Pseudo-label Data Construction Method and Syntax-enhanced Model for Chinese Semantic Error Recognition

Hongyan Wu[1], Nankai Lin[2,✉], Shengyi Jiang[3], Lianxi Wang[2], Aimin Yang[4,5]
[1] College of Computer, National University of Defense Technology
[2] School of Information Science and Technology, Guangdong University of Foreign Studies
[3] School of Information Technology and Engineering, Guangzhou College of Commerce
[4] School of Computer Science and Technology, Guangdong University of Technology
[5] School of Computer Science and Intelligence Education, Lingnan Normal University
Correspondence: neakail@outlook.com

Abstract

Chinese Semantic Error Recognition (CSER) has always been a weak link in Chinese language processing due to the complexity and obscureness of Chinese semantics. Existing research has gradually focused on leveraging pre-trained models to perform CSER. Although some researchers have attempted to integrate syntax information into the pre-trained language model, it requires training the models from scratch, which is time-consuming and laborious. Furthermore, despite the existence of datasets for CSER, the constrained size of these datasets impairs the performance of the models. Thus, in order to address the difficulty posed by a limited sample set and the need of annotating samples with semantic-level errors, we propose a **P**seudo-label **D**ata **C**onstruction method for **CSER** (**PDC-CSER**), generating pseudo-labels for augmented samples based on perplexity and model respectively, which overcomes the difficulty of constructing pseudo-label data containing semantic-level errors and ensures the quality of pseudo-labels. Moreover, we propose a **CSER** method with the **D**ependency **S**yntactic **A**ttention mechanism (**CSER-DSA**) to explicitly infuse dependency syntactic information only in the fine-tuning stage, achieving robust performance, and simultaneously reducing substantial computing power and time cost. Results demonstrate that the pseudo-label technology PDC-CSER and the semantic error recognition method CSER-DSA surpass the existing models.

1 Introduction

As an essential phase in the text proofreading task, text error recognition is extremely crucial to improve the performance of error correction. Chinese text errors are divided into three categories: spelling errors, grammatical errors, and semantic errors. Existing research concerning Chinese text error detection mainly pays attention to Chinese Spelling Check (CSC) as well as Chinese Grammatical Error Diagnosis (CGED) (Wu et al., 2023; Yue et al., 2022; Sun et al., 2023b). Nevertheless, it is still relatively weak in Chinese text error recognition on the semantic level due to Chinese semantics' complexity. Semantic error recognition is conducive to many downstream applications, for instance, automatic speech recognition (Zhao et al., 2021) automated essay scoring (Uto et al., 2020). Hence, in this paper, we focus on Chinese Semantic Error Recognition (CSER), determining whether a Chinese sentence has semantic errors[1].

Large-scale and high-quality annotation data is vital for achieving Chinese text error detection tasks. The research on Chinese spelling errors and grammatical errors has proposed some effective data synthesis methods and provided a large number of available datasets (Wang et al., 2018; Zhang et al., 2022b). Despite the initiation of research efforts into CSER, the inadequacy of pertinent training datasets persists. Sun et al. (2022) proposed the first high-quality annotated dataset for CSER, namely Corpus of Chinese Linguistic Semantic Acceptability (CoCLSA), bridging the gap in the field. Notwithstanding the availability of the dataset for CSER, its size is limited. We conduct experiments on the dataset to explore the performance of the CSER model at 50%, 60%, 70%, 80%, 90%, and 100% of the dataset, and find that the model performance keeps growing without reaching a plateau value. The result in Fig. 1 indicates that the model performance for CSER may further improve with a sustained increase in the training dataset scale. Furthermore, compared with the corpus for CSC and CGED, it is more challenging to annotate the datasets for CSER because the semantic-level errors are complex.

For the CSER task, the traditional CSER meth-

[1]Detailed examples for various Chinese text errors are present in Appendix A.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5391–5402
January 19–24, 2025.

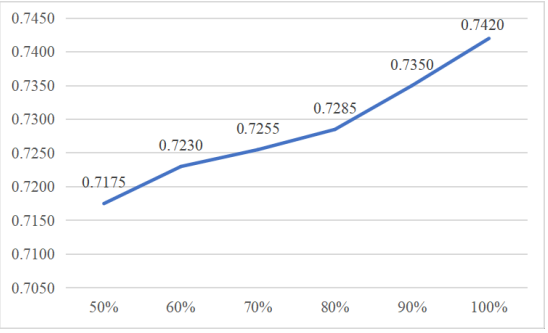


Figure 1: Experimental results in different scale raw data.

ods are mainly based on the combination of rules and statistical models (Luo et al., 2002; Wu et al., 2015). The leverage of inference methods such as description logic reasoning machines further improves the effect of CSER (Ying et al., 2017). However, the establishment of rules and knowledge bases is labor-intensive and the performance of CSER is limited by the quality of the knowledge bases. With the development of deep learning technology, it is a potential research avenue to facilitate the CSER task with the rich semantic knowledge embedded in large-scale pre-trained language models. Existing research has attempted to integrate dependency syntactic information into pre-trained language models for CSER, but requires training the pre-trained model from scratch, consuming a huge waste of energy and time (Sun et al., 2022).

To tackle the above challenges, we propose a **P**seudo-label **D**ata **C**onstruction method for **CSER** (**PDC-CSER**) to extend samples, compensating for the lack of data. Our method is based on two effective strategies to jointly screen the pseudo-labels of augmented samples, which ensures the quality of pseudo-labels. Furthermore, we put forth a **D**ependency **S**yntactic **A**ttention mechanism (**CSER-DSA**) that explicitly integrates dependency syntactic information, exclusively utilized in the fine-tuning stage, which considerably reduces both computational resources and time requirements. Our main contributions are summarized as follows:

1) We introduce a novel method to construct pseudo-label data for CSER. Our method effectively addresses the difficulty posed by constructing pseudo-label data containing semantic-level errors, while ensuring the quality of pseudo-labels.

2) We propose a novel approach of utilizing a dependency syntactic attention mechanism for CSER. This approach explicitly incorporates dependency syntactic information and operates solely in the fine-tuning phase, achieving strong performance and saving significant computational resources and time costs.

3) Results on the CoCLSA dataset demonstrate that the proposed pseudo-label technology PDC-CSER and the semantic error recognition method CSER-DSA outperform the existing models.

2 Related Work

2.1 Chinese semantic error recognition

CSER has long been regarded as a challenging issue in Chinese text error detection. Although existing methods for CSC and CGED have achieved notable success, they are not equipped to tackle the complexity and obscurity of Chinese semantic errors, which are more intricate compared to spelling and grammatical errors (Yunhan et al., 2022). Early CSER approaches mainly relied on rules and statistical models (Luo et al., 2002; Wu et al., 2015). Thereafter, some researchers have attempted to adopt knowledge bases and semantic reasoning methods to perform Chinese semantic error detection (Zhang et al., 2021; Ying et al., 2017).

Nevertheless, the traditional CSER methods rely on manually established rules and knowledge bases, and the performance is limited by the scale and quality of the knowledge bases, making them unsuitable for large-scale semantic error recognition. Pre-trained language models have learned rich prior linguistic, syntactic, and lexical information for downstream tasks through unsupervised training on a large corpus in the pre-training stage. Thus, scholars have conducted investigations about leveraging the rich semantic information in the pre-trained language models to solve semantic errors. Sun et al. (2022) made the first attempt to introduce a pre-trained language model into the CSER task and provided the first annotated datasets for CSER. Whereas small-scale datasets limit the model performance. To our knowledge, there is no investigation into pseudo-label data construction oriented to semantic error recognition, given the difficulty of annotating texts with semantic errors.

2.2 Syntax-enhanced model

Previous research has demonstrated the great potential of syntactic information in various natural language processing tasks (Zhang et al., 2022c; Strubell et al., 2018). Simultaneously, some researchers (Min et al., 2020) stated that pre-trained

language models cannot capture the deep syntactic structure and syntactic information in terms of dependency weights. Therefore, an increasing number of studies have designed syntax-related pre-training tasks to inject dependency syntactic information into pre-trained language models, enhancing the performance of the models Zhang et al. (2022a); Wang et al. (2021).

Yet it is labor-intensive and time-consuming to inject dependency syntactic information in the pre-training phase. Thus, some work focuses on incorporating dependency syntactic information into the pre-trained language model only in the fine-tuning stage, likewise, achieving remarkable improvement (Nguyen et al., 2020). Considering the effectiveness of dependency syntactic information and the fact that most error types of Chinese semantic errors are related to dependency syntax, it is feasible to exploit dependency syntactic information to assist in recognizing Chinese semantic errors. Despite the available method of integrating syntactic information for semantic error recognition (Li et al., 2013; Sun et al., 2022), this method requires training the model from scratch, consuming high energy and time costs. Hence, we infuse more effective dependency syntactic information via a self-attention mechanism, which can also improve the performance of semantic error recognition when employed only in the fine-tuning phase.

3 Pseudo-label data construction method oriented to semantic error recognition

As demonstrated by the empirical experiment in Fig. 1, expanding the scale of the datasets can further enhance the performance of the CSER model. The workflow of the pseudo-label data construction method is shown in Fig. 2. Initially, we adopt a relatively easy yet effective Easy Data Augmentation (Wei and Zou, 2019) method to generate augmented samples. Specifically, we expand the raw training set by randomly replacing synonyms based on the off-the-shelf Synonym Tool[2]. For each sample, synonyms are leveraged to randomly replace 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40% tokens of the raw sentence to obtain the augmented samples. Our initial objective is to minimize the introduction of noise during the data augmentation process. Hence, we exclusively select the top three synonyms with the highest similarity yield by Synonym Tool to randomly replace tokens in the original text. Additionally, we generate high-confidence pseudo-labels for each augmented sample perplexity-based and model-based prediction, respectively. Ultimately, we retain only augmented samples that exhibit consistent pseudo-labels generated by both strategies for training the semantic error recognition model for Chinese text.

3.1 Pseudo-perplexity-based pseudo-label generation

Motivated by Salazar et al. (2020), based on the perplexity scoring mechanism of the pre-trained language models, we utilize the pseudo-perplexity (PPPL) to perform unsupervised acceptability judgment on the augmented samples and assign them pseudo-labels. Pre-trained language models, such as BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019), have achieved great success in contextual language representation due to the use of masked language modeling object, where a token w_t is replaced with [MASK] and predicted via all past and future tokens $S_{\setminus t} := \left(w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{|S|}\right)$. Subsequently, in scoring the fluency of the sample, we initially calculate the relevant pseudo-log-likelihood scores originating from the pre-trained model, which involves the summation of the conditional log probabilities $\log P_{MLM}\left(w_t \mid S_{\setminus t}\right)$ of individual sentence tokens. Specifically, we construct copies with each token masked out and sum the log probability for each masked token over copies to compute the pseudo-log-likelihood score. For each sentence S, the pseudo-log-likelihood score is calculated as follows:

$$PLL(\mathrm{S}) := \sum_{t=1}^{|S|} \log P_{MLM}\left(w_t \mid S_{\setminus t}\right). \quad (1)$$

We further compute the PPPL of each raw sentence and the augmented sentence based on the pseudo-log-likelihood score, which is used in lieu of perplexity. The PPPL is defined as follows:

$$PPPL(\mathrm{S}) := \exp\left(-\frac{1}{|S|} PLL(S)\right). \quad (2)$$

When generating pseudo-labels for augmented samples via pseudo-perplexity, we deem that if the PPPL score of the corresponding augmented sample exceeds that of the raw sentence without semantic errors, the augmented sample will be assigned a pseudo-label indicating that there are semantic errors. Conversely, provided that the PPPL score of

[2]https://github.com/chatopera/Synonyms

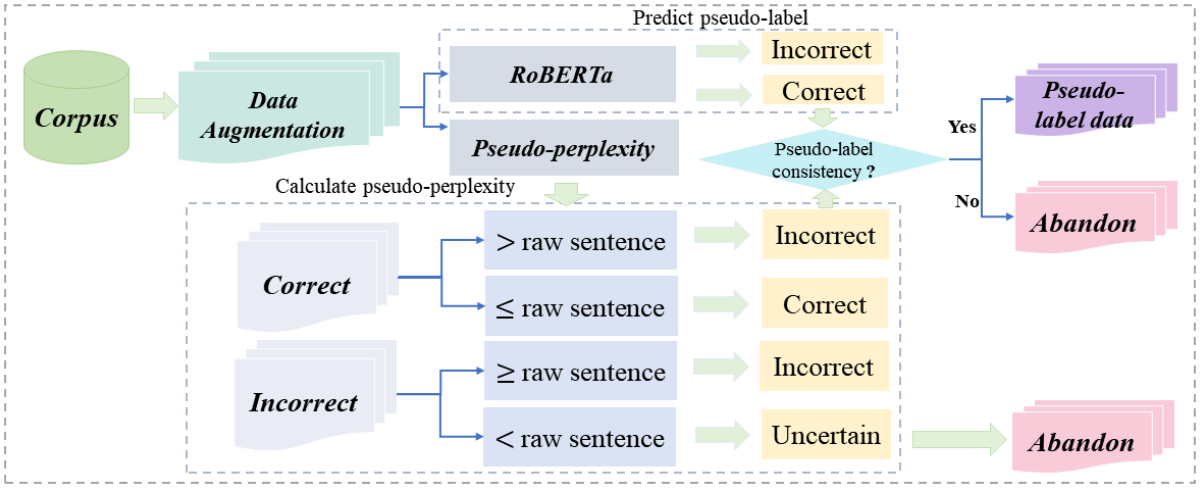


Figure 2: The process of pseudo-label corpus construction.

its corresponding augmented sample is less than or equal to the raw sentence, the augmented sample will be assigned a pseudo-label representing no semantic errors. In the same vein, the corresponding augmented sample with a PPPL score greater than or equal to the raw sentence containing semantic errors will be treated as existing semantic errors, generating relevant pseudo-label. Whereas if the PPPL score for an augmented sample is lower than that of the raw sentence, it is not necessarily considered correct, thus preventing an accurate pseudo-label from being assigned. Therefore, we eliminate the augmented data that cannot accurately determine the pseudo-label to ensure the quality of pseudo-labels.

3.2 Model-based prediction pseudo-label generation

Aiming to predict the probability distribution of pseudo-label of unlabeled texts, we train a semantic error recognition model based on RoBERTa with softmax activation function layer using the raw data. Concretely, concerning each augmented sample, RoBERTa is employed to encode unlabeled sentences to obtain a feature vector, which is then fed into a linear classifier, yielding pseudo-label "0" or "1" depending on a large score, where "0" indicates that augmented sample contains semantic errors and "1" indicates that the augmented sample has no semantic errors.

4 Infused-syntax Chinese semantic error recognition model

As shown in Fig. 3, we propose a syntax-infused model, which incorporates syntax information into pre-trained language models only in the fine-tuning phase. The model consists of four modules: text encoding (TE) module, correlation matrix construction (CMC) module, dependency syntactic information infusion (DSIN) module, and Chinese semantic error recognition (CSER) module. In the text encoding module, we obtain the representation of the sentence exploiting a pre-trained language model. In the correlation matrix construction module, dependency distance has been proposed to effectively capture global syntactic information. Subsequently, we utilize a distance matrix to encode the dependency syntactic tree structure, followed by the normalization of each element to generate a correlation matrix. Moreover, to infuse dependency syntactic information into context representation to obtain syntax-aware representation, we construct a dependency syntactic information infusion module based on the dependency syntactic attention mechanism. Finally, the updated text representation is input into the Chinese semantic error recognition module to determine whether a sentence has a semantic error.

4.1 Text encoding module

In the text encoding module, we exploit a RoBERTa pre-trained model with outstanding performance in text semantic representation to encode sentences. The pre-trained model learns a large amount of prior linguistic, syntactic, and lexical information for downstream tasks through unsupervised training on a mass of corpus in the pre-trained stage. Considering the lack of annotated corpus for the CSER task, we regard the RoBERTa as the backbone of the semantic error recognition model to make full use of the abundant linguistic informa-

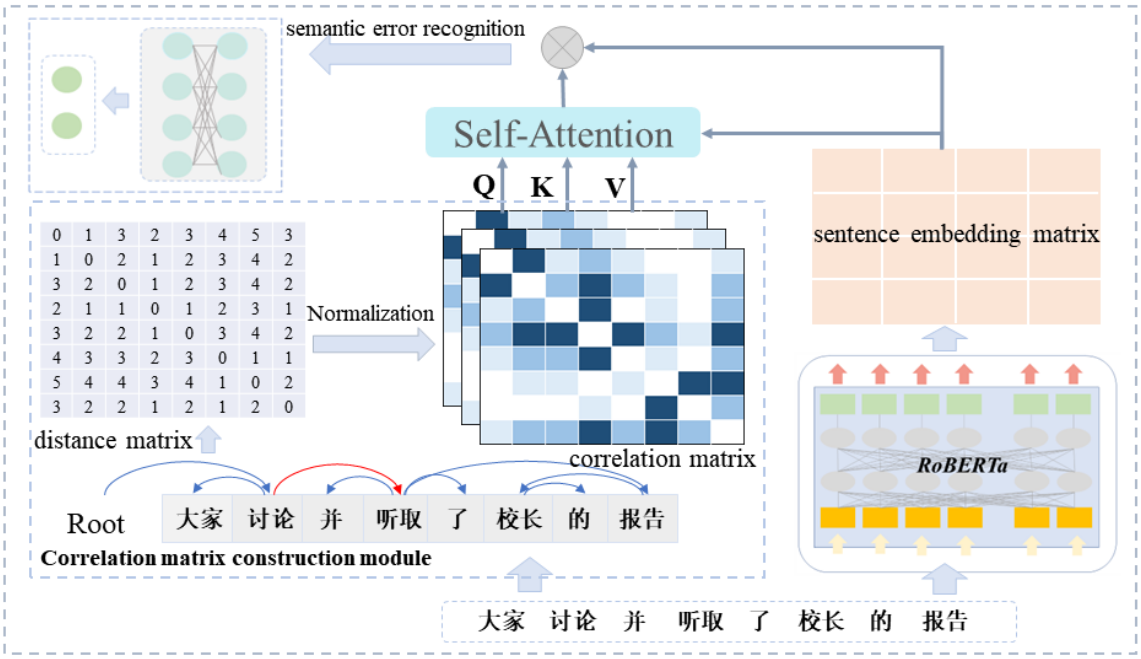


Figure 3: The architecture of infused-syntax model for Chinese semantic error recognition.

tion contained in it. For given input sequence S_i corresponding to subscript i, the process of computing the semantic feature H_i is:

$$H_i = RoBERTa(a_i, b_i, c_i), \tag{3}$$

where a_i, b_i, and c_i are the token, segment, and position embedding respectively.

4.2 Correlation matrix construction module

When constructing the correlation matrix, we adopt the LTP tool[3] (Che et al., 2021) to extract the dependency syntactic information of all sentences and generate the dependency syntactic tree. Furthermore, we define the dependency distance based on the dependency syntactic tree to construct the distance matrix, which is further normalized to obtain the correlation matrix for encoding the syntax tree structure. Our definition of dependency distance and the syntax structure encoding are described in detail in the following.

Dependency Distance over Dependency Syntactic Tree. From an intuitive perspective, the distance separating two tokens in the dependency syntactic tree is indicative of their semantic relatedness. The proximity of the two tokens implies a stronger linguistic association between them. Consequently, we define the dependency distance between two tokens as the number of edges contained in the path from one node to another over the dependency syntactic tree. Specifically, assuming that token v_p is the head of the token v_q, the dependency distance between token v_p and token v_q is $d(v_p, v_q) = 1$. If there is no edge directly connected between token v_p and token v_q, the dependency distance between them is the sum of the edges on the path in the dependency graph. Note that in order to fully employ the syntactic structure, the dependency syntactic tree is simplified to an undirected graph, such that equal significance is attributed to the strength of syntactic dependency between two tokens.

Syntax Tree Structure Encoding. We encode the structure of the dependency syntactic tree via distance matrix D, which can accurately reflect the distance between two provided tokens. Given a sentence corresponding to a dependency syntactic tree, the distance from the p-th token v_p to the q-th token v_q, namely, the element $D_{p,q} \in D$ of the p-th row and q-th column in the distance matrix D is defined as:

$$D_{p,q} = \begin{cases} d(v_p, v_q), & \text{if exist a path from } v_p \text{ to } v_q, \\ 0, & \text{if } p = q. \end{cases} \tag{4}$$

Based on the fact that the dependency distance is inversely proportional to the correlation strength, the distance matrix D is normalized to obtain the correlation matrix $\widetilde{D_{p,q}}$ as follows:

$$\widetilde{D_{p,q}} = \frac{1}{D_{p,q} + 1}. \tag{5}$$

[3]https://github.com/HIT-SCIR/ltp

4.3 Dependency syntactic information infusion module

The correlation matrix constructed for each sentence based on dependency syntactic information is invariant. Nevertheless, the importance of dependency syntactic information varies from token to token. We expect the model to pay more attention to the essential dependency syntactic information in the training phase. Therefore, we leverage the attention mechanism to process the correlation matrix to obtain a learnable correlation matrix:

$$\text{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

$$M = \text{Attention}\left(\widetilde{D}W^Q, \widetilde{D}W^K, \widetilde{D}W^V\right), \tag{7}$$

where Q, K, and V denote the matrix of query, key, and value respectively, which can be calculated by the input representation $\widetilde{D}$. W^Q, W^K, W^V represent for the trainable parameters for linear projections. Meanwhile, M stands for attention-based correlation matrix.

For a sentence S_i, given the contextual representation H_i and the learnable correlation matrix M_i, we inject the dependency syntactic information into the contextual representation to obtain syntax-aware representation $\widetilde{H_i}$ as follows:

$$\widetilde{H_i} = H_i M_i. \tag{8}$$

The syntax-aware attention mechanism is capable of capturing crucial information along the syntactic tree structure. Notably, the attention weight increases with the proximity over the dependency syntactic tree, resulting in greater propagation of syntax information between the corresponding tokens. Conversely, tokens that are further apart on the dependency syntactic tree will receive lower attention weights and, consequently, less syntax information will be propagated.

4.4 Chinese semantic error recognition module

Chinese semantic error recognition module is constructed to identify whether the sentence has semantic errors. In the module, in order to retain pre-training knowledge while integrating syntactic information, we combined the raw text representation H_i with syntax-aware representation $\widetilde{H_i}$ to obtain a new text representation ${H_i}^{new}$, where mainly position-wise add and concatenate two operations are considered.

Position-wise add operation is formulated as follows:

$${H_i}^{new} = H_i + \widetilde{H_i}. \tag{9}$$

Concatenate operation is formulated as follows:

$${H_i}^{new} = \text{concat}(H_i, \widetilde{H_i}). \tag{10}$$

Thereafter, the new feature vector is put into a linear classifier with the softmax function, which is formulated as follows:

$$y_i = \text{softmax}\left((W_3)^T {H_i}^{new} + b\right), \tag{11}$$

where W_3 and b are learnable parameters, and y_i is the predicted probability. Since the CSER task is treated as a binary classification task, cross-entropy loss is employed to calculate the loss that penalizes the predicted class probability $\widehat{P}_j$ based on how far it is from the actual expected value P_j. The cross-entropy loss function L_{ce} for a batch with size N is defined as follow:

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} P_j \log \widehat{P}_j. \tag{12}$$

5 Experiment

Dataset	Correct	Incorrect Semantic	Total
Train	11,488	33,760	45,248
Dev	1,080	1,080	2,060
Test	1,000	1,000	2,000

Table 1: Distribution of experimental data.

5.1 Datasets and Metrics

We conduct our experiment on the dataset CoCLSA constructed by Sun et al. (2022), which is collected from multiple choice questions related to incorrect semantic sentences from the high school examination online resources. The data in CoCLSA contains two types of labels, where "1" indicates the sentence with semantic errors and "0" represents the sentence without semantic errors. The dataset contains a training set, a validation set, and a test set. The statistical information of the dataset is shown in the Table 1. In terms of evaluation metrics, we regard the CSER task as a binary classification task, so precision, recall, F1 score, and accuracy score are utilized for model evaluation as in previous work (Sun et al., 2023a).

Method	P	R	F	ACC
SLA (Li et al., 2021)	72.8	73.0	72.9	72.9
Syntax-RoBERTa (Bai et al., 2021)	73.3	74.3	73.8	73.6
K-adapter (Wang et al., 2021)	72.6	73.7	73.2	-
RoBERTa+DSRP (Sun et al., 2023a)	74.2	74.4	74.3	74.3
RoBERTa+$DSRP^+$ (Sun et al., 2023a)	73.2	75.8	74.8	74.1
SLA+DSRP (Sun et al., 2023a)	72.1	77.1	74.5	73.6
SLA+$DSRP^+$ (Sun et al., 2023a)	72.0	76.9	74.4	73.5
Syntax-RoBERTa+DSRP (Sun et al., 2023a)	73.7	75.9	74.8	74.4
Syntax-RoBERTa+$DSRP^+$ (Sun et al., 2023a)	73.6	76.1	74.8	74.4
ChatGPT (gpt-3.5-turbo-0613)	56.9	38.8	46.1	54.7
RoBERTa (Liu et al., 2019)	73.0	76.7	74.8	74.2
Our Model	**75.6**	**74.8**	**75.2**	**75.3**

Table 2: Experiment results on the CoCLSA dataset.

5.2 Baselines

We leverage following baselines, including **SLA** (Li et al., 2021), **Syntax-RoBERTa** (Bai et al., 2021), **K-adapter** (Wang et al., 2021), **DSRP** (Sun et al., 2023a), $\boldsymbol{DSRP^+}$ (Sun et al., 2023a), **ChatGPT (gpt-3.5-turbo-0613)** and **RoBERTa** (Liu et al., 2019) to verify the effectiveness of our CSER-DSA. More detailed descriptions of baselines are provided in Appendix B.

5.3 Implementation

We leverage LTP[4] for dependency syntactic information extraction and complete all experiments on Chinese semantic error recognition based on PyTorch and RTX 3090 GPU. We fine-tune RoBERTa model for 10 epochs with a batch size of 32, where AdamW(Kingma and Ba, 2015)(Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5 and weight decay of 0.01 is employed.

5.4 Main Result

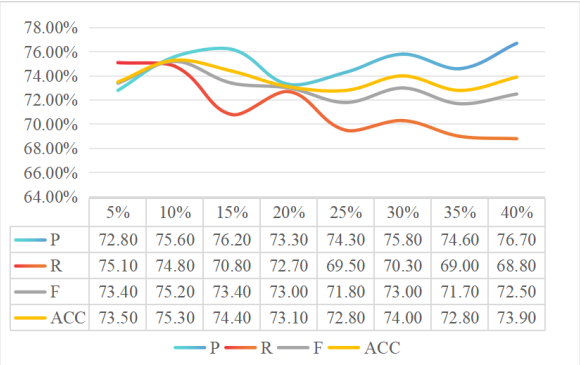
We conduct CSER experiments on the CoCLSA dataset and compare them with the results that are also tested on the benchmark. The results are shown in Table 2. Our model equipped with syntax-aware attention exceeds the RoBERTa results by 0.4% and 1.1% in the F1 score and accuracy score, respectively. Our method achieves state-of-the-art results in the F1 score and accuracy score for the CSER task compared with other methods infusing dependency syntax. Specifically, compared with the model using DSRP or $DSRP^+$ pre-training tasks to implicitly incorporate dependency syntactic information in the pre-trained phase, our model achieves an improvement of 0.9%-1.8% and 0.4%-0.9% in accuracy score as well as F1 score, respectively, indicating that explicitly incorporating dependency syntactic information solely in the fine-tuning phase can effectively improve model performance. Our model consistently outperforms the K-adapter infusing syntax knowledge in the pre-training phase in all metrics. Moreover, our model achieves consistent gains over SLA and Syntax-RoBERTa, which integrate dependency syntactic information leveraging attention mechanism. The results reveal that our proposed method of integrating dependency syntactic information is more effective.

Our investigation of the performance of the large-scale language model ChatGPT on CSER reveals a significant gap between the metrics of ChatGPT and those of existing CSER models. Qu and Wu (2023) discovered that ChatGPT exhibits poor performance in Chinese grammatical error correction task. Our findings demonstrate that ChatGPT is deficient in discerning semantic errors within Chinese text, consequently impeding its proficiency in error correction. We posit that enhancing ChatGPT's capacity for error identification is pivotal for ameliorating its error correction performance.

	P	R	F	ACC
Our Model	75.6	74.8	75.2	75.3
w/o Pseudo-label data	75.6	69.8	72.6	73.7
w/o DSIN module	74.2	70.2	72.1	72.9

Table 3: Experiment results of ablation study.

[4]https://github.com/HIT-SCIR/ltp



	5%	10%	15%	20%	25%	30%	35%	40%
P	72.80	75.60	76.20	73.30	74.30	75.80	74.60	76.70
R	75.10	74.80	70.80	72.70	69.50	70.30	69.00	68.80
F	73.40	75.20	73.40	73.00	71.80	73.00	71.70	72.50
ACC	73.50	75.30	74.40	73.10	72.80	74.00	72.80	73.90

Figure 4: Experiment results of data augmentation ratio.

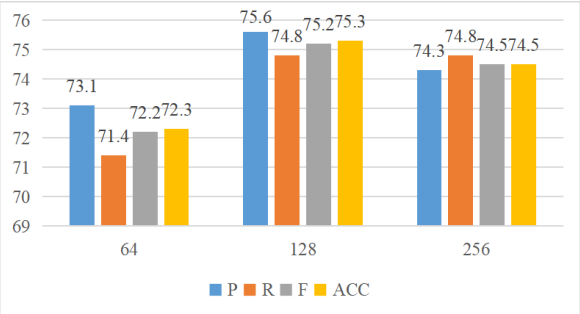


Figure 5: Experiment results of correlation matrix dimension.

5.5 Ablation study

We conduct ablation experiments to investigate the impact of the pseudo-label data construction method oriented to semantic error recognition and dependency syntactic information infusion module. Specifically, we perform experiments based on our proposed model without introducing pseudo-label data or removing dependency syntactic information infusion module, respectively. From the results in Table 3, it can be seen that our proposed model benefits from applying a pseudo-label data construction method and dependency syntactic information infusion. When discarding the pseudo-label data, the precision stays the same. The observation indicates that the integration of pseudo-label data enhances the model's generalization capacity, mainly contributing to an improvement in recall performance. It is remarkable that removing the dependency syntactic information infusion module results in a more substantial decrease in the model's performance, indicating the significant contributions of dependency syntactic information to CSER. Thus our proposed methods are effective.

5.6 Analysis

Data Augmentation Ratio. Owing to the different proportions of tokens replaced in the data augmentation process, the quality of the augmented samples will be uneven. Therefore, we explore the effect of the data augmentation ratio on the experimental results, as shown in Fig. 4. When the augmentation ratio is 10%, the accuracy score and F1 score of the model for CSER are optimal, and the corresponding precision and recall also achieve good performance. With the augmentation ratio increasing, the recall, F1 score, and accuracy score of the model decrease significantly, which may be attributed to the noise introduced by the token excessively replaced. Thus, in the data augmentation phase, we opt to use the most appropriate 10% ratio to construct 33,023 pseudo-label data based on our PDC-CSER method, training the optimal model.

Correlation Matrix Dimension. Additionally, we explore the effect of the correlation matrix dimension on the performance of the model. The results in Fig. 5 intuitively demonstrate that the model has achieved the best performance for semantic error recognition in all metrics when the dimension of the correlation matrix is set to 128.

6 Conclusion

CSER poses significant challenges due to semantic errors' complexity and ambiguity, which has always been a weak link in Chinese language processing. To tackle the challenge of a small sample set and annotating corpus containing semantic errors, we present a novel pseudo-label data construction method for CSER, generating pseudo-labels via two strategies to ensure the quality of pseudo-labels. Furthermore, we propose a method to explicitly incorporate the information into the pre-trained language model for CSER, focusing on more efficient dependency syntax that captures global syntactic information. Unlike previous studies, our proposed method can be employed solely in the fine-tuning phase, resulting in significant improvements while simultaneously saving resources and time cost. Experimental results on the CoCLSA show that our proposed approach outperforms existing models. In the future, we will consider incorporating rich information such as the part-of-speech and dependency types in dependency syntax trees, where connections between part-of-speech tags that do not conform to normal semantic patterns may have semantic errors.

Acknowledgments

This work was supported by the National Social Science Fund of China (No. 22BTQ045) and the 14th Five-Year Plan for Philosophy and Social Sciences Project of Guangdong (No.GD21CTS02).

Limitations

In this section, we discuss the limitations of this work. Initially, our work is mainly for Chinese, although semantic errors exist in other languages, such as English. Additionally, our experimental analysis did not extend to large language models. The quality of the pseudo-label data still has some room for improvement. Although pseudo-labels are jointly screened through two strategies, there may still be some inaccurate or noisy pseudo-labels.

Ethics Statement

The datasets and pre-trained language models used in our study come from open-access repositories. This ensures that we comply with all relevant ethical standards and authorizations. We strictly follow established research ethics throughout our research.

References

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-bert: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3011–3020. Association for Computational Linguistics.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: an open-source neural language technology platform for chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 42–49. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

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*.

Jiayuan Li, Yangsen Zhang, Jinjin Zhu, and Zewei Zhang. 2013. Semantic automatic error-detecting for chinese text based on semantic dependency relationship. In *Chinese Lexical Semantics - 14th Workshop, CLSW 2013, Zhengzhou, China, May 10-12, 2013. Revised Selected Papers*, volume 8229 of *Lecture Notes in Computer Science*, pages 406–415. Springer.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving BERT with syntax-aware local attention. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 645–653. 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. *CoRR*, 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.

Weihua Luo, Zhensheng Luo, and Xiaojin Gong. 2002. Semantic error checking in automatic proofreading for chinese texts. In *IEEE International Conference on Systems, Man and Cybernetics, Yasmine Hammamet, Tunisia, October 6-9, 2002 - Volume 2*, page 5. IEEE.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2339–2352. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq R. Joty, Steven C. H. Hoi, and Richard Socher. 2020. Tree-structured attention with hierarchical accumulation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Fanyi Qu and Yunfang Wu. 2023. Evaluating the capability of large-scale language models on chinese grammatical error correction task. *Preprint*, arXiv:2307.03972.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2699–2712. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages

5027–5038. Association for Computational Linguistics.

Bo Sun, Baoxin Wang, Wanxiang Che, Dayong Wu, Zhigang Chen, and Ting Liu. 2022. Improving pre-trained language models with syntactic dependency prediction task for chinese semantic error recognition. *CoRR*, abs/2204.07464.

Bo Sun, Baoxin Wang, Yixuan Wang, Wanxiang Che, Dayong Wu, Shijin Wang, and Ting Liu. 2023a. Csed: A chinese semantic error diagnosis corpus. *arXiv preprint arXiv:2305.05183*.

Jingbo Sun, Weiming Peng, Zhiping Xu, Shaodong Wang, Tianbao Song, and Jihua Song. 2023b. Incorporating syntactic cognitive in multi-granularity data augmentation for chinese grammatical error correction. In *International Conference on Neural Information Processing*, pages 370–382. Springer.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6077–6088. International Committee on Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2517–2527. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. *arXiv preprint arXiv:2305.17721*.

Yefan Wu, Runbo Zhuang, Ying Jiang, and Fan Li. 2015. Research and realization of chinese text semantic correction based on rule. In *2015 3rd International Conference on Education, Management, Arts, Economics and Social Science*, pages 1394–1404. Atlantis Press.

Jiang Ying, Zhuang Runbo, Wu Yefan, and Zhu Lingxuan. 2017. Semantic level chinese proofreading method based on description logics ontology reasoning. *Computer Systems and Applications*, pages 224–229.

Tianchi Yue, Shulin Liu, Huihui Cai, Tao Yang, Shengkang Song, and Tinghao Yu. 2022. Improving chinese grammatical error detection via data augmentation by conditional error generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2966–2975. Association for Computational Linguistics.

Li Yunhan, Shi Yunmei, Li Ning, and Tian Ying'ai. 2022. A survey of automatic error correction for chinese text. *Journal of Chinese Information Processing*, pages 1–18+27.

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

Shuai Zhang, Lijie Wang, Xinyan Xiao, and Hua Wu. 2022a. Syntax-guided contrastive learning for pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2430–2440. Association for Computational Linguistics.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022b. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3118–3130. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022c. Syngec: Syntax-enhanced grammatical error correction with a tailored gec-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2518–2531. Association for Computational Linguistics.

Yun Zhao, Xuerui Yang, Jinchao Wang, Yongyu Gao, Chao Yan, and Yuanfu Zhou. 2021. BART based semantic correction for mandarin automatic speech recognition system. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2017–2021. ISCA.

A Comparison of Different Chinese Text Correction

Unlike spelling and grammatical errors, semantic errors concentrate on complex semantic and syntax, leading to syntactic violations and even comprehension problems. Table 4 presents examples of different Chinese text proofreading tasks. CSC aims to detect word-level errors caused by misspelled words. As shown in Table 4, the spelling error occurs due to the confusion caused by the phonetic resemblance between the characters "及 (jí)" and "急 (jí)" in CSC task. CGED mainly refers to syntax-level error detection due to local or global grammar exceptions, namely missing words, multiple words, word order or misusing words. More precisely, the error type of the CGED task is word order in Table 4. The correct grammar is that "不能 (can not)" as an adverb should be placed in front of "实现 (achieve)". Different from CSC and CGED, CSER is oriented to more complex semantic-level errors, including collocation, missing, redundant, confusion, fuzziness, word order or illogical errors, requiring stronger capability for the model to understand the semantic information of the context for the sentence with semantic errors. The example of the CSER task in Table 4 exists a semantic error caused by semantic repetition. It is remarkable that the sentence is relatively fluent compared with sentences with grammatical errors owing to the word order, posing a significant obstacle to the model to recognize the semantic errors. As shown in Table 4, humans can easily identify Chinese grammatical errors and Chinese spelling errors. Nevertheless, semantic errors are even challenging for native speakers to recognize because semantic errors typically entail judgment of the syntactic structure of words.

B Baselines

We compare our CSER-DSA with the following baselines:

SLA (Li et al., 2021): The syntax-aware local attention (SLA) can capture the information of important local regions on the syntactic structure, which harnesses syntax-based masking to compute the dot-product of queries and keys and incorporates local attention with standard global attention to calculate the final attention scores.

Syntax-RoBERTa (Bai et al., 2021): The framework is designed to infuse syntactic information, which is applied to an arbitrary Transformer-based pre-trained checkpoint. The method disentangles the self-attention network into various subnetworks via sparse masks reflecting different connections and distances of tokens in a syntax tree and adopts topical attention to aggregate task-specific representations from distinct sub-networks.

K-adapter (Wang et al., 2021): K-adapter allows multiple kinds of knowledge to infuse large pre-trained models in the pre-training phase while maintaining the original representation of a pre-trained model fixed.

DSRP (Sun et al., 2023a): Dependency Structure and Relation Prediction(DSRP) performs multitask training based on the DSP pre-training task and DRP pre-training task. Specifically, the DSP pre-training task only considers two dependency structures, namely child and parent, aiming to learn the directionality of the dependency structure. The DRP pre-training task considers 12 dependency relations from the dependency parser of LTP, enabling the model to learn the diversity of dependency relations.

DSRP$^+$ (Sun et al., 2023a): $DSRP^+$ is consistent with DSRP in terms of DRP pre-training task, while the only difference is the introduction of DSP^+, a variant of DSP, which involves three dependency structures, including child, parent, and others, considering all the dependency structures.

ChatGPT (gpt-3.5-turbo-0613)[5]: Owing to the impressive performance demonstrated by LLMs in zero-shot or few-shot prompting scenarios, we evaluate the performance of ChatGPT with OpenAI's official API in zero-shot setting. We design prompts to activate their capabilities. The prompt template is shown in Figure 6. The gpt-3.5-turbo is opted to be the evaluated model, which stands out as the most advanced and specifically optimized for chat functionality.

RoBERTa (Liu et al., 2019) : RoBERTa is an improved variant of BERT, which utilizes dynamic masks and cancels the NSP task. Nonetheless, the full sentence mechanism is considered. Meanwhile, the size of RoBERTa's training data is ten times the size of BERT's training data. RoBERTa is fine-tuned on the training set containing pseudo-label samples directly in our experiments.

C Case Study

We conduct a case study to demonstrate the effectiveness of integrating dependency syntactic in-

[5]https://chat.openai.com/

CSC	source	接到信后，他迫不急 (jí) 待的连夜赶到县城 After receiving the letter, he rushed at once to the town that night.
	target	接到信后，他迫不及 (jí) 待的连夜赶到县城 After receiving the letter, he rushed at once to the town that night.
CGED	source	没有解决这个问题，不能人类实现更美好的将来。 Without addressing this problem, cannot humanity achieve a better future.
	target	没有解决这个问题，人类不能实现更美好的将来。 Without addressing this problem, humanity cannot achieve a better future.
CSER	source	为了防止不再发生意外，老师们做了很多练习。 To prevent not accidents, teachers have done many drills.
	target	为了防止发生意外，老师们做了很多练习。 To prevent accidents, teachers have done many drills.

Table 4: Examples of Chinese Spelling Correction, Chinese Grammatical Error Diagnosis and Chinese Semantic Error Recognition. Misspelled words are highlighted in red and the corresponding answers are in blue.

{'role' : 'system', 'content' : '你是一个中文文法错误识别系统，可以识别输入句子是否存在语义错误。' }

{'role' : 'system', 'content' : 'You are a Chinese Grammatical Error Recognition System that identifies Chinese semantic errors in input sentences.'}

{'role' : 'user', 'content' : '我将提供给你一个可能存在语义错误的句子，请你判断这个句子是否存在语义错误，并将判断结果以{"是否存在语义错误?" : "是"}的格式输出。输入的句子是{sen}。'}

{'role' : 'user', 'content' : 'I will provide you with a sentence that may have semantic errors. Please determine whether there are semantic errors in this sentence, and output the result as the format { ' Are there semantic errors' : ' yes ' }. The input sentence is { sen }. '}

Figure 6: Prompt Template **(in Chinese)** for Chinese Semantic Error Recognition. The Chinese sentences are the actual prompts used for ChatGPT in the experiment, while the English sentences are the corresponding translations.

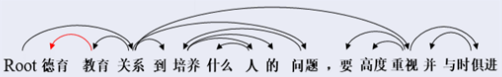
Example	Dependency Syntax	Gold	P_{base}	P_{syntax}
这个问题是值得我们所重视的 *(This issue is worthy of our attention.)*		1	0	1
德育教育关系到培养什么人的问题，要高度重视并与时俱进 *(Moral education is related to the question of what kind of people to cultivate, and it should be highly valued and keep pace with the times.)*		1	0	1

Figure 7: Experiment results of case study.

formation only in the fine-tuning stage for CSER. The results in Fig. 7 show that for two cases with semantic errors, our model provides correct predictions, while RoBERTa does not. The outstanding result indicates that our model effectively captures incorrect dependency syntactic information during the fine-tuning phase, thereby enhancing the performance of the model in recognizing semantic errors related to dependency syntax. It is remarkable that our model gives a good judgment for case 2 containing the semantic repetition error, which is challenging even for humans to recognize due to the high fluency of sentences.