Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis

Lung-Hao Lee[1], Gaoqi Rao[2], Liang-Chih Yu[3,4], Endong Xun[5], Baolin Zhang[6], Li-Ping Chang[7]

[1]Graduate Institute of Library and Information Studies, National Taiwan Normal University
[2]Center for Studies of Chinese as a Second Language, Beijing Language and Culture University
[3]Department of Information Management, Yuan Ze University
[4]Innovative Center for Big Data and Digital Convergence, Yuan Ze University
[5]College of Information Science, Beijing Language and Culture University
[6]Faculty of Language Sciences, Beijing Language and Culture University
[7]Mandarin Training Center, National Taiwan Normal University

lhlee@ntnu.edu.tw, raogaoqi@blcu.edu.cn,lcyu@saturn.yzu.edu.tw
edxun@126.com, zhangbl@blcu.edu.cn, lchang@ntnu.edu.tw

Abstract

This paper presents the NLP-TEA 2016 shared task for Chinese grammatical error diagnosis 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 15 teams registered for this shared task, 9 teams developed the system and submitted a total of 36 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). In conjunction with the COLLING 2016, the third NLP-TEA features a shared task for Chinese grammatical error diagnosis 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.

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 40–48, Osaka, Japan, December 12 2016.

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 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 sentence and the position at which they occur. Each input sentence is given a unique sentence number "sid". If the inputs contain no grammatical errors, the system should return: "sid, correct". If an input sentence 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 follows.

TOCFL (Traditional Chinese)	**HSK (Simplified 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: "參加"is a redundant word) • Example 2 Input: (sid=A2-0007-3) 我要送給你一個慶祝禮物。要是兩、三天晚了，請別生氣。 Output: A2-0007-3, 15, 20, W (Notes: "兩、三天晚了"should be "晚了兩、三天") • Example 3 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 4 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") • 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 two sources: the writing section of the computer-based Test Of Chinese as a Foreign Language (TOCFL) (Lee et al., 2016) and 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 sentences in this set were used to train the grammatical error diagnostic systems. Each sentence with annotated grammatical errors and their corresponding corrections is represented in SGML format, as shown in Fig. 1. For the TOCFL track, we provide 10,693 training sentences with a total of 24,492 grammatical errors, categorized as redundant (4,472 instances), missing (8,739), word selection (9,897) and word ordering (1,384). For the HSK track, we provide 10,071 training sentences with a total of 24,797 grammatical errors, categorized as redundant (5,538 instances), missing (6,623), word selection (10,949) and word ordering (1,687).

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="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>
```

Figure 1: A training sentence denoted in SGML format.

(2)Test Set: This set consists of testing sentences used for evaluating system performance. Table 2 shows statistics for the testing set for both tracks. 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 3) are similar with that of the training set.

Track	#Sentences	#Correct	#Erroneous
TOCFL	3,528 (100%)	1,703 (48.27%)	1,825 (51.73%)
HSK	3,011 (100%)	1,539 (51.11%)	1,472 (48.89%)

Table 2: The statistics of testing set for both tracks.

Track	#Error	#R	#M	#S	#W
TOCFL	4,103 (100%)	782 (19.06%)	1,482 (36.12%)	1613 (39.31%)	226 (5.51%)
HSK	3,695 (100%)	802 (21.71%)	991 (26.82%)	1620 (43.84%)	282 (7.63%)

Table 3: The distributions of error types for both tracks.

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. 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 4: 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 5 summarizes the submission statistics for the 15 participating teams including 8 from universities and research institutes in P.R.C (ANO, BFSU-TZT, BISTU, CCNU, HIT, PKU, SKY and YUN-HPCC), 4 from Taiwan, R.O.C. (CYUT, NCTU+NTUT, NCYU and NTOU), 2 from European (including Dublin with NTU (TWIRL) and Saarland with Harvard (MAZA) and 1 private firm (Sogou Inc.). In the official testing phase, each team could opt to participate in either one or both of the TOCFL and HSK tracks. Each participating team was allowed to submit at most three runs for each track. Of the 15 registered teams, 9 teams submitted their testing results, for a total of 36 runs including 15 TOCFL runs (denoting as #TRuns) and 21 HSK runs (#HRuns).

Table 6 shows the testing results for the TOCFL track. The NCTU+CYUT team achieved the lowest false positive rate (denoted as "FPR") of 0.1362. 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 0.4827. All systems performed above the baseline. The system result submitted by CYUT achieved the best detection accuracy of 0.5955. We use the F1 score to reflect the tradeoffs between precision and recall. The NCYU provided the best error detection results, providing a high F1 score of 0.6779. For identification-level evaluations, the systems need to identify the error types in a given sentences. The system developed by CYUT provided the highest F1 score of 0.3666 for grammatical error identification. For position-level evaluations, CYUT achieved the best F1 score of 0.1248. 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.

Table 7 shows the testing results for the HSK track. The CCNU team did not submit the result by the due date. The SKY team achieved the lowest false positive rate of 0.0481. At the detection-level, the accuracy baseline is 0.5111. Eight runs from 5 teams failed to pass the baseline. The system result submitted by SKY achieved the best detection accuracy of 0.6659. For the F1 score, HIT provided the best error detection results, as high as 0.6628. In both the identification-level and position-level evaluations, HIT achieved the best F1 scores of 0.5215 and 0.3855, in different runs. At the position-level, system performance varied considerably among the teams, from 0.0007 to 0.3855. For the HSK track, better F1 scoresat the identification-level and position-level are achieved than in the TOCFL track. Note that, for teams participating in both two tracks, system performances didn't simply increase from TOCFL to HSK, indicating that differences in data sets had a complex impact on system performance.

Participant (Ordered by abbreviations of names)	**#TRuns**	**#HRuns**
NLP Lab, Zhengzhou University (**ANO**)	0	2
Beijing Foreign Studies University (**BFSU-TZT**)	0	0
Beijing Information Science and Technology University (**BISTU**)	0	0
Central China Normal University (**CCNU**)	0	1
Chaoyang University of Technology (**CYUT**)	3	3
Harbin Institute of Technology (**HIT**)	0	3
Institute of Computational Linguistics, Peking University (**PKU**)	3	3
Saarland University & Hardvard Medical School (**MAZA**)	0	0
National Chiao Tung University & National Taipei University of Technology (**NCTU+NTUT**)	3	0
National Chiayi University (**NCYU**)	3	3
National Taiwan Ocean University (**NTOU**)	0	0
NLP Lab, Zhengzhou University (**SKY**)	0	3
Beijing Sogou Inc. (**Sogou**)	0	0
Dublin City University & National Taiwan University (**TWIRL**)	0	0
School of Information Science and Engineering, Yunnan University (**YUN-HPCC**)	3	3

Table 5: Submission statistics for all participants.

TOCFL Submis-sion	FPR	Detection-level				Identification-level				Position-level			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
CYUT-Run1	0.3470	**0.5955**	0.6259	0.5419	0.5809	**0.5154**	0.4600	0.3021	0.3647	0.3113	0.1461	0.1089	**0.1248**
CYUT-Run2	0.3558	**0.5955**	0.6236	0.5501	0.5846	0.5133	0.4567	0.3061	**0.3666**	0.3061	0.1432	**0.1092**	0.1239
CYUT-Run3	0.3635	0.5941	0.6205	0.5545	0.5856	0.5078	0.4472	0.3001	0.3592	0.3088	0.1196	0.0768	0.0935
NCTU+NTUT-Run1	**0.1362**	0.5442	**0.6593**	0.2460	0.3583	0.5110	**0.4892**	0.1224	0.1958	**0.4603**	**0.2542**	0.0483	0.0811
NCTU+NTUT-Run2	0.2913	0.5530	0.6000	0.4077	0.4855	0.4793	0.4036	0.1982	0.2659	0.3784	0.1644	0.0639	0.0920
NCTU+NTUT-Run3	0.3200	0.5612	0.6013	0.4504	0.5150	0.4773	0.3993	0.2185	0.2824	0.3613	0.1521	0.0668	0.0928
NCYU-Run1	0.5602	0.5507	0.5559	0.6542	0.6011	0.3577	0.2749	0.2862	0.2805	0.1728	0.0074	0.0056	0.0064
NCYU-Run2	0.9612	0.5218	0.5202	**0.9726**	**0.6779**	0.2328	0.2265	**0.4744**	0.3066	0.0231	0.0129	0.0195	0.0155
NCYU-Run3	0.8491	0.5363	0.5307	0.8959	0.6665	0.2653	0.2384	0.4134	0.3024	0.0580	0.0130	0.0163	0.0145
PKU-Run1	0.2284	0.5210	0.5739	0.2871	0.3828	0.4575	0.3418	0.1173	0.1747	0.3844	0.0996	0.0263	0.0416
PKU-Run2	0.7205	0.5258	0.5292	0.7556	0.6224	0.3242	0.2792	0.3712	0.3187	0.1381	0.0680	0.0824	0.0745
PKU-Run3	0.5250	0.5349	0.5467	0.5907	0.5678	0.3705	0.2729	0.2192	0.2431	0.2331	0.0872	0.0651	0.0745
YUN-HPCC-Run1	0.6289	0.5420	0.5444	0.7014	0.6130	0.2211	0.1588	0.3196	0.2122	0.0886	0.0002	0.0002	0.0002
YUN-HPCC-Run2	0.5931	0.5026	0.5167	0.5918	0.5517	0.2322	0.1675	0.3136	0.2184	0.0991	0	0	null
YUN-HPCC-Run3	0.3382	0.4847	0.5030	0.3195	0.3908	0.4023	0.2810	0.1359	0.1832	0.2797	0.0012	0.0005	0.0007

Table 6: Testing results of TOCFL track.

HSK Submission	FPR	Detection-level				Identification-level				Position-level			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
ANO-Run1	0.5601	0.5473	0.5297	0.6596	0.5876	0.4723	0.4244	0.4292	0.4268	0.3687	0.2910	0.2460	0.2666
ANO-Run2	0.6517	0.4779	0.4738	0.6135	0.5346	0.2977	0.2243	0.2535	0.2380	0.1157	0.0046	0.0046	0.0046
*CCNU-Run1	0.3294	0.4988	0.4811	0.3193	0.3838	0.4012	0.2425	0.1324	0.1713	0.2806	0.0187	0.0089	0.0121
CYUT-Run1	0.4016	0.6141	0.6003	0.6304	0.6150	0.5714	0.5306	0.4376	0.4797	0.3202	0.2037	0.2138	0.2086
CYUT-Run2	0.4191	0.6118	0.5951	0.6440	0.6186	0.5662	0.5238	0.4509	0.4846	0.3143	0.2034	0.2225	0.2125
CYUT-Run3	0.4016	0.6141	0.6003	0.6304	0.6150	0.5715	0.5306	0.4352	0.4782	0.3304	0.1814	0.1440	0.1605
HIT-Run1	0.4334	0.6377	0.6111	0.7120	0.6577	0.5683	0.5146	0.5219	0.5182	0.4781	0.4034	0.3691	**0.3855**
HIT-Run2	0.4327	0.6370	0.6108	0.7099	0.6566	0.5744	0.5224	0.5094	0.5158	0.4756	0.3970	0.3483	0.3711
HIT-Run3	0.4516	0.6370	0.6071	0.7296	**0.6628**	0.5565	0.5002	**0.5447**	**0.5215**	0.4475	0.3695	**0.3697**	0.3696
NCYU-Run1	0.2820	0.5526	0.5629	0.3798	0.4535	0.4554	0.3259	0.1877	0.2382	0.3301	0.0244	0.0095	0.0136
NCYU-Run2	0.9467	0.5042	0.4964	**0.9755**	0.6580	0.2687	0.2588	0.5263	0.3470	0.0312	0.0158	0.0217	0.0183
NCYU-Run3	0.9818	0.4846	0.4864	0.9721	0.6484	0.2227	0.2195	0.3578	0.2721	0.0148	0.0081	0.0089	0.0085
PKU-Run1	0.7706	0.4972	0.4910	0.7772	0.6018	0.3104	0.2717	0.3991	0.3233	0.1106	0.0523	0.0674	0.0589
PKU-Run2	0.8070	0.5022	0.4945	0.8254	0.6185	0.3144	0.2765	0.3594	0.3125	0.1016	0.0595	0.0923	0.0724
PKU-Run3	0.8213	0.5058	0.4968	0.8478	0.6265	0.3062	0.2694	0.3586	0.3076	0.0896	0.0520	0.0863	0.0649
SKY-Run1	0.0695	0.6523	0.8326	0.3614	0.5040	0.6605	0.8235	0.2732	0.4102	0.6073	0.6153	0.1783	0.2765
SKY-Run2	**0.0481**	0.6579	**0.8746**	0.3505	0.5005	0.6765	**0.8821**	0.2972	0.4446	0.6376	0.7054	0.2217	0.3373
SKY-Run3	0.0559	**0.6659**	0.8652	0.3750	0.5232	**0.6849**	0.8744	0.3185	0.4669	**0.6477**	**0.7144**	0.2430	0.3627
YUN-HPCC-Run1	0.5608	0.5191	0.5069	0.6026	0.5506	0.3485	0.2800	0.3879	0.3252	0.0654	0.0024	0.0062	0.0035
YUN-HPCC-Run2	0.7122	0.4949	0.4886	0.7113	0.5793	0.3092	0.2681	0.4565	0.3378	0.0373	0.0022	0.0070	0.0034
YUN-HPCC-Run3	0.2710	0.5058	0.4902	0.2724	0.3502	0.4306	0.2886	0.1448	0.1928	0.2701	0.0010	0.0005	0.0007

Table 7: Testing results of HSK track.

Table 8 summarize the approaches and resources for each of the submitted systems. ANO and CCNU did not submit reports on their develop systems. Though neural networks achieved goodperformances 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 CYUT, HIT, NCTU+NTUT and SKY in their system pipe-lines. The CRF based systems model with carefully designed feature templates could maintain the performance with neural networks at the same level in the HSK track. The HIT systems using CRF model and LSTM networks achieved the best F1 scores in the three levels. Moreover, CYUT system is simply based onthe CRF model with multiple feature templates in the TOCFL track.

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, the CYUT, NCTU+NTUT, HIT and SKY teams achieved relatively better performances.

Team	**Approach**	**Word/Character Embedding**	**Additional Resources**
CYUT	CRF	---	NLP-TEA-1&NLP-TEA-2
HIT	CRF+LSTM networks	Character Embedding	---
NCTU+NTUT	W2V+CRF	Word Embedding	Sinica Balanced Corpus v4.0 LDC Chinese Gigaword v2 CIRB0303 Taiwan Panorama Magazine TCC300 Wikipedia(ZH_TW) NLP-TEA-1&NLP-TEA-2
NCYU	RNN+LSTM networks	Word Embedding	NLP-TEA-1&NLP-TEA-2
PKU	Bi-LSTM networks	Word Embedding	NLP-TEA-1&NLP-TEA-2
SKY	Ngram+CRF	---	---
YUN-HPCC	CNN/LSTM networks	Word Embedding	Wikipedia(ZH)

Table 8: Summary of approaches and additional resources used by the submitted systems.

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://ir.itc.ntnu.edu.tw/lre/nlptea16cged.htm.

Acknowledgements

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.

This study was partially supported by the Ministry of Science and Technology, under the grant MOST 103-2221-E-003-013-MY3, MOST 103-2410-H-003-043-MY2, MOST 105-2221-E-003-020-MY2, and MOST 105-2221-E-155-059-MY2, and the "Aim for the Top University Project" and "Center of Learning Technology for Chinese" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan, R.O.C.

Following grants and projects from P.R.C also supported the study in this paper: Social Science Funding China (11BYY054, 12&ZD173, 16AYY007), Social Science Funding Beijing (15WYA017), National Language Committee Project (YB125-42, ZDI135-3), 863 Key Project (SQ2015AA0100074), MOE Annual Project of Key Research Institutes in Univs "Push Platform in Resources of CSL".

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.

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.