Correcting the Misuse: A Method for the Chinese Idiom Cloze Test

Xinyu Wang[1], **Hongsheng Zhao**[2], **Tan Yang**[2], **Hongbo Wang**[1]
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2]School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications
[1,2]{xinyu.wang, zhaohs, tyang, hbwang}@bupt.edu.cn

Abstract

The cloze test for Chinese idioms is a new challenge in machine reading comprehension: given a sentence with a blank, choosing a candidate Chinese idiom which matches the context. Chinese idiom is a type of Chinese idiomatic expression. The common misuse of Chinese idioms leads to error in corpus and causes error in the learned semantic representation of Chinese idioms. In this paper, we introduce the definition written by Chinese experts to correct the misuse. We propose a model for the Chinese idiom cloze test integrating various information effectively. We propose an attention mechanism called Attribute Attention to balance the weight of different attributes among different descriptions of the Chinese idiom. Besides the given candidates of every blank, we also try to choose the answer from all Chinese idioms that appear in the dataset as the extra loss due to the uniqueness and specificity of Chinese idioms. In experiments, our model outperforms the state-of-the-art model.

1 Introduction

The Chinese idiom comprehension requires the ability to understand Chinese idioms. Chinese idiom, which is called "成语" (*chengyu*) in Chinese, consists of four characters. Chinese idioms are mostly derived from stories in ancient literature from Chinese history, and often reflect the moral behind the stories. To measure the ability of understanding Chinese idioms, the Chinese idiom cloze test dataset was proposed (Zheng et al., 2019): given a sentence with a blank, an examinee is required to choose an idiom which best matches the context surrounding the blank. Table 1 shows an example of the Chinese idiom cloze test.

The misuse of Chinese idioms is prevalent among Chinese native speakers who did not receive a professional Chinese education. Due to the metaphorical meaning of Chinese idioms, even Chinese native speakers who do not major in Chinese would use a Chinese idiom with its literal meaning, which causes misuse. Table 2 shows some common misuses of Chinese idioms. The misuse meaning is often related to the literal meaning.

The misuse of Chinese idiom appears in various social media and text such as Weibo and Zhihu. The Chinese word embeddings and Chinese language models are pretrained on these corpora that contain the misuse of Chinese idioms and learn the incorrect meaning of Chinese idioms. For example, in Table 3, we use Google Translate to translate Chinese idioms finding that some results are incorrect, and the incorrect meanings happen to be the common misuses of these Chinese idioms. In this paper, we introduce the definition of Chinese idiom, which is written by the Chinese experts, to correct the misuse. The complete definition describes the accurate interpretation and usage of Chinese idioms. Besides, because the misuse often comes from the literal meaning of the Chinese idiom, we propose an attention mechanism called Attribute Attention that extracts the relationships between the character-level and word-level representations.

Moreover, using the definition to correct the misuse does not mean that the non-misuse part would be dropped. Take 七月流火 in Table 2 as an example. The common misuse of 七月流火 is not totally incorrect. 七月流火 referring to the weather is correct, but the weather turning hot is incorrect. Therefore, we propose Attribute Attention to make use of other representations of 七月流火 even if they contain incorrect information.

In addition, Chinese idioms are derived from stories in ancient literature and contain abundant information. Chinese idioms contain more information so they are more likely to be used in a more specific context than common words. For example, 美 means "beautiful", 轮 means "wheel",

Sentence with a blank	他们希望能___再进一步 They hope that they can ___ and achieve greater success.
A candidate idiom	百尺竿头 Literal translation: at the top of a hundred-foot pole. Free translation: make still further progress.
Definition	比喻到了极高的境地，仍须继续努力，求更大的进步。 When one has achieved great success, one should continue to work hard to make greater progress.

Table 1: An example of the Chinese idiom cloze test that contains a sentence, one of the candidate idioms, and the definition of the idiom.

Chinese idiom	Literal meaning	Misuse meaning	Correct meaning
翻云覆雨	A huge change for clouds and rain	Magnificent	Skillful
七月流火	Fire in July	The weather turned hot	The weather turned cold
三人成虎	Three persons become a tiger	Cooperation lead to great strength	Spread rumors

Table 2: Some common misuses of Chinese idioms.

and 奂 means "magnificent". The Chinese idiom 美轮美奂 means "a building is beautiful". 美轮美奂 can be used only when describing a building, whereas those four characters are not related to building. When those four characters are combined, the meaning becomes narrow. It is more difficult to find two similar Chinese idioms than normal words. In this paper, besides choosing the answer from the given candidates, our model tries to choose the answer from the whole vocabulary of candidate Chinese idioms that appear in the dataset and calculate its loss as a part of the final loss. In this way, relationships between much more idioms can be captured every time. It costs very few extra computing resources but provides significant improvement.

In experiments, our model outperforms the state-of-the-art model. Our main contributions are summarized as follows:

- We introduce the definition and propose Attribute Attention to balance the importance of different representations of the Chinese idiom.
- We add an extra loss obtained by choosing the answer from all Chinese idioms that appear in the dataset, which costs very few extra computing resources but provides significant improvement.

2 Related Work

The cloze test is a classic task of reading comprehension and many methods were proposed (Hermann et al., 2015; Chen et al., 2016; Wang et al., 2018; Zhang et al., 2018; Fu et al., 2019; Fu and Zhang, 2019). The Chinese idiom cloze test is more challenging because Chinese idioms convey the metaphorical meaning and are misused sometimes. Most works related to idioms focused on English idioms identification (Gedigian et al., 2006; Katz and Giesbrecht, 2006; Fazly et al., 2009; Shutova et al., 2010; Salton et al., 2016; Do Dinh et al., 2018b; Flor and Beigman Klebanov, 2018; Do Dinh et al., 2018a; Liu and Hwa, 2018). Some works have tried to use definitions: Spasic et al. (2017) analyzed the sentiment of definitions; Fathima Shirin and Raseek (2018) used the similarity between different definitions. However, these methods introduced definitions but did not try to understand them. Liu et al. (2017) used CharLSTM to encode the meaning of idioms, which has a similar idea to (Jiang et al., 2018). Only a few works have been done with Chinese idioms such as building Chinese emotion lexicons (Xu et al., 2010) and improving Chinese word segmentation (Chan and Chong, 2008; Sun and Xu, 2011; Wang and Xu, 2017). Chengyu Reader (CR) (Jiang et al., 2018) is proposed for the Chinese idiom cloze test, which used the def-

Chinese idiom	Common misuse meaning	Google Translate	Correct translation
空穴来风	Groundless	Groundless	Grounded and justified
危言危行	Dangerous words and behavior	Dangerous words	Upright words and behavior
差强人意	Unsatisfactory	Unsatisfactory	Generally satisfactory

Table 3: Some incorrect translations of Chinese idioms form Google Translate.

initions and the attention mechanism of Attentive Reader (AR) (Hermann et al., 2015; Chen et al., 2016).

3 Approach

Formally, the Chinese idiom cloze test requires the model to choose the correct answer from a number of the candidate idioms given a sentence with a blank. The sentence is defined as a sequence of characters with a blank, which is also called context in the following. The candidate Chinese idiom is defined as a sequence of four characters, which is called idiom in the following. The definition is defined as a sequence of characters interpreting the corresponding idiom. In this paper, the term "BERT" refers to the BERT-like models (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019; Sanh et al., 2019), because any one of them and even the new BERT-like model in the future can be used in our model. Figure 1 is an overview of our model. The following sections will introduce every part of our model one by one.

3.1 Integrating Context and Definition

The definition is not the next sentence of the context. The context and definition do not belong to the same document. It is inappropriate to set the context as the first sentence and set the definition as the second sentence separated by `[SEP]` for BERT. In this section, as shown in Figure 2, we propose a way to integrate the context and definition with BERT, which lets the model "know" that the definition is mainly related to the idiom.

We input the context, the candidate idiom, and definition together. For example, we input the context "他们希望能___再进一步 *(they hope they can ___ and achieve greater success)*", the candidate idiom "百尺竿头 *(make still further progress)*", and the definition "比喻高的成就 *(an outstanding achievement)*" together as "他们希望能 `[MASK]` 再进一步 `[SEP]` 百尺竿头:比喻高的成就 `[SEP]`". The context is defined as $\boldsymbol{v}$. The candidate idiom and the definition are defined as $\boldsymbol{d}$ here.

The Multi-Head Attention is applied to the context and definition in different ways. Formally, the Multi-Head Attention for the context is:

$$\begin{aligned}\boldsymbol{v}_i^{(l)} =&\text{MultiHeadAttention}(\boldsymbol{m}^{(l-1)},\\ &\boldsymbol{v}_1^{(l-1)},\boldsymbol{v}_2^{(l-1)},\ldots,\boldsymbol{v}_{|\boldsymbol{v}|}^{(l-1)})\end{aligned} \tag{1}$$

where $\boldsymbol{v}_i^{(l)}$ denotes the i-th character of the context at the l-th layer, and $\boldsymbol{m}^{(l)}$ denotes the `[MASK]` token at the l-th layer; $|\boldsymbol{v}|$ denotes the number of characters of the context. The context only can "see" itself and the `[MASK]`.

The Multi-Head Attention for the definition is:

$$\begin{aligned}\boldsymbol{d}_i^{(l)} =&\text{MultiHeadAttention}(\boldsymbol{m}^{(l-1)},\\ &\boldsymbol{v}_{[SEP]}^{(l-1)},\boldsymbol{d}_1^{(l-1)},\boldsymbol{d}_2^{(l-1)},\ldots,\boldsymbol{d}_{|\boldsymbol{d}|}^{(l-1)})\end{aligned} \tag{2}$$

where $\boldsymbol{d}_i^{(l)}$ denotes the i-th character of the definition $\boldsymbol{d}$ at the l-th layer, and $\boldsymbol{v}_{[SEP]}^{(l-1)}$ denotes the first `[SEP]` token at the l-th layer; $|\boldsymbol{d}|$ denotes the number of characters of the definition. The definition is inaccessible to the context, which avoids that the BERT regards the definition as the next sentence of the context.

The Multi-Head Attention for the `[MASK]` is:

$$\begin{aligned}\boldsymbol{m}^{(l)} =&\text{MultiHeadAttention}(\boldsymbol{m}^{(l-1)},\\ &\boldsymbol{v}_1^{(l-1)},\boldsymbol{v}_2^{(l-1)},\ldots,\boldsymbol{v}_{|\boldsymbol{v}|}^{(l-1)},\\ &\boldsymbol{d}_1^{(l-1)},\boldsymbol{d}_2^{(l-1)},\ldots,\boldsymbol{d}_{|\boldsymbol{d}|}^{(l-1)})\end{aligned} \tag{3}$$

The `[MASK]` can pay attention to the characters of both the context and definition. On the one hand, `[MASK]` "knows" what kind of idiom could match the context as the correct answer. On the other hand, `[MASK]` "knows" the candidate idiom definition. `[MASK]` integrates the information from context $\boldsymbol{v}$ definition and $\boldsymbol{d}$ in the character-level.

In this way, the relation between the context and the definition is built through the `[MASK]`. The output of the `[MASK]` is defined as $\boldsymbol{h}_m$.

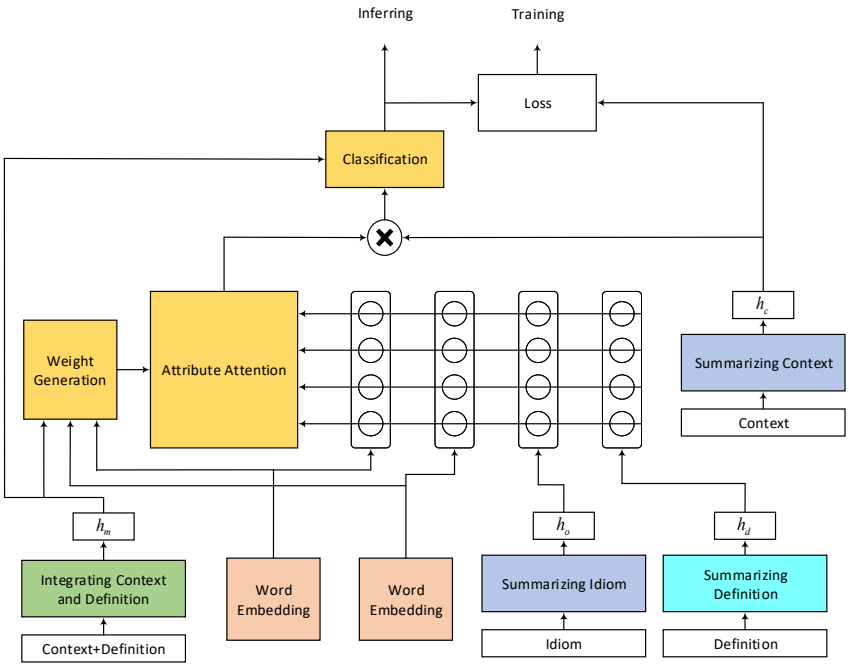


Figure 1: Architecture of our model.

3.2 Attribute Attention

This section is about how to do Attribute Attention and the preparations. In the beginning, we extract the summaries of the context, idiom, and definition. Then we calculate the weight of Attribute Attention with $\boldsymbol{h}_m$ from Section 3.1. After that, Attribute Attention will be done with these summaries and the weight.

3.2.1 Summarizing Context

Summarizing context is to predict what kind of idiom would be the correct answer for the blank based on the contextual information. For example, in Figure 3a, the sentence is "他们希望能___再进一步 *(they hope they can ___ and achieve greater success)*". The input is "他们希望能[MASK]再进一步". The output of [MASK] is defined as $\boldsymbol{h}_c$ as shown in Figure 3a.

3.2.2 Summarizing Idiom

We use BERT to extract and summary character-level information of Chinese idiom. The output is defined as $\boldsymbol{h}_o$, as shown in Figure 3b.

The context and candidate idioms are from the same corpus and share a similar contextual representation. Besides, the [CLS] is not used when summarizing context. Therefore, we use one BERT to model both the context and idiom and use the [CLS] to summarize idioms. In the example of Figure 3b, the candidate idiom is "百尺竿头 *(achieve great achievement)*". The input is "[CLS]百尺竿头"

3.2.3 Summarizing Definition

Introducing the definitions can correct the misuse of idioms. We use [CLS] to summary definition. In the example of Figure 3c, the definition is "比喻高的成就 *(an outstanding achievement)*". The input is "[CLS]比喻高的成就". The output of [CLS] is defined as $\boldsymbol{h}_d$.

3.2.4 Word Embedding of Idiom

We use word embeddings to extract word-level information in this section. To utilize more information from various corpora, more than one word embedding can be introduced. Different attributes of different word embeddings will be assigned different weights in Attribute Attention. The word embeddings from different sources of one idiom are defined as $\{\boldsymbol{e}_i\}_{i=1}^{|\boldsymbol{e}|}$, where $|\boldsymbol{e}|$ is the number of word embeddings.

3.2.5 Weight Generation

As shown in Figure 1, this section is about generating the weight with $\boldsymbol{h}_m$ and $\{\boldsymbol{e}_i\}_{i=1}^{|\boldsymbol{e}|}$. For the standard attention mechanism, the attention weight is a series of scalars, whereas the attention weight is a series of vectors in Attribute Attention.

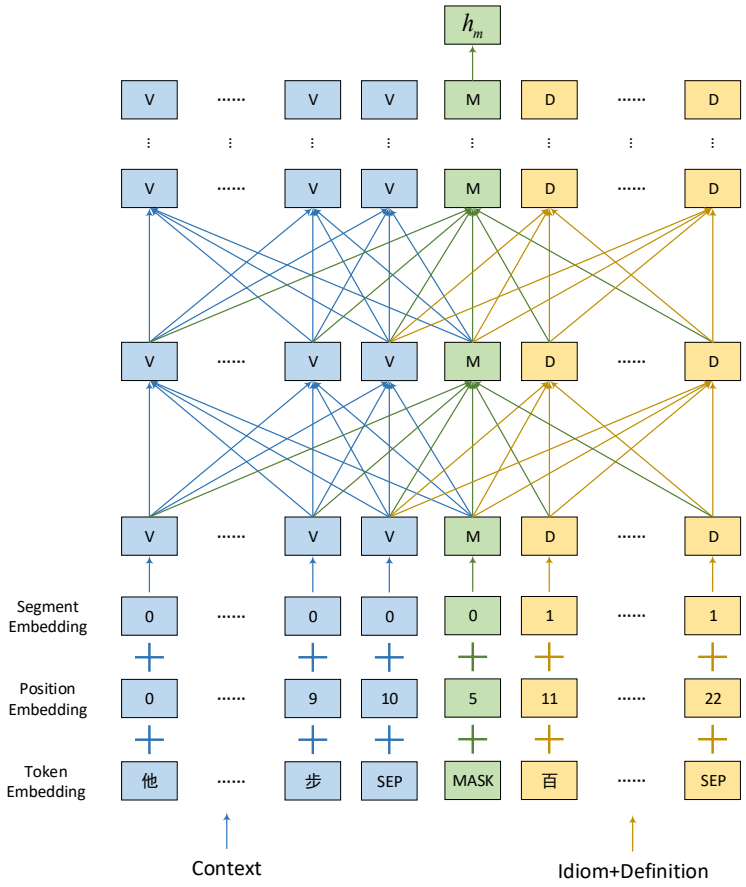


Figure 2: Integrating the context and definition with BERT, where "V" denotes the context, "D" denotes the idiom and definition, and "M" denotes [MASK]. The Multi-Head Attention is applied to the context, definition, and [MASK] in different ways. The input is "他们希望能 [MASK] 再进一步 [SEP] 百尺竿头:比喻高的成就 [SEP]".

$\boldsymbol{h}_m$ contains information about the context and idiom. A Chinese idiom may not be misused in all contexts. $\boldsymbol{h}_m$ can tell the importance of different attributes of an idiom under a certain context. The attention weight vectors for $\boldsymbol{h}_m$ are defined as $\{\boldsymbol{a}_m^{<i>}\}_{i=1}^{|\boldsymbol{e}|+2}$:

$$\boldsymbol{a}_m^{<i>} = \boldsymbol{W}_m^{<i>} \boldsymbol{h}_m \tag{4}$$

where $\boldsymbol{W}_m^{<i>} \in \mathbb{R}^{m \times b}$ is a learnable parameter; m denotes the hidden size of attention, and b denotes the hidden size of BERT such as 768 or 1024.

$\boldsymbol{h}_m$ generates the weight based on the context, which is more accurate but also more likely to overfit. The weight $\{\boldsymbol{a}_m^{<i>}\}_{i=1}^{|\boldsymbol{e}|+2}$ may "remember" every context-idiom pair in the training set. $|\boldsymbol{e}|$ is the number of word embeddings. In this case, we also introduce word embeddings here. The word embedding cannot provide context information but will have stronger generalization ability because it is hard to overfit the training set unless an idiom only appears several times. The attention weight vectors for word embeddings are defined as $\{\boldsymbol{a}_e^{<i>}\}_{i=1}^{|\boldsymbol{e}|+2}$:

$$\boldsymbol{a}_e^{<i>} = \frac{1}{|\boldsymbol{e}|} \sum_{j}^{|\boldsymbol{e}|} \boldsymbol{W}_{e_j}^{<i>} \boldsymbol{e}_j \tag{5}$$

where $\boldsymbol{W}_{e_j}^{<i>} \in \mathbb{R}^{m \times d}$ is a learnable parameter; d denotes the size of word embedding such as 300.

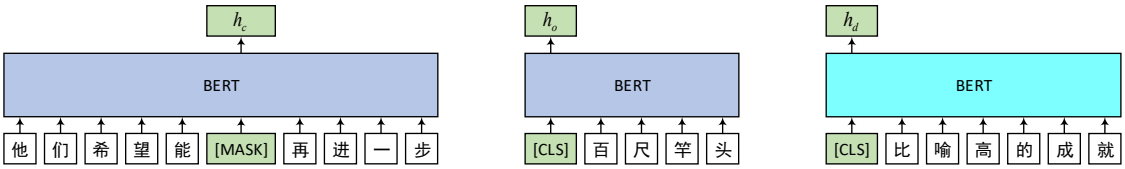
$\boldsymbol{a}_m^{<i>} \in \mathbb{R}^m$ gives more accurate weight but may overfit, whereas $\boldsymbol{a}_e^{<i>} \in \mathbb{R}^m$ is more generalized but lacks the context. We add them up to get the final weight $\{\boldsymbol{a}^{<i>}\}_{i=1}^{|\boldsymbol{e}|+2}$:

$$\boldsymbol{a}^{<i>} = \boldsymbol{a}_m^{<i>} + \boldsymbol{a}_e^{<i>} \tag{6}$$

where $\boldsymbol{a}^{<i>} \in \mathbb{R}^m$. In this way, we can have accuracy and generalization from the two weights.

3.2.6 Attention Calculation

We define $a_j^{<i>}$ as the j-th element of $\boldsymbol{a}^{<i>}$. In other words $a_j^{<i>}$ is the j-th element of the i-th



(a) Context (b) Idiom (c) Definition

Figure 3: Summarizing the context, idiom, and definition.

vector of $\{\boldsymbol{a}^{<i>}\}_{i=1}^{|e|+2}$. Then the softmax function is applied as:

$$\overline{a}_j^{<i>} = \frac{e^{a_j^{<i>}}}{\sum_{k=1}^{|e|+2} e^{a_j^{<k>}}} \tag{7}$$

After that, before applying the attention:

$$\begin{aligned} \boldsymbol{h}_o &\leftarrow \boldsymbol{W}_{ao}\boldsymbol{h}_o \\ \boldsymbol{h}_d &\leftarrow \boldsymbol{W}_{ad}\boldsymbol{h}_d \\ \boldsymbol{e}_i &\leftarrow \boldsymbol{W}_{ae_i}\boldsymbol{e}_i \end{aligned} \tag{8}$$

where $\boldsymbol{W}_{ao} \in \mathbb{R}^{m\times b}$, $\boldsymbol{W}_{ad} \in \mathbb{R}^{m\times b}$, and $\boldsymbol{W}_{ae_i} \in \mathbb{R}^{m\times d}$ are learnable parameters; m denotes the hidden size of attention, b denotes the hidden size of BERT, d denotes the size of word embedding.

As shown in Figure 1, the attention goes through as:

$$\overline{h}_j = \overline{a}_j^{<|e|+1>} h_{o_j} + \overline{a}_j^{<|e|+2>} h_{d_j} + \sum_{i=1}^{|e|} \overline{a}_j^{<i>} e_{i_j} \tag{9}$$

where $\overline{h}_j$ is the j-th element of the output which is defined as $\overline{\boldsymbol{h}} \in \mathbb{R}^m$; h_{o_j} is the j-th element of $\boldsymbol{h}_o$, h_{d_j} is the j-th element of $\boldsymbol{h}_d$, and e_{i_j} is the j-th element of $\boldsymbol{e}_i$;

$\overline{\boldsymbol{h}}$ contains an accurate and correct description of an idiom under a certain context by choosing information from the idiom, definition, and word embeddings. The correct and important part of every representation remains, and the incorrect and unimportant part is dropped.

The final output of Attribute Attention is:

$$u_a = \overline{\boldsymbol{h}}^T \boldsymbol{W}_{ua} \boldsymbol{h}_c \tag{10}$$

where $\boldsymbol{W}_{ua} \in \mathbb{R}^{m\times b}$ is a learnable parameter. $u_a \in \mathbb{R}^1$ is the score to describe whether a candidate idiom is the correct answer.

3.3 Classification

This section will introduce the classification part in Figure 1. One reason for Attribute Attention summarizing the context and definition is to make use of word embedding. Using $\boldsymbol{h}_m$ for classification can provide more details about the relationship between characters of the context and characters of the definition.

Formally, the classification for $\boldsymbol{h}_m$ is:

$$u_m = \boldsymbol{W}_{cm}\boldsymbol{h}_m + b_{cm} \tag{11}$$

where $\boldsymbol{W}_{cm} \in \mathbb{R}^{1\times b}$ and $b_{cm} \in \mathbb{R}^1$ are learnable parameters. $u_m \in \mathbb{R}^1$ is the score describing whether a candidate idiom is the correct answer.

u_a and u_m denote the score of one candidate idiom. We further define the $\{u_{ai}\}_{i=1}^n$ and $\{u_{mi}\}_{i=1}^n$ as the scores of all candidate idioms, where n denotes the number of candidate idioms. Then we add them up:

$$u_{si} = u_{ai} + u_{mi} \tag{12}$$

and pass u_{si} through softmax function:

$$p_i = \frac{e^{u_{si}}}{\sum_{k=1}^n e^{u_{sk}}} \tag{13}$$

p_i is the possibility for the i-th candidate idiom to be the correct answer. This is the end of inferring but not training.

3.4 Extra Loss

Because Chinese idioms are used in more unique and specific context than common words, we choose the answer from all Chinese idioms that appear in the whole cloze test dataset as an extra loss for training. Formally, we use $\boldsymbol{h}_c$ to predict the correct answer from the whole vocabulary of candidate Chinese idioms:

$$\boldsymbol{u}_c = \boldsymbol{W}_{cv}\boldsymbol{h}_c + \boldsymbol{b}_v \tag{14}$$

where $\boldsymbol{W}_{cv} \in \mathbb{R}^{v \times b}$ and $\boldsymbol{b}_v \in \mathbb{R}^v$ are learnable parameters; v denotes the number of all candidate Chinese idioms which is much larger than n.

$$\boldsymbol{q} = \text{softmax}(\boldsymbol{u}_c) \tag{15}$$

$\boldsymbol{q} \in \mathbb{R}^v$ are possibilities for all candidate idioms being the correct answer. In this way, the model can learn relationships between much more idioms every time. Due to the uniqueness and specificity of Chinese idioms, this will not cause limited noises but improve the performance significantly. Without Extra Loss, relationships between only given candidate idioms are considered every time.

When inferring, the max possibility of $\{p_i\}_{i=1}^n$ is the final result. For training, the cross entropy loss of $\{p_i\}_{i=1}^n$ is defined as l_p, and the cross entropy loss of $\boldsymbol{q}$ is defined as l_q. The final loss is:

$$l = l_p + \beta l_q \tag{16}$$

where β is a hyper-parameter to determine the weight of the loss l_q. Empirically, we suggest setting the value of β as 0.5. l is the final loss for training.

4 Experiment

4.1 Training Details

In this section, we will introduce the details and hyper-parameters for training our model.

Dataset ChID dataset (Zheng et al., 2019) is used in experiments. Table 1 shows a simple example of the dataset. Given a sentence with a blank and several candidate Chinese idioms, an examinee is required to choose a Chinese idioms which best matches the context surrounding the blank. The corpus of ChID contain news, novels, and essays. News and novels are treated as in-domain data, which contains a training set, a development set **Dev**, and a test set **Test**. Essays are reserved for out-of-domain test **Out**, which can evaluate the generalization ability. In this way, the model is trained on news and novels but evaluated on essays. **Ran** and **Sim** are two test sets which have the same sentences as **Test**. In **Ran**, candidate idioms are not similar to the golden answer. In **Sim**, candidate idioms are similar idioms to golden answer.

BPretrained Model Pretrained RoBERTa-base (Liu et al., 2019) for Chinese with 12 layers and word embeddings from (Song et al., 2018; Li et al., 2018; Qiu et al., 2018) are used.

Hyper-parameters n is 7 because there are seven candidate idioms for every blank in ChID dataset (Zheng et al., 2019). v is 3848 because ChID dataset (Zheng et al., 2019) contains 3848 candidate idioms in total. The hidden size of attention m is 100. β as 0.5.

Optimizer The optimizer is Adam (Kingma and Ba, 2014) for BERT with linear schedule and a warm-up ratio of 0.05. The learning rate for RoBERTa is 2e-5, and for other parameters is 1e-3.

Parameters number The number of parameters of our model for experiments is 322M. The learnable parameters are initialized by (He et al., 2015).

GPU & Environment The model is running on a GPU of NVIDIA GeForce RTX 2080 Ti. Due to the limited GPU RAM, we use gradient accumulation for training. The operating system is Ubuntu 18.04. We use PyTorch 1.4.0 (Paszke et al., 2019) and Transformers 2.4.1 (Wolf et al., 2019) to implement our model. We also use mixed precision training with NVIDIA Apex 0.1 (Micikevicius et al., 2017) to accelerate our model. It takes an average of 42 hours per epoch, and the model achieves the best result within 10 epochs.

Metrics The metric for evaluation is the accuracy, which is implemented by Scikit-learn (Pedregosa et al., 2011).

4.2 Comparison

The description of other models are as follows:

AR Attentive Reader (AR) (Hermann et al., 2015). AR uses an attention mechanism to read the sentence.

SAR Stanford Attentive Reader (SAR) (Chen et al., 2016). SAR is a improvement based on AR.

CR Chengyu Reader (CR) (Jiang et al., 2018). CR extracts the summary of definition and adopts a similar attention mechanism of AR.

EAR Enhanced Attentive Reader (EAR) (Fu and Zhang, 2019) EA Reader contains a method called Multi-Space Context Fusion and integrates the method with the attention mechanism of AR.

X-RoBERTa We design AR-RoBERTa, SAR-RoBERTa, CR-RoBERTa, and EAR-RoBERTa to make a fair comparison. The LSTMs of them are all replaced by RoBERTa-base which has 12 layers,

	Dev	Test	Ran	Sim	Out
Human	-	87.1	97.6	82.2	86.2
AR (Hermann et al., 2015)	72.7	72.4	82.0	66.2	62.9
SAR (Chen et al., 2016)	71.7	71.5	80.0	64.9	61.7
CR (Jiang et al., 2018)	74.1	73.5	82.8	68.5	65.2
EAR (Fu and Zhang, 2019)	74.6	74.5	84.4	67.9	65.5
AR-RoBERTa (Hermann et al., 2015; Liu et al., 2019)	77.1	77.1	89.0	68.9	70.9
SAR-RoBERTa (Chen et al., 2016; Liu et al., 2019)	76.3	76.7	88.5	68.0	69.8
CR-RoBERTa (Jiang et al., 2018; Liu et al., 2019)	78.0	78.3	89.9	70.0	71.7
EAR-RoBERTa (Fu and Zhang, 2019; Liu et al., 2019)	78.7	79.2	90.5	71.7	72.3
Our model	**83.0**	**83.1**	**92.3**	**76.1**	**77.6**

Table 4: Comparison of accuracies of different models on ChID dataset.

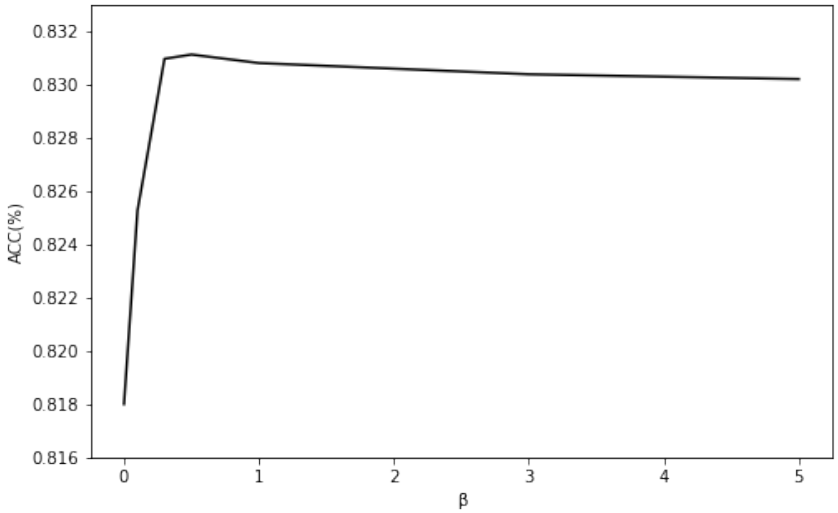


Figure 4: Performance of our model with different β on **Test**.

which is the same as our model. Both LSTM and RoBERTa provides contextual information.

Table 4 shows the accuracies of all methods. The result of human is given by (Zheng et al., 2019). Our model outperforms all other models in **Dev**, **Test**, **Ran**, **Sim**, and **Out**. Besides, our model has much better generalization ability. For example, comparing with EAR-RoBERTa, our model has a 3.9% improvement on **Test** but 5.3% on **Out**.

4.3 Extra Loss Studies

This section explores how β influence the accuracy of our model on **Test**. Figure 4 shows the results. When $\beta = 0$, the Extra Loss is not used, which shows the performance of our model that does not use Extra Loss. The accuracy increase very quickly when $\beta < 0.3$. The accuracy reaches the highest point when $\beta = 0.5$. The accuracy start decreasing slowly when $\beta > 1$. A larger β makes the extra loss l_q too important and overshadow the normal loss l_p, which makes the model deviate from its purpose. Extra Loss gives a significant improvement and costs very few computing resources.

5 Conclusion

In this paper, we propose a model for the Chinese idiom cloze test. We introduce the definition and propose Attribute Attention to balance the importance of different representations of the Chinese idiom. We add Extra Loss calculated by choosing the answer from the whole vocabulary of Chinese idioms to improve the performance further, which costs very few computing resources. In experiments, our model outperforms state-of-the-art method.

References

Samuel W.K. Chan and Mickey W.C. Chong. 2008. An agent-based approach to Chinese word segmentation. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

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

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018a. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018b. One size fits all? a simple LSTM for non-literal token and construction-level classification. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 70–80, Santa Fe, New Mexico. Association for Computational Linguistics.

A. Fathima Shirin and C. Raseek. 2018. Replacing idioms based on their figurative usage. In *2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)*, pages 1–6.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Michael Flor and Beata Beigman Klebanov. 2018. Catching idiomatic expressions in EFL essays. In *Proceedings of the Workshop on Figurative Language Processing*, pages 34–44, New Orleans, Louisiana. Association for Computational Linguistics.

Chengzhen Fu, Yuntao Li, and Yan Zhang. 2019. Atnet: Answering cloze-style questions via intra-attention and inter-attention. In *Advances in Knowledge Discovery and Data Mining*, pages 242–252, Cham. Springer International Publishing.

Chengzhen Fu and Yan Zhang. 2019. Ea reader: Enhance attentive reader for cloze-style question answering via multi-space context fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6375–6382.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City, New York. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Zhiying Jiang, Boliang Zhang, Lifu Huang, and Heng Ji. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–158, New Orleans, Louisiana. Association for Computational Linguistics.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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. *arXiv preprint arXiv:1909.11942*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.

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

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. Springer.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.

I. Spasic, L. Williams, and A. Buerki. 2017. Idiom—based features in sentiment analysis: Cutting the gordian knot. *IEEE Transactions on Affective Computing*, pages 1–1.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 970–979, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 163–172, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Liang Wang, Sujian Li, Wei Zhao, Kewei Shen, Meng Sun, Ruoyu Jia, and Jingming Liu. 2018. Multi-perspective context aggregation for semi-supervised cloze-style reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 857–867, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Ge Xu, Xinfan Meng, and Houfeng Wang. 2010. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1209–1217, Beijing, China. Coling 2010 Organizing Committee.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. 2018. Subword-augmented embedding for cloze reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1802–1814, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.