Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis

Lung-Hao Lee[1], Liang-Chih Yu[2,3], Li-Ping Chang[4]
[1]Information Technology Center, National Taiwan Normal University
[2]Department of Information Management, Yuan Ze University
[3]Innovative Center for Big Data and Digital Convergence, Yuan Ze University
[4]Mandarin Training Center, National Taiwan Normal University
lhlee@ntnu.edu.tw, lcyu@saturn.yzu.edu.tw, lchang@ntnu.edu.tw

Abstract

This paper introduces the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. We describe the task, data preparation, performance metrics, and evaluation results. The hope is that such an evaluation campaign may produce more advanced Chinese grammatical error diagnosis techniques. All data sets with gold standards and evaluation tools are publicly available for research purposes.

1 Introduction

Human language technologies for English grammatical error correction have attracted more attention in recent years (Ng et al., 2013; 2014). In contrast to the plethora of research related to develop NLP tools for learners of English as a foreign language, relatively few studies have focused on detecting and correcting grammatical errors for use by learners of Chinese as a foreign language (CFL). A classifier has been designed to detect word-ordering errors in Chinese sentences (Yu and Chen, 2012). A ranking SVM-based model has been further explored to suggest corrections for word-ordering errors (Cheng et al., 2014). Relative positioning and parse template language models have been proposed to detect Chinese grammatical errors written by US learners (Wu et al., 2010). A penalized probabilistic first-order inductive learning algorithm has been presented for Chinese grammatical error diagnosis (Chang et al. 2012). A set of linguistic rules with syntactic information was manually crafted to detect CFL grammatical errors (Lee et al., 2013). A sentence judgment system has been further developed to integrate both rule-based linguistic analysis and n-gram statistical learning for grammatical error detection (Lee et al., 2014).

The ICCE-2014 workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA) organized a shared task on CFL grammatical error diagnosis (Yu et al., 2014). Due to the greater challenge in identifying grammatical errors in CFL leaners' written sentences, the NLP-TEA 2015 shared task features a Chinese Grammatical Error Diagnosis (CGED) task, providing an evaluation platform for the development and implementation of NLP tools for computer-assisted Chinese learning. The developed system should identify whether a given sentence contains grammatical errors, identify the error types, and indicate the range of occurred errors.

This paper gives an overview of this shared task. The rest of this article is organized as follows. Section 2 provides the details of the designed task. Section 3 introduces the developed data sets. Section 4 proposes evaluation metrics. Section 5 presents the results of participant approaches for performance comparison. Section 6 summarizes the findings and offers futures research directions.

2 Task Description

The goal of this shared task is to develop NLP tools for identifying the grammatical errors in sentences written by the CFL learners. Four PADS error types are included in the target modification taxonomy, that is, mis-ordering (Permutation), redundancy (Addition), omission (Deletion), and mis-selection (Substitution). For the sake of simplicity, the input sentence is selected to contain one defined error types. The developed tool is expected to identify the error types and its position at which it occurs in the sentence.

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 1–6,
Beijing, China, July 31, 2015.

The input instance is given a unique sentence number *sid*. If the inputs contain no grammatical errors, the tool should return "sid, correct". If an input sentence contains a grammatical error, the output format should be a quadruple of "sid, start_off, end_off, error_type", where "start_off" and "end_off" respectively denote the characters at which the grammatical error starts and ends, where each character or punctuation mark occupies 1 space for counting positions. "Error_type" represents one defined error type in terms of "Redundant," "Missing," "Selection," and "Disorder". Examples are shown as follows.

- Example 1

Input: (sid=B2-0080) 他是我的以前的室友
Output: B2-0080, 4, 4, Redundant

- Example 2

Input: (sid=A2-0017) 那電影是機器人的故事
Output: A2-0017, 2, 2, Missing

- Example 3

Input: (sid=A2-0017) 那部電影是機器人的故事
Output: A2-0017, correct

- Example 4

Input: (sid=B1-1193) 吳先生是修理腳踏車的拿手
Output: B1-1193, 11, 12, Selection

- Example 5

Input: (sid=B2-2292) 所以我不會讓失望她
Output: B2-2292, 7, 9, Disorder

The character "的" is a redundant character in Ex. 1. There is a missing character between "那" and "電影" in Ex. 2, and a missed character "部" is shown in the correct sentence in Ex. 3. In Ex. 4, "拿手" is a wrong word. One of correct words may be "好手". "失望她" is a word ordering error in Ex. 5. The correct order should be "她失望".

3 Data Preparation

The learner corpus used in our task was collected from the essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL), administered in Taiwan. Native Chinese speakers were trained to manually annotate grammatical errors and provide corrections corresponding to each error. The essays were then split into three sets as follows.

(1) Training Set: This set included 2,205 selected sentences with annotated grammatical errors and their corresponding corrections. Each sentence is represented in SGML format as shown in Fig. 1. Error types were categorized as redundant (430 instances), missing (620), selection (849), and disorder (306). All sentences in this set were collected to use for training the grammatical diagnostic tools.

```
<DOC>
<SENTENCE id="B1-1120">
我的中文進步了非常快
</SENTENCE>
<MISTAKE start_off="7" end_off="7">
<TYPE>
Selection
</TYPE>
<CORRECTION>
我的中文進步得非常快
</CORRECTION>
</MISTAKE>
</DOC>
```

Figure 1. An sentence denoted in SGML format

(2) Dryrun Set: A total of 55 sentences were distributed to participants to allow them familiarize themselves with the final testing process. Each participant was allowed to submit several runs generated using different models with different parameter settings of their developed tools. In addition, to ensure the submitted results could be correctly evaluated, participants were allowed to fine-tune their developed models in the dryrun phase. The purpose of dryrun is to validate the submitted output format only, and no dryrun outcomes were considered in the official evaluation

(3) Test Set: This set consists of 1,000 testing sentences. Half of these sentences contained no grammatical errors, while the other half included a single defined grammatical error: redundant (132 instances), missing (126), selection (110), and disorder (132). The evaluation was conducted as an open test. In addition to the data sets provided, registered research teams were allowed to employ any linguistic and computational resources to identify the grammatical errors.

4 Performance Metrics

Table 1 shows the confusion matrix used for performance evaluation. In the matrix, TP (True Positive) is the number of sentences with grammatical errors that are correctly identified by the

developed tool; FP (False Positive) is the number of sentences in which non-existent grammatical errors are identified; 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 for which no errors are identified.

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 occurred range of grammatical error. That is to say, the system results should be perfectly identical with the quadruples of gold standard.

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

- False Positive Rate (FPR) = 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 Result	
		Positive (Erroneous)	Negative (Correct)
Gold Standard	Positive	**TP**	**FN**
	Negative	**FP**	**TN**

Table 1. Confusion matrix for evaluation.

For example, given 8 testing inputs with gold standards shown as "B1-1138, 7, 10, Disorder", "A2-0087, 12, 13, Missing", "A2-0904, correct", "B1-0990, correct", "A2-0789, 2, 3, Selection", "B1-0295, correct", "B2-0591, 3, 3, Redundant" and "A2-0920, correct", the system may output the result shown as "B1-1138, 7, 8, Disorder", "A2-0087, 12, 13, Missing", "A2-0904, 5, 6, Missing", "B1-0990, correct", "A2-0789, 2, 5, Disorder", "B1-0295, correct", "B2-0591, 3, 3, Redundant" and "A2-0920, 4, 5, Selection". The evaluation tool will yield the following performance.

- False Positive Rate (FPR) = 0.5 (=2/4) Notes: {"A2-0904, 5, 6, Missing", "A2-0920, 4, 5, Selection"} /{"A2-0904, correct", "B1-0090, correct", "B1-0295, correct", "A2-0920, correct"}
- Detection-level
  - Accuracy =0.75 (=6/8)

    Notes: {"B1-1138, Disorder", "A2-0087, Missing", "B1-0990, correct", "A2-0789, Disorder", "B1-0295, correct", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0904, Missing", "B1-0990, correct", "A2-0789, Disorder", "B1-0295, correct", "B2-0591, Redundant", "A2-0920, Selection".}
  - Precision = 0.67 (=4/6)

    Notes: {"B1-1138, Disorder", "A2-0087, Missing", "A2-0789, Disorder", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0904, Missing", "A2-0789, Disorder", "B2-0591, Redundant", "A2-0920, Selection".}
  - Recall = 1 (=4/4).

    Notes: {"B1-1138, Disorder", "A2-0087, Missing", "A2-0789, Disorder", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0789, Selection", "B2-0591, Redundant"}
  - F1=0.8 (=2*0.67*1/(0.67+1))
- Identification-level
  - Accuracy =0.625 (=5/8)

    Notes: {"B1-1138, Disorder", "A2-0087, Missing", "B1-0990, correct", "B1-0295, correct", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0904, Missing", "B1-0990, correct", "A2-0789, Disorder", "B1-0295, correct", "B2-0591, Redundant", "A2-0920, Selection"}
  - Precision = 0.5 (=3/6)

    Notes: {"B1-1138, Disorder", "A2-0087, Missing", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0904, Missing", "A2-0789, Disorder", "B2-0591, Redundant", "A2-0920, Selection".}

- Recall = 0.75 (=3/4)

  Notes: {"B1-1138, Disorder", "A2-0087, Missing", "B2-0591, Redundant"} / {"B1-1138, Disorder", "A2-0087, Missing", "A2-0789, Selection", "B2-0591, Redundant"}

- F1=0.6 (=2*0.5*0.75/(0.5+0.75))

- Position-level

  - Accuracy =0.5 (=4/8)

    Notes: {"A2-0087, 12, 13, Missing", "B1-0990, correct", "B1-0295, correct", "B2-0591, 3, 3, Redundant"} / {"B1-1138, 7, 8, Disorder", "A2-0087, 12, 13, Missing", "A2-0904, 5, 6, Missing", "B1-0990, correct", "A2-0789, 2, 5, Disorder", "B1-0295, correct", "B2-0591, 3, 3, Redundant", "A2-0920, 4, 5, Selection"}

  - Precision = 0.33 (=2/6)

    Notes: {"A2-0087, 12, 13, Missing", "B2-0591, 3, 3, Redundant"} / {"B1-1138, 7, 8, Disorder", "A2-0087, 12, 13, Missing", "A2-0904, 5, 6, Missing", "A2-0789, 2, 5, Disorder", "B2-0591, 3, 3, Redundant", "A2-0920, 4, 5, Selection"}

  - Recall = 0.5 (=2/4)

    Notes: {"A2-0087, 12, 13, Missing", "B2-0591, 3, 3, Redundant"} / {"B1-1138, 7, 10, Disorder", "A2-0087, 12, 13, Missing", "A2-0789, 2, 3, Selection", "B2-0591, 3, 3, Redundant"}

  - F1=0.4 (=2*0.33*0.5/(0.33+0.5))

5 Evaluation Results

Table 2 summarizes the submission statistics for the participating teams. Of 13 registered teams, 6 teams submitted their testing results. In formal testing phase, each participant was allowed to submit at most three runs using different models or parameter settings. In total, we had received 18 runs.

Table 3 shows the task testing results. The CYUT team achieved the lowest false positive rate of 0.082. 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 errors are correct without errors. According to the test data distribution, the baseline system can achieve an accuracy level of 0.5. All systems achieved results slightly better than the baseline. The system result submitted by NCYU achieved the best detection accuracy of 0.607. We used the F1 score to reflect the tradeoff between precision and recall. In the testing results, NTOU provided the best error detection results, providing a high F1 score of 0.6754. For correction-level evaluations, the systems need to identify the error types in the given sentences. The system developed by NCYU provided the highest F1 score of 0.3584 for grammatical error identification. For position-level evaluations, CYUT achieved the best F1 score of 0.1742. Note that it is difficult to perfectly identify the error positions, partly because no word delimiters exist among Chinese words.

Participant (Ordered by abbreviations of names)	**#Runs**
Adam Mickiewicz University on Poznan (**AMU**)	0
University of Cambridge (**CAM**)	0
Chinese Academy of Sciences (**CAS**)	0
Confucius Institute of Rutgers University (**CIRU**)	0
Chaoyang University of Technology (**CYUT**)	3
Harbin Institute of Technology Shenzhen Graduate School (**HITSZ**)	3
Lingage Inc. (**Lingage**)	0
National Chiayi University (**NCYU**)	3
National Taiwan Ocean University (**NTOU**)	3
National Taiwan University (**NTU**)	0
South China Agriculture University (**SCAU**)	3
Tokyo Metropolitan University (**TMU**)	3
University of Leeds (**UL**)	0
Total	**18**

Table 2. Submission statistics for all participants

Submission	False Positive Rate	Detection Level				Identification Level				Position Level			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
CYUT-Run1	0.096	0.584	0.7333	0.264	0.3882	0.522	0.5932	0.14	0.2265	0.504	0.52	0.104	0.1733
CYUT-Run2	**0.082**	0.579	**0.7453**	0.24	0.3631	**0.525**	**0.6168**	0.132	0.2175	**0.505**	**0.5287**	0.092	0.1567
CYUT-Run3	0.132	0.579	0.6872	0.29	0.4079	0.505	0.5182	0.142	0.2229	0.488	0.45	0.108	**0.1742**
HITSZ-Run1	0.956	0.509	0.5047	0.974	0.6648	0.173	0.2401	0.302	0.2675	0.031	0.0185	0.018	0.0182
HITSZ-Run2	0.938	0.505	0.5027	0.948	0.657	0.149	0.201	0.236	0.2171	0.036	0.0105	0.01	0.0103
HITSZ-Run3	0.884	0.51	0.5056	0.904	0.6485	0.188	0.2273	0.26	0.2425	0.068	0.0221	0.02	0.021
NCYU-Run1	0.48	0.53	0.5294	0.54	0.5347	0.354	0.2814	0.188	0.2254	0.274	0.0551	0.028	0.0371
NCYU-Run2	0.396	0.567	0.5724	0.53	0.5504	0.423	0.3793	0.242	0.2955	0.343	0.1715	0.082	0.111
NCYU-Run3	0.374	**0.607**	0.6112	0.588	0.5994	0.463	0.4451	0.3	**0.3584**	0.374	0.246	0.122	0.1631
NTOU-Run1	1	0.5	0.5	**1**	0.6667	0.117	0.1896	0.234	0.2095	0.005	0.0099	0.01	0.01
NTOU-Run2	0.914	0.531	0.5164	0.976	**0.6754**	0.225	0.2848	**0.364**	0.3196	0.123	0.149	**0.16**	0.1543
NTOU-Run3	0.948	0.519	0.5098	0.986	0.6721	0.193	0.2605	0.334	0.2927	0.093	0.1238	0.134	0.1287
SCAU-Run1	0.62	0.505	0.504	0.63	0.56	0.287	0.2383	0.194	0.2139	0.217	0.0801	0.054	0.0645
SCAU-Run2	0.636	0.503	0.5023	0.642	0.5637	0.279	0.2337	0.194	0.212	0.209	0.0783	0.054	0.0639
SCAU-Run3	0.266	0.503	0.5056	0.272	0.3537	0.416	0.2692	0.098	0.1437	0.385	0.1192	0.036	0.0553
TMU-Run1	0.478	0.516	0.5162	0.51	0.5131	0.313	0.1787	0.104	0.1315	0.27	0.0363	0.018	0.0241
TMU-Run2	0.134	0.524	0.5759	0.182	0.2766	0.479	0.4071	0.092	0.1501	0.449	0.1928	0.032	0.0549
TMU-Run3	0.35	0.546	0.5581	0.442	0.4933	0.42	0.3519	0.19	0.2468	0.362	0.1745	0.074	0.1039

Table 3. Testing results of our Chinese grammatical error diagnosis task.

In summary, none of the submitted systems provided superior performance. It is a really difficult task to develop an effective computer-assisted learning tool for grammatical error diagnosis, especially for the CFL uses. In general, this research problem still has long way to go.

6 Conclusions and Future Work

This paper provides an overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis, including task design, data preparation, evaluation metrics, and performance evaluation results. Regardless of actual performance, all submissions contribute to the common effort to produce an effective Chinese grammatical diagnosis tool, and the individual reports in the shared task proceedings provide useful insight into Chinese language processing.

We hope the data sets collected for this shared task can facilitate and expedite the future development of NLP tools for computer-assisted Chinese language learning. Therefore, all data sets with gold standards and evaluation tool are publicly available for research purposes at http://ir.itc.ntnu.edu.tw/lre/nlptea15cged.htm.

We plan to build new language resources to improve existing techniques for computer-aided Chinese language learning. In addition, new data sets with the contextual information of target sentences obtained from CFL learners will be investigated for the future enrichment of this research topic.

Acknowledgments

We thank all the participants for taking part in our task. We would like to thank Bo-Shun Liao for developing the evaluation tool. This research is partially supported by the "Aim for the Top University" and "Center of Learning Technology for Chinese" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, ROC and is also sponsored in part by the "International Research-Intensive Center of Excellence Program" of NTNU and Ministry of Science and Technology, Taiwan, ROC under the Grant no. MOST 104-2911-I-003-301, and MOST 102-2221-E-155-029-MY3.

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.

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese word ordering errors detection and correction for non-native Chinese language learners. *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING-14*), pages 279-289.

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. *Work-in-Progress Poster Proceedings of the 21st International Conference on Computers in Education* (*ICCE-13*), pages 27-29.

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. *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING-14*), pages 67-70.

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. *Proceedings of the 18th Conference on Computational Natural Language Learning* (*CoNLL-14*): *Shared Task*, pages 1-12.

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

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. *Proceedings of the 24th International Conference on Computational Linguistics* (*COLING-12*), pages 3003-3017.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications* (*NLPTEA-14*), pages 42-47.