Global Attention Decoder for Chinese Spelling Error Correction

Zhao Guo[1], **Yuan Ni**[1], **Keqiang Wang**[1], **Wei Zhu**[1], **Guotong Xie**[1,2,3]
[1]Ping An Healthcare Technology
[2]Ping An Healthcare and Technology Company Limited
[3]Ping An International Smart City Technology Co.,Ltd
{GUOZHAO385,NIYUAN442,WANGKEQIANG265,ZHUWEI972,
XIEGUOTONG}@pingan.com

Abstract

Recent progress has been made in using BERT framework for Chinese spelling error correction (CSC). However, most existing methods correct words based on local contextual information, without considering the influence of error words in sentences. Imposing attention on error contextual information could mislead and decrease the overall performance of CSC. To address this issue, we propose a **G**lobal **A**ttention **D**ecoder (GAD) approach for CSC. Specifically, the proposed method learns the global relationship of the potential correct input characters and the candidates of potential error characters. Rich global contextual information is obtained to alleviate the impact of the local error contextual information. In addition, a BERT with **C**onfusion set guided **R**eplacement **S**trategy (BERT_CRS) is designed to narrow the gap between BERT and CSC. The candidates generated by BERT_CRS covering the correct character with more than 99.9% probability. To demonstrate the effectiveness of our proposed framework, we test our method on three human-annotated datasets. The experimental results show that our approach outperforms all competitor models by a large margin of up to 6.2%, achieving state-of-the-art methods on all datasets.

1 Introduction

Spelling error correction plays an important role in NLP domain. A good spelling error system is the key to improve the performance of upper-layer applications. Spelling error correction aims to detect and correct erroneous characters/words. These spelling errors are mainly from human writing, speech recognition and optical character recognition (OCR) (Afli et al., 2016) systems. In Chinese, erroneous type is usually from character/word's phonological, visual and semantic similarity. According to (Cheng et al., 2020), about 83% and 48% of errors are related to phonological and visual similarity respectively. Although lots of researches have made great progress, Chinese spelling error correction (CSC) still remains a challenging task. Moreover, because the Chinese is composed of pictographic characters without word delimiters, methods from the languages like English can hardly be used for the Chinese. In addition, the meaning of same character in different contexts may change greatly.

Input	餐厅的换经费产适合约会 The restaurant's swap property is suitable for dates
BERT_CRS	餐厅的环经非常适合约会 The restaurant's ring is perfect for dates
+GAD	餐厅的环境非常适合约会 The restaurant environment is perfect for dates

Table 1: A sample data from SIGHAN 2014 (Yu et al., 2014), the incorrect and correct characters marked in red and green color respectively. Since "经" is highly related to "费" in its context, BERT_CRS is difficult to correct. GAD method learns the global relationship between "环" and "境" in candidates of input error characters "换" and "经" respectively (see Fig.1). Rich global contextual information is learned to alleviate the impact of the local noisy contextual information.

Many methods have been proposed for CSC task, which are mainly divided into two categories: 1) based on language models (Yeh et al., 2013; Yu and Li, 2014; Xie et al., 2015); 2) based on seq2seq model (Wang et al., 2019, 2018). Specially, with the emerge of the pre-trained BERT model, many methods (Hong et al., 2019; Zhang et al., 2020; Cheng et al., 2020) are proposed and made great progress. Almost all methods leveraged a confusion set, which contains a set of similar character group in terms of phonological and visual. Specifically, (Yu and Li, 2014) proposed to generate candidates based on confusion set and find the best

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1419–1428

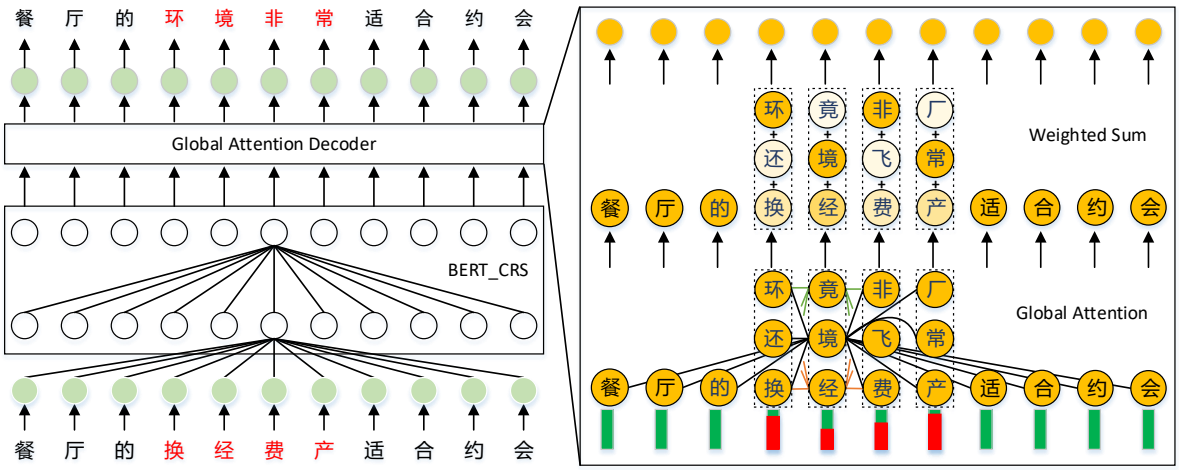

Figure 1: The framework of our proposed global attention decoder method. To illustrate the effectiveness of our model, error words and detection probability are marked with red. For instance "换经费产" and corresponding error detection probability in bottom right. Attention weights are represented with color shade in right.

candidate with highest language model probability. (Cheng et al., 2020) introduced a convolutional graph network that captures similarity and prior dependencies among characters using confusion set. (Wang et al., 2019) proposed a pointer network to generate a character from the confusion set. Previous methods predict each character or word based on its local context that may has noisy information (other errors). So far, no method has been proposed to alleviate the impact of this noisy information.

In this paper, we firstly introduced a BERT with confusion set guided replacement strategy (BERT_CRS), that narrows the gap between BERT and CSC task. Then, we proposed a novel global attention decoder (GAD) based on our BERT_CRS model (see Fig.1), which learns rich global contextual representations to alleviate the influence of the error contextual information during correction. Specifically, in order to solve the impact of the local error contextual information, we introduce additional candidates of potential error characters and hidden state generated by BERT_CRS. Next, global attention component learns the relationships of candidates to obtain the global hidden state and latent global attention weights of candidates. Then, weighted sum operator is adopted among candidates of each character to generate a rich global contextual hidden state. Finally a fully-connected layer to generate the correct characters. As shown in Table.1, Our proposed method is able to correct all spelling errors correctly. It is worthwhile to highlight the following aspects for the proposed approach:

- To narrows the gap between BERT and CSC, we introduce a BERT with confusion set guided replacement strategy, that contains a decision network and a fully-connected layer to simulate the detection and correction subtasks of CSC respectively.
- We proposed a global attention decoder model, which learns the global relationships of the potential correct input characters and the candidates of potential error characters. Rich global contextual information is learned to effectively alleviate the influence of local error contextual information.
- Experiments on the three benchmark datasets demonstrate that our method outperforms the state-of-the-art methods by a large margin of up to 6.2%.

2 Related Work

There is a vast prior research on Chinese spelling error correction (CSC) task so far. Next, We will discuss the algorithms in different periods.

N-gram period. Early research in CSC follow the pipeline of error detection, candidate generation and candidate selection. Almost all proposed methods (Yeh et al., 2013; Yu and Li, 2014; Xie et al., 2015; Tseng et al., 2015) employed an unsupervised n-gram language model to detect errors. Next, a confusion set which is an external knowledge of

the similarity between characters is introduced to confine the candidates. Finally, the best candidate with highest n-gram language model probability is considered as correction character. Specifically, (Yeh et al., 2013) proposed an inverted index based n-gram to map the potential spelling error character to the corresponding characters. (Xie et al., 2015) utilizes the confusion set to replace the characters and then evaluates the modified sentence via a joint bi-gram and tri-gram language model. In (Jia et al., 2013; Xin et al., 2014), a graph model is used to represent the sentence and a single source shortest path (SSSP) algorithm is performed on the graph to correct spell errors. The others viewed it as a sequential labeling problem and employed conditional random fields or hidden Markov models (Tseng et al., 2015; Wang et al., 2018).

Deep learning period. With the development of deep learning methods (Vaswani et al., 2017; Zhang et al., 2020; Hong et al., 2019; Wang et al., 2019; Song et al., 2017; Guo et al., 2016), great progress has been made in all NLP tasks. BERT (Devlin et al., 2018), XLNET (Yang et al., 2019), and Roberta (Liu et al., 2019), and ALBERT (Lan et al., 2019) achieve superior performance in almost all NLP task. Confusion set is still an important part in recent research for CSC task, but more upgrades have been made. Specifically, in (Hong et al., 2019), a pre-trained masked language model is employed as encoder. A confidence-similarity decoder utilizes similarity score to select candidates instead of the confusion set. (Vaswani et al., 2017) proposed a specialized graph convolutional network to incorporate phonological and visual similarity knowledge into BERT model. In (Zhang et al., 2020), a GRU based detection network is introduced and connected with BERT based correction network by a soft-masking technique. The others (Wang et al., 2019) employed a Seq2Seq model with copy mechanism, which generates a new sentence considering the extra candidates from confusion set.

3 The Proposed Approach

In this section, firstly, the problem formulation is elaborated. Then, we briefly describe how to narrow the gap between BERT (Devlin et al., 2018) and Chinese spelling error correction (CSC) using our BERT_CRS model. Finally, we introduce our novel global attention decoder (GAD) framework.

3.1 Problem Formulation

CSC aims to detect and correct errors in Chinese text. Given a sequence $X = \{x_1, x_2, \cdots, x_n\}$, n denotes the number of characters, our BERT_CRS model encodes it into a continuous representation space $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$, $\mathbf{v}_i \in \mathbf{R}^d$ is the contextual level feature of the i-th character, and it is d-dimensional. Here a decision network Φ_d models $\mathbf{V}$ to fit a sequence $Z = \{z_1, z_2, \cdots, z_n\}$, where z_i denotes the detection label of the i-th character, and z_i=1 means the character is incorrect and z_i=0 means it is correct. A fully-connected layer on the top of BERT_CRS as correction network Φ_c models $\mathbf{V}$ to fit a sequence $Y = \{y_1, y_2, \cdots, y_n\}$, where y_i is the correction label of the i-th character. Instead of a simple fully-connected layer as a decoder, our GAD models the additional candidates $c = \{c_1, c_2, \cdots, c_n\}$ to alleviate the impact of local error contextual information, where c represents the potential correct input characters and the candidates of potential error characters and:

$$c_i = \begin{cases} c_{i1}, c_{i2}, \cdots, c_{ik}, & if\ P(z_i = 1) \geq t \\ x_i, & if\ P(z_i = 1) < t \end{cases} \quad (1)$$

where k is the number of candidates. t is the threshold of error probability for characters.

3.2 BERT_CRS approach for CSC

In this section, we take advantage of previous models (Devlin et al., 2018; Liu et al., 2019; Cui et al., 2020) and introduce a replacement strategy using confusion set that narrows the gap between BERT and CSC model. There we call this model as BERT_CRS (BERT with **C**onfusion set guided **R**eplacement **S**trategy). Unlike BERT tasks, BERT_CRS has several modifications.

- We drop NSP task and adopt a decision network for detecting error information, that is similar to detection sub-task of CSC.

- As MacBERT (Cui et al., 2020), instead of masking with [MASK] token, we introduce confusion set guided replacement strategy by replacing phonological and visual similar character for masking purpose. Rarely, when there is no confusion character, we will maintain [MASK] token. The strategy similar to correction sub-task of CSC

- We use 23% of input characters for masking. To keep the balance of detection targets (0 for un-replacement, 1 for replacement), we set 35%, 30%, 30%, 5% probability for un-masking, replacing with confusion character, masking with [MASK] token and replacing with random word respectively. Calculated, the replacing and masking probabilities are approximately the same as masking probabilities of BERT.

With model trained by confusion set guided replace strategy, the top-k candidate characters are almost from the confusion set. That prepares for our GAD model.

Learning. Similar to RoBERTa (Liu et al., 2019), confusion set guided replace strategy uses a dynamic approach during training. Error detection and correction is optimized simultaneously in the learning process.

$$L_d = -\sum_{i=1}^{n} logP(z_i|\Phi_d(\mathbf{V})) \quad (2)$$

$$L_c = -\sum_{i=1}^{n} logP(y_i|\Phi_c(\mathbf{V})) \quad (3)$$

$$L = L_c + \lambda * L_d \quad (4)$$

where L_d and L_c is the objective of detection and correction loss respectively, L is the overall objective that linearly combines L_d and L_c, and $\lambda \in [0, 1]$ denotes the coefficient of detection loss L_d. Specially, $\lambda = 0$ represents that detection loss is not considered.

3.3 Global Attention Decoder

In this paper, we propose an global attention decoder (GAD) model to alleviate the impact of the local error contextual information. Our GAD is an extension of transformer layer (Vaswani et al., 2017), shown in Fig.2.

Self Attention. Relatively, the self-attention mechanism is part of the transformer layer, which takes the output of previous transformer layer or input embedding layer as input to obtain the hidden states with higher semantic representation, as shown in left part of Fig.1. Token representation $\mathbf{VA}_i^l$ at i-th position of l-th layer in self-attention method is defined as below:

$$\mathbf{VA}_i^l = \sum_{p=1}^{n} a_i^p \mathbf{V}_p^{l-1} \mathbf{W}^V \quad (5)$$

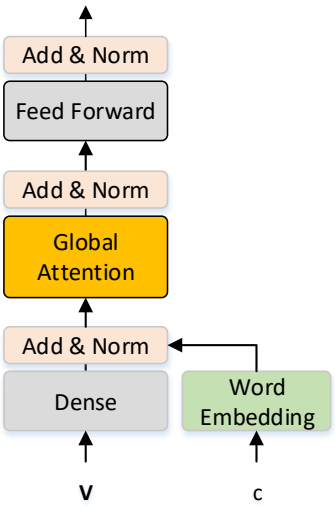

Figure 2: The global attention decoder architecture.

where a_i^p is the attention weight from i-th to p-th token, and $\sum_{p=1}^{n} a_i = 1$, $\mathbf{V}_p^{l-1}$ is the p-th token representation of (l-1)-th layer, $\mathbf{W}^V$ is a learnable projection matrix. This strategy could effectively encode rich token and sentence-level features. However, spelling error information also encoded into hidden states for CSC. Then, Imposing attention on error contextual information could mislead and decrease the overall performance of CSC.

Global Attention. Instead of using only local input information (see Eq.5), we consider potential correct inputs and the candidates of potential error characters to learn their latent relationships, that alleviate the influence caused by local error context. Specifically, as shown in Fig.2, we consider two input sources:

- Contextual representation $\mathbf{V}$, that contains rich semantic information
- Top-k candidates c generate by Φ_c correction network. To reduce the confusion of our GAD during learning, we only generate candidates for the potential error characters (see Eq.1).

To model the two different information, we first embed candidates into continuous representation using the word embedding $\mathbf{E}$ from BERT_CRS. Then, dense and layer-norm layers are introduced to model $\mathbf{V}$ and $\mathbf{E}(c)$ into input state $\mathbf{GI}$:

$$\mathbf{GI} = LayerNorm(\mathbf{Dense}(\mathbf{V}) + \mathbf{E}(c)) \quad (6)$$

Our global attention is introduced to learn the latent relationships between candidates c. Token representation $\mathbf{GA}_{i,j}$ at j-th candidate of i-th token of global attention component is defined as below:

$$\mathbf{GA}_{i,j} = \sum_{p=1}^{n}\sum_{q=1}^{k} a_{i,j}^{p,q}\mathbf{GI}_{p,q}\mathbf{W}_g^V \tag{7}$$

where $\mathbf{W}_g^V$ is a learnable projection matrix and $a_{i,j}^{p,q}$ is the attention weight from j-th candidate of i-th token to q-th candidate of p-th token, $\mathbf{GI}_{p,q}$ denotes the input state of q-th candidate from p-th token. Masking strategy is adopted between candidates from the same token.

$$a_{i,j}^{p,q} = 0,\ if\ i = p\ \&\ j \neq q \tag{8}$$

and $\sum_{p=1}^{n}\sum_{q=1}^{k} a_{i,j}^{p,q} = 1$. Finally, the global attention state $\mathbf{GA}_i$ at i-th position of of global attention component is defined as below:

$$\begin{aligned}
\mathbf{GA}_i &= \sum_{j=1}^{k} \beta_{i,j}\mathbf{GA}_{i,j} \\
\epsilon_{i,j} &= \sum_{p=1}^{n}\sum_{q=1}^{k} \epsilon_{i,j}^{p,q} \\
\beta_{i,j} &= \frac{exp(\epsilon_{i,j})}{\sum_{q=1}^{k} exp(\epsilon_{i,q})}
\end{aligned} \tag{9}$$

where $\beta_{i,j}$ is the global attention weight at j-th candidate of i-th token which quantifies the global relevance of feature $\mathbf{GA}_{i,j}$, $\epsilon_{i,j}^{p,q}$ and $\epsilon_{i,j}$ denote the unnormalized relevant scores of $a_{i,j}^{p,q}$ and $\beta_{i,j}$ respectively. Similar to standard transformer layer, feed forward and layer normalization to encode $\mathbf{GA}$ into final global continuous representation. Moreover, We adopt the multi-head technique used in the transformer layer in our global attention.

Learning. Given hidden states $\mathbf{V}$ and candidates c generated by our BERT_CRS, our GAD model fit the correct sequence Y in the learning process.

$$L_g = -\sum_{i=1}^{n} logP(y_i|\Phi_g(\mathbf{V})) \tag{10}$$

where Φ_g is our GAD network and L_g denotes our overall objective of GAD

4 Experiments

In this section, we evaluate our algorithm on the task of Chinese spelling error correction (CSC).

Training Data	**# Sent**	**Avg.Len**
UnLabeled corpus	3 million	-
(Wang et al., 2018)	271,329	44.4
SIGHAN 2013	350(350)	49.2
SIGHAN 2014	6,526(3432)	49.7
SIGHAN 2015	3,174(2339)	30.0
Total Labeled	281,379	44.4
Test Data	**# sent**	**Avg.Len**
SIGHAN 2013	1000(996)	74.1
SIGHAN 2014	1062(529)	50.1
SIGHAN 2015	1100(550)	30.5

Table 2: Statistics of datasets. The number in the bracket in #Sent column is the count of erroneous sentences

We first present the training data, test data and the evaluation metrics. Secondly we introduce our main results compared with previous state-of-the-art baselines. Then we conduct ablation studies to analyze the effectiveness of the proposed components. Finally, case study are explored.

4.1 Datasets

We consider three publicly available SIGHAN datasets from the 2013 (Wu et al., 2013), 2014 (Yu et al., 2014) and 2015 (Tseng et al., 2015) Chinese Spell Check Bake-offs. Following (Cheng et al., 2020), we adopted the standard split of training and test data of SIGHAN. We also follow the same data pre-processing, that converted the characters in dataset from traditional Chinese to simple Chinese using OpenCC[1].

For training dataset, we also collect 3 million unlabeled corpus from news, wiki and encyclopedia QA domains to pre-train our BERT_CRS model. Following (Wang et al., 2019), we also include additional 271K samples as the labeled training data, which are generated by an automatic method (Wang et al., 2018). The statistics of the data is showed in Table.2

4.2 Baselines

To evaluate the performance of our proposed algorithm, we compare it with following baseline methods.

- JBT (Xie et al., 2015): This method utilizes the confusion set to replace the characters and then evaluates the modified sentence via a **J**oint **B**i-gram and **T**ri-gram LM.
- Hybird (Wang et al., 2018): This method proposes a pipeline where a bidirectional LSTM

[1]https://github.com/BYVoid/OpenCC

Test Set	Model	Detection Level			Correction Level		
		Pre.(%)	**Rec.(%)**	**F1(%)**	**Pre.(%)**	**Rec.(%)**	**F1(%)**
SIGHAN13	JBT (Xie et al., 2015)	79.8	50.0	61.5	77.6	22.7	35.1
	Hybird (Wang et al., 2018)	54.0	69.3	60.7	-	-	52.1
	Seq2Seq (Wang et al., 2019)	56.8	91.4	70.1	79.7	59.4	68.1
	SpellGCN (Cheng et al., 2020)	82.6	88.9	85.7	98.4	88.4	93.1
	BERT (Cheng et al., 2020)	80.6	88.4	84.3	98.1	87.2	92.3
	BERT_CRS	85.5	89.2	87.3	98.9	88.5	93.4
	+GAD	**85.8**	**89.5**	**87.6**	**99.0**	**88.6**	**93.5**
SIGHAN14	JBT (Xie et al., 2015)	56.4	34.8	43.0	71.1	50.2	58.8
	Hybird (Wang et al., 2018)	51.9	66.2	58.2	-	-	56.1
	Seq2Seq (Wang et al., 2019)	63.2	82.5	71.6	79.3	68.9	73.7
	SpellGCN (Cheng et al., 2020)	83.6	78.6	81.0	97.2	76.4	85.5
	BERT (Cheng et al., 2020)	82.9	77.6	80.2	96.8	75.2	84.6
	BERT_CRS	84.6	**81.2**	82.9	97.4	**79.3**	87.4
	+GAD	**85.1**	80.9	**82.9**	**98.0**	79.2	**87.6**
SIGHAN15	JBT (Xie et al., 2015)	83.8	26.2	40.0	71.1	50.2	58.8
	Hybird (Wang et al., 2018)	56.6	69.4	62.3	-	-	57.1
	Seq2Seq (Wang et al., 2019)	66.8	73.1	69.8	71.5	59.5	69.9
	SpellGCN (Cheng et al., 2020)	**88.9**	87.7	**88.3**	95.7	83.9	89.4
	BERT (Cheng et al., 2020)	87.5	85.7	86.6	95.2	81.5	87.8
	BERT_CRS	88.1	**87.9**	88.0	96.1	84.4	89.9
	+GAD	88.6	87.8	88.2	**96.3**	**84.6**	**90.1**

Table 3: The character level performance on both detection and correction level. Our BERT_CRS model achieves similar performance compared with previous state-of-the-art models. Our GAD model achieves better performance.

based sequence labeling model is adopted for detection.

- Seq2Seq (Wang et al., 2019): This method introduces a Seq2Seq model with a copy mechanism to consider the extra candidates from the confusion set.
- FASpell (Hong et al., 2019): This model changes the paradigm by utilizing the similarity metric to select candidate instead of a pre-defined confusion set.
- Soft-Masked BERT (Zhang et al., 2020): This method proposes a detection network, which connected error correction model by a soft-masking technique.
- SpellGCN (Cheng et al., 2020): This model incorporate phonological and visual similarity knowledge into language models for CSC via a specialized graph convolutional network.
- BERT (Devlin et al., 2018): The word embedding on the top of BERT as correction decoder for the CSC task.

4.3 Implementation Details

Training Details. Our code is based on the repository of Transformers[2]. We first fine-tune our BERT_CRS model in 3 million unlabeled corpus based on the pre-trained whole word masking BERT[3]. The procedure runs 5 epochs with a batch size of 1024, learning rate of 5e-5 and max sequence length of 512. Then, we performed the fine-tuning process for our BERT_CRS model in all labeled training data with 6 epochs, a batch size of 32 and a learning rate of 2e-5. Next we fix our BERT_CRS model, and set the number of candidates k and error detection probability t as 4 and 0.25 respectively. Finally we fine-tune our GAD model with 3 epochs, a batch size of 32 and a learning rate of 5e-5. For SIGHAN 13 dataset, we performed additional fine-tune steps for 6 epochs as the data distribution in SIGHAN13 differs from other datasets, e.g. ”的”, ”得” and ”地 are rarely distinguished.

Evaluation Metrics. To evaluate the performance, we employ character and sentence-level accuracy, precision, recall and F1 followed by (Cheng et al., 2020), which are commonly used in the CSC task. In addition, we introduce the official evaluation metrics tool[4], which gives False Positive Rate (FRT), precision, recall, F1 and accuracy in sentence level.

[2]https://github.com/huggingface/transformers

[3]https://github.com/ymcui/Chinese-BERT-wwm

[4]http://nlp.ee.ncu.edu.tw/resource/csc.html

Test Set	Model	Detection Level			Correction Level		
		Pre.(%)	**Rec.(%)**	**F1(%)**	**Pre.(%)**	**Rec.(%)**	**F1(%)**
SIGHAN13	FASpell (Hong et al., 2019)	76.2	63.2	69.1	73.1	60.5	66.2
	SpellGCN (Cheng et al., 2020)	80.1	74.4	77.2	78.3	72.7	75.4
	BERT (Cheng et al., 2020)	79.0	72.8	75.8	77.7	71.6	74.6
	BERT_CRS	84.8	79.5	82.1	83.9	78.7	81.2
	+GAD	**85.7**	**79.5**	**82.5**	**84.9**	**78.7**	**81.6**
SIGHAN14	FASpell (Hong et al., 2019)	61.0	53.5	57.0	59.4	52.0	55.4
	SpellGCN (Cheng et al., 2020)	65.1	69.5	67.2	63.1	67.2	65.3
	BERT (Cheng et al., 2020)	65.6	68.1	66.8	63.1	65.5	64.3
	BERT_CRS	65.4	72.7	68.9	63.4	70.4	66.7
	+GAD	**66.6**	**71.8**	**69.1**	**65.0**	**70.1**	**67.5**
SIGHAN15	FASpell (Hong et al., 2019)	67.6	60.0	63.5	66.6	59.1	62.6
	Soft-Masked BERT (Zhang et al., 2020)	73.7	73.2	73.5	66.7	66.2	66.4
	SpellGCN (Cheng et al., 2020)	74.8	80.7	77.7	72.1	77.7	74.8
	BERT (Cheng et al., 2020)	73.7	78.2	75.9	70.9	75.2	73.0
	BERT_CRS	74.0	80.2	77.2	72.2	77.8	74.8
	+GAD	**75.6**	**80.4**	**77.9**	**73.2**	**77.8**	**75.4**

Table 4: The sentence level performance on both detection and correction level. Specially, SpellGCN (Cheng et al., 2020) reports correction level F1 as 75.9 in SIGHAN15. However, 74.8 is calculated by corresponding precision and recall. There the latter value is reported in the table.

4.4 Main Results

We compare our model with the state-of-the-art methods on the three test datasets, and the results are shown in Tab.3 and Tab.4, that compared the results in character-level and sentence level respectively. BERT_CRS outperforms almost all methods in three datasets, and combined with GAD achieving the best performance. Specifically, under the same amount labeled training data, for character level metric, our method gains the improvement against previous best results (SpellGCN) are 0.4%, 2.1%, 0.7% respectively for correction level F1 metric. For sentence level score, our model outperform SpellGCN by a margin of 6.2%, 2.2%, 0.6% respectively for correction level F1 metric. In addition, Soft-Masked BERT uses 5 million examples that generate by replaced strategy for extra training data. our method outperforms it by a large margin in SIGHAN15 test dataset.

We further consider the official evaluation results of BERT_CRS and GAD to compete with BERT and SpellGCN in SIGHAN15, shown in Tab.6. Our proposed BERT_CRS+GAD achieving better performance than SpellGCN by a margin of 0.2% for correction level F1 metric. In addition, the FPR are 13.1% (BERT_CRS+GAD) v.s. 13.2% (SpellGCN).

4.5 Ablation Studies

In this sub-experiment, we explore the impact of several components, including the coefficient λ and learning rate lr in BERT_CRS and the effective parameter k that is the number of candidates in GAD

Model	Parameters	Value	**F1(%)**
BERT_CRS	λ	1	72.0
		0.5	73.4
		0.1	74.8
	lr	2e-5	74.8
		5e-5	74.6
GAD	k	3	75.4
		4	75.1
		5	74.7

Table 5: The effect of parameters in BERT_CRS and GAD for correction level F1 metric on SIGHAN15.

The Effect of BERT_CRS. Our BERT_CRS introduces confusion set guided replacement strategy using BERT model. Compared with BERT model, for character level metric in Table.3, BERT_CRS improves the performance by a margin 6.6%, 2.5%, 1.8% respectively for correction level F1. For sentence level metric in Table.4, we achieving the scores 81.2% (BERT_CRS) v.s. 74.6% (BERT) on SIGHAN 13, 66.7% (BERT_CRS) v.s. 64.3% (BERT) on SIGHAN 14 and 74.8% (BERT_CRS) v.s. 73.0% (BERT) on SIGHAN 15.

We also show the effect of coefficient λ and learning rate during fine-tuning in all labeled datas, shown in Tabel.5. First we fix learning rate as 2e-5 and tune $\lambda \in [0.1, 0.5, 1]$ on SIGHAN15. When λ=0.1, the best performance is achieved. In addition, big variation is shown with different λ, That is to say, if more attention of detection loss, the performance is unsatisfactory. The reason of the situation may be caused by the imbalance of detection label during the fine-tuning process. In the following

Model	FPR	Detection Level				Correction Level			
		Acc.	Pre.(%)	Rec.(%)	F1(%)	Acc.	Pre.(%)	Rec.(%)	F1(%)
SpellGCN (Cheng et al., 2020)	13.2	**83.7**	85.9	**80.6**	83.1	82.2	85.4	77.6	81.3
BERT (Cheng et al., 2020)	13.6	83.0	85.9	78.9	82.3	81.5	85.5	75.8	80.5
BERT_CRS	14.0	83.1	85.1	80.2	82.6	81.9	84.8	77.8	81.1
+GAD	**13.1**	83.6	**86.0**	80.4	**83.1**	**82.4**	**85.6**	**77.8**	**81.5**

Table 6: The sentence level performance evaluated by official tools on SIGHAN 2015. The smaller FPR score indicates the better performance.

experiments, we set λ=0.1. We tune learning rate from [2e-5, 5e-5]. When 2e-5 is adopted, the better performance is achieved. We set learning rate equal to 2e-5 during experiments.

The Effect of GAD When combined with GAD in BERT_CRS model, the performance is improved under character and sentence level metric, shown in Tabel.3 and Tabel.4. Specifically, for sentence level metric, BERT_CRS+GAD outperform BERT_CRS by a margin of 0.4%, 0.8% and 0.6% respectively for correction level F1 metric.

We also study the impact of candidate number k. Since k is the key parameter which determines the coverage of correct character in candidates, it affects the performance of our algorithm. We study the performance variance with different $k \in [3, 4, 5]$ on SIGHAN15. Shown in Tabel.5, more candidates may degrade the performance. According to statistics, there are 161,365 error characters in all test data and 106, 75, 64 not in candidates for k equal to 3, 4, 5 respectively. The candidates generated by BERT_CRS model have 99.9% probability covering the correct character. Consider the trade-off between cover rate of candidates and performance, We set $k = 4$ in our experiments.

4.6 Case Study

To further analyze our approach, we show some correction results on test data (see Table.7). In Table.7, three categories of spelling error are selected: 1) Continuous characters error; 2) Single character error; 3) No character error. For Continuous characters error instance, ”介绍” (introduce) was misspelled as ”借少” (borrow less). Due to the influence of error characters, BERT_CRS is difficult to correct them all correctly. However, BERT_CRS+GAD alleviates the impact of the local error contextual to correct them all correctly. For single character error instance, ”抱” (pick up) was misspelled as ”包” (pack). Our BERT_CRS+GAD can also learns richer global contextual information to correct it than BERT_CRS. Here it has the same meaning of ”提议” (suggestion) and ”建议” (suggestion) in no character instance, BERT_CRS miscorrects it. These cases prove that our GAD can learn rich global contextual information to alleviate the impact of the local error contextual for CSC.

Continuous characters error
...语言。去外国可以认识很多的人，就可以借少
...语言。去外国可以认识很多的人，就可以借绍
...语言。去外国可以认识很多的人，就可以介绍
... you can meet a lot of people abroad, and introduce these languages.
Single character error
我把小猫抱起来，赶紧包出去到马路边求救...
我把小猫抱起来，赶紧跑出去到马路边求救...
我把小猫抱起来，赶紧抱出去到马路边求救...
I picked up the kitten and hurried out to the side of the road for help.
No character error
...课堂之前可以先有一些提议或许参考的资料...
...课堂之前可以先有一些建议或许参考的资料...
...课堂之前可以先有一些提议或许参考的资料...
Some suggestions or reference materials can be available before the class.

Table 7: Examples of CSC results, the incorrect and correct characters marked in red and green respectively. The first line in the block is input sentence. The second and third line is corrected by BERT_CRS and BERT_CRS+GAD respectively. And the rest is the English translation of the correct sentence.

We also showed some incorrect case to further analyze our model. For example, for the sentence ”希望您帮我素取公平，得到他们适当的赔偿”(I hope you can help me x for justice and get proper compensation from them) where the incorrect word ’x’ is not comprehensible, our GAD changes ”素取”(x) to ”争取”(strive for) that is appropriate in the context, but ground-truth ”诉取”(sue for) is more suitable because the context contains the meaning of litigation. Our GAD model also lacks the inference ability of context strong correlation as described in (Zhang et al., 2020).

5 Conclusions

In this paper, we propose a novel global attention decoder (GAD). Condition on the potential correct input characters and the candidates of po-

tential error characters, GAD reforms the self attention mechanism to learn their global relationships and obtains the rich global contextual information to alleviate the influence caused by error context. In addition, a BERT with **C**onfusion set guided **R**eplacement **S**trategy (BERT_CRS) is designed to narrow the gap between BERT and spelling error correction. Experimental results on three datsets show that our BERT_CRS outperform almost all previous state-of-the art methods, and higher performance is obtained by combining with our GAD.

References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966, Portorož, Slovenia. European Language Resources Association (ELRA).

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. 2016. Attention-based lstm with semantic consistency for videos captioning. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 357–361, New York, NY, USA. Association for Computing Machinery.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169, Hong Kong, China. Association for Computational Linguistics.

Zhongye Jia, Peilu Wang, and Hai Zhao. 2013. Graph model for Chinese spell checking. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 88–92, Nagoya, Japan. Asian Federation of Natural Language Processing.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

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.

Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. 2017. Hierarchical lstm with adjusted temporal attention for video captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2737–2743.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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*, pages 2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785, Florence, Italy. Association for Computational Linguistics.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42, Nagoya, Japan. Asian Federation of Natural Language Processing.

Weijian Xie, Peijie Huang, Xinrui Zhang, Kaiduo Hong, Qiang Huang, Bingzhou Chen, and Lei Huang. 2015. Chinese spelling check system based on n-gram model. In *Proceedings of the*

Eighth SIGHAN Workshop on Chinese Language Processing, pages 128–136, Beijing, China. Association for Computational Linguistics.

Yang Xin, Hai Zhao, Yuzhu Wang, and Zhongye Jia. 2014. An improved graph model for Chinese spell checking. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 157–166, Wuhan, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Jui-Feng Yeh, Sheng-Feng Li, Mei-Rong Wu, Wen-Yi Chen, and Mao-Chuan Su. 2013. Chinese word spelling correction based on n-gram ranked inverted index list. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 43–48, Nagoya, Japan. Asian Federation of Natural Language Processing.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223, Wuhan, China. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132, Wuhan, China. Association for Computational Linguistics.

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.