned us fewer segments than were actually in the training corpus. We unfortunately did not have enough time to perform error analysis on the segmentation and pre-processing tools available nor did we have time to use the most recent Chinese segmenters (Tan and Bond, 2014; Wang et al., 2014) before the shared task submission deadline. Given these difficulties, we had to search for new strategies to approach the task with limited training data that could still achieve results comparable to the state-of-the-art systems. Inspired by existing related work, we considered three alternatives to approach the task.

a) Use an external Mandarin Chinese as a foreign language corpus preferably containing similar tags to those of the training and test data.
b) Generate a list of artificial errors to increase the amount of instances in the training corpus as in Felice and Yuan (2014).
c) Use a frequency-based approach to compare the learner corpus to a standard general language corpus . The assumption is that this comparison would level out non-standard features of the training/test data that are more likely to be errors.

Given the time and resources we had, we decided to go with option (c) and leave the other two for future work. Option (a) seemed to be promising and straightforward in terms of performance, but we did not have suitable training data at our disposal. Acquiring and annotating new data is expensive and time consuming which made option (a) infeasible. As to option (b) we regard it to be a suitable and interesting alternative in cases where training data is not available. However, it is not currently possible to say much about the performance of these methods for Chinese. To our knowledge, previous work has only been done for Indo-European languages.

Option (c) proved to be the most adequate solution for our submission. A frequency-based approach, like the one used in our submission, requires only a large reference corpus (a general standard contemporary language one). We had a couple of suitable resources at our disposal (Chen et al., 1996 Graff and Chen, 2003) and we therefore decided to test this method.

The method works under a similar assumption to the keyword lists widely used in corpus linguistics (Scott, 1997; McEnery, 2009) and also applied on a similar scenario by Zampieri et al. (2013) on Internet data. Keyword lists are produced by comparing two corpora (a study corpus and a reference corpus) using association metrics such as log-likelihood, chi-square or mutual information. These keywords usually reflect salient features of the study corpus. In the case of the present comparison (standard corpus versus learner corpus), it is safe to assume that a reasonable amount of salient features from the learner corpus will be infrequent distributions of words which are very likely to be errors. This is the basic assumption of our approach.

### *3.1 Algorithm*

If one assumes that a reference corpus is a portrait of standard language, lexical items that stand out in the study corpus in comparison to the reference corpus should deviate from what is considered to be 'the norm'. This is a relatively naive assumption and thematic bias may still occur when using unbalanced data. To avoid that, we used a large balanced journalistic corpus (Graff and Chen, 2003) as our standard corpus. From the reference corpus we sampled the first 50,000 sentences and extracted n-grams (1 to 5) using the KenLM Language Model Toolkit (Heafield, 2011).

We pre-processed the training, test and standard corpora using the Stanford tokenizer (Chang et al. 2008). As Chinese is a logographic language we treat every character in isolation. As previously mentioned, the Stanford segmenter yielded a number of errors in segmentation that worsened our system's performance. However, we were not able to evaluate the exact segmenter's performance for our dataset before this submission.

From the training and test corpora provided by the shared task organizers we proceeded to extract a list of ungrammatical n-grams that were not present in the subset of the reference corpus and treated them as key expressions. This calculation returned us a list of 35,000 ungrammatical n-grams not present in the reference corpus.

It is important to note here that the main difference betweem our approach and what is commonly used in corpus linguistics is that the latter uses the lexicon in the form of bag-of-words (or less often bigrams). In these experiments we used the complete set of n-grams (1 to 5) extracted from the corpus thus increasing the coverage of our method.

With these n-gram lists, we trained two classifiers to identify grammatical and ungrammatical instances: 1) a simple n-gram-based classifier to identify correct (grammatical) sentences using the formula below and 2) a Multinomial Naive Bayes (MNB) classifier to identify ungrammatical sentences along with their labels using the Scikit-learn package (Pedregosa et al., 2011).

$$\frac{\sum_{i}^{n} p(ng_i \mid ng_i \notin ng_{giga} \ , \ \dot{ng}_i \in ng_{traintest})}{\sum_{i}^{n} p(ng_i \mid ng_i \in ng_{giga})} > X \tag{1}$$

In the formula above, we tuned the *X* parameter value to optimize the results obtained by the first classifier. After a number of tests we found that the optimal value lies between 0.10 and 0.20. We therefore produced four submissions (runs) using different *X* values: 0.20 for the 1st run, 0.16 for the 2nd run, 0.15 for the 3rd run and 0.10 for the 4th run. The best results were obtained in our first run, using *X* = 0.20 and these are the results that will be reported and discussed next.

## 4. Results

According to the information provided by the organizers, 13 teams registered for the shared task and 6 of them submitted their final results. The results were calculated using standard metrics in text classification, namely: precision, recall, accuracy and F-measure as well as a false positive rate score. No limitation regarding the number of runs was set. The test set provided by the organizers contained 1,750 unlabeled test instances.

The UDS team submitted four runs changing the *X* parameter of our correct sentence classifier as explained in the previous section. In table 1 we present the best results obtained by all 6 groups at the error detection level. At this level, our approach was the fourth best with results reaching 60.37% F1 score and 49.14% accuracy.

Table 1: Error Detection Level: Results.

| Team | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| CIRU | 0.6446 | 0.6128 | 0.7851 | 0.6884 |
| NTOU | 0.5000 | 0.5000 | 1 | 0.6667 |
| KUAS&NTNU | 0.5006 | 0.5003 | 0.9051 | 0.6444 |
| **UDS** | **0.4914** | **0.4945** | **0.7749** | **0.6037** |
| TMU | 0.5171 | 0.5399 | 0.232 | 0.3245 |
| NCYU | 0.4983 | 0.4927 | 0.1154 | 0.187 |

The top four systems obtained F1 scores between 60% and 69%; the 5th and 6th best system, however, obtained significantly lower F-scores. Our results were lower than the 3 best systems but still above the expect 50% baseline. In terms of recall, our system was also ranked as the 4th best and as to the accuracy results, our system was the 5th best. It obtained performance comparable to the 2nd, 3rd and 4th best systems: 49.14% against 50.00%, 50.06% and 51.71% accuracy. The best system obtained significantly higher accuracy scores compared to all other systems, 64.46% accuracy.

In table 2 we present the best results obtained by the six systems at the error diagnosis level.

Table 2: Error Diagnosis Level: Results.

| Team | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| CIRU | 0.4589 | 0.4548 | 0.4137 | 0.4333 |
| NTOU | 0.2074 | 0.2932 | 0.4149 | 0.3436 |
| KUAS&NTNU | 0.2149 | 0.2696 | 0.3337 | 0.2983 |
| **UDS** | **0.2337** | **0.2467** | **0.2594** | **0.2529** |
| TMU | 0.4554 | 0.3545 | 0.1086 | 0.1662 |
| NCYU | 0.4594 | 0.2409 | 0.0377 | 0.0652 |

The error diagnosis level is more difficult than the error detection step. This is due to the multiple tags (e.g. missing words, word order) that could be attributed to each instance. At this stage, the performance of all systems was substantially lower than the error detection step. Once again our system was ranked 4th in terms of both F-score and accuracy achieving 23.37% F1 and 25.29% accuracy. The best system achieved 43.33% f-measure and 45.89% accuracy.

The dataset itself was to our understanding very challenging for the frequency-based methods we proposed. We found that some instances were virtually impossible to correctly tag. Examples of instances that were difficult to classify include single words: 老師 (EN 'teacher'), short expressions: 又很貴 (EN 'also very expensive') and instances that without context were difficult to understand even for native speakers: 姓本多 (EN literally: 'nature', 'by itself', 'many')

The results we obtained were consistently ranked in the middle of the table and they are, to our understanding, comparable to the state-of-the-art performance for the task. By looking at the performance obtained by the CIRU team, we see, however, room for improvement, as will be discussed in the next section.

## 5. Conclusion

This paper described the UDS submission to the shared task on Grammatical Error Diagnosis for Chinese as Foreign Language. We approached the task using frequency information and report results comparable to other state-of-the-art systems. The task is by no means trivial and the almost 9 percentage points behind the best system (CIRU team) showed us that there is still room for improvement. Even so, considering the lack of suitable training data, we believe that the results we obtained are still interesting to report.

We believe that better results can be obtained, for example, by integrating spell checkers (Lin and Chu, 2013) to our algorithm, particularly those that take phonetics into account (Zampieri and de Amorim, 2014). Another issue that should be taken into account in future experiments is the question of segmentation. Very good performance in tokenization is paramount when dealing with logographic languages and this was unfortunately not obtainable with the methods we used.

Finally, in the future we would like to perform experiments to increase the size of the training corpus using artificial errors as proposed by Felice and Yuan (2014). We believe that this is an effective way of producing more data for this task. The performance of these methods when applied to Chinese is still an open question.

## Acknowledgements

We would like to thank the shared task organizers, especially Lung-Hao Lee, for providing prompt replies to all our inquiries. We acknowledge that no *LaTeX* template was available for this submission and we did our best to improve the quality of the formula and tables presented in this paper.

## References

Brockett, C., Dolan, W. B., Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. p. 249-256.

Chang, P. C., Galley, M., Manning, C. (2008) Optimizing Chinese Word Segmentation for Machine Translation Performance. In: Proceedings of WMT.

Chang, R. Y., Wu, C. H., Prasetyo, P. K. (2012). Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism. ACM Transactions on Asian Language Information Processing, 11(1).

Chen, K. J., Huang, C. R., Chang, L. P., Hsu, H. L. (1996). Sinica corpus: Design methodology for balanced corpora. In: Proceedings of PACLIC. p. 167-176.

Dale, R., Anisimoff, I., Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 54-62). Association for Computational Linguistics.

Felice, M., & Yuan, Z. (2014). Generating artificial errors for grammatical error correction. Proceedings of the EACL Student Workshop.

Felice, M., Yuan, Z., Andersen, Ø. E., Yannakoudakis, H., Kochmar, E. (2014). Grammatical error correction using hybrid systems and type filtering. In: Proceedings of CoNLL-2014.

Foster, J., Andersen, Ø. E. (2009). GenERRate: generating errors for use in grammatical error detection. In Proceedings of the fourth workshop on innovative use of NLP for building educational applications. Association for Computational Linguistics. p. 82-90.

Graff, D., Chen, K. (2003) Chinese Gigaword. Philadelphia: Linguistic Data Consortium.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics. p. 187-197.

Heilman, M., Cahill, A., Tetreault, J. (2012). Precision isn't everything: a hybrid approach to grammatical error detection. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 233-241). Association for Computational Linguistics.

Leacock, C., Chodorow, M., Gamon, M., Tetreault, J. (2014). Automated grammatical error detection for language learners – Second Edition. Synthesis lectures on human language technologies. Morgan & Claypool.

Lee, L. H., Yu, L. C., Lee, K. C., Tseng, Y. H., Chang, L. P., Chen, H. H. (2014). A Sentence Judgment System for Grammatical Error Detection. Proceedings of COLING 2014.

Lin, C. J., Chu, W. C. (2013). NTOU Chinese Spelling Check System in SIGHAN Bake-off 2013. In Sixth International Joint Conference on Natural Language Processing.

McEnery, T. (2009) Keywords and moral panics: Mary Whitehouse and media censorship. in D. Archer (ed.) What's in Word-list? Investigating Word Frequency and Keyword Extraction. Oxford.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C.,Tetreault, J. (2013). The CoNLL-2013 Shared Task on Grammatical Error Correction. In: Proceedings of CoNLL-2013.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In: Proceedings of CoNLL-2014.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. (2011) Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 12. p. 2825-2830

Scott, M. (1997) PC Analysis of key words and key key words. System. n. 25. Elsevier. p. 233-245.

Tan, L., Bond, F. (2014). NTU-MC Toolkit: Annotating a Linguistically Diverse Corpus. In: Proceedings of COLING 2014. Dublin, Ireland.

Tetreault, J., Blanchard, D., Cahill, A. (2013). A report on the first native language identification shared task. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. p. 48-57.

Tetreault, J. R., & Chodorow, M. (2008) The ups and downs of preposition error detection in ESL writing. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. Association for Computational Linguistics. p. 865-872.

Wang, M., Voigt, R., Manning, C. (2014) Two Knives Cut Better Than One: Chinese Word Segmentation with Dual Decomposition. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA.

Yu, C. H., Chen, H. T. (2012). Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp. 3003-3017.

Yuan, Z., Felice, M. (2013). Constrained grammatical error correction using Statistical Machine Translation. In: Proceedings of CoNLL-2013.

Zampieri, M., de Amorim, R. C. (2014) Between Sound and Spelling: Combining Phonetics and Clustering Algorithms to Improve Target Word Recovery. Proceedings of the 9th International Conference on Natural Language Processing (PolTAL). Lecture Notes in Computer Science (LNCS). Springer.

Zampieri, M., Hermes, J., Schwiebert, S. (2013) Identification of Patterns and Document Ranking of Internet Texts: A Frequency-based Approach. Non-Standard Data Sources in Corpus-based Research. ZSM-Studien Series - Vol. 5. Shaker. p. 25-39.