Ling@CASS Solution to the NLP-TEA CGED Shared Task 2018

Qinan Hu[12], **Yongwei Zhang**[12], **Fang Liu**[32], **Yueguo Gu**[2]
[1]Institute of Linguistics, Chinese Academy of Social Sciences
[2]China Multilingual and Multimodal Corpora and Big Data Research Centre
[3]School of Software & Microelectronics, Peking University
qinan.hu@qq.com, zhangyw@cass.org.cn, liu_fang@pku.edu.cn, gyg@beiwaionline.com

Abstract

In this study, we employ the sequence to sequence learning to model the task of grammar error correction. The system takes potentially erroneous sentences as inputs, and outputs correct sentences. To breakthrough the bottlenecks of very limited size of manually labeled data, we adopt a semi-supervised approach. Specifically, we adapt correct sentences written by native Chinese speakers to generate pseudo grammatical errors made by learners of Chinese as a second language. We use the pseudo data to pre-train the model, and the CGED data to fine-tune it. Being aware of the significance of precision in a grammar error correction system in real scenarios, we use ensembles to boost precision. When using inputs as simple as Chinese characters, the ensembled system achieves a precision at 86.56% in the detection of erroneous sentences, and a precision at 51.53% in the correction of errors of Selection and Missing types.

1 Introduction

An inter-language is an idiolect developed by a learner of a second language (or L2). It is characteristic that it preserves some features of the first language (or L1), and can overgeneralize some L2 linguistic rules. An investigation on the grammatical errors made by L2 learners will disclose the error patterns, which are beneficial to the teaching and learning process. On the other hand, it will promote the development of systems which can correct grammatical errors made by L2 learners automatically.

The rest of this paper is organized as follows: Section 2 briefly introduces the definition of the NLP-TEA CGED Shared Task 2018. Section 3 gives a quick review on previous studies. Section 4 describes the generation of pseudo data in detail. Section 5 introduces the modeling of the correction task using sequence to sequence learning. Section 6 analyses the experimental results. Finally, conclusions and prospects are drawn in Section 7.

2 NLP-TEA CGED Shared Task 2018

The goal of Chinese Grammar Error Diagnosis (CGED) Shared Task in NLP Tech for Education Application (NLP-TEA) is to develop NLP techniques to automatically correct grammatical errors in Chinese sentences written by L2 learners. The shared task facilitate researchers using different linguistic knowledges and computational techniques to compare their results on the basis of common datasets and evaluation frameworks.

Grammatical errors made by speakers as a second language consist of different types. In CGED, the errors are defined as four types: Missing words ("M"), Redundant words ("R"), word Selection errors ("S"), and Word ordering errors ("W"). It is noticeable that this categorization is different from that of a traditional linguistic point of view, in which the errors are typically categorized into mis-usages of determiners, prepositions, noun forms, verb forms and subject-verb agreement etc. The categorization of errors in CGED tasks correspond to the four operations, i.e. insertions, deletions, substitutions, and transpositions, as defined in Damerau–Levenshtein dis-

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 70–76
Melbourne, Australia, July 19, 2018.

tance (Bard, 2006), respectively. These operations are used to edit a sequence into another.

A developed system should indicate types and positions of the errors, and propose corrections for the errors of S and M types. A system is to be evaluated using four tasks, including the detection of errors, the identification of error types, the identification of positions, and the corrections.

3 Previous Solutions: A Quick Review

Lee et al. (2013) employed handcrafted linguistic rules to detect grammatical errors made by learners of Chinese as a second language. Their system is further integrated with N-gram models to detect the errors (Lee et al., 2014). Most previous studies take the diagnosis of grammatical errors as a sequence labeling problem. They generally assign a B/I/O tag to each word in an input sentence, or each character in a word, to detect the errors. Yu and Chen (2012) proposed to use Conditional Random Field (CRF) (Lafferty et al., 2001) to detect Chinese word ordering errors. In 2014, Cheng et al. (2014) adopted a Support Vector Machine (SVM) (Hearst et al., 1998) to identify Chinese word ordering errors. In recent years, Long-short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) has been a popular neural network model used for this task (Zheng et al., 2016; Yang et al., 2017).

Various features have been taken as the inputs into sequence labeling models, including characters, words, Part-of-Speech (POS) tags (Zheng et al., 2016), dependency information, and Point-wise Mutual Information (Yang et al., 2017), among many others.

4 Pseudo Labeling

The manually labeled dataset for the task of grammar error correction is of very limited size. Since manual labeling is both labor and time consuming, the size of the data set has been a bottleneck for the performances of automatic error correction systems. There have been several approaches to tackling this problem. Cahill et al. (2013) and Grundkiewicz and Junczys-Dowmunt (2014) use the error corrections extracted from Wikipedia revision history as training corpora. Furthermore, many studies adopt a semi-supervised approach to automatically generating a large scale pseudo data set and have reported promising results (Foster and Andersen, 2009; Rozovskaya and Roth, 2010; Dickinson, 2010; Imamura et al., 2012; Felice and Yuan, 2014; Rozovskaya et al., 2017).

4.1 Error Types

In our study, the pseudo data are generated based on a close observation on the errors collected from the manually labeled dataset.

4.1.1 Missing

It is observed that missing words are often functional words. As shown in Sentence 1, in which a particle and a preposition are missing. (The erroneous sentence is represented with E; and the correct sentence, C. The erroneous phrases are in bold.) Sentences 2 and 3 show another type of missing errors, which are caused by improper uses of ellipses.
(1-E) 认识到结婚 **Ø** 过程不满六个月，也可以说 **Ø** 我的故事中我是主动的。
(1-C) 认识到结婚的过程不满六个月，也可以说在我的故事中我是主动的。
(2-E) 所以家长会在孩子很小的时候就让其接受各种各样的学校教育，使 **Ø** 还很脆弱的心理和生理都受到很多压力。
(2-C) 所以家长会在孩子很小的时候就让其接受各种各样的学校教育，使孩子还很脆弱的心理和生理都受到很多压力。
(3-E) 在韩国最近很流行不允许 **Ø** 的电视节目，这节目说公共场所抽烟是不道德的行为。
(3-C) 在韩国最近不允许抽烟的电视节目很流行，这些节目说在公共场所抽烟是不道德的行为。

4.1.2 Redundant

Of all the redundant errors in CGED data set, functional words are among the most frequent. For instance, the particle and the conjunction in Sentences 4-5 are redundant.
(4-E) 如何处理现在在**做**香烟**的**工厂的人的以后的生活。
(4-C) 如何处理现在在香烟工厂工作的人的以后的生活。
(5-E) 大家在手术间里，合作无间**而**救了那位病人。
(5-C) 大家在手术间里，合作无间救了那位病人。

4.1.3 Selection

Selection errors often occur when near-synonyms are misused, as shown in Sentences 6-7. The differences of the usages between these near-synonyms are subtle.
(6-E) 他们知不道吸烟对未成年年的影响会造成的各种**害处**。
(6-C) 他们不知道吸烟对未成年人会造成的各种**伤害**。
(7-E) 从此，父母亲就会教**咱们**爬行、走路、叫爸爸妈妈。
(7-C) 从此，父母亲就会教**我们**爬行、走路、叫爸爸妈妈。

4.1.4 Word Order

Word ordering errors are typically related to the modification of verbs. For instance, the modifiers of the verbs, the auxiliary verb and the adverbs, are misplaced in Sentences 8-10.
(8-E) **采取几种方法应该**帮助他们。
(8-C) **应该采取几种方法**帮助他们。
(9-E) ……但**还是年轻的学生**需要大人的支持和指导……
(9-C) ……但**年轻的学生还是**需要大人的支持和指导……
(10-E) 我走路时常常想抽烟，可能另外抽烟者也**想这样**。
(10-C) 我走路时常常想抽烟，可能别的抽烟者也**这样想**。

4.2 Data Generation

Based on the above observations, we adapt the sentences written by native Chinese speakers to generate ungrammatical sentences. The canonical sentences come from 12 serials of textbooks for students learning Chinese as a second language, 7 serials of textbooks for native Chinese students, and People's Daily newspapers. The sentences are filtered with a length threshold and the controlled vocabularies for teaching Chinese as a second language (Hanban, 2001, 2010). These sentences are tokenized using LTP (Che et al., 2010). And then, the errors of redundant words, missing words, word selection errors and word ordering errors are generated using the operations of insertions, deletions, substitutions, and transpositions, respectively. All adaptations are done in terms of words. 2 millions sentences are adapted in this way.

4.2.1 Missing

(1) To make erroneous sentences with missing words, we randomly select a position in the input sentence. (2) If the word in that position is a functional word, or it is a content word with an antecedent in that sentence, drop this word. Example sentences are shown below.
(11-E) 一天，庙里来 **Ø** 一个瘦和尚。
(11-C) 一天，庙里来**了**一个瘦和尚。
(12-E) 他不仅爱收集动植物标本，还阅读了许多描写 **Ø** 的书。
(12-C) 他不仅爱收集动植物标本，还阅读了许多描写**动植物**的书。

4.2.2 Redundant

(1) Randomly select a position in the input sentence. (2) Randomly select a word according to word frequencies. (3) Insert the word into that position.
(13-E) 达尔文妈妈喜欢种花**的**。
(13-C) 达尔文妈妈喜欢种花。

4.2.3 Selection

(1) Randomly select a position in the input sentence. (2) Select a near-synonym of the word in that position based on their similarities computed using word embeddings. (3) Replace the word in that position with the near-synonym.
(14-E) 老鼠又去咬蜡烛，蜡烛倒了，庙里**爆炸**了。
(14-C) 老鼠又去咬蜡烛，蜡烛倒了，庙里**着火**了。

4.2.4 Word Order

(1) Randomly select a position in the input sentence. (2) Swap the word in that position with its neighbor.
(15-E) 这些剪纸的技艺，都是人们世代一代手把手地**下来传**的。
(15-C) 这些剪纸的技艺，都是人们世代一代手把手地**传下来**的。

5 Ling@CASS Solution: Methodology and System Development

A new task, the corrections of the errors of missing and selection types, has been intro-

duced to CGED 2018. We accordingly need a reconsideration of the appropriateness of using sequence labeling models (Sakaguchi et al., 2017). Unlike the B/I/O tag set which is close, the corrections of the missing, and selection types of errors form an open set. In addition, the corrections generally give rise to output sentences with lengths different from input ones. Therefore, the correction task has gone beyond the capabilities of sequence labeling models.

Sequence to sequence learning (seq2seq) maps an input sequence to an output sequence of varying lengths. It has been the mainstream model for machine translation nowadays (Klein et al., 2017). The correction task can be modeled as a translation task, in which the ungrammatical sentences are from an original language, and the corrections are from a target language. The translation model has been used in several previous studies on grammar error corrections (Schmaltz et al., 2016; Chaitanya, 2017; Yuan and Felice, 2013).

The state-of-the-art performances on machine translation are presented by FairSeq in terms of both accuracy and speed (Gehring et al., 2017). FariSeq significantly differs from previous seq2seq models in that its architecture is based entirely on Convolutional Neural Networks (CNN), instead of the prevalent Recurrent Neural Networks (RNN), so that computations can be fully parallelized during training and optimization.

In our study, we employ the FairSeq model. The Fairseq models are pre-trained with the pseudo labeled data, and fine-tuned with the manually labeled data delivered in CGED. The inputs to Fairseq models are as simple as Chinese characters and POS tags of characters. The POS are tagged using LTP (Che et al., 2010). We use the default settings of FairSeq, except that we use 512 dimensions of character embeddings. The embeddings are randomly initialized and we do NOT use any other resources.

6 Ling@CASS Solution: the Outcome

6.1 Evaluation on Corrections

As shown in Table 1, we have four basic system configurations. These configurations are different in the use of pseudo corpus and POS tags. The evaluation in Table 1 reveals that the use of pseudo data has improved both precision and recall in the correction task of the word selection errors and missing errors, while that of POS tags does not make a significant contribution.

In real scenarios of grammar error diagnoses, the evaluation metrics of precision, recall and F1 are not of the same importance. A teacher would always prefers a grammar error correction system with high precision, even if it has a low recall, than a system returns lots of noises. Being aware of the significance of precision in a grammar error correction system in practice, we further use ensembles to boost precisions. The tag "(>1)" indicates that the correction has been confirmed by at least two basic systems; and "(>2)", at least three. The ensembled systems steadily achieve a precision greater than 50%, with a recall greater than 8%. These performances are much higher than the best in CGED 2018 submissions, where the precision is 29.32%, and recall is 1.58%.

The official submission of our team to CGED 2018 is the result of an ensemble of the systems 3 and 4, where the results are simply merged.

6.2 Evaluation on Detections, Identifications of Error Types, and Positions

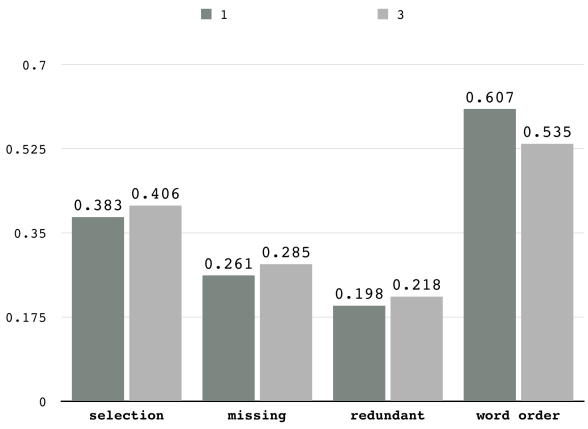

Figure 1: Impacts of Pseudo Data

We also evaluated the systems on the detections, and the identifications of error types and positions. Figure 2 shows a detailed analysis on the precision of the identification of error positions for all four types of errors. It reveals

ID	Pseudo corpus	CGED corpus	Character	POS	P	R	F1
1		Y	Y		0.2678	0.0984	0.1439
2		Y	Y	Y	0.2657	0.1060	0.1515
3	Y	Y	Y		0.2830	0.1153	0.1638
4	Y	Y	Y	Y	0.2672	0.1139	0.1597
1+3 (>0)					0.2149	0.1313	0.1631
3+4 (>0) Submission					0.2126	0.1395	0.1685
1+2 (>1)					**0.5153**	0.0806	0.1394
3+4 (>1)					0.5056	0.0896	0.1523
1+3 (>1)					0.5105	0.0823	0.1417
1+2+3+4 (>2)					0.5080	0.0896	0.1524

Table 1: Performances on Corrections

	FPR	Detection				Identification			Position		
		Acc.	P	R	F1	P	R	F1	P	R	F1
3+4 (>0) Submission	0.3470	0.6630	0.7109	0.6709	0.6903	0.4853	0.4096	0.4442	0.2482	0.1814	0.2096
1+2 (>1)	0.064	0.5342	0.8127	0.2184	0.3443	0.6653	0.1436	0.2362	0.4861	0.0906	0.1528
3+4 (>1)	0.0512	0.5599	0.8632	0.2542	0.3927	0.7015	0.1663	0.2688	0.5024	0.1006	0.168
1+3 (>1)	**0.0448**	0.5475	**0.8656**	0.227	0.3596	0.6853	0.1463	0.2411	0.5104	0.0932	0.1577
1+2+3+4 (>2)	0.0544	0.5545	0.8524	0.2471	0.3831	0.6819	0.1600	0.2592	0.5030	0.0996	0.1663

Table 2: Performances on Detections, Identifications of Error Types & Positions

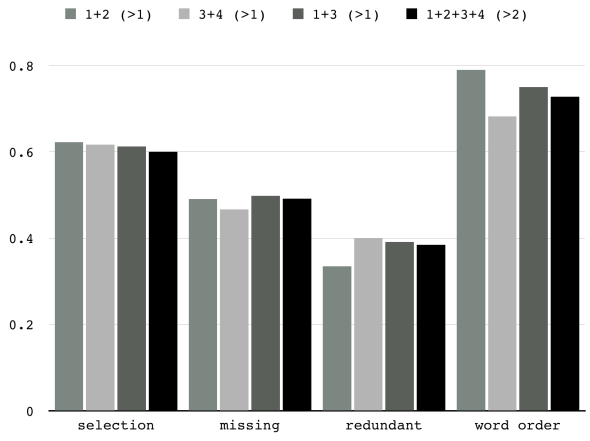

Figure 2: Difficulties of Error Types

that the current pseudo data has a positive impact on the precision of all error types, except for the word ordering errors. It indicates the word ordering pseudo data has much room for improvements.

Figure 1 shows that the identification of the positions of these errors is of different difficulties to the systems. While the ensembled systems are proficient in handling word ordering errors, they have the most difficulties in handling redundant errors.

Table 2 shows the ensembled system 1+3 (>1) achieves a False Positive Rate (FPR) at 4.48% and a precision of 86.56% the detection of erroneous sentences, which are better than the best FPR 4.99% and the best precision 82.76% in CGED 2018 submissions, respectively.

7 Conclusion and Future Work

In CGED 2018, we employ the sequence to sequence learning to model the task of grammar error correction. We adopt a semi-supervised approach to breakthrough the bottlenecks of very limited size of manually labeled data. Specifically, we adapt correct sentences written by native Chinese speakers to generate pseudo grammatical errors made by learners of Chinese as a second language. The pseudo data is used to pre-train the model and gives rise to improvements in both precision and recall. Being aware of the significance of precision in a grammar error correction system in real scenarios, we use ensembles to boost precision. The use of pseudo data has a positive impact on the identification of missing errors, redundant errors, and word selection errors.

In the future work, we will use multi-task to jointly optimize the four tasks all together (Luong et al., 2015). In addition, we will investigate more sophisticated techniques for the generation of pseudo data.

References

Gregory V. Bard. 2006. Spelling-error tolerant, order-independent pass-phrases via the

damerau-levenshtein string-edit distance metric. In *Australasian Symposium on Acsw Frontiers*, pages 117–124.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517.

G Krishna Chaitanya. 2017. *GRAMMATICAL ERROR CORRECTION*. Ph.D. thesis, Indian Institute of Technology Bombay Mumbai 400076 (India) 14.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: a chinese language technology platform. In *International Conference on Computational Linguistics: Demonstrations*, pages 13–16.

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

Markus Dickinson. 2010. Generating learner-like morphological errors in russian. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 259–267. Association for Computational Linguistics.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126.

Jennifer Foster and Øistein E Andersen. 2009. Generrate: generating errors for use in grammatical error detection. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*, pages 82–90. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing*, pages 478–490. Springer.

Hanban. 2001. 汉语水平词汇与汉字等级大纲 *The Syllabus of the Graded Words and Characters for Chinese Proficiency Test*. 经济科学出版社 Economic Science Press.

Hanban. 2010. 新汉语水平考试大纲 *New Chinese Proficiency Test Syllabus*. 商务印书馆 The Commercial Press, China.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 388–392. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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 *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. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 67–70.

Minh Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *Computer Science*.

Alla Rozovskaya and Dan Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 154–162. Association for Computational Linguistics.

Alla Rozovskaya, Dan Roth, and Mark Sammons. 2017. Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4):723–760.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network.

Allen Schmaltz, Yoon Kim, Alexander M Rush, and Stuart M Shieber. 2016. Sentence-level grammatical error identification as sequence-to-sequence correction. *arXiv preprint arXiv:1604.04677*.

Yi Yang, Pengjun Xie, Jun Tao, Guangwei Xu, Linlin Li, and Si Luo. 2017. Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into lstms for chinese grammatical error diagnosis task. In *Proceedings of the IJCNLP 2017, Shared Tasks, Taipei, Taiwan, November 27 - December 1, 2017, Shared Tasks*, pages 41–46.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting word ordering errors in chinese sentences for learning chinese as a foreign language. *Proceedings of COLING 2012*, pages 3003–3018.

Zheng Yuan and Mariano Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61.

Bo Zheng, Wanxiang Che, Jiang Guo, and Ting Liu. 2016. Chinese grammatical error diagnosis with long short-term memory networks. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 49–56, Osaka, Japan. The COLING 2016 Organizing Committee.