A Generalized Idiom Usage Recognition Model Based on Semantic Compatibility

Changsheng Liu, Rebecca Hwa
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260, USA
{changsheng, hwa}@cs.pitt.edu

Abstract

Many idiomatic expressions can be used figuratively or literally depending on the context. A particular challenge of automatic idiom usage recognition is that idioms, by their very nature, are idiosyncratic in their usages; therefore, most previous work on idiom usage recognition mainly adopted a "per idiom" classifier approach, i.e., a classifier needs to be trained separately for each idiomatic expression of interest, often with the aid of annotated training examples. This paper presents a transferred learning approach for developing a *generalized model* to recognize whether an idiom is used figuratively or literally. Our work is based on the observation that most idioms, when taken literally, would be somehow semantically at odds with their context. Therefore, a quantified notion of **semantic compatibility** may help to discern the intended usage for any arbitrary idiom. We propose a novel *semantic compatibility model* by adapting the training of a Continuous Bag-of-Words (CBOW) model for the purpose of idiom usage recognition. There is no need to annotate idiom usage examples for training. We perform evaluative experiments on two corpora; results show that the proposed generalized model achieves competitive results compared to state-of-the-art per-idiom models.

1 Introduction

Idioms appear frequently in languages. Many idioms can be interpreted figuratively or literally depending on the context (Fazly, Cook, and Stevenson 2009). For example, the idioms "play with fire" and "get wind" are used differently in the instances below:

*#1 [*lit.*]Kids* ***playing with fire****: experts warn parents to look out for danger signs.*
*#2[*fig.*]The UN is* ***playing with fire*** *over North Korea crisis.*
*#3[*lit.*]Here in Portland we're just gonna get rain, the coast is gonna* ***get wind****. Stay safe!*
*#4[*fig.*]FAA will* ***get wind*** *of that crooked airways' shady dealings.*

The ability to automatically distinguish whether a potential idiomatic phrase is used literally or figuratively is beneficial to many natural language processing (NLP) applications such as machine translation and sentiment analysis (Salton, Ross, and Kelleher 2014; Williams et al. 2015). A particular challenge for automatic idiom usage recognition is that idioms, by their very nature, are idiosyncratic in their usages. For example, the proposition "of" following "get wind" often indicates the idiom is used figuratively (as in instance #4), while for idiom "play with fire", one might need more complicated linguistic clues to infer its usage, such as a violation of selectional preference (as in instance # 2). Therefore, the majority of previous work on idiom usage detection adopted a "per idiom" classifier approach; i.e., a classifier needs to be trained separately for each idiomatic expression of interest, often with the aid of annotated training examples (Rajani, Salinas, and Mooney 2014; Peng, Feldman, and Vylomova 2014; Liu and Hwa 2017). Since there are a large number of idioms in the text, we aim to build an efficient generalized idiom usage recognizer without supervision.

The insight underlying the method we propose is that when the literal interpretation of a potential idiomatic expression is not compatible with the context, it typically indicates that the idiom is used figuratively. For instance, in example #4 above, the word "wind" is semantically far away from most surrounding words; the literal sense of "get wind" does not fit well with the context. In general, this *semantic incompatibility* is a strong indicator that the idiom has a non-literal interpretation in the context. This paper presents a method for building a general idiom usage recognizer by determining the semantic compatibility between the literal meanings of idioms and their contexts.

This notion of semantic compatibility is reminiscent of the training objective of negative sampling in word2vec, which is originally used for learning low dimensional word embeddings (Mikolov et al. 2013b; 2013a). Its Continuous Bag-of-Words (CBOW) variant internally tries to maximize the probability of positive (compatible) context-word pairs and minimize the probability of randomly sampled negative (incompatible) pairs. Thus, if CBOW can successfully capture the semantic compatibility feature in text, it is also possible that we can apply it to determine the semantic compatibility between an idiom and its context.

However, the CBOW model mainly uses semantic compatibility as a roundabout way to learn useful vectors for words. The post-hoc evaluations of the model concentrate on the learned embeddings of words (Mikolov et al. 2013a; Levy, Goldberg, and Dagan 2015); whether the learned

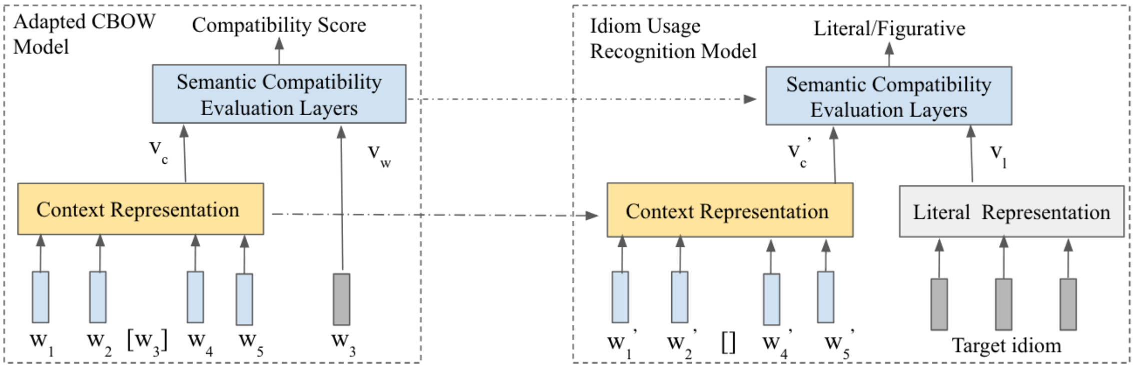
Adapted CBOW Model
Compatibility Score
Semantic Compatibility Evaluation Layers
v_c
v_w
Context Representation
w_1 w_2 $[w_3]$ w_4 w_5
w_3
Idiom Usage Recognition Model
Literal/Figurative
Semantic Compatibility Evaluation Layers
v_c'
v_l
Context Representation
Literal Representation
w_1' w_2' [] w_4' w_5'
Target idiom

Figure 1: The overview of our idiom usage recognition model in a transfer learning fashion: the CBOW is adapted for semantic compatibility measurement which can be trained on raw large corpus; the learned representations and parameters are then used for idiom usage recognition. [] indicates target word or idiom.

model can be directly applied to measure semantic compatibility is understudied. In this work, we analyze the potential limitations of the standard CBOW model in terms of semantic compatibility measurement (see Section 3.1). We further propose a novel semantic compatibility model by adapting the standard CBOW in two ways. First, we introduce several alternatives for context representation. We exploit bidirectional LSTM (Graves, Jaitly, and Mohamed 2013) to model the sequential information in context and two self-attention mechanisms (Vaswani et al. 2017) to capture the critical context words. Second, we add a multilayer perceptron layer to relax CBOW's constraint on contextual similarity and tailor it for capturing semantic compatibility.

The overview of our method is shown in Figure 1. Here, the semantic compatibility model is used in a **transfer learning fashion**: (1) the model is first trained on large raw text corpora (such as Wikipedia) with the aim of predicting the semantic compatibility between context and *a single word*; (2) the learned model can then be applied to determine an idiom's intended usage by measuring the semantic compatibility between the idiom's literal sense and the context. Since most idioms are multi-word expressions, we treat each idiom as a single semantic unit and build a literal representation for it; this enables a seamless reuse of the semantic compatibility model for usage recognition. The advantages of our model are: (1) there is no need for annotated idiom usage examples since the core component of our usage recognition model (i.e., the semantic compatibility model) is trained on raw text corpora; (2) the model is general; i.e., it can be applied to different idioms without further parameter tuning. We conduct experiments on two benchmark idiom corpora; results suggest that the proposed generalized model achieves competitive results compared to state-of-the-art per-idiom models.

2 Continuous Bag-of-Words Revisited

The CBOW training procedure using negative sampling is presented in (Mikolov et al. 2013a). It defines two sets of embeddings: the "official" word embeddings and a second set of context embeddings for each word in the vocabulary. The embeddings in the two sets are K-dimensional vectors which are tuned iteratively by scanning huge amounts of texts by a sliding window. The model internally tries to predict a target word using context words in the window based on a heavily trimmed neural network. For each observed pair of context and target word, the model samples several "negative" words which are not compatible with the context. The training objective is to maximize the probability of positive (compatible) context-word pairs and minimize the probability of negative (incompatible) pairs generated from a known noise distribution.

Specifically, the loss function used in CBOW is:

$$\log \sigma(v_w^{'} v_c) + \sum_{W_j \in W_{neg}} \log \sigma(-v_{w_j}^{'} v_c) \qquad (1)$$

where v_c is the context embedding, v_w and v_{w_j} are the word embeddings of positive and negative target words respectively. Since the sliding window usually contains more than one words, v_c is represented as the average of context embeddings of words within the window. The sigmoid function $\sigma(v_w^{'} v_c)$ can be considered as a semantic compatibility measurement; the model will update the context embeddings and word embeddings iteratively so as to assign high score to positive (compatible) pairs and lower score the negative (incompatible) pairs.

3 A Generalized Idiom Usage Recognition Model

We want to develop a generalized model for idiom usage recognition based on semantic compatibility. In this section, we first analyze the potential limitations of CBOW for semantic compatibility measurement. Then we present how we adapt the CBOW for semantic compatibility. Finally, we describe how we exploit the adapted model for idiom usage recognition.

3.1 Limitations of CBOW for Semantic Compatibility

CBOW uses semantic compatibility as an auxiliary task to learn useful vectors for words to capture their similarity in a hidden semantic space. An interesting question is whether the learned context embeddings and the word embeddings, together with the sigmoid function, can be directly applied as a measurement of semantic compatibility. Although it seems plausible at first glance, we argue that there are three potential limitations of CBOW that impede its use as a semantic compatibility measurement.

1) A lack of sequential information To represent the context, CBOW simply uses the average of all the context embeddings, thus the order information is not preserved.

2) Not all words are equal In CBOW, all words contribute equally to the context representation. This limitation might not significantly impact the quality of the learned word embeddings, but could be problematic for semantic compatibility. In many cases, a few key context words are critical clues to determine the semantic compatibility between the context and a word.

3) A paradox of transitivity In CBOW, the **direct** dot product between context representation and target word embedding is used to model their semantic compatibility. However, a dot product operation is not appropriate for encoding semantic compatibility relation; the dot product aims to capture a similarity relation ($\approx$) between two embeddings, which could lead to a paradox of transitivity in the case of semantic compatibility because a word often appears in very different contexts. For example, in *John Lennon wrote a [song] called "Working Class Hero"* and *I like to listen to the same [song] on repeat*, the semantics of the two contexts of "song" are very different. Let C_1 and C_2 denote embeddings of two different contexts, i.e., $C_1 \not\approx C_2$. A target word T could be compatible with both $C1$ and $C2$ (as shown in the above example). If we use the direct dot product to model their compatibility, we can get $T \approx C_1$ and $T \approx C_2$ in the embedding space since T is compatible with both C_1 and C_2. Based on the transitive property of similarity relation, $C_1 \approx C_2$ can be inferred, which contradicts with the premise $C_1 \not\approx C_2$.

3.2 Adapting CBOW for Semantic Compatibility

We have discussed the potential limitations of CBOW for semantic compatibility. The first two limitations are related to context representations, while the third limitation is about the dot product operation. We propose to adapt the CBOW model to better capture semantic compatibility relation. In terms of context representation, we additionally use a special bidirectional Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber 1997) to encode sequence information. Meanwhile, we exploit self-attention mechanism (Lin et al. 2017; Vaswani et al. 2017; Li et al. 2016) to give more weight to important words when encoding context. Finally, instead of the simple dot product, a semantic evaluation layer is used to overcome the aforementioned paradox of transitivity.

Context Representation In standard CBOW, the context representation is the average of the embeddings of context words (denoted as ACE). Apart from ACE, we also exploit bidirectional LSTM for context representation, which has been shown to be effective for modelling sequential data (Graves, Jaitly, and Mohamed 2013; Melamud, Goldberger, and Dagan 2016; Peters et al. 2018). The overview of our architecture is illustrated in Fig. 2

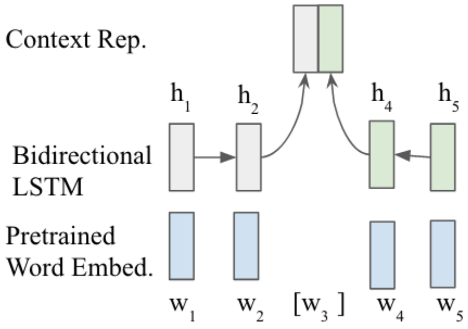

Figure 2: Bidirectional LSTM for context representation

Our architecture is not the same as standard Bidirectional LSTM (Graves, Jaitly, and Mohamed 2013). In our model, the two LSTMs gravitate toward the target words: a forward LSTM will generate a hidden representation for each word before the target word and a reversed LSTM will generate a hidden representation for each word following the target word; we do not feed the LSTMs with the target word itself. Let h be the hidden representation of word w (i.e., the output of the LSTMs), the context representation of the target word at position i is the concatenation of the hidden representations of the two neighboring words, i.e.,

$$c_i = [h_{i-1}; h_{i+1}] \tag{2}$$

Attention Layer In both ACE and the LSTM based context representation, we do not explicitly consider the importance of words. In this paper, we exploit attention mechanism to enable our model to automatically identify those important words for semantic compatibility.

Attention mechanisms have generally been used to allow for an alignment of the input and output sequences, e.g. the source and target sentence in machine translation (Bahdanau, Cho, and Bengio 2014), or for an alignment between two input sentences as in question answering (Santos et al. 2016; Xiong, Zhong, and Socher 2016). In our work, we apply the idea of attention to a rather different kind of scenario, in which we only have the raw input sentence. We propose two self-attention (or intra-attention) models: global attention and local attention. The first one uses a vector to capture all the words that are important globally. As semantic compatibility usually involves the local interaction between words, our second attention model captures those words that have strong semantic relation with the other words in the context.

Global Attention Figure 3 illustrates the global attention architecture when using bidirectional LSTM for context encoding. Assume v is the attention vector. The attention layer will generate an importance score g_i for each word w_i based on the dot product between v and its hidden representation h_i:

$$g_i = v \cdot h_i + b \tag{3}$$

Here the attention vector v is a parameter to be learned in the training process, which can be considered as a global variable trying to "memorize" those critical words in a sentence based on the current context. The importance score is then normalized using softmax:

$$a_i = \frac{\mathrm{e}^{g_i}}{\sum_{p=1}^{n} \mathrm{e}^{g_p}}. \tag{4}$$

The attention-based context representation is a weighted sum of hidden states of LSTMs:

$$v_c = \sum_{i=1}^{n} h_i a_i. \tag{5}$$

Note that this global attention models can also be applied to the ACE for context representation. The only difference is the input to the attention layer: we only need to replace h_i in Equation 3 and 5 with the word embedding w_i.

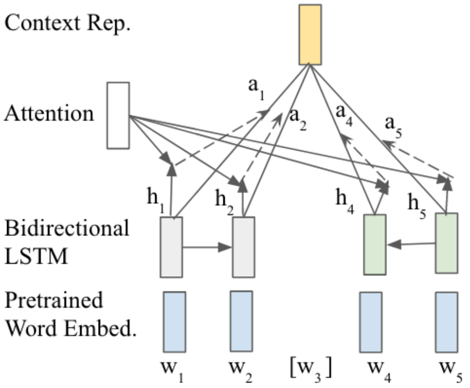

Figure 3: The global attention architecture when using bidirectional LSTM for sequential encoding

Local Attention While global attention is useful, we argue that it might not fully capture the semantic compatibility information in a sentence. A word that is important for semantic compatibility globally or in other sentences might not be important for the target sentence. Semantic compatibility usually involves the interactions among words within a sentence. We introduce a diagonal relevance matrix A with values $A_{i,j} = f(w_i, w_j)$ to characterize the strength of semantic interaction between words w_i and w_j. The scoring function f is computed as the inner product between the embeddings of w_i and w_j. If a word has a strong semantic relation with another word, it is highly possible that this word is important. So we apply a max operation over the row of A (excluding the value in the diagonal because it is the relevance score between a word and itself) to select the largest value as the importance score for each word; i.e.,

$$l_i = \max_j A_{i,j} \tag{6}$$

Following the global attention, a softmax layer is applied to normalize the raw score l_i; the final context representation is a weighted sum of hidden states of LSTMs. The overview of local attention is illustrated in Figure 4. Similarly, when applying local attention to ACE, the final context representation is a weighted sum of word embeddings.

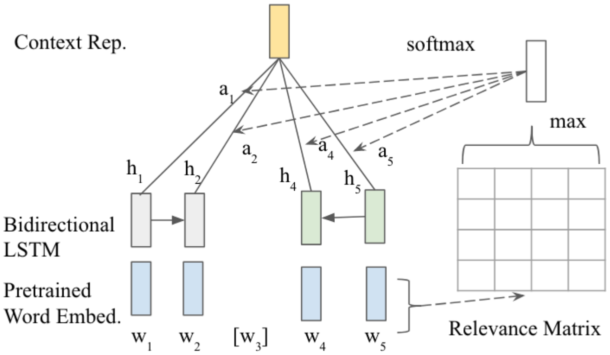

Figure 4: The local attention architecture when using bidirectional LSTM for sequential encoding

Semantic Compatibility Evaluation Layer To quantify the semantic compatibility between a context and a target word, standard CBOW uses the direct dot product between context embedding and target word embedding as the metric. We have argued that the direct dot product operation may lead to a paradox of transitivity. To address this limitation, we feed the context representation into a multilayer network of perceptrons with a ReLu nonlinearity activation function:

$$\mathrm{L}(v_c) = f_2(relu(f_1(v_c))) \tag{7}$$

where f_1 and f_2 denote fully connected layer. Then we use the following formula to measure the semantic compatibility between a context and a word:

$$\sigma(v_l^{'} L(v_c)) \tag{8}$$

Recall the main reason of paradox of transitivity is that a word can appear in very different contexts; the direct dot product between word embedding and context representation would, however, force these different contexts being similar to each other. This paradox is avoided by the multilayer perceptron network L since it relaxes the contextual similarity constraints, i.e., it can map the context representations that are different originally to similar embeddings which are close to the target word. We refer the whole mapping and measuring schema as the semantic compatibility evaluation layer.

Training We train our adapted CBOW on the Wikipedia corpus [1] using negative sampling. The loss function is:

$$\log \sigma(v_w^{'} L(v_c)) + \sum_{W_j \in W_{neg}} \log \sigma(-v_{w_j}^{'} L(v_c)) \tag{9}$$

The model is trained end-to-end using the Adam optimizer (Kingma and Ba 2015). Standard CBOW scans the whole corpus using a sliding window of a fixed size. Alternatively, we train the model sentence by sentence because using all the context words in a sentence can yield more precise context representation, which is essential for semantic compatibility.

[1] https://dumps.wikimedia.org

3.3 Idiom Usage Recognition based on Semantic Compatibility

We have described how we adapt the standard CBOW for semantic compatibility measurement and train it on a large corpus. Given a context representation and a word embedding, the learned model is expected to tell us whether they are compatible. However, we want to measure the semantic compatibility between a context and an idiom, which is usually a multi-word expression. To reuse the learned model, we first build a representation of the literal sense of the idiom. Then we use the semantic compatibility layer to evaluate whether the literal representation is compatible with the context.

Literal Representation of Idiom We experiment with the following two representations of the literal sense of idiom:

AWE, the average of the embeddings of words forming the idiom. The intuition is that the literal sense of idiom is compositional.

AKWE, the average of the embeddings of key words in the idiom. This representation might lose partial information of the literal interpretation of idiom, but we hypothesize it could benefit our task. The intuition is that one or two words in idiom will be the crucial clue that indicates whether a figurative or literal sense was intended. Consider the figurative example of "get wind" in the Introduction section, the word "wind" does not fit well with the context and this incompatibility serves as a strong signal of the intended usage, while the word "get" provides less information. In this paper, for verb-noun combinations, we only choose the noun as the keyword; for noun-noun combination, we choose both nouns as the keywords; for the other types of idiom, the non-stop words are selected as the keywords.

Usage Classification Given a context representation v_c and the literal representation of idiom v_l, we calculate their compatibility score using the following formula:

$$\sigma(v_l^{'} L(v_c) + b_u) \tag{10}$$

where b_u is a bias term, which is tuned based on a development dataset. If the score is larger than 0.5, the instance will be classified as literal usage. Otherwise, it will be labeled as figurative usage.

4 Evaluation

We conduct experiments to address the following questions:

1. How effective is our overall approach? How does it compare against previous work?

2. How effective is the standard CBOW for idiom usage recognition?

3. Does our model effectively address the limitations of CBOW?

4.1 Experimental Setup

Baselines We compare our models with three unsupervised models: Sporleder and Li (2009), Li and Sporleder (2009)[2] and Fazly, Cook, and Stevenson (2009). For supervised models, we compare our models with Rajani, Salinas, and Mooney (2014) and Liu and Hwa (2017) (using 5-fold cross validation). All these models are per-idiom models except the one presented in (Sporleder and Li 2009).

Our models We experiment with two base context representations: ACE and bidirectional LSTM, over which we additionally propose two attention models: local and global attention. Therefore we have four variants for context representations. In terms of the representation of literal sense of idiom, we experiment with AWE and AKWE. So our full models have eight variants.

Parameter setting To train the adapted CBOW, we follow the standard training procedure in word2vec using negative sampling. To increase the training speed, we uniformly sampled a set of sentences from Wikipedia to build a corpus of 100M tokens. We find using a corpus of this size is sufficient to train a reliable model so we do not use the full corpus. All tokens with a frequency of less than 50 are trimmed. The hyperparameters are summarized in Table 1.

When applying the adapted CBOW model to idiom usage recognition, we need to set the bias term b_u in Equation 10 with value in a reasonable range. We picked 10 idioms that are different from the evaluation set, collected 50 instances from the web for each idiom, and labeled them ourselves. We find that b_u in the range of $[0.06, 0.15]$ yield good results.

Parameter	Value
word embedding size	200
context embedding size	200
LSTM hidden size	200
f_1 input/output size	200/400
f_2 input/output size	400/200
negative samples	15
epoch	10
batch size	500
learning rate	0.001

Table 1: Hyperparameters of our network.

Evaluative Data We compare all the methods using two publicly available corpora of idiomatic usages: SemEval 2013 Task 5B corpus (Korkontzelos et al. 2013) and Verb-Noun Combination (VNC) dataset (Cook, Fazly, and Stevenson 2008). Some idioms from the VNC dataset have very few figurative (or literal) instances; this presents a problem for supervised baselines. To facilitate full comparisons, we select the subset of idioms from the VNC corpus whose number of literal and figurative instances are both higher than 10.

4.2 Experimental Result

Table 2 provides a detailed comparison of our models with previous approaches. We can observe that ACE+LocalAtt+AKWE gets an F-score of 0.76 (accuracy of 0.75) on SemEval corpus and 0.75 (accuracy of 0.73) on VNC corpus, which outperforms the per-idiom models from Rajani, Salinas, and Mooney (2014), Li and

[2] Due to the query frequency restriction on the API of Normalized Google Distance (NGD), we replace NGD with word embeddings to measure the semantic relatedness among words.

		SemEval		VNC	
Type	Model	Avg. F_{fig}	Avg.Acc	Avg. F_{fig}	Avg.Acc
Per-Idiom	Rajani et al., 2014	0.71*	0.75	0.69*	0.7
	Li and Sporleder, 2009	0.64*	0.62*	0.67*	0.66*
	Fazly et al., 2009	-	-	0.73	0.74
	Liu and Hwa, 2017	0.77	0.77	0.75	0.75
Generalized	Sporleder & Li	0.58*	0.52*	0.61*	0.57*
Our Model	ACE + GlobalAtt + AWE	0.72	0.69	0.71	0.7
	ACE + GlobalAtt + AKWE	0.74	0.7	0.73	0.7
	ACE + LocalAtt + AWE	0.74	0.73	**0.76**	**0.73**
	ACE + LocalAtt + AKWE	**0.76**	**0.75**	0.75	**0.73**
	Bidirectional LSTM + GlobalAtt + AWE	0.68	0.68	0.67	0.67
	Bidirectional LSTM + GlobalAtt + AKWE	0.72	0.72	0.69	0.7
	Bidirectional LSTM + LocalAtt + AWE	0.69	0.68	0.7	0.69
	Bidirectional LSTM + LocalAtt + AKWE	0.73	0.72	0.72	0.71

Table 2: The performances of different models. Avg. F_{fig} denotes average figurative F-score, Avg.Acc denotes average accuracy. * indicates the difference is significant with our model ACE+LocalAtt+AKWE at the 95% confidence level. Since the method from Fazly, Cook, and Stevenson (2009) restricted their experiment to VNC type, we only report their performance on the VNC corpus.

Sporleder (2009) and the generalized model from Sporleder and Li (2009). Moreover, the model is competitive to the supervised per-idiom model from Liu and Hwa (2017), which is state-of-the-art in this task.

4.3 Detailed Analysis

Using Standard CBOW for Idiom Usage Recognition In this study, we experiment with using standard CBOW for idiom usage recognition, in which ACE is used as the context representation and the direct dot product between context representation and target word representation is used as a measurement of semantic compatibility. The training and evaluation procedures are the same as those used for our full models.

Model	Avg. F_{fig}	Avg.Acc
CBOW+AWE	0.63	0.62
CBOW+AKWE	0.65	0.63

Table 3: The results of CBOW for idiom usage recognition. Results are averaged across all the idioms in the two corpora.

Table 3 shows the performance of CBOW for idiom usage recognition, which is significantly worse than our adapted models. Arguably, CBOW is insufficient to capture the semantic compatibility information in text. To illustrate this point, we compare the CBOW and our adapted model (we use the bidirectional LSTM + Local Attention for context representation) to select the most compatible words based on a given context. We find the results of CBOW remains to be of wildly varying quality. Considering the example "can you see the [] i try to make?", the top 10 most compatible words to fill in the bracket predicted by the two models are shown in Table 4.

As we can see, CBOW has a fairly poor semantic compatibility measurement; all the words tend to make little sense in the context. In contrast, the adapted model has much better results. Since our idiom usage recognition heavily relies on the underlying model's ability of measuring semantic compatibility, this could potentially explain why CBOW has a worse performance in the downstream task.

CBOW	Adapted CBOW
please	stuff
want	positives
you	ripples
hear	ones
how	things
try	changes
sure	figures
wish	pictures
know	dilema
do	negatives

Table 4: Top 10 most compatible words in "can you see the [] i try to make?"

To better understand the effectiveness of sequential information, the attention mechanism and the semantic compatibility layer, we did an ablation study. The results are shown in Table 5. Since AKEW tends to outperform AWE as shown in Table 2, we only experimented with AKEW as the literal representation of idiom.

Sequential Information We find that the importance of sequential information is closely related to the attention model. In Table 2, we observe that our full non-sequential models (ACE variants) generally outperform the sequential models (Bidirectional LSTM variants). Without attention, however, sequential information can significantly boost the performance of our model; Bidirectional LSTM + AKEW achieves F-score of 0.7 while ACE + AKEW only gets 0.66 as shown in Table 5. Intuitively, with the aid of attention, our model can identify those critical words, which enhances the expressiveness of context representation by simple weighted

Model	Avg. F_{fig}	Avg.Acc
ACE+GlobalAtt+AKEW	0.74	0.7
- w/o Semantic Layer	0.66	0.64
ACE+LocalAtt+AKEW	0.76	0.74
- w/o Semantic Layer	0.67	0.66
- w/o attention	0.66	0.67
Bidirectional LSTM+GlobalAtt+AKEW	0.71	0.71
- w/o Semantic Layer	0.65	0.64
Bidirectional LSTM+LocalAtt+AKEW	0.73	0.72
- w/o Semantic Layer	0.66	0.66
- w/o attention	0.7	0.69

Table 5: The results of ablation study. Results are averaged across all the idioms in the two corpora.

averaging.

Attention In Table 5, we observe that the removal of the attention layer can result in a performance drop for both the ACE and Bidirectional LSTM variants. This shows the effectiveness of our attention model in terms of context representations. Moreover, global attention is not as competitive as local attention. For example, the Bidirectional LSTM+LocalAtt+AKEW model achieves an averaged F-score of 0.73 on the two corpora while the Bidirectional LSTM+GlobalAtt+AKEW model gets 0.71. This observation aligns with our intuition that semantic compatibility usually involves local interactions among words within the sentence. In Figure 5 we visualize the attention layer using the first example in the Introduction section. The global attention tends to assign higher weights to non-stop words such as "kids", "experts" and "sign", while the local attention tends to assign higher weights to words with strong semantic relation, such as "warn" and "danger".

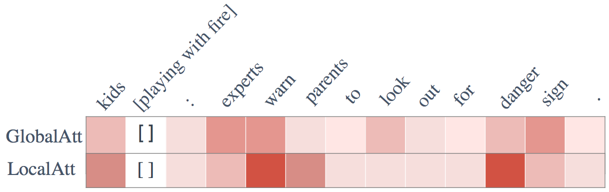

Figure 5: Visualization of attention layer

The semantic compatibility layer We have argued that a direct dot product between context representation and target word embedding could lead to a paradox of transitivity. To address this problem, we add a multilayer perceptron network over the context representation so as to map different contexts to embeddings that are close to the target word.

In Table 5, we observe that the performances of our models decrease significantly without the semantic compatibility layer. Among all the full models, the ACE+LocalAtt+AKEW has the most severe performance drop: from 0.76 to 0.67 in terms of F-score and 0.74 to 0.66 in terms of accuracy. This suggests that the semantic compatibility layer is essential to our model.

5 Related Work

5.1 Idiom Usage Recognition

Most previous work on idiom usage detection adopted a "per idiom" classifier approach (Birke and Sarkar 2006; Fazly, Cook, and Stevenson 2009; Li and Sporleder 2009; Peng, Feldman, and Vylomova 2014; Liu and Hwa 2017). For example, Fazly, Cook, and Stevenson(2009) hypothesized that idiomatic usages of an expression tend to occur in a small number of canonical form(s). For each idiom, they proposed a probabilistic method to automatically extract the canonical forms from large corpus for idiom usage recognition. Rajani, Salinas, and Mooney (2014) extracted all non-stop words as features and used them to train a L2 regularized Logistic Regression (L2LR) classifier (Fan et al. 2008). While previous works generally ignored the linguistic properties of idiomatic expression and their interaction with context representations, Liu and Hwa (2017) presented an adaptive method that applies supervised ensemble learning to select representations for different idioms.

Sporleder and Li (2009) proposed a generalized method by building a cohesion graph to include all content words in the context; if removing the idiom improves cohesion, they assumed the instance is figurative. Continuing on this work, Li and Sporleder (2009) used their cohesion graph method to label a subset of the test data with high confidence. This subset is then used as the training data for a downstream supervised classifier based on a set of linguistic features.

5.2 Attention

In recent years, there has been a growing research interest in the attention mechanism. Instead of using all available information, the attention mechanism aims at softly selecting the most important information in the learning process. It has been successfully applied to tasks such as machine translation (Bahdanau, Cho, and Bengio 2014), sentence summarization (Rush, Chopra, and Weston 2015) , question answering (Santos et al. 2016) and image captioning (Xu et al. 2015). In these models, attentions have been typically used for alignment between two sources of information. Cheng, Dong, and Lapata (2016) introduced a self-attention (or intra-attention) to induce relations among tokens in a single sequence. Vaswani et al. (2017) showed that self-attention could also be applied directly on raw word embeddings for machine translation. The other applications of self-attention include question answering (Li et al. 2016) and sentiment analysis(Lin et al. 2017).

6 Conclusion

This paper presents a generalized model to recognize whether an idiom is used figuratively or literally based on the idea of semantic compatibility. We analyze the limitations of CBOW in terms of semantic compatibility measurement and propose a novel semantic compatibility model by adapting the training of CBOW for the purpose of idiom usage recognition. Experiments on two benchmark idiom usage corpora show that the proposed generalized model achieves competitive results compared to state-of-the-art per-idiom models.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Birke, J., and Sarkar, A. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

Cheng, J.; Dong, L.; and Lapata, M. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Cook, P.; Fazly, A.; and Stevenson, S. 2008. The vnc-tokens dataset.

Fan, R. E.; Chang, K. W.; Hsieh, C. J.; Wang, X. R.; and Lin, C. J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9:1871–1874.

Fazly, A.; Cook, P.; and Stevenson, S. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1):61–103.

Graves, A.; Jaitly, N.; and Mohamed, A.-r. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 273–278. IEEE.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Korkontzelos, I.; Zesch, T.; Zanzotto, F. M.; and Biemann, C. 2013. Semeval-2013 task 5: Evaluating phrasal semantics.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Li, L., and Sporleder, C. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 315–323. Association for Computational Linguistics.

Li, P.; Li, W.; He, Z.; Wang, X.; Cao, Y.; Zhou, J.; and Xu, W. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *ICLR*.

Liu, C., and Hwa, R. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of The 31st AAAI Conference on Artificial Intelligence*.

Melamud, O.; Goldberger, J.; and Dagan, I. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 51–61.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Peng, J.; Feldman, A.; and Vylomova, E. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *EMNLP* 2019–2027.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 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)*, volume 1, 2227–2237.

Rajani, N. F.; Salinas, E.; and Mooney, R. 2014. Using abstract context to detect figurative language.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 379–389.

Salton, G.; Ross, R.; and Kelleher, J. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese.

Santos, C. d.; Tan, M.; Xiang, B.; and Zhou, B. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Sporleder, C., and Li, L. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 754–762. Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Williams, L.; Bannister, C.; Arribas-Ayllon, M.; Preece, A.; and Spasić, I. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications* 42(21):7375–7385.

Xiong, C.; Zhong, V.; and Socher, R. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.