IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis

Gaoqi Rao[1], Baolin Zhang[2], Endong Xun[3]
{[1]Center for Studies of Chinese as a Second Language, [2]Faculty of Language Sciences,
[3]College of Information Science} Beijing Language and Culture University
raogaoqi-fj@163.com, zhangbl@blcu.edu.cn, edxun@126.com

Abstract

This paper presents the IJCNLP 2017 shared task for Chinese grammatical error diagnosis (CGED) which seeks to identify grammatical error types and their range of occurrence within sentences written by learners of Chinese as foreign language. We describe the task definition, data preparation, performance metrics, and evaluation results. Of the 13 teams registered for this shared task, 5 teams developed the system and submitted a total of 13 runs. We expected this evaluation campaign could lead to the development of more advanced NLP techniques for educational applications, especially for Chinese error detection. All data sets with gold standards and scoring scripts are made publicly available to researchers.

1 Introduction

Recently, 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) and COLING-2016 (Lee et al., 2016). In conjunction with the IJCNLP 2017, the shared task for Chinese grammatical error diagnosis is organized again. 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.

The rest of this paper is organized as follows. Section 2 describes the task in detail. Section 3 introduces the constructed datasets. Section 4 proposes evaluation metrics. Section 5 reports the results of the participants' approaches. Conclusions are finally drawn in Section 6.

2 Task Description

The goal of this shared task is to develop NLP techniques to automatically diagnose grammatical errors in Chinese sentences written by CFL learners. Such errors are defined as redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 1–8,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

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 and corresponding notes are shown as Table 1 shows. This year, we only have one track of HSK.

HSK (Simplified Chinese)

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”)

Example 4
Input: (sid=00038801320)饥饿的问题也是应该解决的。世界上每天由于饥饿很多人死亡。
Output: 00038801320, 19, 25, W
(Notes: “由于饥饿很多人” should be “很多人由于饥饿”)

Table 1: Example sentences and corresponding notes.

3 Datasets

The learner corpora used in our shared task were taken from the writing section of the Hanyu Shuiping Kaoshi(HSK, Test of Chinese Level)(Cui et al, 2011; Zhang et al, 2013).

Native Chinese speakers were trained to manually annotate grammatical errors and provide corrections corresponding to each error. The data were then split into two mutually exclusive sets as follows.

(1) Training Set: All units in this set were used to train the grammatical error diagnostic systems. Each unit contains 1 to 5 sentences with annotated grammatical errors and their corresponding corrections. All units are represented in SGML format, as shown in Table 2. We provide 10,449 training units with a total of 26,448 grammatical errors, categorized as redundant (5,852 instances), missing (7,010), word selection (11,591) and word ordering (1,995).

In addition to the data sets provided, participating research teams were allowed to use other public data for system development and implementation. Use of other data should be specified in the final system report.

```
<DOC>
<TEXT id="200307109523200140_2_2x3">
因为养农作物时不用农药的话，生产率较低。那肯定价格要上升，那有钱的人想吃
多少，就吃多少。左边的文中已提出了世界上的有几亿人因缺少粮食而挨饿。
</TEXT>
<CORRECTION>
因为种植农作物时不用农药的话，生产率较低。那价格肯定要上升，那有钱的人想
吃多少，就吃多少。左边的文中已提出了世界上有几亿人因缺少粮食而挨饿。
</CORRECTION>
<ERROR start_off="3" end_off="3" type="S"></ERROR>
<ERROR start_off="22" end_off="25" type="W"></ERROR>
<ERROR start_off="57" end_off="57" 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>
```

Table 2: A training sentence denoted in SGML format.

(2) Test Set: This set consists of testing sentences used for evaluating system performance. Table 3 shows statistics for the testing set for this year. About half of these sentences are correct and do not contain grammatical errors, while the other half include at least one error. The distributions of error types (shown in Table 4) are similar with that of the training set. The proportion of the correct sentences is sampled from data of the online Dynamic Corpus of HSK[1].

#Units	**#Correct**	**#Erroneous**
3,154 (100%)	1,173 (48.38%)	1,628 (51.62%)

Table 3: The statistics of correct sentences in testing set.

Error Type	
#R	1,062 (21.78%)
#M	1,274 (26.13%)
#S	2,155 (44.20%)
#W	385

[1] http://202.112.195.192:8060/hsk/login.asp

	(7.90%)
#Error	4,876 (100%)

Table 4: The distributions of error types in testing set.

4 Performance Metrics

Table 5 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. All error types will be regarded as incorrect.

(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.

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)

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 5: Confusion matrix for evaluation.

For example, for 4 testing inputs with gold standards shown as "00038800481, 6, 7, S", "00038800481, 8, 8, R", "00038800464, correct", "00038801261, 9, 9, M", "00038801261, 16, 16, S" and "00038801320, 19, 25, W", the system may output the result as "00038800481, 2, 3, S", "00038800481, 4, 5, S", "00038800481, 8, 8, R", "00038800464, correct", "00038801261, 9, 9, M", "00038801261, 16, 19, S" and "00038801320, 19, 25, M". The scoring script will yield the following performance.

- False Positive Rate (FPR) = 0 (=0/1)
- Detection-level
  - Accuracy = 1 (=4/4)
  - Precision = 1 (=3/3)
  - Recall = 1 (=3/3)
  - F1 = 1 (=(2*1*1)/(1+1))
- Identification-level
  - Accuracy = 0.8333 (=5/6)
  - Precision = 0.8 (=4/5)
  - Recall = 0.8 (=4/5)
  - F1 = 0.8 (=(2*0.8*0.8)/(0.8+08))
- Position-level
  - Accuracy = 0.4286 (=3/7)
  - Precision = 0.3333 (=2/6)
  - Recall = 0.4 (=2/5)
  - F1 = 0.3636 (=(2*0.3333*0.4)/(0.3333+0.4))

5 Evaluation Results

Table 6 summarizes the submission statistics for the 13 participating teams including 10 from universities and research institutes in China (NTOUA, BLCU, SKY, PkU-Cherry, BNU_ICIP, CCNUNLP, CVTER, TONGTONG, AL_I_NLP), 1 from the U.S. (Harvard University) and 1 private firm (Lingosail Inc.). In the official testing phase, each participating team was allowed to submit at most three runs. Of the

13 registered teams, 5 teams submitted their testing results, for a total of 13 runs.

Participant (Ordered by abbreviations of names)	#Runs
ALI_NLP	3
BLCU	0
BNU_ICIP	3
CCNUNLP	0
Cherry	0
CVTER	2
Harvard University	0
NTOUA	2
PkU	0
SKY	0
TONGTONG	0
YNU-HPCC	3
Lingosail	0

Table 6: Submission statistics for all participants.

Table 7 shows the testing results of CGED2017. The BNU team achieved the lowest false positive rate (denoted as "FPR") of 0.098. Detection-level evaluations are designed to detect whether a sentence contains grammatical errors or not. A neutral baseline can be easily achieved by always reporting all testing sentences as correct without errors. According to the test data distribution, the baseline system can achieve an accuracy of 0.5162. However, not all systems performed above the baseline. The system result submitted by ALI_NLP achieved the best detection accuracy of 0.6465. We use the F1 score to reflect the tradeoffs between precision and recall. The ALI_NLP provided the best error detection results, providing a high F1 score of 0.8284. For identification-level evaluations, the systems need to identify the error types in a given sentences. The system developed by YNU-HPCC provided the highest F1 score of 0.7829 for grammatical error identification. For position-level evaluations, ALI_NLP achieved the best F1 score of 0.2693. Perfectly identifying the error types and their corresponding positions is difficult in part because no word delimiters exist among Chinese words in the given sentences.

NTOUA, CVTE and ALI_NLP submit reports on their develop systems. Though neural networks achieved good performances in various NLP tasks, traditional pipe-lines were still widely implemented in the CGED task. LSTM+CRF has been a standard implementation. Unlike CGED2016, though CRF model in pipeline were only equipped with simple designed feature templates.

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 CFL contexts. From organizers' perspectives, a good system should have a high F1 score and a low false positive rate. Overall, ALI_NLP, YNU-HPCC and CVTE achieved relatively better performances.

TEAM	RUNs	False Positive Rate	Accuracy	Precision	Recall	F1	Accurac y	Precision	Recall	F1	Accuracy	Precision	Recall	F1
				Detection Level				Identification Level				Position Level		
YNU-HPCC	run3	0.6104 (716/1173)	0.5311	0.6298	0.6148	0.6222	0.3979	0.4086	0.3298	0.365	0.1702	0.0981	0.0698	0.0816
	run2	0.7383 (866/1173)	0.5891	0.6417	0.7829	0.7053	0.3879	0.3825	0.4575	0.4167	0.1426	0.1056	0.1191	0.112
	run1	0.6513 (764/1173)	0.5796	0.65	0.7163	0.6816	0.4218	0.4219	0.4217	0.4218	0.1778	0.1262	0.1191	0.1225
NTOUA	run2	1 (1173/1173)	0.6281	0.6281	**1**	**0.7716**	0.3889	0.3889	0.506	0.4398	0.018	0.018	0.082	0.0295
	run1	1 (1173/1173)	0.6281	0.6281	**1**	**0.7716**	0.3211	0.3211	**0.6099**	0.4207	0.0212	0.0212	0.0958	0.0348
CVTE	run2	0.3154 (370/1173)	0.539	0.708	0.4528	0.5523	0.4711	0.5391	0.2057	0.2978	0.2602	0.1093	0.0465	0.0653
	run1	0.1441 (169/1173)	0.4756	0.7459	0.2504	0.3749	0.4461	**0.606**	0.1214	0.2023	0.3314	0.118	0.0204	0.0348
BNU	run3	0.1893 (222/1173)	0.5181	0.7547	0.3448	0.4733	0.4696	0.5707	0.1786	0.2721	0.3798	0.2968	0.0715	0.1152
	run2	0.1355 (159/1173)	0.4794	0.758	0.2514	0.3776	0.4412	0.5527	0.131	0.2118	0.3735	0.2818	0.0515	0.0871
	run1	**0.098 (115/1173)**	0.4721	**0.7894**	0.2176	0.3411	0.4337	0.5474	0.106	0.1776	0.3775	0.2773	0.0418	0.0727
AL_I_NLP	run3	0.3052 (358/1173)	0.6173	0.7597	0.5714	0.6523	**0.5513**	0.6007	0.3756	0.4622	**0.4121**	**0.3663**	0.213	**0.2693**
	run2	0.6607 (775/1173)	**0.6465**	0.6792	0.8284	0.7464	0.4654	0.453	0.6006	0.5164	0.2264	0.1949	**0.2941**	0.2344
	run1	0.6172 (724/1173)	0.6439	0.686	0.7986	0.738	0.488	0.4791	0.5657	**0.5188**	0.2547	0.2169	0.2752	0.2426

Table 7: Testing results of CGED2017.

6 Conclusions

This study describes the NLP-TEA 2016 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 http://www.cged.science.

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 proceeding of NLPTEA2016). Gong Qi, Tang Peilan, Luo Ping and Chang Jie contributed in the proofreading of data.

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).

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 Interlanguage Corpus of Chinese from Global Learners". Language Teaching and Linguistic Study, 2013(5), pages 27-34.