A BERT-based Dual Embedding Model for Chinese Idiom Prediction

Minghuan Tan
School of Information Systems
Singapore Management University
`mhtan.2017@phdcs.smu.edu.sg`

Jing Jiang
School of Information Systems
Singapore Management University
`jingjiang@smu.edu.sg`

Abstract

Chinese idioms are special fixed phrases usually derived from ancient stories, whose meanings are oftentimes highly idiomatic and non-compositional. The Chinese idiom prediction task is to select the correct idiom from a set of candidate idioms given a context with a blank. We propose a BERT-based dual embedding model to encode the contextual words as well as to learn dual embeddings of the idioms. Specifically, we first match the embedding of each candidate idiom with the hidden representation corresponding to the blank in the context. We then match the embedding of each candidate idiom with the hidden representations of all the tokens in the context thorough context pooling. We further propose to use two separate idiom embeddings for the two kinds of matching. Experiments on a recently released Chinese idiom cloze test dataset show that our proposed method performs better than the existing state of the art. Ablation experiments also show that both context pooling and dual embedding contribute to the improvement of performance.

1 Introduction

In this paper, we study Chinese idiom prediction, a language understanding problem that has not been extensively explored before in computational linguistics. Chinese idioms, mainly Chengyu (成语) (set phrases) (Wang and Yu, 2010; Wang, 2019), have fixed forms in structure; the component characters (mostly four) cannot be changed. Chinese idioms are characterized by rich contents, concise forms and frequent use (Wang, 2019) with properties of structural regularity, semantic fusion, and functional integrity (Shao, 2018; Wang, 2019). Chinese idioms are commonly used in both written and spoken Chinese, and understanding Chinese idioms is important for learning Chinese as a second language.

The meaning of each Chinese idiom may not be literally understood through the composition of its characters, especially for those which are derived from historical stories or formulated using ancient Chinese grammars. For example, "一定不易" is literally interpreted as *"it must be not easy"* in modern Chinese. However, the idiom is constructed from grammars and word senses of ancient Chinese. Its idiomatic meaning is *"once decided, never change"*, which is not even close to the literal meaning. As a result, the usage of Chinese idioms poses a challenge on language understanding not only for humans but also for artificial intelligence. Due to their pervasive usage, Chinese idiom prediction is an important task in Chinese language understanding.

There have been several studies focusing on representing Chinese idioms using neural network models (Jiang et al., 2018; Liu et al., 2019b), but they were limited by the amount of data available for training. Recently, Zheng et al. (2019) released a large-scale Chinese IDiom Dataset (ChID) to facilitate machine comprehension of Chinese idioms. The ChID dataset contains more than 500K passages and 600K blanks, making it possible for researchers to train deep neural network models. The dataset is in cloze test style that target Chinese idioms in passages are replaced by blanks. For each blank, a set of candidate Chinese idioms is provided and the task is to pick the correct one based on the context. Table 1

Passage: 戴尔克·施特略夫把自己的工作全部撂下，整天服侍病人，又体贴，又关切。他的手脚非常利索，把病人弄得舒舒服服。大夫开了药，他总是连哄带骗地劝病人按时服用，我从来没想到他的手段这么巧妙。无论做什么事他都不嫌麻烦。尽避他的收入一向只够维持夫妻两人的生活，从来就不宽裕，现在他却_____，购买时令已过、价钱昂贵的美味，想方设法叫思特里克兰德多吃一点东西（他的胃口时好时坏，叫人无法捉摸）。

Dirk Stroeve, giving up his work entirely, nursed Strickland with tenderness and sympathy. He was dexterous to make him comfortable, and he exercised a cunning of which I should never have thought him capable to induce him to take the medicines prescribed by the doctor. Nothing was too much trouble for him. Though his means were adequate to the needs of himself and his wife, he certainly had no money to waste; but now he was _____ in the purchase of delicacies, out of season and dear, which might tempt Strickland's capricious appetite.

Candidates:
○ 月明星稀 The moon is bright and stars are few; with a clear moon and few stars
○ 苦尽甘来 bitterness ends and happiness begins
○ 坐吃山空 even a great fortune can be depleted by idleness
● 大手大脚 extravagant or wasteful
○ 斤斤计较 haggle over every ounce
○ 不见天日 a world of darkness; total absence of justice
○ 好吃懒做 be fond of eating and averse to work; be gluttonous and lazy

Table 1: An example showing a passage with a blank and seven candidate idioms. The idiom with the solid circle is the ground truth idiom. The passage is from a Chinese translation of *The Moon and Sixpence*. Translations of idioms are extracted from online dictionary `http://dict.cn`.

shows an example from the testing set of ChID. We can see that among the seven candidates, most can fit into the local context "现在他却_____" ("but now he was _____") well grammatically, but to select the best answer we need to understand the entire passage.

In this paper, we propose a BERT-based dual embedding model for the Chinese idiom prediction task. We first present two baseline models that use BERT to process and match passages and candidate answers in order to rank the candidates. Observing that these baselines do not explicitly model the global, long-range contextual information in the given passage for Chinese idiom prediction, we propose a context-aware pooling operation to force the model to explicitly consider all contextual words when matching a candidate idiom with the passage. Furthermore, we propose to split the embedding vector of each Chinese idiom into two separate vectors, one modeling its local properties and the other modeling its global properties. We expect the embedding for local properties to capture the syntactic properties of an idiom, while the embedding for global properties to capture its topical meaning. In addition, using idiom embeddings makes it possible for us to consider the entire Chinese idiom vocabulary as the candidate set, which is computationally intractable compared to pretrained BERT models with multiple-sequence classification. we apply this enlarged candidates heuristic to all the models with idiom embeddings to further strengthen the performance.

To evaluate the effectiveness of the BERT-based dual embedding model, we conduct experiments on the ChID dataset. Our experiments show that our method can outperform several existing methods tested by Zheng et al. (2019) as well as our baseline methods. We also find that both context-aware pooling and dual embedding contribute to the performance improvement. To prove the effectiveness of our model, we also evaluate it against a public leaderboard of ChID Competition. The results show that our model is competitive compared to the top-ranked systems. We can also achieve better performance with a large margin compared with several methods using pretrained language models. We also conduct further analysis using a gradient-based attribution method to check if our model can indeed capture global information to make correct predictions. Some case studies show that indeed our method makes use of more global contextual information to make predictions.

2 Related Work

2.1 Cloze-style Reading Comprehension

Cloze-style reading comprehension is an important form in assessing machine reading abilities. Researchers created many large-scale cloze-style reading comprehension datasets like CNN/Daily Mail (Hermann et al., 2015), Children's Book Test (CBT) (Hill et al., 2015) and RACE (Lai et al., 2017). These datasets have inspired the design of various neural-based models (Hermann et al., 2015; Chen et al., 2016) and some become benchmarks for machine reading comprehension. The dataset ChID used in this paper is also a large scale cloze-style dataset but focuses on Chinese idiom prediction.

2.2 Pre-trained Language Models

Language model pre-training has been proven to be effective over a list of natural language tasks at both sentence level (Bowman et al., 2015) and token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016). Existing strategies of using pre-trained language models include feature-based methods like ELMO (Peters et al., 2018) and fine-tuning methods such as OpenAI GPT (Radford et al., 2018) and BERT (Devlin et al., 2019). BERT-based fine-tuning strategy and its extensions (Cui et al., 2019; Yang et al., 2019; Liu et al., 2019a) are pushing performance of neural models to near-human or super-human level. In this paper, we use pre-trained Chinese BERT with Whole Word Masking (Cui et al., 2019) as text sequence processor.

2.3 Modelling Figurative Language

Figurative (or non-literal) language is different from literal language where words or characters in literal language act in accordance with conventionally accepted meanings or denotation. In figurative language, meaning can be detached from the words or characters while a more complicated meaning or heightened effect is reattached. As a special type of figurative language, idioms have been actively researched in tasks like Idiom Identification (Muzny and Zettlemoyer, 2013), Idiom Recommendation (Liu et al., 2019b) and Idiom Representation (Gutiérrez et al., 2016; Liu et al., 2017; Jiang et al., 2018; Zheng et al., 2019). In this paper, we will focus on the representations of Chinese idioms using a BERT-based approach.

3 Method

3.1 Task Definition and Dataset

We formally define the Chinese idiom prediction task as follows. Given a passage P, represented as a sequence of tokens $(p_1, p_2, \ldots, p_n)$, where each token is either a Chinese character or the special "blank" token `[MASK]`, and a set of K candidate Chinese idioms denoted as $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$, our goal is to select an idiom $a^* \in \mathcal{A}$ that best fits the blank in P. See the example in Table 1.

We assume that a set of training examples in the form of triplets, each containing a passage, a candidate set and the ground truth answer, is given. We denote the training data as $\{(P_i, \mathcal{A}_i, a_i^*)\}_{i=1}^N$. We use $\mathcal{V}$ to denote the vocabulary of all Chinese idioms observed in the training data, i.e., $\mathcal{V} = \cup_{i=1}^N \mathcal{A}_i$.

To facilitate the study of Chinese idiom comprehension using deep learning models, Zheng et al. (2019) released the ChID dataset. The dataset was created in the "cloze" style. The authors collected diverse passages from novels and essays on the Internet and news articles from THUCTC (Guo et al., 2016). The authors then replaced target Chinese idioms found in these passages with the blank token. To construct the candidate answer set for each blank, the authors considered synonyms, near-synonyms and other idioms either irrelevant or opposite in meaning to the ground truth idiom (Zheng et al., 2019).

3.2 BERT Baselines

Previous methods applied to the ChID dataset are not based on BERT (Devlin et al., 2019) or Transformer (Vaswani et al., 2017) architecture. Because of the success of BERT for many NLP tasks, here we first present two BERT baselines. The first one treats a Chinese idiom as a sequence of characters.

It combines the passage with each candidate idiom into a single sequence and processes multiple sequences, one for each candidate, using BERT. The second one treats a Chinese idiom as a single token that has its own embedding vector. The method uses BERT to process the passage and then matches the encoded passage with each candidate idiom's embedding. These baselines can be regarded as standard ways to solve the Chinese idiom prediction problem using BERT.

For the second baseline that uses idiom embeddings, we also present a heuristic that uses an enlarged candidate set to improve learning. This heuristic is only applicable to the second baseline because it would be computationally too expensive for the first baseline.

BERT Baseline with Idioms as Character Sequences: A straightforward way to apply BERT for Chinese idiom prediction is as follows. Given a passage $P = (p_1, p_2, \ldots, \texttt{[MASK]}, \ldots, p_n)$ and a candidate answer $a_k \in \mathcal{A}$, we first concatenate them into a single sequence $(\texttt{[CLS]}, p_1, p_2, \ldots, p_n, \texttt{[SEP]}, a_{k,1}, a_{k,2}, a_{k,3}, a_{k,4}, \texttt{[SEP]})$, where $a_{k,1}$ to $a_{k,4}$ are the four Chinese characters that idiom a_k is composed of. We can then directly use BERT to process this sequence and obtain the hidden representation for $\texttt{[CLS]}$ on the last (L-th) layer, denoted by $\mathbf{h}_{k,0}^L \in \mathbb{R}^d$. To select the best answer idiom, we first use a linear layer to process $\mathbf{h}_{k,0}^L$ for $k = 1, 2, \ldots, K$ and then use standard softmax to obtain the probabilities of each candidate. To train the model, we use standard negative log likelihood as the loss function.

BERT Baseline with Idiom Embeddings: Many Chinese idioms are non-compositional and therefore their meanings should not be directly derived from the embeddings of its four individual characters, as the baseline above does. E.g., "狐假虎威" literally means a fox assuming the majesty of a tiger, but it is usually used to describe someone flaunting his powerful connections. Therefore, we hypothesize that learning a single embedding vector for the entire idiom can help the understanding of idioms.

In this second BERT baseline, instead of concatenating the passage and a candidate answer into a single sequence for BERT to process, we keep them separated. We only use BERT to process the passage sequence $(\texttt{[CLS]}, p_1, p_2, \ldots, \texttt{[MASK]}, \ldots, p_n, \texttt{[SEP]})$. Afterwards, we use the hidden representation of $\texttt{[MASK]}$ at the last (L-th) layer, denoted as $\mathbf{h}_b^L$, to match each candidate answer. In this way, no matter how many candidate answers there are, BERT is used to process the passage only once. On the other hand, each Chinese idiom has a hidden embedding vector, which is to be learned.

We use $\mathbf{a}_k$ to denote the embedding vector for candidate $a_k \in \mathcal{A}$. The hidden representation $\mathbf{h}_b^L$ is fused with each candidate idiom via element-wise multiplication. Then the probability of selecting a_k among all the candidates $\mathcal{A}$ is defined as follows:

$$p_k = \frac{\exp(\mathbf{w} \cdot (\mathbf{a}_k \odot \mathbf{h}_b^L) + b)}{\sum_{k'=1}^{K} \exp(\mathbf{w} \cdot (\mathbf{a}_{k'} \odot \mathbf{h}_b^L) + b)}, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are model parameters, and $\odot$ is element-wise multiplication. To train the model, we again use negative log likelihood as the loss function.

Heuristic with Enlarged Candidate Set: The ChID dataset uses only a small set of negative answers in each candidate set and these negatives are fixed for each example during training. It is reasonable to expect that most of the remaining Chinese idioms not in the candidate set are also negative answers and including them in the training data may help. We therefore use a heuristic that considers an enlarged candidate set to further boost the performance.

To apply this heuristic, we define a candidate set $\mathcal{A}'$ to be the same as $\mathcal{V}$ (i.e., the vocabulary containing all Chinese idioms observed in the training data), and then define a second term in the loss function that is the negative log likelihood of selecting the correct answer from this enlarged candidate set.

Note that because $\mathcal{A}'$ is large, this heuristic is not feasible to be applied to the character sequence-based BERT baseline, because it would require inserting each candidate into the passage for BERT to process, which would be computationally too expensive. Therefore, this enlarged candidate set heuristic is only applied to the idiom embedding-based BERT baseline. Specifically, we can define the probability

of selecting answer $a \in \mathcal{A}'$ as follows:

$$q_a = \frac{\exp(\mathbf{a} \cdot \mathbf{h}_b^L)}{\sum_{c \in \mathcal{A}'} \exp(\mathbf{c} \cdot \mathbf{h}_b^L)}. \tag{2}$$

Let q_i^* denote the probability of selecting the ground truth idiom among all candidates in $\mathcal{A}'$ for the i-th training example, and p_i^* denote the probability of selecting the correct answer among the original candidate set $\mathcal{A}$ for the i-th training example. Our training loss function is then defined as follows:

$$L = -\sum_{i=1}^{N} (\log(p_i^*) + \log(q_i^*)). \tag{3}$$

3.3 Our Dual Embedding Model

The BERT baselines presented above are reasonable baselines, but they have a potential problem. We observe that in order for an idiom to fit into a passage well, it has to not only grammatically (i.e., syntactically) fit into the local context surrounding the [MASK] token but also show semantic relevance to the whole passage. In the example shown in Table 1, a correct answer has to first be an adjective rather than, say, a noun or a verb. In addition, given the global context of the entire passage, it is understood that the correct answer should convey the meaning of "extravagant."

Based on the observation above, we introduce the following two changes to the second BERT baseline, i.e., the idiom embedding-based BERT baseline, introduced in Section 3.2.

3.3.1 Context-aware Pooling

As we have pointed out earlier, oftentimes Chinese idioms have non-compositional meanings, and to evaluate whether a Chinese idiom is suitable in a passage, we need to understand the semantic meaning of the entire passage. Therefore, it is important for us to not only try to match an idiom with the local context it is to be placed in (which can roughly be modeled by $\mathbf{h}_b^L$) but also to match it with the entire passage. Let us use $\mathbf{a}_k$ to denote the embedding for idiom a_k. Recall that $\mathbf{H}^L = (\mathbf{h}_0^L, \mathbf{h}_1^L, \ldots, \mathbf{h}_n^L)$ represents the hidden states of the last layer of BERT after it processes the passage sequence. Our method with context-aware pooling can be represented as follows:

$$p_k = \frac{\exp(\mathbf{a}_k \cdot \mathbf{h}_b^L + \max_{i=0}^{n}(\mathbf{a}_k \cdot \mathbf{h}_i^L))}{\sum_{k'=1}^{K} \exp(\mathbf{a}_{k'} \cdot \mathbf{h}_b^L + \max_{i=0}^{n}(\mathbf{a}_{k'} \cdot \mathbf{h}_i^L))}. \tag{4}$$

3.3.2 Dual Embeddings

Because we need to match an idiom with both $\mathbf{h}_b^L$ and the entire passage, the second idea we propose is to split the embedding of an idiom into two "sub-embedding" vectors, which we refer to as "dual embeddings." Let us use $\mathbf{a}_k^u$ and $\mathbf{a}_k^v$ to denote the two embeddings for idiom a_k.

We then calculate the probability of selecting candidate a_k as follows:

$$p_k = \frac{\exp(\mathbf{a}_k^u \cdot \mathbf{h}_b^L + \max_{i=0}^{n}(\mathbf{a}_k^v \cdot \mathbf{h}_i^L))}{\sum_{k'=1}^{K} \exp(\mathbf{a}_{k'}^u \cdot \mathbf{h}_b^L + \max_{i=0}^{n}(\mathbf{a}_{k'}^v \cdot \mathbf{h}_i^L))}. \tag{5}$$

We also adopt the heuristic of enlarged candidate set from Section 3.2. With the candidate set $\mathcal{A}'$ to be the same as $\mathcal{V}$, we still use dual embeddings to represent each idiom, but when we match the dual embeddings with the passage, we use both $\mathbf{a}^u$ and $\mathbf{a}^v$ to match $\mathbf{h}_b^L$ only. This is because it would be too expensive to match $\mathbf{a}^v$ of each candidate with the entire sequence of hidden states $\mathbf{H}^L$ as we now have many candidates. So we define the probability of selecting answer $a \in \mathcal{A}'$, i.e., selecting the ground truth answer from the entire vocabulary of Chinese idioms, as follows:

$$q_a = \frac{\exp(\mathbf{a}^u \cdot \mathbf{h}_b^L + \mathbf{a}^v \cdot \mathbf{h}_b^L)}{\sum_{c \in \mathcal{A}'} \exp(\mathbf{c}^u \cdot \mathbf{h}_b^L + \mathbf{c}^v \cdot \mathbf{h}_b^L)}. \tag{6}$$

Similarly, to train the model, we use negative log likelihood as shown before.

1316

4 Experiments

In this section, we evaluate our proposed dual embedding method using the ChID dataset. We also use an attribution method to visualize how each proposed method works on some selected cases.

4.1 Evaluation on ChID-Official

	In-domain				Out-of-domain	Total
	Train	Dev	Test	Total	Out	Total
Passages	520,711	20,000	20,000	560,711	20,096	580,807
Distinct idioms	3,848	3,458	3,502	3,848	3,626	3,848
Total blanks	648,920	24,822	24,948	698,690	30,023	728,713

Table 2: Some statistics of the **ChID** dataset.

Data Split: In the first set of experiments, We use the official release of ChID[1], denoted as **ChID-Official**. The data has a training set, a development set and a few different test sets. Besides the standard test set **Test**, the authors also constructed the following test sets: **Ran**: In this test set, the candidate idioms are randomly sampled from the vocabulary $\mathcal{V}$. No synonyms or near-synonyms were intentionally added as candidates. **Sim**: In this test set, the candidates are sampled from the top-10 similar idioms and are more challenging than the Ran test dataset. The only difference of **Test**, **Ran** and **Sim** is the candidate sets. **Out**: This is an out-of-domain test dataset. The passages come from essays (whereas the training and development data comes from news and novels). Statistics of the data can be found in Table 2.

Methods Compared: We compare the following different methods. Performance of the first three baselines are directly taken from (Zheng et al., 2019). It is worth noting that the three baselines use BiLSTM as their backbones while our methods use BERT (Transformer) as our backbones. Although BiLSTM with attention can also capture the global contextual information in the passages, our experiments below will show that empirically our BERT-based methods are more effective.

Language Model (LM): This method is based on standard bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997; Zhou et al., 2016). It uses BiLSTM to encode the given passage and obtain the hidden state of the blank. Then it compares the blank state with the embedding vector of each candidate idiom to choose the best idiom.

Attentive Reader (AR): This method also uses BiLSTM but augments it with attention mechanism. It is based on the Attentive Reader model by (Hermann et al., 2015).

Standard Attentive Reader (SAR): This is an altered version of Attentive Reader, where attention weights are computed using a bilinear matrix (Chen et al., 2016).

BL-CharSeq: This is the first BERT baseline treating idioms as character sequences.

BL-IdmEmb (w/o EC): This is the second BERT baseline using idiom embeddings. In this version, we do not use enlarged candidate set.

BL-IdmEmb: This baseline is the same as BL-IdmEmb (w/o EC) but incorporates the heuristic of enlarged candidate set.

Ours-CP: This is our method with contextual pooling (CP) as presented in Section 3.3.1. This method also incorporates the enlarged candidate set heuristic.

Ours-Full (CP+DE): This is our method with both context pooling (CP) and dual embedding (DE), as presented in Section 3.3.2. This method also uses the enlarged candidate set heuristic.

Evaluation Metrics: A standard metric for the task of Chinese idiom prediction is accuracy, which is the percentage of test examples where our predicted idiom is the same as the ground truth idiom. Here besides accuracy, we also consider another setting where we do not have a pre-defined set of candidate idioms, or in other words, we consider *all* Chinese idioms in our vocabulary as candidates. For this

		Dev		Test		Ran		Sim		Out	
		ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR	ACC	MRR
Human	(Zheng et al., 2019)	-	-	87.1	-	97.6	-	82.2	-	86.2	-
LM	(Zheng et al., 2019)	71.8	-	71.5	-	80.7	-	65.6	-	61.5	-
AR	(Zheng et al., 2019)	72.7	-	72.4	-	82.0	-	66.2	-	62.9	-
SAR	(Zheng et al., 2019)	71.7	-	71.5	-	80.0	-	64.9	-	61.7	-
BL-CharSeq		79.33	-	79.42	-	88.84	-	72.93	-	73.11	-
BL-IdmEmb (w/o EC)		73.59	0.017	73.31	0.017	81.05	0.017	68.13	0.017	63.82	0.012
BL-IdmEmb		80.24	0.433	79.76	0.429	91.87	0.429	71.93	0.429	72.17	0.332
Ours-CP		82.03	0.436	81.86	0.434	92.46	0.434	74.71	0.434	74.82	0.328
Ours-Full (CP+DE)		**82.58**	**0.450**	**82.40**	**0.447**	**92.73**	**0.447**	**75.02**	**0.447**	**75.73**	**0.354**

Table 3: The experiment results on ChID. We only compute MRR for methods that have idiom embeddings.

setting, we use Mean Reciprocal Rank (MRR) (Voorhees, 1999; Radev et al., 2002), a well-established metric for ranking problems, as the evaluation metric.

Other Settings: We use pre-trained BERT for Chinese with Whole Word Masking (WWM) (Cui et al., 2019)[2]. To reduce computational cost, we choose 128 as the maximum length for the input sequence, and we truncate passages longer than this limit by keeping only the 128 characters surrounding [MASK], with [MASK] in the middle.

We use 4 Nvidia 1080Ti GPU cards and a batch size of 10 per card with a total 5 training epochs. The initial learning rate is set to $5e^{-5}$ with 1000 warm-up steps. We use the optimizer *AdamW* in accordance with a learning rate scheduler *WarmupLinearSchedule*. Our code has been made available online[3].

Results: We show the comparison of the performance of the various methods together with the human performance in Table 3. For Human, LM, AR and SAR, the performance shown in the table is taken directly from ChID (Zheng et al., 2019).

We can observe the following from the table. (1) In general, methods using BERT (including both the baselines and our methods) perform substantially better than previous methods based on BiLSTMs. This is not surprising and confirms the general observation that pre-trained BERT is generally very effective for many NLP tasks. (2) Our two methods that use context pooling to explicitly incorporate more contextual information consistently work better than the BERT-based baselines that do not perform context pooling. This shows the importance of using context pooling to encode long-range contextual information for the task of Chinese idiom prediction. (3) Comparing **Ours-Full (CP+DE)** with **Ours-CP**, we can see that **Ours-Full (CP+DE)** consistently outperforms **Ours-CP**, for all evaluation splits in terms of both accuracy and MRR. This shows that our full model using dual embeddings coupled with context-aware pooling makes the model more expressive and captures the underlying meanings of Chinese idioms better. It is also worth noting that on the **Out** split, **Ours-Full (CP+DE)** achieves significant improvement over **Ours-CP**, showing better generalization ability of the dual embeddings.

It is interesting to observe that although we hypothesize that the meanings of Chinese idioms are oftentimes not compositional, **BL-CharSeq** performs better than **BL-IdmEmb (w/o EC)**. We suspect that this is because the **BL-CharSeq** method allows cross attention between the passage and the characters in each candidate idiom, whereas **BL-IdmEmb (w/o EC)** encodes both the passage and a candidate as a vectors without allowing any cross attention between them. However, the design of **BL-IdmEmb (w/o EC)** allows a large number of candidates to be considered, and when we use the enlarged candidate set, we see that **BL-IdmEmb** performs similarly to **BL-CharSeq**. When we subsequently incorporate context pooling and dual embedding, we are able to achieve better performance than **BL-CharSeq**.

Model	Dev	Test	Out
Top-1 (wssb)	88.35	90.57	**85.54**
Top-2 (On The Road)	**90.59**	**91.35**	84.93
Top-3 (Beenle)	81.94	89.27	84.72
BERT-base	82.20	82.04	-
ERNIE-base	82.46	82.28	-
RoBERTa-large	85.31	84.50	-
RoBERTa-wwm-large-ext	85.81	85.37	-
Ours-Full	89.68	89.55	84.43

Table 4: Experiment results on ChID-Competition.

Overall, we can see that the experiment results demonstrate that both context-aware pooling and dual embeddings are effective, and our proposed full method generally can outperform all the other methods we consider that represent the state of the art.

4.2 Evaluation on ChID-Competition

In the second set of experiments, we use **ChID-Competition**[4], which is the data for an online competition[5] on Chinese idiom comprehension. Different from ChID, for each entry in ChID-Competition, a list of passages are provided with the same candidate set, and therefore some heuristic strategies can be used (for instance, the exclusion method). The challenge is that ground truth answers will be similar in semantic meanings, and prediction models need to focus on their differences while comparing similar contexts to make the correct predictions. ChID-Competition is divided into *Train*, *Dev*, *Test* and *Out* splits for different evaluation stages.

To further test the competency of our model, we evaluate the full model **Ours-Full** on **ChID-Competition**. Considering the differences between ChID-Official and ChID-Competition, we use some heuristic methods to postprocess the predictions in order to optimize the results globally for a candidate set. Without changing the training paradigm, we treat this problem an assignment problem during postprocessing and use Linear Sum Optimization to optimize the assignment. The linear sum assignment problem is also known as minimum weight matching in bipartite graphs. The method we used is the Hungarian algorithm, also known as the Munkres or Kuhn-Munkres algorithm. Suppose for each blank, we get a probability distribution over the candidate set C. Then define a cost matrix Z where $Z_{i,j}$ represents the log probability of the i-th blank choosing c_j. Formally, let X be a boolean matrix where $X_{i,j}$ is 1 if the i-th blank chooses the candidate j. Our optimization problem can be written as

$$\min \sum_i \sum_j Z_{i,j} X_{i,j}, \tag{7}$$

so that each candidate is assigned to at most one blank, and each blank to at most one candidate.

The comparison between our method and previous methods is listed in Table 4. In the first section of the table, we list the top-ranked competitors from the competition leaderboard. It is worth noting that these systems are used for competition purposes and may not be publicly available. We then show the results using several pre-trained language models, where the results are found on the CLUE leaderboard[6]. Finally, we list our own full model **Ours-Full**, which used a larger pre-trained RoBERTa for Chinese[7]. The experiment results show that our full model achieves competitive results compared with the top ranked systems of the competition.

---

[4] https://github.com/zhengcj1/ChID-Dataset/tree/master/Competition

[5] https://biendata.com/competition/idiom/

[6] We show representative systems on the leaderboard as of the submission date of this paper. https://github.com/CLUEbenchmark/CLUE.

[7] https://github.com/brightmart/roberta_zh

4.3 Further Analysis Through Attribution Method

To better understand how our models achieve consistent improvement, we adopt the gradient based attribution method, Integrated Gradients (IG) (Sundararajan et al., 2017), to visualize how each character contributes to the final prediction. To make the visualization more readable, we first perform Chinese word segmentation to merge characters into words. The attribution value of a word is the highest absolute value of all merged characters.

We show some cases in Figure 1, where red color represents positive correlation with the prediction and blue color represents negative correlation with the prediction. For the example on the left, both "供不应求" (in great demand) and "大名鼎鼎" (famous) are positive idioms with a sense of "being abundant in", but the correct answer is "大名鼎鼎" based on the context, because the context suggests that this idiom serves as an adjective to modify a person, and only "大名鼎鼎" is used to describe a person. On the one hand, we hypothesize that **BL-IdmEmb** may have learned the correlation between "多年" (for many years) and "供不应求," and thus makes a wrong prediction solely based on this signal. On the other hand, **Ours-CP** chooses "大名鼎鼎", likely because it is consistent with the word "顾问" (consultant), which is a person, together with the conjunction word "以及" (and), suggesting that context-aware pooling may have helped the understanding of the context.

For the example on the right hand side of the figure, the two candidates "斤斤计较" (to haggle over every ounce) and "大手大脚" (extravagant) are antonyms and represent different attitudes towards spending money. Both idioms suit the context well syntactically. However, the context has the word "却" (but) and the word "价钱昂贵" (expensive), suggesting the person is extravagant with money, making "大手大脚" the correct candidate. This example shows that for more complex contextual understanding, **Ours-Full** has advantages over **Ours-CP**.
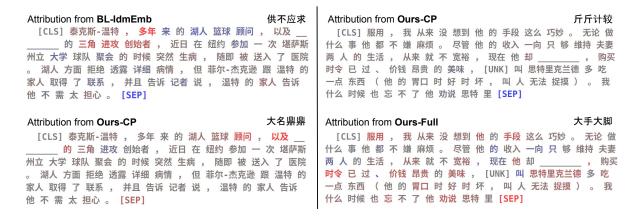

Figure 1: Example cases with attribution values of words shown in red and blue. Red indicates positive correlation with the prediction while blue indicates negative correlation with the prediction.

5 Conclusion

In this paper, we proposed a BERT-based dual embedding method to study Chinese idiom prediction. We used a dual-embedding to not only capture local context information but also match the whole context passage. Our experiments showed that our dual-embedding design can improve the performance of the base model, and both the idea of context-aware pooling and the idea of dual embedding can help improve the idiom prediction performance compared to the baseline methods on the ChID dataset.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August. Association for Computational Linguistics.

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

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

Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu, and Maosong Sun. 2016. Thuctc: An efficient chinese text classifier.

E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany, August. Association for Computational Linguistics.

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.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

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, June. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. 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, September. 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. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yuanchao Liu, Bo Pang, and Bingquan Liu. 2019b. Neural-based Chinese idiom recommendation for enhancing elegance in essay writing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5522–5526, Florence, Italy, July. Association for Computational Linguistics.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA, October. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Jianmin Shao. 2018. *General introduction to modern Chinese*. Shanghai, China: Shanghai Educational Publishing House.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *In Proceedings of TREC-8*, pages 77–82.

Lei Wang and Shiwen Yu. 2010. Construction of Chinese idiom knowledge-base and its applications. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 11–18, Beijing, China, August. Coling 2010 Organizing Committee.

S. Wang. 2019. *Chinese Multiword Expressions: Theoretical and Practical Perspectives*. Springer Singapore.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

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, July. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany, August. Association for Computational Linguistics.