Learning and Evaluating Chinese Idiom Embeddings

Minghuan Tan and **Jing Jiang**
School of Computing and Information Systems
Singapore Management University
mhtan.2017@phdcs.smu.edu.sg jingjiang@smu.edu.sg

Abstract

We study the task of learning and evaluating Chinese idiom embeddings. We first construct a new evaluation dataset that contains idiom synonyms and antonyms. Observing that existing Chinese word embedding methods may not be suitable for learning idiom embeddings, we further present a BERT-based method that directly learns embedding vectors for individual idioms. We empirically compare representative existing methods and our method. We find that our method substantially outperforms existing methods on the evaluation dataset we have constructed.

1 Introduction

Chengyu (成语, literally meaning "set phrases") are a type of idiomatic expressions in Chinese that usually consist of four Chinese characters. They are mostly derived from ancient Chinese literature and many of them are based on historical stories. The semantic meanings of Chengyu are often non-compositional and sometimes metaphoric. For example, the Chengyu 瓜田李下 literally means "melon field, beneath the plums," but its idiomatic meaning is to warn people to avoid situations where a person may be easily suspected of wrongdoing. Chengyu are commonly used in modern Chinese language, and using computational methods to understand Chengyu plays an important role in Chinese language understanding. For example, a recent work studied how to improve essay writing with recommending Chinese idioms (Liu et al., 2019), and others studied how to improve reading comprehension by correcting usage of Chinese idioms (Wang et al., 2020) and differentiating synonyms of Chinese idioms (Long et al., 2020). In this paper, we refer to Chengyu as Chinese idioms, although there are also other types of idioms in Chinese.

Recent years have witnessed the success of deep neural networks for many NLP tasks. A central idea behind deep neural networks for NLP is to use dense embedding vectors to represent language units including words, phrases and sentences, and such embeddings have been shown to be useful for many tasks such as sentiment analysis (Yu et al., 2017), question answering (Hao et al., 2017) and machine translation (Zhou et al., 2016). We therefore believe that it is also desirable to derive embedding vectors for Chinese idioms that can accurately capture their semantic meanings. However, it is not clear whether existing methods for Chinese word embeddings are effective in deriving good Chinese idiom embeddings, and there are at least two reasons for this.

First, existing Chinese word embedding evaluation datasets do not have sufficient coverage of idioms. For example, in the commonly used WordSim-240 (Wang et al., 2011) and WordSim-296 (Chen et al., 2015) datasets for Chinese word relatedness, no idiom is found. More recently, Huang et al. (2019) released a COS960 dataset with similarities of Multiword Expressions (MWEs). Although COS960 covers 150 Chinese idioms, this is still a relatively small number, and only 20 MWE pairs in COS960 consist of both idioms. For the word analogy task, another commonly used evaluation task, Chen et al. (2015) created the first Chinese dataset with 1,125 analogies, but no idiom is included. Li et al. (2018) released a large and balanced dataset CA8 for word analogy. Although CA8 has 400 entries that contain idioms, they only cover 32 unique idioms and no idiom pairs are included. With this lack of coverage of idioms in existing evaluation datasets, we cannot judge whether existing Chinese word embedding methods work well for Chinese idioms.

Second, it is reasonable to suspect that existing word embedding methods for Chinese have

Proceedings of Recent Advances in Natural Language Processing, pages 1387–1396
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_155

limitations that make them less suitable for Chinese idioms. For non-contextualized word embedding methods such as Continuous-Bag-Of-Words (CBOW) and Skip-Gram with Negative Sampling (SGNS), they treat contexts as bags of words, but given the complex meanings of Chinese idioms, learning their embeddings from bag-of-word representations of contextual words without considering the order and interactions between these contextual words may not be sufficient. Existing pre-trained non-contextualized Chinese word embeddings are also usually trained with a relatively small context window, but the semantic meaning of a Chinese idiom is often based on a larger context where the idiom appears. In fact, it has been observed that larger context windows result in more topicality (Levy and Goldberg, 2014; Bansal et al., 2014), and we suspect that for learning Chinese idiom embeddings a larger context window helps. Therefore, existing pre-trained non-contextualized Chinese word embeddings may not capture the semantic meanings of Chinese idioms well. On the other hand, recent contextualized word embedding methods such as BERT (Devlin et al., 2019) and its variants (e.g., ERNIE (Zhang et al., 2019)) consider longer contexts and use attention mechanism to model interactions between words, but since they do not focus on learning word embeddings, they do not learn a single embedding vector for each Chinese idiom. Although we can aggregate the character-level representations of the characters inside an idiom and treat the aggregated representation as the idiom embedding, since many Chinese idioms' semantics are non-compositional, this simplified approach is likely not ideal.

In this paper, we study the problem of learning and evaluating Chinese idiom embeddings. To overcome the first challenge stated above, i.e., the lack of suitable evaluation dataset for Chinese idiom embeddings, we construct an evaluation dataset that contains Chinese idiom synonyms and antonyms. We also define two evaluation metrics to measure how close the ground truth idiom synonyms are in an embedding space in order to quantify the quality of the embedding space. To overcome the second challenge stated above, i.e., the potential limitations of existing word embedding methods for Chinese idioms, we propose to adapt a method (Tan and Jiang, 2020) for Chinese idiom recommendation to learn idiom embeddings. This method learns a single embedding vector directly for each idiom and encodes the contextual information using BERT.

With the evaluation dataset we have created, we empirically compare a SGNS-based non-contextualized word embedding method for Chinese, two variants of BERT for Chinese, and our Chinese idiom embedding method. We find that based on the two metrics we have defined to measure closeness of synonyms in an embedding space, our method performs substantially better than existing methods. We also find that our method can better distinguish idiom antonyms from idiom synonyms than existing embedding methods. We also conduct further analysis to demonstrate that embedding methods that rely more on Chinese character information show advantages only when the synonyms share many common characters.

The contributions of our work are twofold: (1) We construct an evaluation dataset to facilitate the evaluation of Chinese idiom embeddings. Code and data are released on github[1]. (2) We present a BERT-based method that directly learns Chinese idiom embeddings, and we empirically compare this method with existing Chinese word embedding methods to demonstrate both the importance of learning a single embedding vector for an entire idiom and the importance of using BERT to encode the context when learning these idiom embeddings.

2 Related Work

Word Embeddings Word embedding is an important technique in NLP. It computes dense meaning representations for discrete words. It is built upon the distributional hypothesis that linguistic items with similar distributions have similar meanings. Several methods have been proposed to learn non-contextualized word embeddings efficiently, including Continuous Bag-Of-Words (CBOW), Skip-Gram with Negative Sampling (SGNS) and GloVe (Pennington et al., 2014). In this paper, we use an SGNS-based Chinese word embedding method as a representative non-contextualized word embeding method for evaluation. Contextualized word embeddings such as ELMO, GPT and BERT have been developed in recent years and shown their high effectiveness for many NLP tasks. In this paper, we use two representative BERT variants, BERT-wwm and ERNIE, to evaluate Chinese idiom embeddings derived from pre-trained Chi-

[1] https://github.com/VisualJoyce/ChengyuBERT

nese BERT models.

Evaluation of Chinese Word Embeddings For word embeddings, existing evaluation methods can be categorized into intrinsic and extrinsic methods (Schnabel et al., 2015). Commonly used intrinsic methods include word similarity and word analogy, while extrinsic methods rely on downstream NLP tasks (Pennington et al., 2014). In this paper, we use an intrinsic method to evaluate Chinese idiom embeddings.

Several benchmark datasets for evaluating Chinese word embeddings have been released (Wang et al., 2011; Finkelstein et al., 2001; Jin and Wu, 2012; Chen et al., 2015; Guo et al., 2014; Huang et al., 2019; Li et al., 2018). But as we pointed out in Section 1, existing datasets have low coverage of Chinese idioms.

Neural Network Models for Chinese Idiom Understanding Despite the importance of Chengyu in Chinese language understanding, there have been only a few pieces of work on Chengyu using neural models (Jiang et al., 2018; Liu et al., 2019; Zheng et al., 2019). Chinese Chengyu Recommendation (CCR) has been addressed in recent years (Liu et al., 2019; Jiang et al., 2018; Zheng et al., 2019). In this paper, we adapt a method for CCR (Tan and Jiang, 2020) to learn Chinese idiom embeddings.

3 Construction of the Evaluation Dataset

A standard intrinsic task for evaluating word embeddings is word similarity (Bakarov, 2018; Wang et al., 2019). For Chinese idioms, a natural choice of idiom pairs that are semantically similar are synonyms or near-synonyms[2]. Although previously Wang et al. (2013) constructed a Chinese idiom knowledge base that contains idiom synonyms, this knowledge base is not publicly available. On the other hand, there exist online resources containing synonyms and near-synonyms of Chinese idioms. We choose two websites, kxue.com (快学网)[3] and Baidu Baike (百度百科)[4], as the sources from which to crawl idiom synonyms and near-synonyms. We also collect idiom antonyms from these two websites because an antonym of an idiom is often topically related to that idiom and therefore may be also close to that idiom in an embedding space. However, we expect a good idiom embedding method to be able to separate antonyms from synonyms.

Idiom Vocabulary: According to Wang et al. (2013), there are in total around 38K Chinese idioms, among which around 3.5K are commonly used. In order to obtain a vocabulary of Chinese idioms with high coverage, we merge the idioms found in the following four resources: (1) Chengyu Daquan[5], (2) Xinhua Chengyu Dictionary[6], (3) Chengyu Cloze Test[7], and (4) ChID. [8]. This gives us a Chinese idiom vocabulary with 33,237 idioms.

ChIdSyn: As we have pointed out earlier, we believe idiom synonyms can help us evaluate idiom embeddings. To construct a large dataset of Chinese idiom synonyms, we crawled synonyms from two websites: (1) Kxue.com is an online Chinese thesaurus. It has a dedicated page where Chinese idiom synonyms are listed. Each entry in this list consists of a key and a value, where the key is a Chinese idiom and the value is one or more other Chinese idioms that are near-synonyms of the key. We crawled all the entries from this idiom synonym page on kxue.come[9]. Baidu Baike is an online encyclopedia in Chinese. For each idiom, there is a section called 成语辨析 (Chengyu Differentiation) that lists its synonyms and antonyms.[10] We crawled the synonyms of those idioms in our vocabulary that can be found on Baidu Baike. In total, we obtained around 30k entries of Chinese synonyms. We then removed those idioms in the data that are not in our idiom vocabulary as described earlier. In the end we obtained a total of around 21K entries in our synonym dataset, where each entry consists of a *query idiom* and a set of other idioms that are the query idiom's synonyms or near-synonyms.

We observe that a significant portion of the synonyms share common characters with the query idioms. For example, 山盟海誓 (oath of eternal love) and 海誓山盟 are treated as near-synonyms

[2]We use near-synonyms to refer to idioms that do not have exactly the same meaning but their meanings are highly similar. It is not common to have Chinese idioms that are complete synonyms, except for those that are variants of the same basic form.

[3]http://chengyu.kxue.com/

[4]https://baike.baidu.com/

[5]www.guoxue.com/chengyu/CYML.htm

[6]github.com/pwxcoo/chinese-xinhua

[7]github.com/bazingagin/chengyu_data

[8]https://github.com/zhengcj1/ChID-Dataset

[9]We crawled the data from http://chengyu.kxue.com/list/jinyici.html before October 19, 2020.

[10]For example, for the idiom "一马平川", see https://baike.baidu.com/item/一马平川.

in our dataset, but these two idioms contain exactly the same set of Chinese characters. In fact, they are variants of the same basic form. Another example is 挨家挨户 (door to door) and 挨门挨户, which share three common characters. In general, it is not uncommon for Chinese idioms to have such variants due to historical reasons such as misuse (including literary malapropism). Although these are valid near-synonyms, we suspect that they may affect the evaluation of idiom embeddings. This is because those idiom embeddings that rely more on character-level information are likely to gain advantages when evaluated on these near-synonym pairs sharing common characters. For example, if an idiom embedding is obtained by averaging the character embeddings of its component characters, then it is very easy for this type of idiom embeddings to recognize that 山盟海誓 and 海誓山盟 are near-synonyms (because they would have the same average character embedding), but we would not be able to know whether such embeddings truly capture the semantic meanings. We also suspect that for those idioms that have near-synonyms sharing common characters, their semantic meanings are more likely to be compositional and thus less idiomatic. For example, for the idiom 挨家挨户, the character 挨 means "in sequence" and both 家 and 户 mean "household." The meaning of the idiom, which is "door to door," can be directly inferred from the meanings of the characters. Therefore, when the character 家 (household) is replaced with the character 门 (door), the meaning of the idiom remains the same.

Consequently, we move those synonyms that share at least two common characters with the query idioms into a separate dataset, which we will not use as the main evaluation dataset. The remaining synonyms always have no more than one common character with their query idioms. We refer to this cleaned synonym dataset as *ChIdSyn*, and the separate dataset containing synonyms sharing two or more common characters is referred to as *ChIdSyn-com*. We will use *ChIdSyn-com* for additional analysis in our experiments. Statistics of *ChIdSyn* and *ChIdSyn-com* can be found in Table 1.

ChIdAnt: From the same two websites, we have also collected around 10K entries in an antonym dataset which we refer to as *ChIdAnt*. Similarly, each entry in this dataset consists of a query idiom and its antonyms. Although antonyms are idioms having opposite meanings, they are often topically closely related. For example, the idiom 饱学之士 means "a scholarly man," and its antonym 胸无点墨 means "uneducated." We can see that their meanings are topically closely related. We therefore suspect that they are still close in an embedding space, but ideally a good idiom embedding method should be able to distinguish the synonyms of a query idiom from its antonyms. Table 1 gives some statistics of *ChIdAnt*.

	Before Filtering		After Filtering	
	#Idioms	#Entries	#Idioms	#Entries
Crawled	33,524	30,354	21,745	20,753
ChIdSyn	11,387	8,897	8,125	6,822
ChIdSyn-com	28,622	24,147	18,498	15,836
ChIdAnt	11,263	9,733	7,939	7,316

Table 1: Statistics of the crawled datasets. *Crawled* refers to synonyms and near-synonyms. We list antonyms separately in the last line of the table.

4 Learning Chinese Idiom Embeddings

Existing Chinese word embedding methods can be used to derive idiom embeddings. However, as we have discussed in Section 1, they may not be ideal for learning Chinese idiom embeddings. In this section, we first briefly review existing Chinese word embedding methods and how we use them to obtain idiom embeddings. We then present a method to learn Chinese idiom embeddings based on BERT. Our proposed method is adapted from a method for Chinese idiom recommendation (Tan and Jiang, 2020).

4.1 Non-contextualized Word Embeddings

Continuous Bag-Of-Words (CBOW) and Skip-Gram with Negative-Sampling (SGNS) (Mikolov et al., 2013) are two most commonly used efficient log-linear prediction models for learning non-contextualized word embeddings. CBOW tries to predict a word based on its context, where the context is represented as the average word embeddings within the contextual window. In contrast, SGNS tries to predict the contextual words of a given word, and negative sampling is used to reduce the computational cost.

Both CBOW and SGNS have been used to learn Chinese word embeddings (Chen et al., 2015; Li et al., 2018). Chinese is an ideographic language with no explicit word delimiter between words (Li and Yuan, 1998). Chinese segmentation tools are

therefore used to identify word boundaries when learning Chinese word embeddings. On the other hand, Chinese words consist of characters, which have their own semantic meanings. Therefore character information has been incorporated to improve Chinese word embeddings (Chen et al., 2015). In addition, inspired by N-gram SGNS for English (Zhao et al., 2017; Bojanowski et al., 2017), which predicts contextual N-grams rather than contextual words, Li et al. (2018) trained Chinese word embeddings using N-gram SGNS and found that both N-gram and character features bring significant and consistent improvement.

However, Chinese idioms are not always treated as words by Chinese segmentation tools. They are sometimes separated into multiple words. Therefore, only a subset of the idioms in our idiom vocabulary can be found as words in existing pre-trained non-contextualized Chinese word embeddings, and we are only able to perform evaluation on this subset of idioms.

4.2 BERT and Its Variants

Recently, contextualized word embeddings have shown to be highly effective for many NLP tasks. BERT (Devlin et al., 2019) is probably the most commonly used contextualized word embedding model. The original BERT model is pre-trained using the Masked Language Model (MLM) task and the Next Sentence Prediction (NSP) task. Since the original BERT was proposed, there have been some variants of it proposed, including BERT with whole word masking (BERT-wwm) (Cui et al., 2019) and ERNIE (Zhang et al., 2019) that incorporates a multi-stage knowledge masking strategy which adds word-level masking, phrase-level masking and entity-level masking.

The original Chinese-BERT starts from embeddings of individual Chinese characters at the bottom layer. When BERT-wwm or ERNIE is applied to Chinese, although words are identified and masked using Chinese segmentation tools, the model still does not learn embedding vectors directly for entire words. Therefore, to obtain an embedding for an idiom, we need to aggregate the component characters' embeddings. In this paper, we take the vector representations of individual characters at the top layer of BERT, and average these character representations as the embedding for the entire idiom.[11]

[11] We have also experimented with another setting where we use the [CLS] token's representation at the top layer as the idiom representation. We found this to perform worse than using average character embedding.

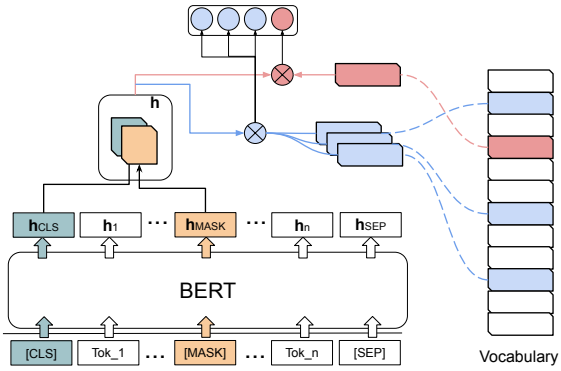

Figure 1: Model structure for BERT with SGNS. The red flow shows the path for the target idiom while the light blue flows show paths for negative sampled idioms used for the learning.

4.3 Learning Idiom Embeddings with BERT

As we have pointed out earlier, existing non-contextualized Chinese word embedding methods model contextual words in a bag-of-word manner, which is suboptimal for encoding the contextual information. Chinese-BERT and its variants can better encode the contextual information using the Transformer architecture, but they do not learn a single embedding vector for an entire Chinese idiom, and therefore they are not ideal either because idioms often have non-compositional semantics. We propose to combine BERT contextual encoding with single embedding vectors for Chinese idioms.

Specifically, to train idiom embeddings, we perform the task of idiom prediction based on its context. Given an idiom v appearing in a context window $c = (w_{-k}, \ldots, w_{-2}, w_{-1}, \texttt{[MASK]}, w_1, w_2, \ldots, w_k)$, where w_i are the contextual words and [MASK] replaces the idiom v in the original text, the task aims to predict v based on c. To do so, our idea is to assume that v has an embedding vector $\mathbf{e}_v$ to be learned. We then use BERT to derive a hidden representation $\mathbf{h}$ that represents c and use $\mathbf{h}$ and $\mathbf{e}_v$ to derive a log-linear score to indicate how likely v fits into the context c.

Note that the task described above is similar to the prediction task used by CBOW, but instead of simply using the average word embedding to represent the context c, our method uses BERT to encode c. The task described above is also similar to the Masked Language Model task of BERT, but we mask and predict whole idioms rather than

individual characters.

Concretely, to use BERT to encode the sequence c, following standard practice, we prepend the token `[CLS]` to the beginning of c and append `[SEP]` to the end of c. We also include position embeddings. For segment embeddings, we treat the sequence c as a single segment. Let $\mathbf{h}_{\texttt{CLS}} \in \mathbb{R}^d$ denote the hidden vector produced by the last layer of BERT representing `[CLS]`, and $\mathbf{h}_{\texttt{MASK}} \in \mathbb{R}^d$ the similarly produced hidden vector representing `[MASK]`. We then define the following vector $\mathbf{h}$ to combine $\mathbf{h}_{\texttt{CLS}}$ and $\mathbf{h}_{\texttt{MASK}}$ into a single vector representation because both are important for representing the context c, $\mathbf{h} = \mathbf{W}[\mathbf{h}_{\texttt{CLS}}; \mathbf{h}_{\texttt{MASK}}; \mathbf{h}_{\texttt{CLS}} \odot \mathbf{h}_{\texttt{MASK}}; \mathbf{h}_{\texttt{CLS}} - \mathbf{h}_{\texttt{MASK}}]$, where $\odot$ is element-wise multiplication between two vectors and $\mathbf{W} \in \mathbb{R}^{d \times 4d}$ is a matrix to be learned.

We then use a standard log-linear model based on the dot product between $\mathbf{h}$ and $\mathbf{e}_v$ to train our model. To use the hidden representation $\mathbf{h}$ of the context to predict the idiom v, we take its idiom embedding $\mathbf{e}_v$, apply Layer Normalization (Ba et al., 2016) LN on it. We also adopt negative sampling to select negative Chengyu. The learning objective is defined as

$$-(\log \sigma(LN(\mathbf{e}_v)^\intercal \mathbf{h}) + \sum_{v' \in \mathcal{N}_v} \log \sigma(-LN(\mathbf{e}_{v'})^\intercal \mathbf{h})),$$

where $\mathcal{N}_v$ contains a fixed number of negative samples for each Chinese idiom, and $\sigma(\cdot)$ is the sigmoid function. Besides the transformation $\mathbf{W}$ and LN, during the training process, the BERT layers will be finetuned and the whole vocabulary will be learned from random initialization. The model structure is illustrated in Figure 1.

5 Experiments

5.1 Experiment Setup

Evaluation metrics: Recall that our main evaluation dataset is the *ChIdSyn* dataset that contains entries of query idioms and their near-synonyms, where these near-synonyms share at most one common character with the query idiom. We design two evaluation metrics to measure whether near-synonyms in *ChIdSyn* are close to each other in an embedding space. (1) ***Recall@K***: Given a query idiom v_n, we rank all idioms based on their idiom embeddings' cosine or Euclidean distances with the query idiom's embedding. Let $\mathcal{R}_{v_n}^{(K)}$ represent the top-K ranked idioms. Let $\mathcal{S}_{v_n}$ denote the set of ground truth near-synonyms of v_n. *Recall@K* is defined as

$$Recall@K = \frac{1}{N} \sum_{n=1}^{N} \frac{|\mathcal{S}_{v_n} \cap \mathcal{R}_{v_n}^{(K)}|}{|\mathcal{S}_{v_n}|},$$

where N is the total number of query idioms in *ChIdSyn*. (2) ***Coherence@K***: However, it is not guaranteed that all near-synonyms of a query idiom v are identified in the online resources we crawled, i.e., some of the top-K ranked idioms may be indeed near-synonyms but are not found in the ground truth near-synonym set. To overcome this limitation, we can measure whether a query idiom and its ground truth near-synonyms share many common "similar" idioms. In this way, even if a real near-synonym u of idiom v is missed from the ground truth, if u is found to be similar to both v and its ground truth near-synonyms, it will contribute positively to the metric. We therefore define the following metric, which we call *Coherence@K*:

$$Coherence@K = \frac{1}{N} \sum_{n=1}^{N} \frac{|\cap_{u \in \mathcal{S}'_{v_n}} \mathcal{R}_u^{(K)}|}{|\cup_{u \in \mathcal{S}'_{v_n}} \mathcal{R}_u^{(K)}|},$$

where v_n is a query idiom, N is the total number of query idioms, $\mathcal{S}'_{v_n} = \{v_n\} \cup \mathcal{S}_{v_n}$ (i.e., v_n together with its ground truth near-synonyms), and $\mathcal{R}_u^{(K)}$ is the top-K similar idioms to u, where similarity can be based on either cosine or Euclidean distance.

Methods to be compared: We empirically compare the following embedding methods: (1) **SGNS** and its variants: We use Chinese word embeddings released by Li et al. (2018), which are trained using the Skip-Gram with Negative Sampling method. There are a few variations of these embeddings. **SGNS+B** uses bigram prediction, **SGNS+C** incorporates character information, and **SGNS+B+C** uses both bigram prediction and character information. Li et al. (2018) also experimented with different genres of text for training. In this paper, we use their pre-trained word embeddings trained on the literature genre because this provides fair comparison with our method, which is also trained on Chinese text in the literature genre. (2) **BERT-wwm**: This refers to averaging the top-layer character representations after using the pre-trained Chinese-BERT-wwm (Cui et al., 2019) to process an idiom. (3) **ERNIE**: This refers to averaging the top-layer character representations after using Chinese ERNIE (Zhang et al., 2019) to process an

	Recall@K								Coherence@K					
	Cosine				Euclidean				Cosine			Euclidean		
	1	3	5	10	1	3	5	10	3	5	10	3	5	10
SGNS	0.054	0.102	0.132	0.178	0.031	0.056	0.071	0.092	0.038	0.043	0.045	0.027	0.031	0.036
SGNS+C	0.030	0.084	0.127	0.198	0.009	0.022	0.030	0.048	0.032	0.038	0.043	0.023	0.029	0.038
SGNS+B	0.067	0.127	0.159	0.210	0.043	0.080	0.101	0.131	0.047	0.051	0.053	0.034	0.038	0.042
SGNS+B+C	0.051	0.128	0.184	0.271	0.017	0.046	0.063	0.089	0.043	0.055	0.059	0.030	0.041	0.047
BERT-wwm	0.031	0.084	0.117	0.170	0.030	0.078	0.111	0.163	0.028	0.034	0.037	0.026	0.030	0.034
ERNIE	0.037	0.109	0.161	0.238	0.036	0.110	0.163	0.244	0.038	0.048	0.058	0.037	0.049	0.060
Ours-16	0.145	0.282	0.357	0.451	0.142	0.275	0.348	0.433	0.107	0.113	0.113	0.105	0.109	0.110
Ours-32	**0.164**	**0.327**	**0.411**	**0.519**	**0.163**	**0.322**	**0.404**	**0.503**	**0.126**	**0.137**	**0.142**	**0.123**	**0.136**	**0.139**

Table 2: *Recall@K* and *Coherence@K* on *ChIdSyn*, where ranking is based on either cosine or Euclidean distance.

idiom. (4) **Ours-16**: This is our method where we set the context window size to be 16 characters. (5) **Ours-32**: This is also our method with a larger context window of 32 characters.

Training data: We collect online ebooks from the literature domain with a size comparable to that of the training corpus used by Li et al. (2018). We extract sentences from our crawled corpus and keep only those sentences containing idioms. Since the average word length for Chinese is around 1.6 characters, we use a window size of 8 characters on each side, i.e., 16 characters in total, which is comparable to the SGNS method that used a window size of 5 words on each side. To test how context length may affect the results, we also train our model using a larger window size of 16 characters on each side, i.e, 32 characters in total. The two versions of our model are named **Ours-16** and **Ours-32**, respectively. To ensure fair comparison, we use only the subset of the entries from *ChIdSyn* where we have idiom embeddings from all methods. This results in a subset of 3,716 entries from *ChIdSyn* for our experiments, which is still a relatively large number. Similarly, for some further analysis we do using *ChIdSyn-com*, we also use only a subset of the data, which contains 2,342 entries. A subset of *ChIdSyn-Ant* with 3940 entries is also used for further analysis.

5.2 Main Results

We first present the results of all the methods we compare using the metrics *Recall@K* and *Coherence@K* on *ChIdSyn*, see Table 2. We can draw the following major conclusions from the table: (1) If we compare **Ours-16** with the **SGNS** methods, we can see that Ours-16 clearly outperforms these SGNS methods. Recall that we use a similar context window size as the SGNS methods. The main difference of Ours-16 from the SGNS methods is that we use Chinese-BERT to encode the context whereas the SGNS methods do not model the interactions between the contextual words. This implies that when learning Chinese idiom embeddings, it is important to model the order of and interactions between the contextual words. (2) Comparing **Ours-16** with **BERT-wwm** and **ERNIE**, we can see that Ours-16 also substantially outperforms these two BERT-based methods. Recall that the main difference of our method and these BERT methods is that we directly learn a single idiom embedding vector whereas for these BERT methods we need to aggregate character embeddings to derive idiom embeddings. The results suggest that many Chinese idioms' semantic meanings cannot be simply derived from their character embeddings and therefore it is important to associate a Chinese idiom with a single embedding vector and to learn this embedding vector from the contexts of this idiom. (3) **Ours-32** performs clearly better than **Ours-16**. This suggests that a larger context window is very useful for learning Chinese idiom embeddings, which have not been found to be the case for word embeddings (Lison and Kutuzov, 2017).

Besides the major conclusions drawn above, we can also see from the two tables that: (1) For the SGNS methods, adding character information may actually either hurt the performance or improve the performance very little. In other words, there is no consistent observation that character information helps for Chinese idiom embeddings, which is not the case for Chinese word embeddings (Chen et al., 2015; Zhao et al., 2017). This verifies our hypothesis that existing conclusions drawn from evaluating Chinese word embeddings may not apply to idiom embeddings. (2) For the two BERT-based methods, we can see that ERNIE performs clearly better than BERT-wwm. It is worth noticing that ERNIE uses Baidu Baike in which most idioms have entries and

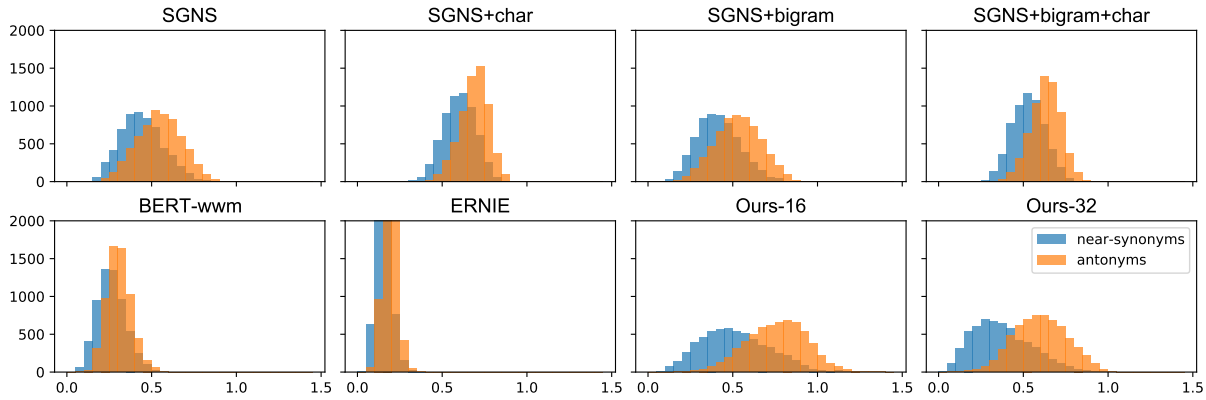

Figure 2: Cosine distance distribution of near-synonym and antonym pairs.

would be treated as entities by the entity-level mask. Intuitively, the embeddings extracted using ERNIE should be better than BERT-WWM, whose CWS tools may not be able to recognize all the idioms.

5.3 Further Analysis

In this section, we conduct some further comparison and analysis using *ChIdSyn-com* and *ChIdAnt*.

Synonyms with Common Characters: Recall that we identified a set of near-synonyms that share two or more common characters. We suspect that these idiom synonyms are easier to be identified if the idiom embeddings rely more on character-level information. To verify this hypothesis, we compare the various methods using *Recall@K* based on cosine distance on *ChIdSyn-com*. The results are shown in Table 3. We can see that indeed those existing methods that rely more on character-level information, namely, SGNS+C, SGNS+B+C, BERT-wwm and ERNIE generally perform better than the other methods, including our methods. This verifies our hypothesis above. Note that because the synonyms in *ChIdSyn-com* share many common characters, being able to identify them does not imply that the embeddings truly capture the semantic meanings of the idioms. Since SGNS+C, SGNS+B+C, BERT-wwm and ERNIE actually do not perform well on *ChIdSyn*, we argue that they are effective only for synonyms sharing many common characters, and this implies that they rely on superficial patterns to encode idioms.

Antonyms: Recall that earlier we raised the hypothesis that good idiom embedding methods should be able to distinguish antonyms from synonyms, although both can be topically related to the query idioms. In fact, a previous study by Samenko et al. (2020) also found that embeddings contain information that distinguishes synonyms and antonyms. Inspired by them, we think that the separability of near-synonyms and antonyms may reflect the quality of the learned embeddings. We therefore visualize the distributions of cosine distances (i.e, 1 minus cosine similarity) of idiom near-synonym pairs and antonym pairs in Figure 2, using *ChIdSyn* and *ChIdAnt*. We can see from the figure that our methods **Ours-16** and **Ours-32** clearly has a distinguishable cosine distance distribution for antonyms compared with synonyms, whereas for the other methods the two distributions are less distinguishable. This again demonstrates the advantage of our idiom embedding methods.

K	1	3	5	10
SGNS	0.130	0.223	0.270	0.334
SGNS+B	0.175	0.287	0.341	0.404
SGNS+C	0.518	0.775	**0.857**	**0.924**
SGNS+B+C	**0.526**	**0.776**	0.846	0.908
BERT-wwm	0.467	0.662	0.714	0.786
ERNIE	0.531	0.760	0.825	0.880
Ours-16	0.380	0.555	0.612	0.675
Ours-32	0.449	0.655	0.722	0.786

Table 3: *Recall@K* on *ChIdSyn-com*.

6 Conclusion

In this paper, we constructed a new evaluation dataset that contains Chinese idiom synonyms and antonyms to facilitate the evaluation of Chinese idiom embeddings. We presented a method that learns Chinese idiom embeddings by predicting idioms based on BERT-encoded contexts. We also propose two metrics to measure closeness of synonyms in the embedding space. Our method performs substantially better than existing methods.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *CoRR*, abs/1801.09536.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, Maryland. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1236–1242. AAAI Press.

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. *CoRR*, abs/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. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 406–414, New York, NY, USA. Association for Computing Machinery.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231, Vancouver, Canada. Association for Computational Linguistics.

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. COS960: A chinese word similarity dataset of 960 word pairs. *CoRR*, abs/1906.00247.

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.

Peng Jin and Yunfang Wu. 2012. SemEval-2012 task 4: Evaluating Chinese word similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 374–377, Montréal, Canada. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Haizhou Li and Baosheng Yuan. 1998. Chinese word segmentation. In *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation*, pages 212–217, Singapore. Chinese and Oriental Languages Information Processing Society.

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, Melbourne, Australia. Association for Computational Linguistics.

Pierre Lison and Andrey Kutuzov. 2017. Redefining context windows for word embedding models: An experimental study. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 284–288, Gothenburg, Sweden. Association for Computational Linguistics.

Yuanchao Liu, Bo Pang, and Bingquan Liu. 2019. 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. Association for Computational Linguistics.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym knowledge enhanced

reader for Chinese idiom reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3684–3695, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. 2020. Synonyms and antonyms: Embedded conflict. *CoRR*, abs/2004.12835.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2020. A BERT-based dual embedding model for Chinese idiom prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1312–1322, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19.

Lei Wang, Shiwen Yu, Xuefeng Zhu, and Yun Li. 2013. Chinese idiom knowledge base for chinese information processing. In *Proceedings of the 13th Chinese Conference on Chinese Lexical Semantics*, CLSW'12, pages 302–310, Berlin, Heidelberg. Springer-Verlag.

Xiang Wang, Yan Jia, Bin Zhou, Zhao-Yun Ding, and Zheng Liang. 2011. Computing semantic relatedness using chinese wikipedia links and taxonomy. *Journal of Chinese Computer Systems*, 32(11):2237–2242.

Xinyu Wang, Hongsheng Zhao, Tan Yang, and Hongbo Wang. 2020. Correcting the misuse: A method for the Chinese idiom cloze test. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):671–681.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhe Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du. 2017. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 244–253, Copenhagen, Denmark. 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.

Guangyou Zhou, Zhiwen Xie, Tingting He, Jun Zhao, and Xiaohua Tony Hu. 2016. Learning the multilingual translation representations for question retrieval in community question answering via non-negative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1305–1314.