Pre-Training with Syntactic Structure Prediction for Chinese Semantic Error Recognition

Anonymous ACL submission

Abstract

Existing Chinese text error detection mainly focuses on spelling errors and simple grammatical errors. These errors have been studied extensively and are relatively simple for humans. Chinese Semantic Error Recognition (CSER) pays attention to more complex semantic errors that humans cannot easily recognize compared with Chinese text error detection. Considering the complex syntactic relation between words, we find that syntactic structure from the syntax tree can help identify semantic errors. In this paper, we consider adopting the pre-trained models to solve the task of CSER. To make the model learn syntactic structure in the pre-training stage, we designed a novel pre-training task to predict the syntactic structure from the syntax tree between different words. Due to the lack of a published dataset for CSER, we build a high-quality dataset for CSER for the first time named Corpus of Chinese Linguistic Semantic Acceptability (CoCLSA), which is extracted from the high school examinations. The experimental results on the CoCLSA show that our pre-trained model based on the new pre-training task has a positive performance compared with existing pre-trained models.

1 Introduction

The recognition of text errors such as Chinese spelling errors (Jiang et al., 2012; Wang et al., 2021) and Chinese grammar errors (Lee et al., 2015) is widely mentioned in previous research. However, there is no sufficient research on semantic errors of Chinese sentences, including improper collocation, improper word order, incomplete or redundant components, confusion in structure, unclear semantics, and illogical errors. The recognition of semantic errors has essential applications in the domain of education, journalism, and publishing. In this paper, we consider the Chinese Semantic Error Recognition (CSER) task, a binary classification task to determine whether a Chinese sentence has semantic errors.

Task	Sentence
CSC	个人~~触须~~（储蓄）卡存款也有利息吗 Is there interest on personal ~~tentacle~~ (debit) card deposits
GCED	从小到大~~为了你们~~（你们为了）照顾我，付出很多 Since childhoood, you have paid a lot to take care of me
CSER	英法帝国主义~~烧毁并洗劫~~（洗劫并烧毁）了北京圆明园 British and French imperialism ~~burned and ransacked~~ (ransacked and burned) Beijing’s Old Summer Palace

Table 1: Examples of different tasks. The incorrect semantics cannot be translated in the sentence in CGED task, so we translate it into the correct semantics.

Unlike Chinese Spelling Check (CSC) and Chinese Grammatical Error Diagnosis (CGED), CSER is oriented to more complex incorrect sentence phenomena and needs to understand the sentence’s semantics to make a judgment. Table 1 shows the examples of text errors for different tasks. As shown in Table 1, the error in the sentence in CSER task is improper word order because “洗劫” (ransacked) should be placed before “烧毁” (burned) due to the time sequence. More examples can be seen in Appendix A, consisting of all types of semantic errors. According to Table 1, we can see that Chinese grammatical errors and Chinese spelling errors can be recognized easily by humans. However, the sentences with semantic errors are relatively fluent and even difficult for humans to recognize because these errors usually require the syntactic structure of words to be judged.

Sentences with semantic errors have a strong correlation with syntactic information. For example, as shown in Figure 1, the mistake of the sentence is that there is a wrong dependency between “烧毁” (burned) and “洗劫” (ransacked), which can be discovered by syntactic parsing. In the incorrect sentence, “烧毁” (burned) is the parent node of “洗劫” (ransacked). However, in the correct sentence, “洗劫” (ransacked) should be the parent node of “烧毁” (burned). Therefore, it is beneficial for the model to learn the relationship of nodes in the syntax tree.

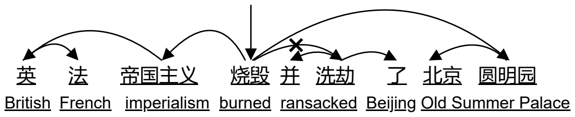

Figure 1: Syntax analysis of semantic incorrect sentences, the incorrect semantic is marked as “×”.

In this paper, we introduce a novel pre-training task allowing the model to learn more syntactic information. We randomly select a couple of words in a sentence and then predict the relationship of a couple of words through the representation of the hidden layer of BERT. Then we perform pre-training on a corpus of one million sentences from Wikipedia. We fine-tune the obtained pre-trained models on a high-quality dataset.

To obtain a high-quality dataset for the CSER task, we use the web crawler to obtain Chinese multiple-choice questions related to incorrect semantic sentences from the online resources of the high school examinations in the past ten years. We extract sentences from the above question bank. Then we organize these data into a dataset with correct sentences and incorrect semantic sentences, named Corpus of Chinese Linguistic Semantic Acceptability (CoCLSA). We fine-tune the pre-trained models on CoCLSA. The experimental results show that our proposed model exceeds the existing pre-trained models.

Our main contributions are summarized as follows:

- We propose a novel pre-training task to learn directional and differentiated syntactic information without adding additional knowledge. Experiments have proved that our pre-training task can solve the CSER task better.

- To solve the problem of vacancy of the dataset for the CSER task, we provide a high-quality dataset for CSER with 24,228 sentences, namely CoCLSA.

To facilitate this research, all the codes and the CoCLSA datasets in this paper will be released at https://XXX.

2 Related Work

Many researchers have made outstanding achievements on CSC (Zhang et al., 2020) and CGED (Fu et al., 2018). Existing CSC and CGED models cannot achieve good results for CSER because semantic errors are often difficult compared to other errors. The existing CGED models do not consider the characteristics of semantic errors; they are related to the syntactic structure. Some researchers try to solve CSER based on rules (Wu et al., 2015) and the Semantic Knowledge-base (Guan and Zhang, 2012; Zhang et al., 2021). However, for some more complex and obscure semantic errors, the traditional method is powerless. As far as we know, there are currently almost no investigators researching Chinese Semantic Error Recognition (CSER) through the pre-trained model.

A lot of pre-training tasks are proved to be effective, such as pre-training tasks in BERT, ERNIE 3.0 (Sun et al., 2021), BERT-wwm (Cui et al., 2019) and RoBERTa (Liu et al., 2019). The pre-training task in SEPREM (Xu et al., 2021) is a dependency prediction task, which predicts the depth of dependence between two tokens. Syntactic information is added explicitly into SEPREM, which is complicated to implement. Meanwhile, the differences and directionality of syntax are not considered. In our paper, we try to make the model learn syntactic information in the pre-training stage with a novel pre-training task to solve the CSER task better.

3 Methodology

3.1 Syntax Tree

Syntactic parsing shows a substantial improvement in the field of NLP. In this paper, we use the dependency parser of LTP (Che et al., 2010) to generate syntactic parsing, which provides a series of Chinese natural language processing tools. Then we generate the syntax tree based on the results of the syntactic parsing. We define the syntax tree as $\mathcal{T} = \{\mathcal{C}, \mathcal{N}, \mathcal{E}\}$, where $\mathcal{C}$ represents the correlation between two nodes, $\mathcal{N}, \mathcal{E}$ represents node and edge set. $\mathcal{D}(\mathcal{N}_i, \mathcal{N}_j)$ is denoted as the distance between node $\mathcal{N}_i$ and $\mathcal{N}_j$, which is the minimal distance from node $\mathcal{N}_i$ along the edge to node $\mathcal{N}_j$. And we consider four types of correlation: child, parent, sibling and indirection as follows.

- $\mathcal{C}_{ij}$=child if $\mathcal{N}_i$ is child node of $\mathcal{N}_j$ and $\mathcal{D}(\mathcal{N}_i, \mathcal{N}_j)$=1.

- $\mathcal{C}_{ij}$=parent if $\mathcal{N}_i$ is parent node of $\mathcal{N}_j$ and $\mathcal{D}(\mathcal{N}_i, \mathcal{N}_j)$=1.

- $\mathcal{C}_{ij}$=sibling if $\mathcal{N}_i$ and $\mathcal{N}_j$ have the same parent $\mathcal{N}_k$ and $\mathcal{D}(\mathcal{N}_i, \mathcal{N}_k)$=1 and $\mathcal{D}(\mathcal{N}_j, \mathcal{N}_k)$=1.

- $\mathcal{C}_{ij}$=indirection if any of the above is not met.

In particular, we consider that the nodes in the syntax tree may contain multiple Chinese characters. Therefore, we set each Chinese character in the node to the same correlation with another node, and the correlation within the node is set to indirection. The difference in syntactic structure is reflected in the fact that we divide them into several different categories. The directionality of the syntactic structure is reflected in the parent node and the child node. On the other hand, we consider $\mathcal{D}(\mathcal{N}_i, \mathcal{N}_j)$=1 for the reason that deeper syntactic structures are more complicated and difficult for the model, and burden the model.

3.2 Syntactic Structure Prediction

We have the following two pre-training tasks as shown in Figure 2. The first one is MLM, the same as BERT, which can learn the semantic relationship of context. Another pre-training task is Syntactic Structure Prediction (SSP), which is proposed to allow the pre-trained model to learn the syntactic structure from the syntax tree explicitly. First of all, we can learn from the syntax tree that there are four correlations between any two tokens, namely child, parent, sibling, and indirection. We select some pairs of Chinese characters and let the model predict the relationship between them. To ensure the balance of labels, 25% of the correlations between pairs of Chinese characters we choose are child, 25% are parent, 25% are sibling, and others are indirection. We use BERT to generate the representation of the last hidden states of the pairs of Chinese characters we selected and then put it into the classifier for classification tasks. In this paper, we select Multilayer Perceptron (MLP) as the classifier consisting of 4 layers. We select Rectified Linear Unit as an activation function in MLP. We predict the type of relationship. Therefore, the direction and difference of the syntactic structure are well learned by the pre-trained model.

4 Experiments

4.1 CoCLSA

We use the web crawler to obtain Chinese multiple-choice questions related to incorrect semantic sentences from the high school examination online resources in the past ten years. Then we organize these data into a dataset with a total of 24,228 sentences with two labels. One of the labels is correct sentences, and the other is incorrect semantic sentences. We choose 20,291 sentences as the train dataset, 1,937 sentences as the validation dataset, and 2,000 as the test dataset. Since most of the multiple-choice questions we crawl are sentences with semantic errors, there are more semantic incorrect sentences in CoCLSA. To ensure reasonableness, we divide the validation and test sets with the same number of correct and incorrect semantic sentences. Therefore, the proportion of incorrect semantic sentences is higher in the training set. We tried to use oversampling and undersampling methods to make the number of correct and incorrect semantic sentences roughly the same in the train set. However, we find that the effect is the same (or even worse) as formal training. Hence, we use the train set in Table 2 directly for experiments. More details can be seen in Table 2.

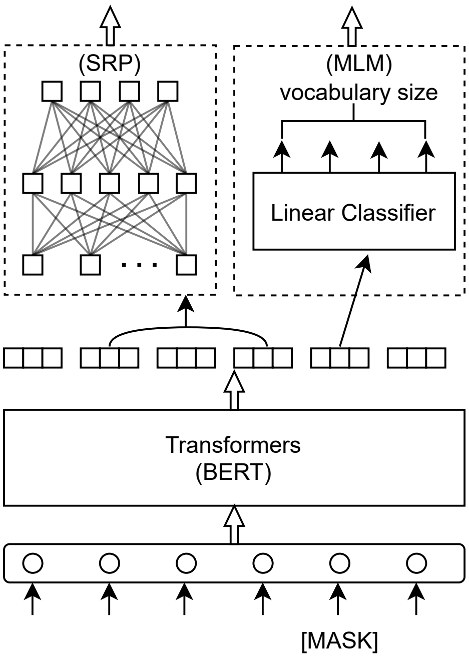

Figure 2: Structure of our pre-trained model.

Model	#Line	Avg.Length	Error Ratio
Train	20,291	56.78	78.34%
Dev	1,937	57.44	49.97%
Test	2,000	57.53	50.00%

Table 2: Details of CoCLSA where Error Ratio means the proportion of semantic incorrect sentences in the total data.

4.2 Experimental setup

In the pre-training stage, we use 1 million Wikipedia data as pre-training dataset. We use

Model	P	R	F_1	ACC
BERT	64.7±0.9	77.3±0.8	70.5±0.3	67.9±0.7
BERT-SSP	65.7±0.2	**78.4**±0.3	71.5±0.2	68.7±0.2
ERNIE	65.7±0.4	77.7±1.4	71.2±0.4	68.5±0.1
ERNIE-SSP	66.3±0.6	77.7±0.4	71.5±0.2	69.1±0.5
BERT-wwm	66.6±0.8	75.8±0.5	70.9±0.4	68.9±0.6
BERT-wwm-SSP	67.1±0.7	77.8±1.8	72.0±0.6	69.8±0.4
RoBERTa	67.8±0.7	77.5±1.6	72.3±0.4	70.4±0.2
RoBERTa-SSP	**68.9**±0.9	77.4±1.2	**72.9**±0.2	**71.2**±0.4

Table 3: We report the average score and standard deviation of 3 independent runs with different seeds.

Model	P	R	F_1	ACC
iFLYTEK-HFL	55.5	24.4	33.9	52.4
Meta Writing Assistant	59.0	28.8	38.7	54.4
BERT-SSP	67.5	76.4	71.7	69.8
ERNIE-SSP	67.9	76.0	71.7	70.0
BERT-wwm-SSP	66.0	**80.0**	**72.3**	69.4
RoBERTa-SSP	**70.8**	73.6	72.2	**71.6**
Human	72.4	78.6	75.1	74.1

Table 4: Comparison of humans and machines based on a small test dataset containing 500 sentences, which has the same number of correct and incorrect sentences.

LTP[1] as a tool for syntactic parsing. We pre-train for 10 epochs with an effective batch size of 256. We use AdamW optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) with a learning rate of 2e-5. We use a learning rate warmup for 2,500 steps. In the fine-tuning stage, we use CoCLSA as fine-tuning dataset. We fine-tune the pre-trained models for 4 epochs with an effective batch size of 32. We use AdamW optimizer with a learning rate of 2e-5 and weight decay of 0.01. The implementation of pre-training and fine-tuning is based on HuggingFace's Transformer (Wolf et al., 2019), which consists of 12-layer, 768-hidden, 12-heads. We have the following baselines: ***BERT-base,Chinese*** (Devlin et al., 2019), ***BERT-wwm*** (Cui et al., 2019), ***ERNIE*** (Sun et al., 2019), ***RoBERTa*** (Liu et al., 2019).

4.3 Results and Analysis

Table 3 demonstrates the results of different models on the CSER task fine-tuned with CoCLSA. Firstly, the best performance in our models is accomplished by RoBERTa-SSP, which has an improvement of 0.6% F1 score and 0.8% accuracy compared with RoBERTa. Meanwhile, our models with a new pre-training task SSP outperform their baseline models with an improvement of 1% approximately. This proves that the pre-training task SSP is adequate for the CSER task.

Table 4 shows the comparison of humans and machines based on a small test dataset containing 500 sentences extracted from the test set in Table 2. To explore the difference between the level of machines and humans, we choose four college students, including undergraduates, postgraduates, and doctoral students. We ask them to label the small test set without external help, such as the network. We test the indicators of each person, such as F1 score and accuracy, and finally, take the average of four people. Although the labeling level of four people cannot represent the average level of the entire human being, it reflects the characteristics of college students on the CSER task to a certain extent. We calculate the kappa correlation coefficient between any two people and find that the results are between 35%-45%, proving that the CSER task is tricky for college students. As we can see in Table 4, there is still a distance for machines to reach the human level of CSER task.

We select publicly available Meta Writing Assistant[2] and iFLYTEK-HFL[3] systems for measurement. Our pre-trained models outperform them by at least 33% F1 score and 17% accuracy score. The results show that the existing Chinese error correction systems can not recognize semantic errors well compared with our pre-trained models.

5 Conclusion

Unlike CGED and CSC, CSER is more difficult for humans to identify semantic errors and more dependent on syntactic information due to the complexity and variety of incorrect semantic sentences. This paper proposes the pre-trained model with a novel pre-training task that can learn syntactic information better. Due to the vacancy of the dataset for CSER, we provide a high-quality dataset named CoCLSA. We fine-tune the pre-trained models on CoCLSA. The experimental results show that our models exceed existing pre-trained models and Chinese error correction systems. Our work could bring machines closer to the human level for the CSER task, and there is much room for improvement in the future.

[1] http://ltp.ai/

[2] https://xiezuocat.com/

[3] http://check.hfl-rc.com/

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16, Beijing, China. Coling 2010 Organizing Committee.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *ArXiv preprint*, abs/1906.08101.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 52–59, Melbourne, Australia. Association for Computational Linguistics.

Jun Guan and Yang-sen Zhang. 2012. Study of automatic semantic errors checking for chinese text based on skcc. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 2983–2986. IEEE.

Ying Jiang, Tong Wang, Tao Lin, Fangjie Wang, Wenting Cheng, Xiaofei Liu, Chenghui Wang, and Weijian Zhang. 2012. A rule based chinese spelling and grammar detection system utility. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437–440. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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*, pages 1–6, Beijing, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv preprint*, abs/2107.02137.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *ArXiv preprint*, abs/1904.09223.

Baoxin Wang, Wanxiang Che, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2021. Dynamic connected networks for chinese spelling check. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2437–2446.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Yefan Wu, Runbo Zhuang, Ying Jiang, and Fan Li. 2015. Research and realization of chinese text semantic correction based on rule. In *J]. Proceedings of 2015 3rd International Conference on Education, Management, Arts, Economics and Social Science (ICEMAESS 2015)*.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online. Association for Computational Linguistics.

Rui Zhang, Yangsen Zhang, Gaijuan Huang, and Ruoyu Chen. 2021. Research on proofreading method of semantic collocation error in chinese. In *International Conference on Artificial Intelligence and Security*, pages 709–722. Springer.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.

ID	Sentence	Error Type
1	我国棉花~~的生产~~，长期以来，一直不能自给，需要适量的进口 The ~~production of~~ cotton in our country has been unable to support itself for a long time, and it needs appropriate imports	improper collocation
2	英法帝国主义~~烧毁并洗劫~~（洗劫并烧毁）了北京圆明园 British and French imperialism ~~burned and ransacked~~ (ransacked and burned) Beijing's Old Summer Palace	improper word order
3	有关部门严肃处理了某些加油站擅自哄抬汽油价格（的行为） Relevant departments have seriously dealt with (the acts of) some gas stations driving up gasoline prices without authorization	incomplete components
4	有一部分网友却对雷锋及雷锋精神提出了各种各样的所谓~~质疑~~（疑问） Some netizens have raised various so-called doubts about Lei Feng and his spirit	redundant components
5	他的家乡是福建省福州市~~人~~ His hometown is Fuzhou, Fujian Province ~~people~~	confusion in structure
6	山上的水宝贵，我们把它留给~~晚上来~~的人喝 The water is precious, we leave it to people who come to drink ~~at night~~ (late)	unclear semantics
7	与会专家一致认为，当前防止煤矿~~不~~出事故的最好办法，就是加强安全工作 The experts at the meeting agreed that the best way to prevent coal mines ~~without~~ accidents is to strengthen safety work	illogical errors

Table 5: Examples of CSER task with seven types of error.

A Appendix

We list all types of semantic errors as shown in Table 5. These semantic errors are often examined in Chinese examinations of junior and senior high schools. In Sentence 1, "不能自给" (unable to support) should modify "棉花" (cotton), not "棉花的生产" (production of cotton). In Sentence 2, "烧毁" (burned) should be placed after "洗劫" (ransacked) according to the chronological order. Sentence 3 lacks the object "的行为" (the acts). In Sentence 4, "质疑" (call into question) contains the meaning of "提出" (raise). Sentence 5 mixes the two complete sentences together. The one is that "他的家乡是福建省福州市" (His hometown is Fuzhou, Fujian Province). The other is that "他是福建省福州市人" (He is a native of Fuzhou, Fujian Province). Sentence 6 is ambiguous because "晚上来" can be interpreted as "晚上/来" (at night) or "晚/上来" (late). There is a logical problem with Sentence 7, that is, inappropriate multiple negative words. We only need to keep one of "防止" (prevent) and "不" (without).