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

Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language

Liang-Chih YU[a,b], Lung-Hao LEE [c,d*] & Li-Ping CHANG [e]
[a]*Department of Information Management, Yuen-Ze University, Taiwan*
[b]*Innovation Center for Big Data and Digital Convergence, Yuen-Ze University, Taiwan*
[c]*Information Technology Center, National Taiwan Normal University, Taiwan*
[d]*Department of Computer Science and Information Engineering, National Taiwan University, Taiwan*
[e]*Mandarin Training Center, National Taiwan Normal University, Taiwan*
*lhlee@nlg.csie.ntu.edu.tw

Abstract: We organize a shared task on grammatical error diagnosis for learning Chinese as a Foreign Language (CFL) in the ICCE-2014 workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA). In this paper, we describe all aspects of this shared task, including task description, data preparation, evaluation metrics, and testing results. The aim is, through such evaluation campaigns, more advanced computer-assisted Chinese learning techniques will be emerged.

Keywords: Computer-assisted language learning, shared task, Mandarin Chinese

1. Introduction

China's growing global influence has prompted a surge of interest in learning Chinese as a foreign language (CFL), and this trend is expected to continue. However, whereas many computer-assisted learning tools have been developed for use by students of English as a Foreign Language (EFL), support for CFL learners is relatively sparse, especially in terms of tools designed to automatically detect and correct Chinese grammatical errors. For example, while Microsoft Word has integrated robust English spelling and grammar checking functions for years, such tools for Chinese are still quite primitive.

In contrast to the plethora of research related to EFL learning, relatively few studies have focused on computer-assisted language learning for CFL learners. Relative position and parse template language models have been adopted to detect Chinese errors written by US learners (Wu et al. 2010). Machine learning models have been applied to detect word-ordering errors in Chinese sentences from the HSK dynamic composition corpus (Yu and Chen, 2012). Ranking SVM based model has been further explored to rank the candidates and suggest the proper corrections of word ordering errors (Cheng et al. 2014). A penalized probabilistic First-Order Inductive Learning (pFOIL) algorithm has been proposed for grammatical error diagnosis (Chang et al. 2012). Linguistic rule based approach has been presented to detect grammatical errors written by CFL learners (Lee et al. 2013). A sentence judgment system has been implemented to integrate rule-based linguistic analysis and n-gram statistical learning for detecting grammatical errors (Lee et al. 2014). SIGHAN 2013 bakeoff on Chinese spelling check evaluation focus on developing automatic checker to detect and correct spelling errors (Wu et al. 2013). In summary, human language technologies for Chinese learning have attracted more attentions in recent years.

In the ICCE-2014 workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA), we organize a shared task on Chinese grammatical error diagnosis that provides an evaluation platform for developing and implementing computer-assisted learning tools. The data sets in our task are collected from the Chinese as the Foreign Language (CFL) learners' written essays. Given a sentence with/without one of grammatical errors, *i.e.*, redundant word, missing word, word disorder, and word selection, the developed system should indicate whether contains grammatical errors and further points out which one of defined error types. The hope is that,

through such evaluation campaigns, more advanced Chinese grammatical error detecting techniques will be emerged.

We give an overview of the shared task on grammatical error diagnosis for learning Chinese as a foreign language. The rest of this article is organized as follows. Section 2 details the designed task. Section 3 introduces the data sets provided in this evaluation. Section 4 proposes the evaluation metrics. Section 5 presents the results of participants' approaches for performance comparison. Finally, we conclude this paper with the findings and future research direction in the Section 6.

2. Shared Task Description

The goal of this shared task is developing the computer-assisted tools to detect several kinds of grammatical errors, that is, redundant word, missing word, word disorder, and word selection. The input sentence contains one of defined error types. The developed tool should indicate which kind of error type is embedded in the given sentence. If the input sentence, which is given a unique sentence number SID, contains no grammatical errors, the tools should return "SID, Correct". If an input sentence contains a defined grammatical error, the output format should be "SID, error_type". We simplify the task that there are only one error type may be in the given sentence. Examples are shown as follows. In example 1, the character "被" is a redundant word. There is a missing word "有" in the example 2 and its correct usage is shown in example 3. The sentence in the example 4 has word disorder error, *i.e.*, the word "很早" should be preceded the word "起床". The word "一個" in the example 5 is an incorrect word selection, the correct word should be "一件".

- Example 1
  Input: (sid=B2-1447-6)　希望沒有人再被食物中毒
  Output: B2-1447-6, Redundant
- Example 2
  Input: (sid=C1-1876-2)　對社會國家不同的影響
  Output: C1-1876-2, Missing
- Example 3
  Input: (sid=C1-1876-2)　對社會國家有不同的影響
  Output: C1-1876-2, Correct
- Example 4
  Input: (sid=A2-0775-2)　我起床很早
  Output: A2-0775-2, Disorder
- Example 5
  Input: (sid=B1-0110-2)　我會穿著一個黃色的襯衫
  Output: B1-0110-2, Selection

3. Data Sets

Mandarin Training Center (MTC) of National Taiwan Normal University (NTNU) was founded in 1956 for teaching Chinese as a foreign language. Currently, MTC is the most renowned Chinese language center in Taiwan, in which around 1700 CFL learners from more than 70 countries enrolled each academic quarter. The learner corpus used in our task is collected from the computer-based writing Test of Chinese as a Foreign Language (TOCFL). The writing test is designed according to the six proficiency levels of the Common European Framework of Reference (CEFR). Test takers have to complete two different tasks for each level. For example, for the A2 (Waystage level) candidates, they will be asked to write a note and describe a story after looking at four pictures. All candidates are asked to complete the writings on line.

We further ask the annotators to label the grammatical errors in CFL learners' written sentences and provide their correct usage. Our prepared data is further divided into three distinct sets. (1) **Training set**: 1,506 CFLs' writings are collected in which 5,607 grammatical errors are annotated. Each CFL learners' writing is represented in SGML format shown in Figure 1. The title attribute is

used to describe the topic of the writing test. There is only one grammatical error in an annotated sentence. The error types are also indicated along with their corresponding correct usages. All sentences in this set can be used to train the developed grammatical error detection tool. (2) **Dryrun set**: Total 33 sentences are given for participants to familiarize themselves with the final testing process. Each participant can submit several runs generated using different models with different parameter settings. In addition to make sure the submitted results can be correctly evaluated, participants can fine-tune their developed models in the dryrun phase. The purpose of dryrun is for output format validation only. No matter which performance can be achieved that will not be included in our official evaluation. (3) **Test set**: In total, there are 1,750 testing sentences. A half of these instances contain no grammatical errors. Another half of testing cases includes one grammatical error per sentence. The number of error type redundant, missing, disorder, and selection is 279, 350, 120, and 126, respectively. The distribution is the same with our given training set. The policy of our evaluation is an open test. In addition to our provided data sets, registered research teams can employ any linguistic and computational resources to detect grammatical errors in the sentences.

```
<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>
```

Figure 1. An essay represented in SGML format.

4. Performance Metrics

Table 1 shows the confusion matrix used for performance evaluation. In the matrix, True Positive (TP) is the number of sentences with grammatical errors that are correctly proposed by the developed tool; False Positive (FP) is the number of sentences without grammatical errors that are incorrectly proposed; True Negative (TN) is the number of sentences without grammatical errors that are identified correctly; False Negative (FN) is the number of sentences with grammatical errors that are incorrectly regarded as correct sentences.

Table 1: The confusion matrix for evaluation.

Confusion Matrix		**System Result**	
		Positive (With grammatical errors)	Negative (Without grammatical errors)
Gold Standard	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The criteria for judging correctness are distinguished into two levels. (1) **Detection level**: all error types are regarded as incorrect. Binary classification of a testing instance, *i.e.*, correct or incorrect, should be completely identical with the gold standard. (2) **Identification level**: this level could be considered as a multi-class categorization problem. In addition to correct instances, all error

types should be clearly identified, *i.e.*, Redundant, Missing, Disorder, and Selection. The following metrics are measured in both levels with the help of the confusion matrix.

- **False Positive Rate** (FPR) = FP / (FP+TN)
- **Accuracy** = (TP+TN) / (TP+FP+TN+FN)
- **Precision** = TP / (TP+FP)
- **Recall** = TP / (TP+FN)
- **F1** = 2 * Precision *Recall / (Precision + Recall)

For example, give 8 testing inputs with gold standards shown as "A2-0802-4, correct", "A2-3344-1, Selection", "B1-3419-8, Missing", "B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant", and "C1-1873-7, correct". The system may output the result shown as "A2-0802-4, Disorder", "A2-3344-1, Redundant", "B1-3419-8, Selection", "B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant", and "C1-1873-7, Missing". The evaluation tool will yield the following performance.

- **False Positive Rate (FPR)** = 0.5 (= 2 / 4).
  Notes: {"A2-0802-4, Disorder", "C1-1873-7, Missing"} /{"A2-0802-4, correct", "B1-3520-3, correct", "B2-1918-7, correct", "C1-1873-7, correct"}
- **Detection Accuracy** = 0.75 (=6/8).
  Notes: {"A2-3344-1, Redundant", "B1-3419-8, Selection", "B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-0802-4, correct", "A2-3344-1, Selection", "B1-3419-8, Missing", "B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant", "C1-1873-7, correct"}
- **Detection Precision** = 0.67 (=4/6).
  Notes: {"A2-3344-1, Redundant", "B1-3419-8, Selection", "B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-0802-4, Disorder", "A2-3344-1, Redundant", "B1-3419-8, Selection", "B2-2231-4, Disorder", "C1-1744-1, Redundant", "C1-1873-7, Missing"}
- **Detection Recall** = 1 (=4/ 4).
  Notes: {"A2-3344-1, Redundant", "B1-3419-8, Selection", "B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-3344-1, Selection", "B1-3419-8, Missing", "B2-2231-4, Disorder", "C1-1744-1, Redundant"}
- **Detection F1**= 0.8024 (=2*0.67*1/(0.67+1))
- **Identification Accuracy** = 0.5 (=4/8).
  Notes: {"B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-0802-4, correct", "A2-3344-1, Selection", "B1-3419-8, Missing", "B1-3520-3, correct", "B2-1918-7, correct", "B2-2231-4, Disorder", "C1-1744-1, Redundant", "C1-1873-7, correct"}
- **Identification Precision** = 0.33 (=2/6).
  Notes: {"B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-0802-4, Disorder", "A2-3344-1, Redundant", "B1-3419-8, Selection", "B2-2231-4, Disorder", "C1-1744-1, Redundant", "C1-1873-7, Missing"}
- **Identification Recall** = 0.5 (=2/ 4).
  Notes: {"B2-2231-4, Disorder", "C1-1744-1, Redundant"} /{"A2-3344-1, Selection", "B1-3419-8, Missing", "B2-2231-4, Disorder", "C1-1744-1, Redundant"}
- **Identification F1**= 0.3976 (=2*0.33*0.5/(0.33+0.5))

5. Evaluation Results

Table 2 shows the participant teams and their testing submission statistics. Our shared task attracted 13 research teams. There are 4 teams that come from Taiwan, *i.e.*, AS, KUAS & NTNU, NCYU, and NTOU. 3 teams originate from China, *i.e.*, HITSZ, PKU, and PolyU. The remaining 6 teams are CIRU from United States of America, MU from New Zealand, SPBU from Russia, TMU from Japan, UL from United Kingdom, and UDS from Germany. Among 13 registered teams, 6 teams submitted their testing results. In total, we had received 13 runs in the formal testing phase.

Table 2: Result submission statistics of all participants.

Participants (Ordered by abbreviations of names)	**#Submissions**
Academia Sinica (**AS**)	0
Confucius Institute of Rutgers University (**CIRU**)	1
Harbin Institute of Technology Shenzhen Graduate School (**HITSZ**)	0
National Kaohsiung University of Applied Sciences & National Taiwan Normal University (**KUAS & NTNU**)	3
Massey University (**MU**)	0
National Chiayi University (**NCYU**)	1
National Taiwan Ocean University (**NTOU**)	2
Peking University (**PKU**)	0
The Hong Kong Polytechnic University (**PolyU**)	0
Saint Petersburg State University (**SPBU**)	0
Tokyo Metropolitan University (**TMU**)	2
Saarland University (**UDS**)	4
University of Leeds (**UL**)	0
Total	**13**

Table 3: Testing results of our shared task.

Submission	**FPR**	**Detection Level**				**Identification Level**			
		Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
CIRU-Run1	0.496	**0.6446**	**0.6128**	0.7851	**0.6884**	**0.4589**	**0.4548**	0.4137	**0.4333**
KUAS & NTNU-Run1	0.904	0.5006	0.5003	0.9051	0.6444	0.2149	0.2696	0.3337	0.2983
KUAS & NTNU-Run2	0.2686	0.5217	0.5374	0.312	0.3948	0.4109	0.2516	0.0903	0.1329
KUAS & NTNU-Run3	0.904	0.5006	0.5003	0.9051	0.6444	0.2074	0.2607	0.3189	0.2869
NCYU-Run1	**0.1189**	0.4983	0.4927	0.1154	0.187	0.4594	0.2409	0.0377	0.0652
NTOU-Run1	1	0.5	0.5	**1**	0.6667	0.16	0.2424	0.32	0.2759
NTOU-Run2	1	0.5	0.5	**1**	0.6667	0.2074	0.2932	**0.4149**	0.3436
TMU-Run1	0.1977	0.5171	0.5399	0.232	0.3245	0.4554	0.3545	0.1086	0.1662
TMU-Run2	0.1691	0.5103	0.5287	0.1897	0.2792	0.4531	0.3084	0.0754	0.1212
UDS-Run1	0.792	0.4914	0.4945	0.7749	0.6037	0.2337	0.2467	0.2594	0.2529
UDS-Run2	0.6286	0.4949	0.4959	0.6183	0.5504	0.2869	0.2435	0.2023	0.221
UDS-Run3	0.5783	0.4949	0.4955	0.568	0.5293	0.3057	0.247	0.1897	0.2146
UDS-Run4	0.2491	0.5046	0.509	0.2583	0.3427	0.428	0.2968	0.1051	0.1553

Table 3 shows the testing results of our shared task. In addition to achieving promising detection effects of grammatical errors, reducing the false positive rate, which is percentage of the correct sentences that are incorrectly reported containing grammatical errors, is also important. The research teams, NCYU and TMU, achieved relatively low false positive rates.

Detection level evaluations are designed to detect whether a sentence contains grammatical errors or not. Accuracy is usually adopted to evaluate the performance, but it is affected by the distribution of testing instance. The baseline can be achieved easily by always guessing without errors. That is accuracy of 0.5 in this evaluation. Some systems achieved slightly better than the baseline, *i.e.*, CIRU, KUAS&NTNU, TMU and UDS. Registered teams may send different runs that aimed at optimizing the recall or precision rates. These phenomena guide us to adopt F1 score to

reflect the tradeoff between precision and recall. In the testing results, CIRU accomplished the best detection effects of indicating grammatical errors, which resulted the best F1 score 0.6884. For identification level evaluations, the systems need to identify the error types in the given sentences. The research team came from CIRU accomplished the best correction accuracy 0.4589. Most systems cannot effectively identify the input sentences to point out possible grammatical errors. Our testing results indicate that the system developed by CIRU accomplished the best identification F1 0.4333.

In summary, it is a really difficult task to develop the computer-assisted learning tool for grammatical error diagnosis, especially learning Chinese as a foreign language, since there are only target sentences without the help of their context. We cannot find a relatively promising system according to our testing results. In general, this research problem still has long way to go.

6. Conclusions and Future Work

This paper describes the overview of NLPTEA 2014 shared task on grammatical error diagnosis for learning Chinese as a foreign language. We introduce the task designing ideas, data preparation details, evaluation metrics, and the results of performance evaluation. This task also encourages researchers to bravely propose various ideas and implementations for possible breakthrough. No matter how well their implementations would perform, they contribute to the community by enriching the experience that some ideas or approaches are promising or impractical, as verified in this shared task. Their reports in the proceeding will reveal the details of these various approaches and contribute to our knowledge about computer-assisted Chinese learning.

All data sets and their accompanying gold standards and evaluation tool are publicly available for research purposes at http://ir.itc.ntnu.edu.tw/lre/nlptea14cfl.htm. We hope our provided data can serve as a benchmark to help developing better Chinese learning tools. This shared task also motivates us to build more language resources in the future to possibly improve the state-of-the-art techniques.

Acknowledgements

This research was supported by the Ministry of Science and Technology, under the grant MOST 102-2221-E-155-029-MY3, 103-2221-E-003-013-MY3, 103-2911-I-003-301 and the “Aim for the Top University Project” and “Center of Learning Technology for Chinese” of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan.

References

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), article 3.

Cheng, S.-M., Yu, C.-H., & Chen, H.-H. (2014). Chinese word ordering errors detection and correction for non-native Chinese language learners. *Proceedings of COLING’14* (pp. 279-289), Dublin, Ireland: ACL Anthology.

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H., & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE’13* (pp. 27-29). Bali, Indonesia: Asia-Pacific Society for Computers in Education.

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’14* (pp. 23-29), Dublin, Ireland: ACL Anthology.

Wu, C.-H., Liu, C.-H., Harris, M., & Yu, L.-C. (2010). Sentence correction incorporating relative position and parse template language models. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 170-1181.

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. *Proceedings of SIGHAN’13* (pp. 35-42), Nagoya, Japan: ACL Anthology.

Yu, C.-H. & Chen, H.-H. (2012). Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. *Proceedings of COLING’12* (pp. 3003-3017), Mumbai, India: ACL Anthology.