Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnosis

Gaoqi Rao Erhong Yang Baolin Zhang
Beijing Language and Culture University
{raogaoqi, yangerhong, zhangbaolin}@blcu.edu.cn

Abstract

This paper presents the NLPTEA 2020 shared task for Chinese Grammatical Error Diagnosis (CGED) which seeks to identify grammatical error types, their range of occurrence and recommended corrections within sentences written by learners of Chinese as a foreign language. We describe the task definition, data preparation, performance metrics, and evaluation results. Of the 30 teams registered for this shared task, 17 teams developed the system and submitted a total of 43 runs. System performances achieved a significant progress, reaching F1 of 91% in detection level, 40% in position level and 28% in correction level. All data sets with gold standards and scoring scripts are made publicly available to researchers.

1 Introduction

Automated grammar checking for learners of English as a foreign language has achieved obvious progress. 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 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), neuro network modelling (Zheng et al., 2016; Fu et al., 2018) and hybrid methods (Lee et al., 2014; Zhou et al., 2017).

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). Its name was fixed from then on: Chinese Grammatical Error Diagnosis (CGED). As a part of IJCNLP 2017, the shared task was organized (Rao et al., 2017). In conjunction with NLP-TEA workshop in ACL 2018, CGED was organized again (Rao et al., 2018). 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 data sets. 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 (and furtherly correct) grammatical errors in Chinese sentences written by CFL learners. Such errors are

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 25–35
Suzhou, China, December 4, 2020.

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

Hanyu Shuiping Kaoshi (HSK)
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 Data Sets

The learner corpora used in our shared task were taken from the writing section of the HSK (Pinyin of *Hanyu Shuiping Kaoshi*, 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 1129 training units with a total of 2,909 grammatical errors, categorized as redundant (678 instances), missing (801), word selection (1228) and word ordering (201).

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.

#Units	**#Correct**	**#Erroneous**
1,457 (100%)	307 (21.07%)	1,150 (78.93%)

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

Test Set: This set consists of testing units used for evaluating system performance. Table 3 shows statistics for the testing set for this year. According to the sampling in the writing sessions in HSK, over 40% of the sentences contain no error. This was simulated in the test set, in order to test the performance of the systems in false positive identification. The distributions of error types (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].

Error Type	
#R	769 (21.05%)
#M	864 (23.65%)
#S	1694 (46.36%)
#W	327 (8.95%)
#Error	3,654 (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.

Besides the traditional criteria in the past share tasks, Correction-level was introduced to CGED since 2018.

(4) Correction-level: For the error types of Selection and Missing, recommended corrections are required. At most 3 recommended corrections are allowed for each S and M type error. In this level the amount of the corrections recommended would influence the precision and F1 in this level. The trust of the recommendation would be test. The sub-track TOP1 count only one recommended correction, while TOP3 count one hit, if one correction in three hits the golden standard, ignoring its ranking.

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)

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: Precision = 1 (=3/3)
Recall = 1 (=3/3)
F1 = 1 (=(2*1*1)/(1+1))
Identification-level: Precision = 0.8 (=4/5)
Recall = 0.8 (=4/5)
F1 = 0.8 (=(2*0.8*0.8)/(0.8+08))
Position-level: Precision = 0.3333 (=2/6)
Recall = 0.4 (=2/5)
F1 = 0.3636 (=(2*0.3333*0.4)/(0.3333+0.4))

[1] http://bcc.blcu.edu.cn/hsk

<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.

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.

5 Evaluation Results

Table 6 summarizes the submission statistics for the 17 participating teams. In the official testing phase, each participating team was allowed to submit at most three runs. Of the 17 teams, 11 teams submitted their testing results in Correction-level, for a total of 43 runs.

Participant (Ordered by names)	**#Runs**	**Correction-level**

Boli	2	√
CYUT	2	-
DumbCat	1	√
Flying	3	√
LDU	3	-
NJU-NLP	3	-
OrangePlus	3	√
PCJG	3	√
SDU_MLA	1	-
SPPD	3	-
TextCC-CloudPoineer	3	√
TMU-NLP	1	√
UNIPUS-Flaubert	3	√
XHJZ	3	√
YD_NLP	3	√
ZZUNLP-HAN	3	√
ZZUNLP-YAN	3	-

Table 6: Submission statistics for all participants.

Table 7 to 11 show the testing results of the CGED2020 in 6 tracks: false positive rate (FPR), detection level, identification level, position level and correction level (in two settings: top1 and top3). All runs of top F1 score are highlighted in the tables. The CYUT achieved the lowest FPR of 0.0163, about one third of the lowest FPR in the CGED 2018. Detection-level evaluations are designed to detect whether a sentence contains grammatical errors or not. A neutral baseline can be easily achieved by reporting all testing sentences containing errors. According to the test data distribution, the baseline system can achieve an accuracy of 0.7893. However, not all systems performed above the baseline. The system result submitted by NJU-NLP achieved the best detection F1 of 0.9122, beating the 0.9 mark for the first time. For identification-level evaluations, the systems need to identify the error types in a given unit. The system developed by Flying and OrangePlus provided the highest F1 score of 0.6736 and 0.6726 for grammatical error identification. For position-level, Flying achieved the best F1 score of 0.4041, crossing the 0.4 mark for the first time. OrangePlus reached 0.394. Perfectly identifying the error types and their corresponding positions is difficult because the error propagation is serious. In correction-level, UNIPUS-Flaubert achieved best F1 of 0.1891 in top1 setting and YD_NLP of 0.1885 top3 setting.

In CGED 2020, the implementation of pre-trained model like BERT achieved significant improvement in many tracks. The "standard pipe-line" biLSTM+CRF in CGED2017 and 2018 is replaced. Hybrid methods based on pre-trained model were proposed by most of the teams. ResNet, graph convolution network and data argumentation appeared for the first time in the solutions. The rethinking the data construction (including pseudo data generation) and feature selection did not attract the attention of the participants. However, the balance of the FPR and other track did not progress a lot. The rough merging strategies implemented in hybrid methods and the over generation of generation models may lead the drop in FPR. From organizers' perspectives, a good system should have a high F1 score and a low false positive rate.

In summary, none of the submitted systems provided a comprehensive superior performance using different metrics, indicating the difficulty of developing systems for effective grammatical error diagnosis, especially in CFL contexts. It is worth noting that in the track of detection, the performance over 0.9 is close to the application of actual scene. In the highly focused track of position and correction, variant teams lead the ranks, unlike the past CGEDs. It's a very exciting phenomena indicating the attraction the task increased quickly.

TEAM Name	Run	FPR	TEAM Name	Run	FPR

Boli	1	0.7590	SPPD	1	0.1498
	2	0.7687		2	0.1107
CYUT	1	**0.0163**		3	0.0749
	2	0.5472	TextCC-CloudPoineer	1	0.2476
DumbCat	1	0.2052		2	0.2834
Flying	1	0.1010		3	0.4104
	2	0.2573	TMU-NLP	1	0.1726
	3	0.3257	UNIPUS-Flaubert	1	0.2508
LDU	1	0.0423		2	0.2443
	2	0.0489		3	0.4756
	3	0.0391	XHJZ	1	0.8762
NJU-NLP	1	0.6124		2	0.7752
	2	0.2378		3	0.7068
	3	0.0554	YD_NLP	1	0.2052
OrangePlus	1	0.2443		2	0.2345
	2	0.2964		3	0.2182
	3	0.2606	ZZUNLP-HAN	1	0.6645
PCJG	1	0.5440		2	0.6775
	2	0.8176		3	0.7394
	3	0.3844	ZZUNLP-YAN	1	0.8078
SDU_MLA	1	0.5179		2	0.7557
				3	0.6938

Table7. Results of CGED 2020 in False Positive Rate (FPR)

TEAM Name	**RUN**	**Detection Level**			**TEAM Name**	**RUN**	**Detection Level**		
		Pre.	**Rec.**	**F1**			**Pre.**	**Rec.**	**F1**
Boli	1	0.8149	0.8922	0.8518	SPPD	1	0.9541	0.8313	0.8885
	2	0.814	0.8983	0.8541		2	0.9649	0.8139	0.8830
CYUT	1	0.9875	0.3443	0.5106		3	0.9743	0.7574	0.8523
	2	0.8117	0.6296	0.7091	TextCC-CloudPoineer	1	0.9265	0.7565	0.8329
DumbCat	1	0.9078	0.5391	0.6765		2	0.9182	0.7809	0.8440
Flying	1	0.9649	0.7409	0.8382		3	0.8784	0.7913	0.8326
	2	0.9273	0.6213	0.6736	TMU-NLP	1	0.9404	0.7270	0.8200
	3	0.9101	0.8800	**0.8948**	UNIPUS-Flaubert	1	0.9214	0.7852	0.8479
LDU	1	0.9851	0.7496	0.8514		2	0.9207	0.7574	0.8311
	2	0.9828	0.7452	0.8477		3	0.8782	0.9157	**0.8966**
	3	0.9851	0.6887	0.8106	XHJZ	1	0.8062	0.9730	0.8818
NJU-NLP	1	0.8565	0.9757	**0.9122**		2	0.8069	0.5874	0.6799
	2	0.9303	0.8478	0.8872		3	0.8180	0.8478	0.8326
	3	0.9739	0.5513	0.7041	YD_NLP	1	0.9387	0.8383	0.8857
OrangePlus	1	0.9282	0.8435	0.8838		2	0.9319	0.8565	**0.8926**
	2	0.9161	0.8643	0.8895		3	0.9357	0.8478	0.8896
	3	0.9252	0.8600	**0.8914**	ZZUNLP-HAN	1	0.8262	0.8435	0.8348
PCJG	1	0.8225	0.6730	0.7403		2	0.8145	0.7939	0.8041
	2	0.8142	0.9565	0.8796		3	0.8136	0.8617	0.8370

	3	0.8698	0.6852	0.7665		1	0.8118	0.9304	0.8671
SDU_MLA	1	0.8138	0.5965	0.6884	ZZUNLP-YAN	2	0.8182	0.9078	0.8607
						3	0.8254	0.8757	0.8498

Table8. Results of CGED 2020 in Detection Level

TEAM Name	RUN	Identification Level			TEAM Name	RUN	Identification Level		
		Pre.	Rec.	F1			Pre.	Rec.	F1
Boli	1	0.5883	0.5347	0.5602	SPPD	1	0.7166	0.5892	0.6467
	2	0.5872	0.5389	0.5620		2	0.7600	0.5676	0.6499
CYUT	1	0.6412	0.166	0.2637		3	0.7843	0.4862	0.6003
	2	0.4902	0.2768	0.3538	TextCC-CloudPoineer	1	0.7090	0.4982	0.5852
DumbCat	1	0.7002	0.3929	0.5034		2	0.7034	0.5285	0.6035
Flying	1	0.7769	0.4738	0.5886		3	0.6751	0.5051	0.5779
	2	0.7356	0.6213	**0.6736**	TMU-NLP	1	0.6980	0.4228	0.5266
	3	0.7320	0.6011	0.6601	UNIPUS-Flaubert	1	0.7415	0.4890	0.5893
LDU	1	0.5714	0.6897	0.6250		2	0.7515	0.4710	0.5791
	2	0.5715	0.6874	0.6241		3	0.6507	0.6420	0.6463
	3	0.75	0.2772	0.4048	XHJZ	1	0.5669	0.6714	0.6147
NJU-NLP	1	0.5571	0.8432	0.6709		2	0.5897	0.6011	0.5953
	2	0.7018	0.5779	0.6339		3	0.6063	0.5873	0.5966
	3	0.7939	0.2975	0.4328	YD_NLP	1	0.7788	0.5503	0.6449
OrangePlus	1	0.7223	0.6121	0.6627		2	0.7623	0.5678	0.6508
	2	0.7188	0.5450	0.6200		3	0.7711	0.5577	0.6473
	3	0.7230	0.6287	**0.6726**	ZZUNLP-HAN	1	0.5856	0.4416	0.5035
PCJG	1	0.6136	0.3154	0.4166		2	0.5053	0.4127	0.4543
	2	0.5926	0.5678	0.5799		3	0.5018	0.5060	0.5039
	3	0.6499	0.3687	0.4705	ZZUNLP-YAN	1	0.5899	0.5126	0.5485
SDU_MLA	1	0.5411	0.2813	0.3701		2	0.6150	0.5076	0.5562
						3	0.64	0.5214	0.5746

Table9. Results of CGED 2020 in Identification Level

TEAM Name	RUN	Position Level			TEAM Name	RUN	Position Level		
		Pre.	Rec.	F1			Pre.	Rec.	F1
Boli	1	0.2284	0.1719	0.1962	SPPD	1	0.3595	0.2671	0.3065
	2	0.2284	0.1755	0.1985		2	0.4225	0.2822	0.3384
CYUT	1	0.0134	0.0033	0.0053		3	0.4673	0.2466	0.3228
	2	0.0136	0.0068	0.0091	TextCC-CloudPoineer	1	0.3612	0.2392	0.2878
DumbCat	1	0.3565	0.1828	0.2417		2	0.3518	0.2518	0.2935
Flying	1	0.4970	0.2529	0.3352		3	0.3577	0.2318	0.2813
	2	0.4320	0.3514	0.3876	TMU-NLP	1	0.3460	0.1639	0.2224
	3	0.4715	0.3536	**0.4041**	UNIPUS-Flaubert	1	0.4758	0.2343	0.3140
LDU	1	0.1397	0.1612	0.1497		2	0.4606	0.2288	0.3057
	2	0.1407	0.1621	0.1506		3	0.3147	0.2739	0.2929

	3	0	0	0.0000	XHJZ	1	0.2368	0.2849	0.2586
NJU-NLP	1	0.2097	0.4648	0.2890		2	0.2610	0.2663	0.2636
	2	0.4008	0.288	0.3351		3	0.2993	0.2655	0.2814
	3	0.5757	0.1519	0.2404	YD_NLP	1	0.5145	0.2965	0.3762
OrangePlus	1	0.4366	0.3372	0.3805		2	0.4822	0.3011	0.3707
	2	0.4241	0.2731	0.3323		3	0.5011	0.2995	0.3749
	3	0.4428	0.361	**0.3977**	ZZUNLP-HAN	1	0.2502	0.1472	0.1854
PCJG	1	0.0885	0.0342	0.0494		2	0.0996	0.0665	0.0798
	2	0.2582	0.2143	0.2342		3	0.067	0.0613	0.0640
	3	0.3282	0.1399	0.1962	ZZUNLP-YAN	1	0.29	0.1941	0.2326
SDU_MLA	1	0.0708	0.0276	0.0398		2	0.2874	0.1892	0.2282
						3	0.2783	0.2042	0.2356

Table10. Results of CGED 2020 in Position Level

TEAM Name	**RUN**	**Correction Level（TOP1）**			**Correction Level（TOP3）**		
		Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Boli	1	0.079	0.0629	0.0700	0.079	0.0629	0.0700
	2	0.0768	0.0629	0.0692	0.0768	0.0629	0.0692
DumbCat	1	0.2502	0.1126	0.1553	0.2502	0.1126	0.1553
Flying	1	0.246	0.1149	0.1567	0.246	0.1149	0.1567
	2	0.2105	0.154	0.1779	0.2105	0.154	0.1779
	3	0.229	0.1575	**0.1867**	0.229	0.1575	**0.1867**
OrangePlus	1	0.1356	0.1095	0.1211	0.0766	0.1837	0.1081
	2	0.1886	0.1247	0.1502	0.0961	0.1767	0.1245
	3	0.178	0.1536	0.1649	0.0934	0.2283	0.1325
PCJG	1	0.0492	0.0233	0.0307	0.0492	0.0223	0.0307
TextCC-CloudPoineer	1	0.1737	0.1247	0.1452	0.0983	0.1454	0.1173
	2	0.1696	0.1341	0.1498	0.0973	0.156	0.1198
TMU-NLP	1	0.2258	0.1032	0.1417	0.2258	0.1032	0.1417
UNIPUS-Flaubert	1	0.2848	0.1415	**0.1891**	0.2276	0.1595	0.1876
	2	0.2587	0.1372	0.1793	0.1582	0.1646	0.1613
	3	0.2014	0.1603	0.1785	0.1339	0.188	0.1564
XHJZ	1	0.1293	0.1763	0.1492	0.1293	0.1763	0.1492
	2	0.1465	0.1646	0.1550	0.1465	0.1646	0.1550
	3	0.1764	0.1646	0.1703	0.1764	0.1646	0.1703
YD_NLP	1	0.3238	0.1290	**0.1845**	0.2982	0.1372	**0.1879**
	2	0.3293	0.1263	**0.1826**	0.3132	0.1337	**0.1874**
	3	0.3386	0.1259	**0.1836**	0.3217	0.1333	**0.1885**
ZZUNLP-HAN	1	0.0027	0.0012	0.0017	0.0018	0.002	0.0019
	2	0.0009	0.0004	0.0006	0.0007	0.0008	0.0007

Table11. Results of CGED 2020 in Correction Level

6 Conclusion

This study describes the NLP-TEA 2020 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. Lung-Hao Lee helped a lot in consultation and bidding. Xiangyu Chi, Mengyao Suo, Yuhan Wang and Shufan Zhou contributed a lot in data reviewing.

This study was supported by the projects from National Language Committee Project (YB135-90).

References

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 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.

Gaoqi Rao, Baolin Zhang, Endong Xun, Lung-Hao Lee. IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis. *In Proceedings of the IJCNLP 2017*, Shared Tasks, Taipei, Taiwan: 1-8

Gaoqi Rao, Qi Gong, Baolin Zhang, Endong Xun. Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis. 2018. *In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'18), pages 42-51,* Melbourne, Australia.

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.

Bo Zheng, Wanxiang Che, Jiang Guo, Ting Liu. 2016. Chinese Grammatical Error Diagnosis with Long Short-Term Memory Networks. *In proceedings of 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'16)*, Osaka, Japan, December 2016, pages 49–56.

Xin Zhou, Jian Wang, Xu Xie, Changlong Sun, Luo Si. Alibaba at IJCNLP-2017 Task 2: A Boosted Deep System for Dimensional Sentiment Analysis of Chinese Phrases. *In proceedings of the IJCNLP 2017*, Shared Tasks, pages 100–110, Taipei, China.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, Ting Liu. Chinese Grammatical Error Diagnosis using Statistical and Prior Knowledge driven Features with Probabilistic Ensemble Enhancement. *In Proceedings of*

the 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'18), pages 52–59, Melbourne, Australia.