AXBERT: AN EXPLAINABLE CHINESE SPELLING CORRECTION METHOD DRIVEN BY ASSOCIATIVE KNOWLEDGE NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has shown promising performance on various machine learning tasks. Nevertheless, the unexplainability of deep learning models severely restricts the usage domains that require feature explanations, such as text correction. Therefore, a novel explainable deep learning model (named AxBERT) is proposed for Chinese spelling correction by aligning with an associative knowledge network (AKN). Wherein AKN is constructed based on the co-occurrence relations among Chinese characters, which denotes the explainable statistic logic contrasted with unexplainable BERT logic. And a translator matrix between BERT and AKN is introduced for the alignment and regulation of the attention component in BERT. In addition, a weight regulator is designed to adjust the attention distributions in BERT to appropriately model the sentence semantics. Experimental results on SIGHAN datasets demonstrate that AxBERT can achieve extraordinary performance, especially upon model precision compared to baselines. Our explainable analysis, together with qualitative reasoning, can effectively illustrate the explainability of AxBERT.

1 INTRODUCTION

Text correction methods serve as essential tools for people in various application scenarios, such as machine translation, office writing assistance, etc (Ghufron & Rosyida, 2018; Napoles et al., 2017; Omelianchuk et al., 2020). Wherein with the development of search engines, speech recognition, and so on, spelling correction is currently the most commonly used text correction method, which aims to optimize the inputting text and improve the prediction performance of the whole framework.

Explainability, much-needed for AI, describes the capacity of the methods to explain the basis of the decision to people (Mittelstadt et al., 2019; Beckh et al., 2021). Regardless of the extraordinary performance achieved by the recent spelling correction methods, the increasing unexplainability constrains the further application of the methods in some specific domains (Holzinger, 2018; Seeliger et al., 2019) and adversely affect the trust of AI methods(Gunning & Aha, 2019).

The fact is that *no explainable method exists for spelling correction with a good performance*. The rule-based correction methods are widely employed in specific domains required for regulatable correction, such as medical and biological (Crowell et al., 2004; Lai et al., 2015). Generally, the rule-based methods conduct the presetting rules based on the character correlations in the correction to realize the explainable decision-making process (Xiong et al., 2015; Yeh et al., 2014). However, the complex semantic context determines that the complete coverage of error cases for the presetting rules is impossible, indicating that rule-based methods are insufficient to handle complex errors.

In recent works, the proposed transformer-based language models significantly improve the performance of spelling correction. Researchers attempt to explain the language model by use of the extraction for the character relations, but *the extracted relations are irregular compared to linguistics and experience*. The irregular relations among the characters in BERT exist for two reasons:

- *The redundancy is widely distributed in the extracted information from attention layers*. While attention is considered as the key component in transformer-based models (Chefer et al., 2021; Vig, 2019), the function of every attention head is still undetermined. Pre-

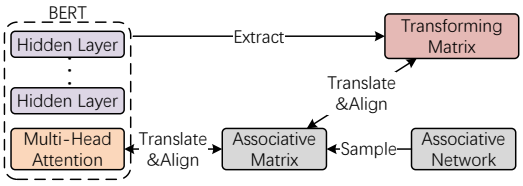


Figure 1: Information flow in AxBERT

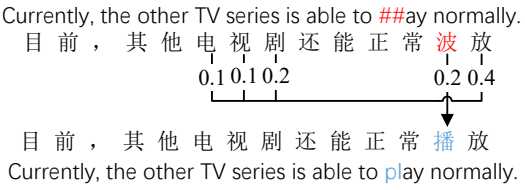


Figure 2: Example of the semantic relation

vious researches demonstrate that the pruning of the redundant attention heads is able to improve the performance of the language model (Voita et al., 2019; Behnke & Heafield, 2020; Wang et al., 2021), which indicates that the redundancy disrupts the modeling process and severely constrains the further analysis.

- *The layers perform different functions in transformer-based language model (BERT).* The layer-wise analysis is conducted in BERT, and the results illustrate that the containing layers serve different functions (van Aken et al., 2019; Lu et al., 2021; Reif et al., 2019). Nevertheless, the existence of various layer functions challenging the comprehensive explaining of BERT, which needs a unification for transferring the inside logic to understandable logic for human beings.

In this work, we address the obstacles by (1) developing an explainable method for regulatable Chinese spelling correction and (2) constructing an alignment and regulation closed-loop circuit (illustrated in Figure 1) to integrate the explainable statistic logic into BERT which effectively reduce the redundancy and unify the logic among the layers. Wherein the alignment and regulation are not stated explicitly, but the explainability of the entire framework is increased respect for the specific explainable component with integrated information, which is demonstrated for effectiveness in previous work (Liu et al., 2019; Chen et al., 2020; Rybakov et al., 2020). Our approach - AxBERT - consist of an **A**ssociative knowledge network (AKN, Li et al. (2022) and (**x**) **BERT** (Devlin et al., 2019) to realize an explainable Chinese spelling correction method with extraordinary performance. More specifically, our main contributions are concluded as:

- An *associative matrix* sampled from *associative knowledge network* is introduced to AxBERT, which reflected explainable statistic logic with context localization. (section 3.1)
- We use the least-squares function to quantify transformation process in BERT as a *transforming matrix* reflected as unexplainable BERT logic. (section 3.2)
- A *translator matrix* is introduced to bridge the gap between statistic logic and BERT logic, which is multiplied with the attention component in BERT to align with statistical distribution. (section 3.3 and 3.4)
- We introduce *weight regulator*, relied on the character-level similarity between attention and AKN, to regulate the attention distributions for a appropriate semantic modeling process and better correction performance. (section 3.5)
- The outstanding correction performance on SIGHAN datasets demonstrates the effectiveness of AxBERT. In Addition, the explainable analysis is designed to exhibit the regulating process and quantitatively verify the explainability of our proposed method (section 4).

2 Related Work

2.1 Explainable Analysis in BERT

The backbones of the BERT mainly consist of unexplainable feature representations. The existing analysis to explain BERT conduct different aspect including self-attention, linguistic knowledge, etc (Rogers et al., 2021). Researchers attempt to obtain the semantic relation among the tokens to reveal the basis of the modeling result (Htut et al., 2019; Goldberg, 2019; Hewitt & Manning, 2019) (e.g., Figure 2). The analysis from the attention perspective exhibit the visualization of the attention distribution and statistical results to quantitatively explain BERT (Clark et al., 2019a; Vig & Belinkov, 2019; Kovaleva et al., 2019; Bian et al., 2021). From the comprehensive perspective, the

overrated of self-attention in the analysis is one-sided, and also disadvantages for further analysis (Li et al., 2019; Pande et al., 2021), because the representation transformation in BERT contains the processing from different layers. Inspired by the previous works in explainable analysis, we comprehensively quantify transformation logic from all the layers in BERT while opting for the attention distribution as the key to integrating the explainable statistic logic into BERT.

2.2 TRANSFORMER-BASED CHINESE SPELLING CORRECTION

Benefits from the proposed transformer network, the transformer-based language models can efficiently capture the semantic information of the given sentences (Zhang et al., 2019; Devlin et al., 2019; Clark et al., 2019b). Based on the extraordinary semantic modeling ability, the transformer-based language models are introduced for spelling correction task, which significantly enhances the correction performance. One of the principle approaches is to consider the masked token prediction task and conduct this task as correcting the inappropriate characters (Zhang et al., 2020; Cui et al., 2020). Besides, the researchers make use of the external features of the characters, such as phonic and shape, to expand the embedding dimensions to achieve reliable correction results (Hong et al., 2019; Cheng et al., 2020; Liu et al., 2022). Furthermore, the correction methods with well-performed decoders have emerged recently, which enables the candidate distributions with precise result (Bao et al., 2020; Li & Shi, 2021).

2.3 ASSOCIATIVE KNOWLEDGE NETWORK

Associative knowledge network (Li et al., 2022), a statistical network based on the co-occurrence among the phrases. We introduce a modified AKN to AxBERT, constructed at character level, to fit the embedding layer of BERT. For the sentences in reference articles, we initial and update AKN in AxBERT, which consist of the element $\boldsymbol{A}_{i,j}$, according to:

$$\boldsymbol{A}_{i,j} = \prod_{\text{sent}} \text{SR} \sum_{\text{sent}} \frac{1}{distance_{\langle i,j \rangle}} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{v \times v}$, v is the length of the character list. The shorter the distance between the characters, the stronger associative relation among the characters. We find that the scores of commonly used characters will accumulate to extremely high in the updating process, the shrink rate SR is introduced to keep balance, which is default as 0.95.

3 METHODOLOGY

3.1 CONSTRUCTION OF ASSOCIATIVE MATRIX

People measure the relations among the characters according to knowledge from experience, which is similar to the statistic logic in AKN. Therefore, based on the associative relations from AKN, we introduce an associative matrix $\boldsymbol{M}_S$ with a contextification process to represent the explainable statistic logic of the given sentence (with length d), which is defined as:

$$\boldsymbol{M}_{S_{i,j}} = \sigma \frac{\boldsymbol{A}_{i,j}}{\text{Avg}(\boldsymbol{M}_{S_i})} - 0.5 \tag{2}$$

which $\boldsymbol{M}_S \in \mathbb{R}^{d \times d}$, $\boldsymbol{M}_{S_{i,j}}$ is the associative score of character pair $\langle i, j \rangle$, and $\boldsymbol{M}_{S_{i:}}$ is the i-th row of $\boldsymbol{M}_S$. Besides, $\sigma(\cdot)$ and $\text{Avg}(\cdot)$ is function of sigmoid and average.

3.2 QUANTIFY BERT LOGIC AS TRANSFORMING MATRIX

BERT (Devlin et al., 2019) consists of embedding layers, attention layers, and linear layers. From the bottom to the top, the hidden layers transform the embedding representation into the semantic representation, where the embedding representation and the semantic representation fit in the same tensor-shape. We define the complex transformation in BERT with the hidden size of H as:

$$\mathbf{F} = \text{Transforming}(\mathbf{E}) \tag{3}$$

where $\mathbf{E} \in \mathbb{R}^{d \times H}$ indicates the embedding representation and $\mathbf{F} \in \mathbb{R}^{d \times H}$ indicates last hidden representation. $\text{Transforming}(\cdot)$ is defined as the transformation process in BERT.

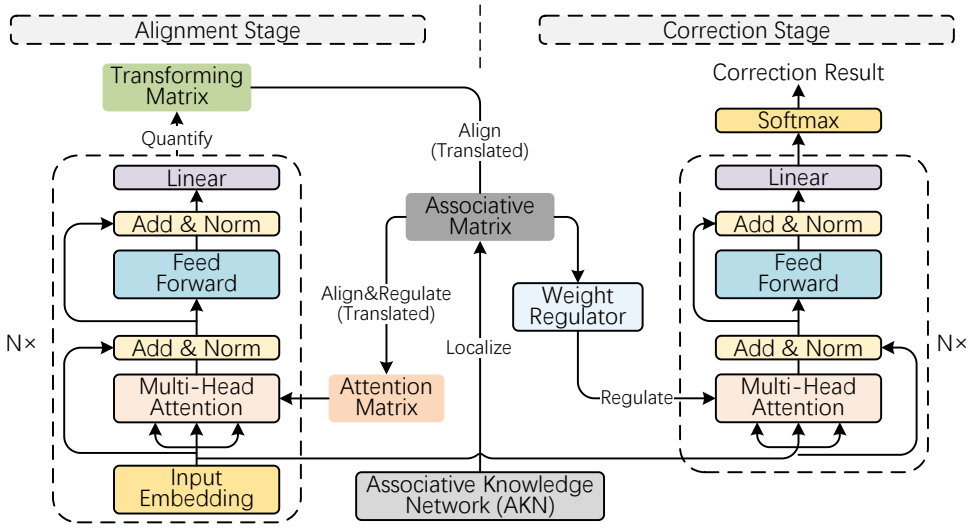


Figure 3: Overview of AxBERT's architecture. The inputting given sentence is firstly modeled by BERT in the alignment stage (Left), we quantify the transformation process of the sentence representations in BERT as the transforming matrix $\boldsymbol{M}_T$. (Middle) $\boldsymbol{M}_T$ is aligned with the associative matrix $\boldsymbol{M}_S$ with the help of the translator matrix $\boldsymbol{M}_F$. Then $\boldsymbol{M}_F$ is also used in the alignment between attention matrix $\boldsymbol{M}_A$ and $\boldsymbol{M}_S$. After the series alignment and regulation process, we compute the character-level similarity between attention and $\boldsymbol{M}_S$ within the weight regulator in the correction stage (Right). Besides, note that parameters of BERT in the alignment stage and correction are real-time sharing.

The transformation function $\mathrm{transforming}(\cdot)$, which represents the processing of the BERT layers, is considered as the comprehensive transformation logic of BERT. By use of the least-squares function, we approximately quantify the transformation process as a transforming matrix $\boldsymbol{M}_T$:

$$\boldsymbol{M}_T = \mathrm{LeastSquares}(\mathbf{E}, \mathbf{F}) \tag{4}$$

where $\boldsymbol{M}_T \in \mathbb{R}^{d \times d}$, and $\mathrm{LeastSquares}(\cdot)$ is least-squares function. Transforming matrix $\boldsymbol{M}_T$, the solution of least-square equation, serves as approximate quantification of the transforming process from the $\mathbf{E}$ to $\mathbf{F}$ and represents the unexplainable BERT logic contrasted to $\boldsymbol{M}_S$.

3.3 Translate BERT Logic to Explainable Statistic Logic

From the perspective of linguistics and experience, the characters that co-occurred in phrases are considered associated. In contrast, the unexplainable language model conducts the modeling process with irregular influence logic of the representations regardless of the relations among the characters. Even though BERT logic is quantified as a transforming matrix, the unexplainability still constrains the direct understanding of people. Inspired by the conception of translation, we introduce a translator matrix $\boldsymbol{M}_F$, which serves as the translator between the flattened transforming matrix $\bar{\boldsymbol{M}}_T$ and associative matrix $\bar{\boldsymbol{M}}_S$. We aim to find the appropriate translator matrix to indirectly fit the unattainable alignment between the unexplainable BERT logic and explainable statistical logic. Note that the translator matrix $\boldsymbol{M}_F$, which is parameter-isolated from the backbone of AxBERT, is trained in parallel with the correction task according to:

$$L_F = 1 - S_{T,S} \tag{5}$$

$$S_{T,S} = \mathrm{CosSim}_{-1}(\boldsymbol{M}_F \times \bar{\boldsymbol{M}}_T, \bar{\boldsymbol{M}}_S) \tag{6}$$

which $\mathrm{CosSim}_{-1}(\cdot)$ is the cosine similarity function at -1-th dimension, $\bar{\boldsymbol{M}}_T \in \mathbb{R}^{d^2 \times 1}$, $\bar{\boldsymbol{M}}_S \in \mathbb{R}^{d^2 \times 1}$, $\boldsymbol{M}_F$ is the translator matrix and $\boldsymbol{M}_F \in \mathbb{R}^{d^2 \times d^2}$, L_F is the objective for training of the translator matrix. Wherein, the flatten operation for $\boldsymbol{M}_T$ and $\boldsymbol{M}_S$ enable the communication among the different semantic representations of the characters.

3.4 ATTENTION REGULATION FOR EXPLAINABILITY VIA TRANSLATOR MATRIX

As we mentioned in section 1, the redundancy distributed in the representations in BERT adversely affects the analysis of BERT. In order to reduce the redundancy in AxBERT, we introduce a regulation on attention to integrating the explainable statistic logic into BERT. Attention layers, the most significant component of BERT, serve as the key to integrating statistical information. While every component in BERT performs under various logic, even attention is unable to comprehensively represent BERT, but the regulation process enables the progressive unification of the different logic in learning. Besides, because the bottom attention layers are more concerned with character structure (Belinkov et al., 2017; Jawahar et al., 2019), a dynamical attention combination method is applied based on the number of attention layers, which realizes a more comprehensive capturing for attention information. The alignment between the flattened attention matrix $\bar{\boldsymbol{M}}_A$ and associative matrix $\bar{\boldsymbol{M}}_S$ is defined as:

$$S_{A,S} = \mathrm{CosSim}_{-1}(\bar{\boldsymbol{M}}_A, {\boldsymbol{M}_F}^{-1} \times \bar{\boldsymbol{M}}_S) \tag{7}$$

$$\boldsymbol{M}_A = \sum_{i=0}^{layer} (1 - \frac{i}{layer}) \sum_{j=0}^{head} \mathbf{AttDis}_{i,j} \tag{8}$$

where $\bar{\boldsymbol{M}}_A \in \mathbb{R}^{d^2 \times 1}$, ${\boldsymbol{M}_F}^{-1} \in \mathbb{R}^{d^2 \times d^2}$ and $\mathbf{AttDis}_{i,j} \in \mathbb{R}^{d \times d}$. $\mathbf{AttDis}_{i,j}$ is the attention distribution from the attention head of i-th layer and j-th head. $layer$ and $head$ are the layer number and attention head number of BERT encoder.

3.5 REGULATION FOR CORRECTION TASK WITH WEIGHT REGULATOR

Generally, in BERT-based correction methods, the semantic representations of the characters are modeled based on the context of the given sentence. However, the semantic influence from the irrelative characters or the error characters will decrease the accuracy of the modeling process. In order to maintain an appropriate modeling process in AxBERT, we apply a weight regulator to regulate the attention distributions in the correction stage. Specifically, the errors are distributed with lower scores compared with the other characters in AKN. By aligning the associative matrix with the combined attention matrix $\boldsymbol{M}_A$ at character level, the sequences of the similarity scores of the given sentence are obtained, where the error positions are computed as lower scores than the other positions. The weight $\mathbf{W}$ in the weight regulator is computed as following:

$$\mathbf{W} = \mathrm{CosSim}_{-2}(\boldsymbol{M}_{Ain} + \boldsymbol{M}_{Aout}, \boldsymbol{M}_S) \tag{9}$$

$$\boldsymbol{M}_{Ain} = \frac{{\boldsymbol{M}_A}^T}{\mathrm{Avg}_{\mathrm{col}}(\boldsymbol{M}_A)^T} \tag{10}$$

$$\boldsymbol{M}_{Aout} = \frac{M_A}{\mathrm{Avg}_{\mathrm{row}}(\boldsymbol{M}_A)} \tag{11}$$

where $\mathbf{W} \in \mathbb{R}^d$, $\boldsymbol{M}_{Ain} \in \mathbb{R}^d$ and $\boldsymbol{M}_{Aout} \in \mathbb{R}^d$. $\mathrm{CosSim}_{-2}(\cdot)$ is the cosine similarity function at -2-th dimension (character-level), $\mathrm{Avg}_{\mathrm{row}}(\cdot)$ and $\mathrm{Avg}_{\mathrm{col}}(\cdot)$ are row and column average functions in the matrix. Note that different from the undirected associative score, the attention between characters is directed, which contains two kinds of degrees as $\boldsymbol{M}_{Ain}$ for the in-degree and $\boldsymbol{M}_{Aout}$ for the out-degree respectively located in columns and rows in extracted attention matrix $\boldsymbol{M}_A$. By accumulating $\boldsymbol{M}_{Ain}$ and $\boldsymbol{M}_{Aout}$, the attention is transferred into the undirected form, which fits the form of the associative matrix $\boldsymbol{M}_S$.

In order to maintain the appropriate semantic influence of characters, based on the obtained character-level weight $\mathbf{W}$, we introduce a weight matrix $\boldsymbol{M}_W$ to regulate it in the correction stage, which fits the shape of the attention distributions. As a result, the regulated out-degrees of inappropriate characters are decreased, and the out-degrees of appropriate characters are increased. The weight matrix $\boldsymbol{M}_W$ in the weight regulator is computed according to:

$$\boldsymbol{M}_W = \mathbf{W} \times (\frac{1}{\mathbf{W}})^T \cdot \mathrm{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_d) \tag{12}$$

$$\mathbf{AttDis}_{R_i:} = (1 - \frac{i}{layer})\mathbf{AttDis}_i \boldsymbol{M}_W \tag{13}$$

where $M_W \in \mathbb{R}^{d\times d}$, $\textbf{AttDis}_R$ and **AttDis** respectively indicates the regulated and original attention distribution, which $\textbf{AttDis}_R \in \mathrm{R}^{d\times d}$ and $\textbf{AttDis} \in \mathrm{R}^{d\times d}$. $\mathrm{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_d)$ is the diagonal matrix value of $\mathbf{W}$. Besides, we introduce a weight decay process for the regulation process to decrease the regulation intensity for the attention layers on the top.

3.6 TRAINING OF AXBERT

The objective L is composed of L_A and L_C corresponding to the alignment and correction stages, which are defined as:

$$L = \lambda(L_C) + (1-\lambda)(L_A) \tag{14}$$

$$L_C = -\sum_{t=1}^{T'} \log P(\mathrm{y}_t|\mathrm{X}) \tag{15}$$

$$L_A = 1 - S_{A,S} \tag{16}$$

where L_C is the objective of the correction task, L_A is the objective of the attention alignment. $S_{A,S}$ defined in equation 7. λ is the combining factor, and we set 0.8 in our training.

Additionally, we design a pre-train process for AxBERT, where the massive pre-train process is applied to improve the generalization ability of AxBERT. Specifically, We randomly replace 13.5% of the characters in the correct sentences with the random tokens. For the replaced sentences, we conduct the alignment task and correction task with the same objectives defined in formula 14-16.

4 EXPERIMENTS

4.1 SETTINGS

The pretrained Chinese BERT is used in AxBERT. We opt streams of 128 tokens, a small batch of size 32, and learning rates of 2e-5 and 4e-5 for the 30-epoch-training of transforming matrix and the correction components. Additionally, the dropout rate of 0.3 is used for the embedding layers, scale-dot product attention, and hidden layers in BERT and ReLU function.

4.2 DATASETS

SIGHAN (Yu et al., 2014; Tseng et al., 2015) , a benchmark for Traditional Chinese spelling check evaluation, is used in our training and experiments, which contains SIGHAN14 and SIGHAN15 datasets. We follow the same pre-processing procedure as (Wang et al., 2019) for SIGHAN to convert the characters to simplified Chinese, which is widely used in the baseline methods. Wherein, SIGHAN14 consists of a 2,339-sentence-TrainSet and a 1,062-sentence-TestSet, and SIGHAN15 consists of a 6,526-sentence-TrainSet and a 1,100-sentence-TestSet. We use the TestSets as the benchmark dataset in our experiment and the TrainSets as part of the training corpus.

HybirdSet (Wang et al., 2018) is a method for automatic corpus generation for Chinese spelling check, which conducts the OCR- (Tong & Evans, 1996) and ASR-based (Hartley & Reich, 2005) methods to generate the visually or phonologically resembled spelling errors. Hybird dataset is composed of a 274,039-sentence-TrainSet and a 3,162-sentence-TestSet. In the previous works, the TrainSet of Hybird datasets is used the training of the correction methods (Wang et al., 2019; Cheng et al., 2020). We adopted the same strategy and constructed the training corpus by mixing the TrainSets of SIGHAN and the Hybird.

CLUE (Xu et al., 2020) CLUE, an open-ended, community-driven project, is the most authoritative natural language understanding benchmark for Chinese including 9 tasks spanning several well-established single-sentence/sentence-pair classification tasks. We use the news dataset in the CLUE to initialize associative knowledge network, which contains 2,439 articles.

4.3 COMPARISON APPROACHES

We evaluate our method on Chinese spelling correction and compare with several approaches as baseline. Besides, we also evaluate original BERT masked language model, which is initialized with the same setting as the BERT encoder in AxBERT.

HanSpeller++ conducts a multi-stepped reranking strategy by Hidden Markov Language Model for correction task, which is a remarkable rule-based spelling correction method (Xiong et al., 2015).

Confusionset introduce the copy strategy into Seq2Seq model for spelling correction task (Wang et al., 2019).

SoftMask is a BERT-based spelling correction method with a soft-mask generator, where the soft-masked strategy is similar to the concept of error detection (Zhang et al., 2020).

FASPell conducted the Seq2Seq prediction by incorporating BERT with additional visual and phonology features (Hong et al., 2019).

SpellGCN incorporated BERT and the graph convolutional network initialized with phonological and visual similarity knowledge for Chinese spelling correction (Cheng et al., 2020).

PLOME integrates the phonological and visual similarity knowledge into a pre-trained masked language model with a large pre-train corpus consisted of one million Chinese Wikipedia pages. And it is the SOTA in previous work (Liu et al., 2021).

HeadFilt is an adaptable filter for Chinese Spell Check, which conducts the domain-shift conditioning problem by introducing a hierarchical embedding according to the pronunciation similarity and morphological similarity (Nguyen et al., 2021).

4.4 Evaluation Method

For the evaluation of the spelling correction performance, we use the same evaluation matrix to assess the precision, recall, and F1-score at sentence-level and character-level with the same evaluation matrix of the previous works (Wang et al., 2019; Zhang et al., 2020; Cheng et al., 2020). Note that because of the difference between traditional Chinese and simplified Chinese, the wrong cases in the given converted TestSet is partly incorrect. Therefore, we directly evaluate the result by comparing the sentences in the TestSet with the predicted sentences from AxBERT and the baselines.

In order to exhibit the regulatability and quantitatively analyze of AxBERT. We design the explainable analysis to evaluate the explainability of our method on SIGHAN-15 contained two sub-analysis as similarity analysis and regulatable analysis. For the similarity analysis, we compute the cosine similarity between the attention and associative distributions. Additionally, the regulatable analysis illustrates the using case indicated the situation that the specific character pairs are not expected to be modified in correction. For a given sentence with errors which is able to be successfully corrected, by changing the associative scores among the related character pairs contained the errors, we assess the ratio between the number of retaining errors and the total changed errors.

4.5 Main Results

The experiment results are illustrated in Table 1, while part of the results of the baseline methods is incomparable with other work. P, R, and F1 denote the precision, recall, and F1 score, respectively.

On the sentence-level, the results demonstrate that our model is especially outstanding in precision. From the results, we find that the BERT-based correction methods (including HeadFilt, BERT, AxBERT) are different with the other methods in correction strategy, where AxBERT serve as the SOTA in precision and F1-score. Compared with the other baselines, the higher precision and lower recall reflect that the AxBERT is ”cautious” in correction. We believe that the main reason of ”cautious” tendency is the introduction of the alignment and regulation in AxBERT. The integrated explainable logic and the additional validation of the character relations from AKN improve the robustness of AxBERT and provide a consistent correction process.

Previous works opted the sentence-level evaluation as the first choice comparing with the character-level evaluation (Hong et al., 2019; Zhang et al., 2020; Nguyen et al., 2021). Even if we think that sentence-level evaluation is more convincing, the result of the character-level evaluation is illustrated. And the different correction strategy also reflects in the character-level evaluation. For the character-level evaluation, AxBERT is better than most of the baselines but still lower than the SOTA (PLOME) even if there is not a significant gap between them cause that the cautious strategy makes the number of corrected samples of AxBERT less than the other methods.

Table 1: The performance of our method and baseline methods.

Method	Sentence Level						Character Level					
	Detection			Correction			Detection			Correction		
SIGHAN14	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ConfusionSet	-	-	-	-	-	-	63.2	82.5	71.6	79.3	68.9	73.7
FASPell	61.0	53.5	57.0	59.4	52.0	55.4	-	-	-	-	-	-
SpellGCN	65.1	**69.5**	67.2	63.1	**67.2**	65.3	83.6	**78.6**	81.0	97.2	**76.4**	**85.5**
HeadFilt	82.5	61.6	70.5	82.1	60.2	69.4	-	-	-	-	-	-
BERT	81.6	64.1	71.8	81.0	62.6	70.6	**89.4**	74.1	81.0	96.9	71.8	82.5
AxBERT	**81.9**	64.4	**72.1**	**81.7**	63.2	**71.2**	87.8	76.2	**81.6**	**98.2**	74.8	84.9
SIGHAN15	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HanSpeller++	80.3	53.3	64.0	79.7	51.5	62.5	-	-	-	-	-	-
ConfusionSet	-	-	-	-	-	-	66.8	73.1	69.8	71.5	59.5	69.9
SoftMask	73.7	73.2	73.5	66.7	66.2	66.4	-	-	-	-	-	-
FASPell	67.6	60.0	63.5	66.6	59.1	62.6	-	-	-	-	-	-
SpellGCN	74.8	80.7	77.7	72.1	77.7	75.9	88.9	87.7	88.3	95.7	83.9	89.4
HeadFilt	84.5	71.8	77.6	84.2	70.2	76.5	-	-	-	-	-	-
PLOME	77.4	**81.5**	79.4	75.3	**79.3**	77.2	**94.5**	**87.4**	**90.8**	**97.2**	**84.3**	**90.3**
BERT	84.1	75.9	79.8	83.6	72.7	77.8	92.1	84.7	88.2	95.3	80.7	87.4
AxBERT	**88.4**	78.7	**83.3**	**88.1**	76.2	**81.7**	92.0	84.9	88.3	96.9	82.3	89.2

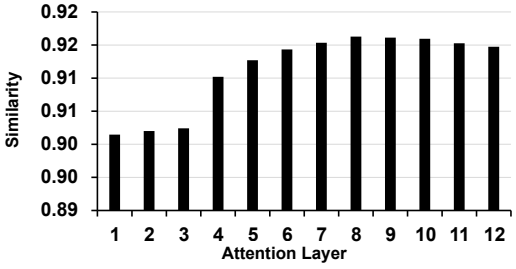


Figure 4: Similarity of association and attention

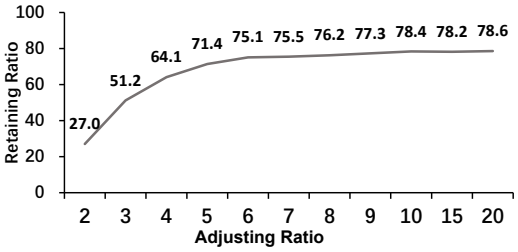


Figure 5: Retaining ratio result

4.6 Explainable Analysis

The detailed similarities between the associative matrix and the attention distributions are presented in figure 4. As the results from the previous works, the low layers prefer to learn the word structure while the top layers prefer to learn the word meanings (Belinkov et al., 2017; Jawahar et al., 2019). The similarity results supported the above conclusion. While associative relations are regarded as character relations based on semantics (meaning), the higher layer in AxBERT is assessed as more similar to the associative matrix.

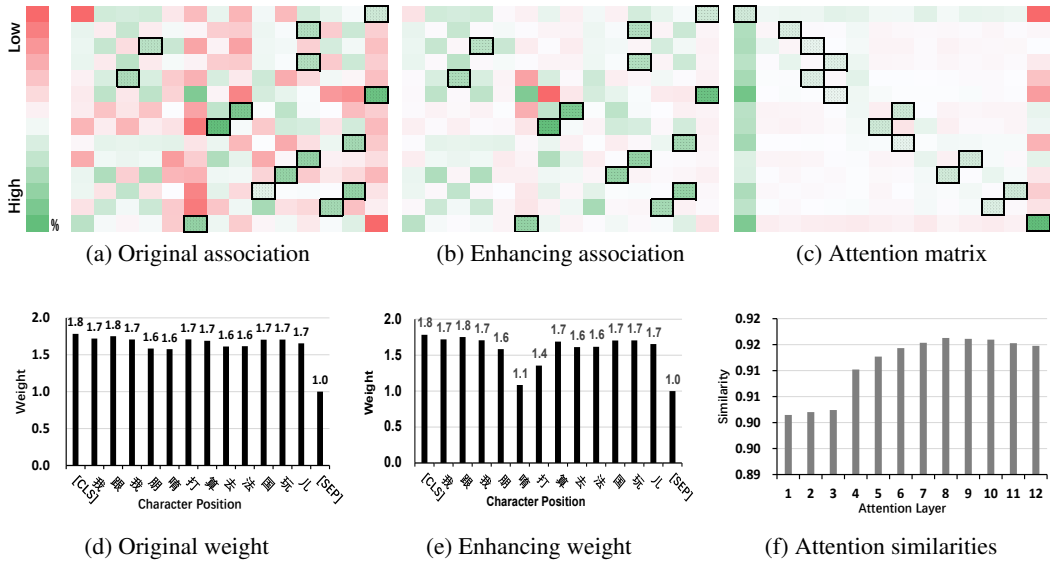
In the regulatable analysis, for the successfully corrected errors, we multiple different adjusting ratio to the corresponding associative scores to implicitly influence the weight regulator according to formula 9. The errors are expected to keep retained, and we calculate the retaining ratio of adjusted errors to the total. The retaining ratios with the different adjusting ratios are shown in figure 5, with the increase of the adjusting ratio, the number of the retaining errors is increasing. The most rapid increase occurs when the adjusting ratio is set from 2 to 6; after that, the growth rate slowed down and stabilized at around 78%. We think that the errors that were not successfully retained at last are so fatal in semantics, so that the correction method has to handle them to keep the semantic fluency.

4.7 Case Study

As shown in table 2, the errors 朋唷/fri*** is successfully corrected to 朋友/friend. Besides, we also present the regulated sentence after we adjust the associative relation among 朋 and 唷. The

Table 2: Correction result of the case study

Wrong Sentence	我跟我朋唷打算去法国玩儿/I plan to travel to France with my fri***
Predicted Sentence	我跟我朋友打算去法国玩儿/I plan to travel to France with my friend
Correct Sentence	我跟我朋友打算去法国玩儿/I plan to travel to France with my friend
Regulated Sentence	我跟我朋唷打算去法国玩儿/I plan to travel to France with my fri***



(a) Original association (b) Enhancing association (c) Attention matrix

(d) Original weight (e) Enhancing weight (f) Attention similarities

Figure 6: Result of the case study

associative and attention distributions of the sentence are respectively presented in figure 6(a), 6(b) and 6(c), where the black-bordered units are the highest score in row. The distributions reflect the relations among the characters. By comparing figure 6(a), 6(b) and 6(c), green we can discover that the positions of the black-bordered units are distributed similarly, which demonstrate the alignment between attention and AKN. Besides, we also show the weight in weight regulator of the given sentence in figure 6(d), 6(e). The corresponding weight of the adjusted characters 朋 and 唷 are decreased to make them less influence on other characters to keep retained.

5 Conclusion

This paper reports an explainable Chinese spelling correction method named AxBERT, which is driven by semantic alignment and regulation. While the alignment and regulation are applied to hidden layers of BERT, the various logic from the components in BERT are unified with clearer semantic relations. Besides, the weight regulator is introduced to regulate the attention distribution to model the sentence in a more appropriate way, which effectively improves the correction performance. In the evaluation of SIGHAN dataset, the effectiveness and explainability of AxBERT are demonstrated.

With the help of explainability and high performance, AxBERT is able to be widely employed in various usage scenarios. Specifically, after the training with the general correction corpus, simply adjusting for the associative relations enable AxBERT to fit the unusual character correlations in specific domains, such as the medical, biology, and legal domains. In the future, we plan to extend the explainable information to the decoder structure as a non-autoregression prediction, which can realize a variable-length correction framework in order to tackle wider correction situations.

REFERENCES

Zuyi Bao, Chen Li, and Rui Wang. Chunk-based Chinese spelling check with global optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2031–2040, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.184. URL https://aclanthology.org/2020.findings-emnlp.184.

Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. Explainable machine learning with prior knowledge: an overview. *arXiv preprint arXiv:2105.10172*, 2021.

Maximiliana Behnke and Kenneth Heafield. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2664–2674, 2020.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 930–945, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.72. URL https://aclanthology.org/2021.naacl-main.72.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 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*, pp. 871–881, 2020.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert‘‘s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019a.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019b.

Jonathan Crowell, Qing Zeng, Long Ngo, and Eve-Marie Lacroix. A frequency-based technique to improve the spelling suggestion rank in medical queries. *Journal of the American Medical Informatics Association*, 11(3):179–185, 2004.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Muhammad Ali Ghufron and Fathia Rosyida. The role of grammarly in assessing english as a foreign language (efl) writing. *Lingua Cultura*, 12(4):395–403, 2018.

Yoav Goldberg. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

David Gunning and David Aha. Darpa''s explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

Matthew W Hartley and David E Reich. Method and system for speech recognition using phonetically similar word alternatives, June 21 2005. US Patent 6,910,012.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. 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)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Andreas Holzinger. From machine learning to explainable ai. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 55–66, 2018. doi: 10.1109/DISA.2018.8490530.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 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)*, pp. 160–169, 2019.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*, 2019.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. 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)*, pp. 4365–4374, 2019.

Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics*, 55:188–195, 2015.

Piji Li and Shuming Shi. Tail-to-tail non-autoregressive sequence prediction for chinese grammatical error correction. 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)*, pp. 4973–4984, 2021.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1293–1303, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1124. URL https://aclanthology.org/P19-1124.

Yulin Li, Zhenping Xie, and Fanyu Wang. An associative knowledge network model for interpretable semantic representation of noun context. *Complex & Intelligent Systems*, pp. 1–21, 2022.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. 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)*, pp. 2991–3000, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.233. URL https://aclanthology.org/2021.acl-long.233.

Yujia Liu, Hongliang Guo, Shuai Wang, and Tiejun Wang. Visual and phonological feature enhanced siamese bert for chinese spelling error correction. *Applied Sciences*, 12(9):4578, 2022.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. Knowledge aware conversation generation with explainable reasoning over augmented graphs. 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)*, pp. 1782–1792, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1187. URL https://aclanthology.org/D19-1187.

Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. Influence patterns for explaining information flow in bert. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4461–4474. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/239f914f30ea3c948fce2ea07a9efb33-Paper.pdf.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 229–234, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2037.

Minh Nguyen, Gia H Ngo, and Nancy F Chen. Domain-shift conditioning using adaptable filtering via hierarchical embeddings for robust chinese spell check. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2027–2036, 2021.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170, Seattle, WA, USA, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.16.

Madhura Pande, Aakriti Budhraja, Preksha Nema, Pratyush Kumar, and Mitesh M. Khapra. The heads hypothesis: A unifying statistical approach towards understanding multi-headed attention in bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13613–13621, May 2021. doi: 10.1609/aaai.v35i15.17605. URL https://ojs.aaai.org/index.php/AAAI/article/view/17605.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 01 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00349. URL https://doi.org/10.1162/tacl_a_00349.

Sergei Rybakov, Mohammad Lotfollahi, Fabian J. Theis, and F. Alexander Wolf. Learning interpretable latent autoencoder representations with annotations of feature sets. *bioRxiv*, 2020. doi: 10.1101/2020.12.02.401182. URL https://www.biorxiv.org/content/early/2020/12/03/2020.12.02.401182.

Arne Seeliger, Matthias Pfaff, and Helmut Krcmar. Semantic web technologies for explainable machine learning models: A literature review. *PROFILES 2019*, pp. 30, 2019.

Xiang Tong and David A Evans. A statistical approach to automatic ocr error correction in context. In *Fourth workshop on very large corpora*, 1996.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pp. 32–37, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3106. URL https://aclanthology.org/W15-3106.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pp. 1823‘‘1832, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358028. URL https://doi.org/10.1145/3357384.3358028.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, 2019.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2517–2527, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1273. URL https://aclanthology.org/D18-1273.

Dingmin Wang, Yi Tay, and Li Zhong. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5780–5785, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1578. URL https://aclanthology.org/P19-1578.

Hanrui Wang, Zhekai Zhang, and Song Han. Spatten: Efficient sparse attention architecture with cascade token and head pruning. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 97–110. IEEE, 2021.

Jinhua Xiong, Qiao Zhang, Shuiyuan Zhang, Jianpeng Hou, and Xueqi Cheng. Hanspeller: a unified framework for chinese spelling correction. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language*, 2015.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.419. URL https://aclanthology.org/2020.coling-main.419.

Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, and Yong-Ting Chen. Chinese word spelling correction based on rule induction. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 139–145, 2014.

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

Computational Linguistics. doi: 10.3115/v1/W14-6820. URL https://aclanthology.org/W14-6820.

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

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL https://aclanthology.org/P19-1139.