

Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis

Gaoqi Rao Qi Gong Baolin Zhang Endong Xun
Beijing Language and Culture University
{raogaoqi , gongqi , zhangbaolin , xunendong}@blcu.edu.cn

Abstract

This paper presents the NLPTEA 2018 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 foreign language. We describe the task definition, data preparation, performance metrics, and evaluation results. Of the 20 teams registered for this shared task, 13 teams developed the system and submitted a total of 32 runs. Progress in system performances was obviously, reaching F1 of 36.12% in position level and 25.27% 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; Zhou et al., 2017) 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). 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 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 (and furtherly correct) grammatical errors in Chinese sentences written by CFL learners. Such errors are 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.

HSK (Simplified Chinese)
<p>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.)</p> <p>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</p> <p>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”)</p> <p>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”)</p>

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 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 402 training units with a total of 1,067 grammatical errors, categorized as redundant (208 instances), missing (298), word selection (474) and word ordering (87).

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
--------	----------	------------

3,549 (100%)	1,562 (44.01%)	1,987 (55.99%)
--------------	----------------	----------------

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 (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 HSK¹.

Error Type	
#R	1,119 (22.20%)
#M	1,381 (27.40%)
#S	2,167 (43.00%)
#W	373 (7.40%)
#Error	5,040 (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 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 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 = $TP / (TP+FN)$

$$F1 = 2 * Precision * Recall / (Precision + Recall)$$

- 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
 - Precision = 0.8 (=4/5)
 - Recall = 0.8 (=4/5)
 - F1 = 0.8 (= (2*0.8*0.8)/(0.8+0.8))
- Position-level
 - Precision = 0.3333 (=2/6)
 - Recall = 0.4 (=2/5)
 - F1=0.3636
(=(2*0.3333*0.4)/(0.3333+0.4))
- Correction-level
 - Precision = 0.125 (=1/8)
 - Recall = 0.3333 (=1/3)
 - F1=0.1818
(=(2*0.3333*0.125)/(0.3333+0.125))

¹ <http://bcc.blcu.edu.cn/hsk>

- Correction-level (Top3)
 - Precision = 0.3333 (=1/3)
 - Recall = 0.3333 (=1/3)
- F1=0.3333
 (= (2*0.3333*0.3333)/(0.3333+0.3333))

```

<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。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 12 participating teams including 10 from universities and research institutes in China (AutoNLP, BUPT, CYUT-III, ECNU, HFL, CMMC,

NCYU, NTOU, PkU_ICL), 1 from the U.S. (UIUC) and 1 from India (IIT). Two teams (HFL and DM_NLP) of enterprises are all from China. In the official testing phase, each participating team was allowed to submit at most three runs. Of the 12 registered teams, 8 teams submitted their testing results in Correction-level, for a total of 32 runs.

Participant (Ordered by names)	#Runs	Correction-level
AutoNLP	3	√
BUPT	3	√
CYUT-III	3	√
DM_NLP	3	√
ECNU	3	-
HFL	3	√
IIT (BHU)	1	√
CMMC-BDRC	3	√
NCYU	3	√
NTOU	1	-
PkU_ICL	3	√
UIUC	2	-
Walker	1	-

Table 6: Submission statistics for all participants.

Table 7 shows the testing results of the CGED2018. The CYUT-III achieved the lowest false positive rate (denoted as “FPR”) of 0.0499, about half of the lowest FPR in the CGED 2017. 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 0.5599. However, not all systems performed above the baseline. The system result submitted by HFL achieved the best detection accuracy of 0.7578 and CMMC-BDRC in F1 of 0.7563. For identification-level evaluations, the systems need to identify the error types in a given unit. The system developed by HFL provided the highest F1 score of 0.5503 for grammatical error identification. For position-level evaluations, HFL achieved the best F1 score of 0.3612. 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.

In correction-level, DM_NLP achieved best precision (0.2932 and 0.3077) in correction and top3 correction track. HFL’s runs reached best F1 of 0.1723 and 0.2527.

10 participants submitted 11 reports on their systems. Though neural networks achieved good performances in various NLP tasks, traditional statistic models and pipe-lines were still widely implemented in the CGED task. LSTM+CRF has been a standard implementation. Unlike CGED 2017, participants began to rethink the importance of the feature selection and statistics.

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, HFL, DM_NLP, and CMMC-BDRC achieved relatively better performances.

TEAM	Runs	FPR	Detection				Identification			Position		
			Acc.	pre	rec	F1	pre	re	F1	pre	rec	F1
AutoNLP	run1	0.3301	0.5131	0.6349	0.4232	0.5079	0.4792	0.1995	0.2817	0.1185	0.0442	0.0644
	run2	0.1642	0.4897	0.6698	0.2494	0.3634	0.5139	0.1323	0.2105	0.1585	0.0331	0.0547
	run3	0.4715	0.4996	0.6346	0.5426	0.5850	0.4735	0.2646	0.3395	0.1129	0.0609	0.0792
BUPT	run1	0.8412	0.5711	0.5752	0.8953	0.7004	0.3506	0.5663	0.4331	0.0482	0.0882	0.0623
	run2	0.5019	0.6005	0.6331	0.6809	0.6562	0.4134	0.3519	0.3802	0.0608	0.0504	0.0551
	run3	0.5480	0.6236	0.6377	0.7584	0.6929	0.4084	0.4161	0.4122	0.0630	0.0609	0.0620
CYUT-III	run1	0.0499	0.4683	0.6953	0.0896	0.1587	0.5426	0.0418	0.0776	0.0586	0.0032	0.0060
	run2	0.1780	0.6016	0.7535	0.4282	0.5461	0.5433	0.2790	0.3687	0.1470	0.0711	0.0959

	run3	1.0000	0.4728	0.5805	0.8448	0.6881	0.2589	0.2640	0.2614	0.0070	0.0173	0.0100
DM_NLP	run1	0.3214	0.6131	0.6897	0.5617	0.6191	0.4038	0.3657	0.3838	0.2924	0.1842	0.2260
	run2	0.2183	0.6174	0.7399	0.4882	0.5882	0.5943	0.3113	0.4086	0.3900	0.1777	0.2441
	run3	0.2279	0.6238	0.7390	0.5073	0.6016	0.5877	0.3242	0.4179	0.3855	0.1850	0.2500
ECNU	run1	0.3470	0.5923	0.6663	0.5445	0.5993	0.4767	0.2836	0.3556	0.1238	0.0667	0.0867
	run2	0.3873	0.5796	0.6452	0.5536	0.5959	0.4452	0.2740	0.3392	0.0901	0.0506	0.0648
	run3	0.1255	0.5762	0.7760	0.3417	0.4745	0.6139	0.1818	0.2805	0.3745	0.0858	0.1397
HFL	run1	0.1613	0.7101	0.8276	0.6090	0.7017	0.7107	0.4173	0.5259	0.5341	0.2729	0.3612
	run2	0.7554	0.6436	0.6171	0.9572	0.7504	0.3931	0.7331	0.5118	0.1441	0.3886	0.2102
	run3	0.1754	0.7278	0.8254	0.6517	0.7283	0.6874	0.4588	0.5503	0.4752	0.2906	0.3606
IIT (BHU)	run1	0.4190	0.4483	0.5668	0.3889	0.4613	0.2737	0.1705	0.2102	0.0071	0.0030	0.0042
CMMC-BDRC	run1	0.5314	0.6889	0.6736	0.8621	0.7563	0.4834	0.5952	0.5335	0.2741	0.3177	0.2943
	run2	0.3547	0.6988	0.7266	0.7408	0.7336	0.5831	0.4955	0.5357	0.3839	0.2966	0.3346
	run3	0.3470	0.6630	0.7109	0.6709	0.6903	0.4853	0.4096	0.4442	0.2482	0.1814	0.2096
NCYU	run1	0.9987	0.5596	0.5598	0.9985	0.7174	0.2381	0.9749	0.3828	0.0030	0.0390	0.0056
	run2	0.9994	0.5599	0.5599	0.9995	0.7177	0.2382	0.9752	0.3828	0.0030	0.0384	0.0056
	run3	0.9994	0.5599	0.5599	0.9995	0.7177	0.2382	0.9752	0.3828	0.0030	0.0380	0.0055
NTOUA	run1	0.9481	0.5323	0.5497	0.9099	0.6854	0.3297	0.5812	0.4207	0.0065	0.0191	0.0096
PKU_ICL	run1	0.5538	0.6388	0.6448	0.7901	0.7101	0.4483	0.4737	0.4607	0.1642	0.1605	0.1624
	run2	0.2298	0.6317	0.7432	0.5229	0.6139	0.5567	0.3018	0.3914	0.2868	0.1309	0.1797
	run3	0.5679	0.6267	0.6359	0.7796	0.7004	0.4433	0.4710	0.4567	0.1615	0.1615	0.1615
UIUC	run1	0.1274	0.5540	0.7519	0.3035	0.4324	0.6311	0.1696	0.2673	0.2385	0.0536	0.0875
	run2	0.1274	0.5540	0.7519	0.3035	0.4324	0.6311	0.1696	0.2673	0.2385	0.0536	0.0875
walker	run1	0.9309	0.5441	0.5562	0.9179	0.6926	0.3144	0.6266	0.4187	0.0078	0.0189	0.0110

Table7. Results of CGED 2018 in Detection-level, Identification-level and Position-level

TEAM	Runs	Correction			Top3 Correction	
		pre	rec	F1	pre	F1
AutoNLP	run1	0.1667	0.0110	0.0206	0.1667	0.0206
	run2	0.1626	0.0113	0.0211	0.1626	0.0211
	run3	0.1626	0.0113	0.0211	0.1626	0.0211
BUPT	run1	0.0046	0.0093	0.0062	0.0046	0.0062
	run2	0.0033	0.0028	0.0030	0.0033	0.0030
	run3	0.0092	0.0087	0.0090	0.0092	0.0090
CYUT-III	run1	0.0040	0.0008	0.0014	0.0040	0.0014
DM_NLP	run1	0.2603	0.0161	0.0303	0.2701	0.0314
	run2	0.2932	0.0158	0.0299	0.3077	0.0314
	run3	0.2700	0.0180	0.0338	0.2832	0.0355
HFL	run1	0.2087	0.1468	0.1723	0.3059	0.2527
	run2	0.0386	0.1696	0.0629	0.0722	0.1177
	run3	0.1509	0.1400	0.1453	0.2391	0.2301

IIT (BHU)	run1	0.0000	0.0000	0.0000	0.0000	0.0000
CMMC-BDRC	run1	0.1364	0.1651	0.1494	0.1432	0.1569
	run2	0.1852	0.1609	0.1722	0.1934	0.1798
	run3	0.2126	0.1395	0.1685	0.2190	0.1735
NCYU	run1	1.2079E-05	0.0003	2.3164E-05	3.6236E-05	6.9493E-05
	run2	3.6235E-05	8.4531E-04	6.9490E-05	1.0870E-04	2.0847E-04
	run3	3.6235E-05	8.4531E-04	6.9490E-05	1.0870E-04	2.0847E-04
PkU_ICL	run1	0.0296	0.0775	0.0429	0.0822	0.1189
	run2	0.0556	0.0662	0.0604	0.1522	0.1655
	run3	0.0316	0.0814	0.0456	0.0881	0.1270

Table 8. Results of CGED 2018 in Correction-level

6 Discussions

Table 9 summarizes the approaches and resources for each of the submitted systems, according to their 1st draft of system reports (some details were not clearly described yet). PkU_ICL, NCYU and IIT(BHU) did not submit reports on their systems. Though neural networks achieved good performances in various NLP tasks, traditional pipe-lines were still widely implemented in the CGED task. CRF, as a sequence labelling model with flexible feature space, was chosen by DM_NLP, CMMC, ECNU, HFL, walker and UIUC in their system pipe-lines. Further, UIUC applied its pipe-line only with CRF and post processing, achieving comparable results. NTOU conducted their runs based on frequent subsentences matching in internet corpus.

For LSTM modelling, feature choice played an important role, influencing the system performance a lot. Besides character and word, part of speech (POS) based on the segmentation, are widely selected. ePMI, cPMI, Adjacent Word Collocation (AWC), Dependent Word Collocation (DWC), Contextualized Char

Representation are newly implemented features in this task.

For LSTM itself, AutoNLP applied policy gradient in modelling. Some participant added additional memory gate in the neuro, a quite normal trick in machine translation, helping their system achieve high F1 score over 50% in position-level and over 40% in correction-level. The submissions were withdrawn, due to the suspected overfitting of testing set. Although it cannot reflect the real achievement in this task, the phenome is still meaningful in particular context, like computer assistant essay correction².

In correction-level, DM_NLP applied rule-based, NMT and SMT models and merge the generated results in hybrid pipe-line. HFL also followed the strategy of multi-model merging, using PMI scoring and a seq2seq network Their pipelines are shown in Fig.1.

More various additional resources appeared in CGED 2018. Besides Gigawords and Wikipedia Corpus, Google Ngram, People’s Daily, Chinese 5gram are newly introduced resources in this task. More impressively, CMMC utilized domain dictionary in L2 teaching to form pseudo writing data for training set enhancement, improving their performances in all aspects.

Team	Approach	Features	Correction Model	Additional Resources
Ali_GM	BiLSTM+CRF	Char, POS, AWC, DWC	Rule-System, NMT, SMT	Gigawords, Lang8
AutoNLP	Policy Gradient LSTM model			

² In the widely existing scenario of large scale examination correction, users may manually correct some submissions for pre-training, then the model with additional memory

mechanism can automatically finish the rest with a high F1 score.

BUPT	bi-LSTM	Contextualized Char Representation		Wiki Corpus
CMMC	LSTM+CRF (Seq2Seq & Seq Label)	Char, POS		People's Daily, Domain Dictionaries in L2 Teaching, Self-generated corpus
CYUT-III	LSTM	Word		
ECNU	LSTM+CRF	Char, POS, Dependency, BOW5		
HFL	BiLSTM+CRF	Gaussian ePMI, POS, PMI, BOW	PMI Scoring, Seq2Seq Networks	external corpus (unclear), Zuowen & Baike (unpublic)
NTOU	Rule-system	Frequent string matching		Chinese Web 5-grams
UIUC	CRF+Rule-system	Word, Char		Google Chinese N-grams
walker	BiLSTM+CRF			

Table 9: Summary of approaches and additional resources used by the submitted systems.

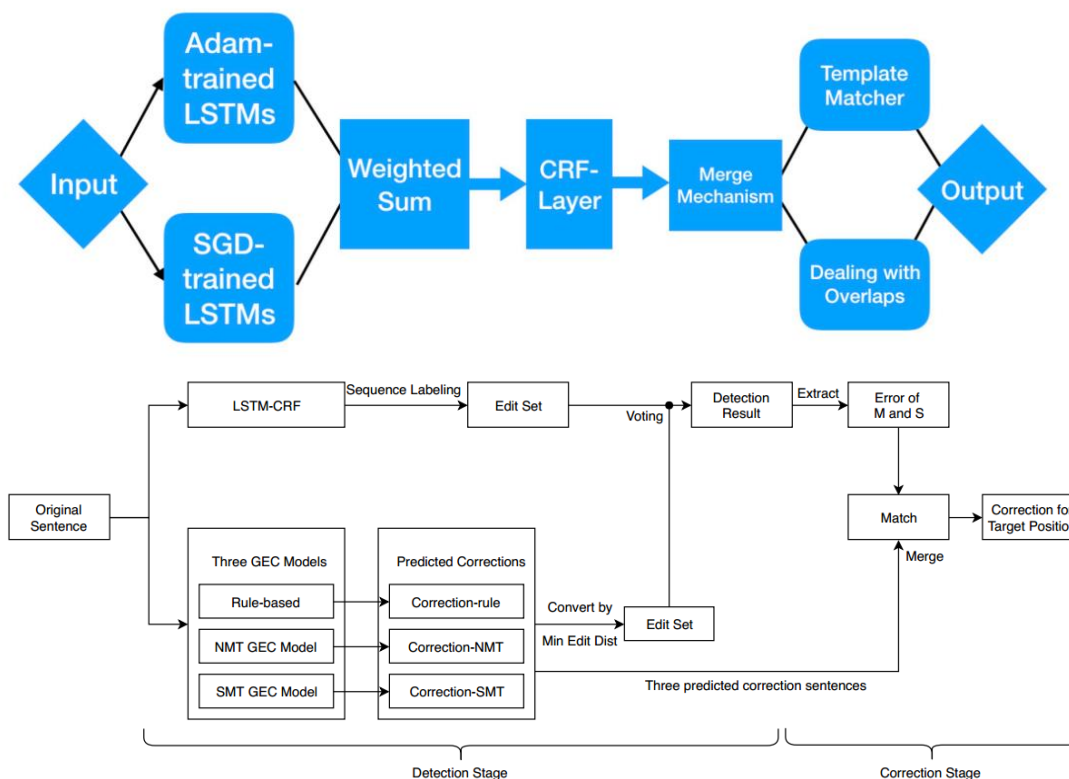


Fig.1 Pipe-lines of the HFL (up) and DM_NLP

7 Conclusion

This study describes the NLP-TEA 2018 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 the CGED 2016). Lung-Hao Lee contributed a lot in consultation and bidding.

This study was supported by the projects from Beijing Advanced Innovation Center for Language Resources (KYD17004), Institute Project of Beijing Language and Culture University (18YJ060001), Social Science Funding of China (16AYY007), Social Science Funding of Beijing (15WYA017), National Language Committee Project (ZDI135-58, ZDI135-3), MOE Project of Key Research Institutes in Universities (16JJD740004).

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.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition 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
- 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 1st Workshop 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 (NLPTEA2016)*, 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*, Taipei, China: 100–10.