NLP for Chinese L2 Writing: Evaluation of Chinese Grammatical Error Diagnosis

Gaoqi Rao[1], Lung-hao Lee[2]
1.Beijing Language and Culture University, 2.National Taiwan Normal University
15. Xueyuan Rd., Beijing, China; 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan
E-mail: raogaoqi@blcu.edu.cn, lhlee@ntnu.edu.tw

Abstract

This paper presents the shared task of Chinese grammatical error diagnosis (CGED) which seeks to identify grammatical error types and their range of occurrence within sentences written by L2 learners of Chinese. We describe the task definition of CGED, and overview the past 4 CGED shared tasks, especially CGED2016 and CGED2017 containing simplified character track of HSK, in data preparation, performance metrics, and evaluation results. Until now, none of the participants has developed an over performed system, showing potential of solving the task, although approaches were significant since the first CGED in 2014. We expected this evaluation campaign could lead to the development of more advanced NLP techniques for educational applications, especially for Chinese error detection and automatic correction. All data sets with gold standards and scoring scripts are made publicly available to researchers.

Keywords: CGED, error detection, L2 Chinese learning

1. Introduction

In recent years, automated grammar checking for learners of English as a foreign language has attracted more attention. For example, Helping Our Own (HOO) is a series of shared tasks in correcting textual errors (Dale and Kilgarriff, 2011; Dale et al., 2012). The shared tasks at CoNLL 2013 and CoNLL 2014 focused on grammatical error correction, increasing the visibility of educational application research in the NLP community (Ng et al., 2013; 2014).

Many of these learning technologies focus on learners of English as a Foreign Language (EFL), while relatively few grammar checking applications have been developed to support Chinese as a Foreign Language(CFL) learners. Those applications which do exist rely on a range of techniques, such as statistical learning (Chang et al, 2012; Wu et al, 2010; Yu and Chen, 2012), rule-based analysis (Lee et al., 2013) and hybrid methods (Lee et al., 2014). In response to the limited availability of CFL learner data for machine learning and linguistic analysis, the ICCE-2014 workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA) organized a shared task on diagnosing grammatical errors for CFL (Yu et al., 2014). A second version of this shared task in NLP-TEA was collocated with the ACL-IJCNLP-2015 (Lee et al., 2015), COLING-2016 (Lee et al., 2016) and IJCNLP 2017 (Rao et al., 2017). In 2018, the shared task for Chinese grammatical error diagnosis is organized again at NLP-TEA workshop in conjunction with ACL2018.

The main purpose of these shared tasks is to provide a common setting so that researchers who approach the tasks using different linguistic factors and computational techniques can compare their results. Such technical evaluations allow researchers to exchange their experiences to advance the field and eventually develop optimal solutions to this shared task.

2. Task Description

The goal of this shared task is to develop NLP techniques to automatically diagnose grammatical errors in Chinese sentences written by L2 learners. Such errors are defined as redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word ordering errors ("W"). The input sentence may contain one or more such errors. The developed system should indicate which error types are embedded in the given unit (containing 1 to 5 sentences) and the position at which they occur. Each input unit is given a unique number "sid". If the inputs contain no grammatical errors, the system should return: "sid, correct". If an input unit contains the grammatical errors, the output format should include four items "sid, start_off, end_off, error_type", where start_off and end_off respectively denote the positions of starting and ending character at which the grammatical error occurs, and error_type should be one of the defined errors: "R", "M", "S", and "W". Each character or punctuation mark occupies 1 space for counting positions. Example sentences, corresponding notes and data in SGML format are shown as Table 1 and Figure 1 show. In 2014 and 2015, we organized one track of TOCFL (Test Of Chinese as a Foreign Language) (Lee et al., 2016). In 2016, two tracks of TOCFL and HSK (Hanyu Shuiping Kaoshi)(Cui et al, 2011; Zhang et al, 2013) were organized, while in 2017 and 2018, only HSK track was and will be organized. We welcome the affiliations constructing data set of traditional characters to join the shared task in organization.

3. Datasets

Native Chinese speakers were trained to manually annotate grammatical errors and provide corrections corresponding to each error. The data were then split into Training Set and Test Set. Each unit (contain at least 1 sentence) with annotated grammatical errors and their corresponding corrections is represented in SGML format. The scale and error type distribution of the Training Set in CGED2016 and CGED2017 are reported in Table2. In test set, correct sentences are contained, in order to test the false positive rate of the systems. The distributions of error types (shown in Table 3) are similar with that of the training set.

4. Performance Metrics

Table 4 shows the confusion matrix used for evaluating system performance. In this matrix, TP (True Positive) is the number of sentences with grammatical errors are correctly identified by the developed system; FP (False Positive) is the number of sentences in which non-existent grammatical errors are identified as errors; TN (True Negative) is the number of sentences without grammatical errors that are correctly identified as such; FN (False Negative) is the number of sentences with grammatical errors which the system incorrectly identifies as being correct.

The criteria for judging correctness are determined at three levels as follows.

(1) Detection-level: Binary classification of a given sentence, that is, 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. All error types should be clearly identified. A correct case should be completely identical with the gold standard of the given error type.

(3) Position-level: In addition to identifying the error types, this level also judges the occurrence range of the grammatical error. That is to say, the system results should be perfectly identical with the quadruples of the gold standard.

(4) Correction-level: In the coming CGED2018 in conjunction with ACL2018 in July 2018, the participant systems are required to offer 0 to 3 recommended corrections to error types of missing and selection. The amount of the correction to recommend depends on the trust computation at each error. More recommendation would increase the recall, but somehow reduce precision, since the gold standard only offers one correction to each error.

The following metrics are measured at all levels with the help of the confusion matrix.

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

5. Evaluation Results and Analysis

Table 5 and Table 6 summarize the submission statistics and best F1 of position-level for the participants in CGED2016 and CGED2017. In summary, none of the submitted systems provided superior performance using different metrics, indicating the difficulty of developing systems for effective grammatical error diagnosis, especially in L2 contexts, although approaches were significant since the first CGED in 2014.

From the proceedings of the 2 shared tasks, we observed the transformation in methods: from traditional statistical modeling to deep neuro networks. About one third of the participants in CGED2016 conduct the system based on Ngram or fined turned CRF, while none of the teams continued to carry out the experiments in these ways. LSTM+CRF has been nearly standard solution to task by each team, similar to other NLP tasks.

Also like what happened in other NLP tasks, deep learning modeling as resource intensive required methods, approached better performance easier in big dataset with high quality. Unfortunately, writing data of L2 Chinese learner are quite limited in both size and quality. Track of HSK as an example, organizers from BLCU digitalized the scored writing section from the exam. Teachers in exam scoring were not required the high consistency, like other annotation task like word segmentation or sentiment analysis. On the other hand, the NLP for Chinese as L2 learning does not have a long history and impact among academia, leading to the relative low resource construction, comparing with other newly appeared task like SQuAD.

These problems in resource aspect partially lead to the limited performance of deep learning modeling. However, this task can be viewed as a low resource NLP task to challenge.

6. Conclusions

This study describes the shared task for Chinese grammatical error diagnosis, including task design, data preparation, performance metrics, and evaluation results. Regardless of actual performance, all submissions contribute to the common effort to develop Chinese grammatical error diagnosis system, and the individual reports in the proceedings provide useful insights into computer-assisted language learning for CFL learners.

We hope the data sets collected and annotated for this shared task can facilitate and expedite future development in this research area. Therefore, all data sets with gold standards and scoring scripts are publicly available online at www.cged.science.

7. Acknowledgments

We thank all the participants for taking part in our shared task. We would like to thank Kuei-Ching Lee for implementing the evaluation program and the usage feedbacks from Bo Zheng (in CGED2016). Gong Qi, Tang Peilan, Luo Ping and Chang Jie contributed in the proofreading of the data in CGED2017/2018.

This study was supported by the projects from P.R.C: High-Tech Center of Language Resource(KYD17004), BLCU Innovation Platform(17PT05), Institute Project of BLCU(16YBB16) Social Science Funding China (11BYY054, 12&ZD173, 16AYY007), Social Science Funding Beijing (15WYA017), National Language

Committee Project (YB125-42, ZDI135-3), MOE Project of Key Research Institutes in Univ(16JJD740004).

TOCFL (Traditional Chinese)	HSK (Standard Chinese)
• Example 1 Input: (sid=A2-0007-2) 聽說妳打算開一個慶祝會。可惜我不能參加。因為那個時候我有別的事。當然我也要參加給你慶祝慶祝。 Output: A2-0007-2, 38, 39, R (Notes: “参加”is a redundant word) • Example 2 Input: (sid=A2-0011-1) 我聽到你找到工作。恭喜恭喜！ Output: A2-0011-1, 2, 3, S A2-0011-1, 9, 9, M (Notes: “聽到”should be “聽說”. Besides, a word “了”is missing. The correct sentence should be “我聽說你找到工作了”. • Example 3 Input: (sid=A2-0011-3) 我覺得對你很抱歉。我也很想去，可是沒有辦法。 Output: A2-0011-3, correct	• Example 1 Input: (sid=00038800481) 我根本不能了解这妇女辞职回家的现象。在这个时代，为什么放弃自己的工作，就回家当家庭主妇? Output: 00038800481, 6, 7, S 00038800481, 8, 8, R (Notes: “了解”should be “理解”. In addition, “这” is a redundant word.) • Example 2 Input: (sid=00038800464)我真不明白。她们可能是追求一些前代的浪漫。 Output: 00038800464, correct • Example 3 Input: (sid=00038801261)人战胜了饥饿，才努力为了下一代作更好的、更健康的东西。 Output: 00038801261, 9, 9, M 00038801261, 16, 16, S (Notes: “能” is missing. The word “作”should be “做”. The correct sentence is “才能努力为了下一代做更好的”)

Table 1: Example sentences and corresponding notes.


```
<DOC>
<TEXT id="A2-0005-1">
我聽說你打算開一個慶祝會。對不起，我要參加，可是沒有空。你開一個慶祝會的時候我不能會參加，是因為我在外國做工作。
</TEXT>
<CORRECTION>
我聽說你打算開一個慶祝會。對不起，我要參加，可是沒有空。你開慶祝會的時候我不能參加，是因為我在外國工作。
</CORRECTION>
<ERROR start_off=”31” end_off=”32” type="R"></ERROR>
<ERROR start_off=”42” end_off=”42” type="R"></ERROR>
<ERROR start_off=”53” end_off=”53” type="R"></ERROR>
</DOC>
<DOC>
<TEXT id="200210543634250003_2_1x3">
对于“安乐死”的看法，向来都是一个极具争议性的题目，因为毕竟每个人对于死亡的观念都不一样，怎样的情况下去判断，也自然产生出很多主观和客观的理论。每个人都有着生存的权利，也代表着每个人都能去决定如何结束自己的生命的权利。在我的个人观点中，如果一个长期受着病魔折磨的人，会是十分痛苦的事，不仅是病人本身，以致病者的家人和朋友，都是一件难受的事。
</TEXT>
<CORRECTION>
对于“安乐死”的看法，向来都是一个极具争议性的题目，因为毕竟每个人对于死亡的观念都不一样，无论在怎样的情况下去判断，都自然产生出很多主观和客观的理论。每个人都有着生存的权利，也代表着每个人都能去决定如何结束自己的生命。在我的个人观点中，如果一个长期受着病魔折磨的人活着，会是十分痛苦的事，不仅是病人本身，对于病者的家人和朋友，都是一件难受的事。
</CORRECTION>
```

```
<ERROR start_off="46" end_off="46" type="M"></ERROR>
<ERROR start_off="56" end_off="56" type="S"></ERROR>
<ERROR start_off="106" end_off="108" type="R"></ERROR>
<ERROR start_off="133" end_off="133" type="M"></ERROR>
<ERROR start_off="151" end_off="152" type="S"></ERROR>
</DOC>
```

Figure 1: Example units in SGML format (in traditional and standard character).

Evaluation	**Track**	**#Units**	**#Error**	**#R**	**#M**	**#S**	**#W**
CGED2016	TOCFL	10,693	24,492 (100%)	4,472 (18.3%)	8,739 (35.7%)	9,897 (40.4%)	1,384 (5.7%)
	HSK	10,071	24,797 (100%)	5,538 (22.3%)	6,623 (26.7%)	10949 (44.2%)	1,687 (6.8%)
CGED2017	HSK	10,449	26,448 (100%)	5,852 (22.1%)	7,010 (26.5%)	11,591 (43.8%)	1,995 (7.5%)

Table 2: The statistics of training set.

Evaluation	**Track**	**#Units**	**#Correct**	**#Erroneous**	**#Error**	**#R**	**#M**	**#S**	**#W**
CGED2016	TOCFL	3,528	1,703 (48.3%)	1,825 (51.7%)	4,103 (100%)	782 (19.06%)	1,482 (36.12%)	1,613 (39.31%)	226 (5.51%)
	HSK	3,011	1,539 (51.1%)	1,472 (48.9%)	3,695 (100%)	802 (21.71%)	991 (26.82%)	1620 (43.84%)	282 (7.63%)
CGED2017	HSK	3,154	1,173 (48.4%)	1,628 (51.6%)	4,876 (100%)	1,062 (21.78%)	1,274 (26.13%)	2,155 (44.20%)	385 (7.90%)

Table 3: The statistics of testing set.

Confusion Matrix		System Results	
		Positive (Erroneous)	Negative(Correct)
Gold Standard	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Table 4: Confusion matrix for evaluation.

Participant (Ordered by abbreviations of names)	**#TRuns**	**F1**	**#HRuns**	**F1**
NLP Lab, Zhengzhou University (**ANO**)	0	-	2	0.2666
Central China Normal University (**CCNU**)	0	-	1	0.0121
Chaoyang University of Technology (**CYUT**)	3	**0.1248**	3	0.2125
Harbin Institute of Technology (**HIT**)	0	-	3	**0.3855**
Institute of Computational Linguistics, Peking University (**PKU**)	3		3	0.0724
National Chiao Tung University & National Taipei University of Technology (**NCTU+NTUT**)	3	0.0745	0	-
National Chiayi University (**NCYU**)	3	0.0155	3	0.0183
NLP Lab, Zhengzhou University (**SKY**)	0	-	3	0.3627
School of Information Science and Engineering, Yunnan University (**YUN-HPCC**)	3	0.0007	3	0.0035

Table 5: Submission statistics for all participants in CGED2016.

Participant (Ordered by abbreviations of names)	#Runs	F1
ALI_NLP	3	**0.2693**
BNU_ICIP	3	0.1152
CVTER	2	0.0653
NTOUA	2	0.0348
YNU-HPCC	3	0.1255

Table 6: Submission statistics for all participants in CGED2017.

8. References

Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo. 2012. Error diagnosis of Chinese sentences usign inductive learning algorithm and decomposition-based testing mechanism. ACM Transactions on Asian Language Information Processing, 11(1), article 3.

Xiliang Cui, Bao-lin Zhang. 2011. The Principles for Building the "International Corpus of Learner Chinese". Applied Linguistics, 2011(2), pages 100-108.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In Proceedings of the 13th European Workshop on Natural Language Generation(ENLG'11), pages 1-8, Nancy, France.

Reobert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposiiton and determiner error correction shared task. In Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications(BEA'12), pages 54-62, Montreal, Canada.

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 18th Conference on Computational Natural Language Learning (CoNLL'14): Shared Task, pages 1-12, Baltimore, Maryland, USA.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In Proceedings of the 17th Conference on Computational Natural Language Learning(CoNLL'13): Shared Task, pages 1-14, Sofia, Bulgaria.

Lung-Hao Lee, Li-Ping Chang, and Yuen-Hsien Tseng. 2016. Developing learner corpus annotation for Chinese grammatical errors. In Proceedings of the 20th International Conference on Asian Language Processing (IALP'16), Tainan, Taiwan.

Lung-Hao Lee, Li-Ping Chang, Kuei-Ching Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2013. Linguistic rules based Chinese error detection for second language learning. In Proceedings of the 21st International Conference on Computers in Education(ICCE'13), pages 27-29, Denpasar Bali, Indonesia.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'15), pages 1-6, Beijing, China.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang, and Hsin-Hsi Chen. 2014. A sentence judgment system for grammatical error detection. In Proceedings of the 25th International Conference on Computational Linguistics (COLING'14): Demos, pages 67-70, Dublin, Ireland.

Lung-Hao Lee, Rao Gaoqi, Liang-Chih Yu, Xun, Eendong, Zhang Baolin, and Chang Li-Ping. 2016. Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. The Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA' 16), pages 1-6, Osaka, Japan.

Chung-Hsien Wu, Chao-Hong Liu, Matthew Harris, and Liang-Chih Yu. 2010. Sentence correction incorporating relative position and parse template language models. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), pages 1170-1181.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. In Proceedings of the 24th International Conference on Computational Linguistics (COLING'12), pages 3003-3017, Bombay, India.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as foreign language. In Proceedings of the 1stWorkshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'14), pages 42-47, Nara, Japan.

Bao-lin Zhang, Xiliang Cui. 2013. Design Concepts of "the Construction and Research of the Inter-language Corpus of Chinese from Global Learners". Language Teaching and Linguistic Study, 2013(5), pages 27-34.