Verb Metaphor Detection via Contextual Relation Learning

Wei Song[1]*, **Shuhui Zhou**[1]*, **Ruiji Fu**[2,3], **Ting Liu**[4], **Lizhen Liu**[1]
[1]College of Information Engineering and Academy for Multidisciplinary Studies,
Capital Normal University, Beijing, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
[4]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
{wsong, shzhou, liz_liu7480}@cnu.edu.cn,
rjfu@iflytek.com, tliu@ir.hit.edu.cn

Abstract

Correct natural language understanding requires computers to distinguish the literal and metaphorical senses of a word. Recent neural models achieve progress on verb metaphor detection by viewing it as sequence labeling. In this paper, we argue that it is appropriate to view this task as relation classification between a verb and its various contexts. We propose the Metaphor-relation BERT (MrBERT) model, which explicitly models the relation between a verb and its grammatical, sentential and semantic contexts. We evaluate our method on the VUA, MOH-X and TroFi datasets. Our method gets competitive results compared with state-of-the-art approaches.

1 Introduction

Metaphor is ubiquitous in our daily life for effective communication (Lakoff and Johnson, 1980). Metaphor processing has become an active research topic in natural language processing due to its importance in understanding implied meanings.

This task is challenging, requiring contextual semantic representation and reasoning. Various contexts and linguistic representation techniques have been explored in previous work.

Early methods focused on analyzing restricted forms of linguistic context, such as subject-verb-object type grammatical relations, based on hand-crafted features (Shutova and Teufel, 2010b; Tsvetkov et al., 2013; Gutiérrez et al., 2016). Later, word embeddings and neural networks were introduced to alleviate the burden of feature engineering for relation-level metaphor detections (Rei et al., 2017; Mao et al., 2018). However, although grammatical relations provide the most direct clues, other contexts in running text are mostly ignored.

Recently, token-level neural metaphor detection draws more attention. Several approaches discovered that wider context can lead to better performance. Do Dinh and Gurevych (2016) considered a fixed window surrounding each target token as context. Gao et al. (2018) and Mao et al. (2018) argued that the full sentential context can provide strong clues for more accurate prediction. Some recent work also attempted to design models motivated by metaphor theories (Mao et al., 2019; Choi et al., 2021).

Despite the progress of exploiting sentential context, there are still issues to be addressed. First of all, a word's local context, its sentential context and other contexts should be all important for detecting metaphors; however, they are not well combined in previous work. More importantly, as shown in Figure 1, most token-level metaphor detection methods formulate metaphor detection as either a single-word classification or a sequence labeling problem (Gao et al., 2018). The context information is mainly used for learning contextual representations of tokens, rather than modeling the interactions between the target word and its contexts (Zayed et al., 2020).

In this paper, we focus on token-level verb metaphor detection, since verb metaphors are of the most frequent type of metaphoric expressions (Shutova and Teufel, 2010a). As shown in Figure 1, we propose to formulate verb metaphor detection as a relation extraction problem, instead of token classification or sequence labeling formulations. In analogy to identify the relations between entities, our method models the relations between a target verb and its various contexts, and determines the verb's metaphoricity based on the relation representation rather than only the verb's (contextual) representation.

We present a simple yet effective model — Metaphor-relation BERT (MrBERT), which is adapted from a BERT (Devlin et al., 2019) based state-of-the-art relation learning model (Bal-

*These authors contributed equally to this work.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4240–4251
August 1–6, 2021.

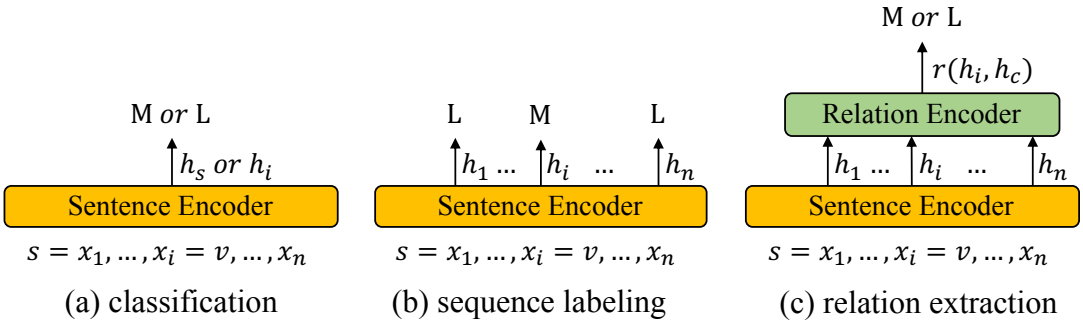

Figure 1: Formulations of verb metaphor detection: (a) a single word classification model; (b) a sequence labeling model; (c) **the proposed relation extraction model**, where h_s, h_i, h_c and $r(h_i, h_c)$ represent the representations of a sentence, a token, the context and the relation between the target verb v and its context components.

dini Soares et al., 2019). Our model has three highlights, as illustrated in Figure 2. First, we explicitly extract and represent context components, such as a verb's arguments as the local context, the whole sentence as the global context, and its basic meaning as a distant context. So multiple contexts can be modeled interactively and integrated together. Second, MrBERT enables modeling the metaphorical relation between a verb and its context components, and uses the relation representation for determining the metaphoricity of the verb. Third, the model is flexible to incorporate sophisticated relation modeling methods and new types of contexts.

We conduct experiments on the largest metaphor detection corpus VU Amsterdam Metaphor Corpus (VUA) (Steen, 2010). Our method obtains competitive results on the large VUA dataset. Detail analysis demonstrates the benefits of integrating various types of contexts for relation classification. The results on relatively small datasets, such as MOH-X and TroFi, also show good performance and model transferability.

2 Formulating Verb Metaphor Detection

This section briefly summarizes the common formulations of token-level verb metaphor detection as a background, and discusses the relation between this paper and previous work.

The task A given sentence contains a sequence of n tokens $\mathbf{x} = x_1, ..., x_n$, and a target verb in this sentence is x_i. Verb metaphor detection is to judge whether x_i has a literal or a metaphorical sense.

Basic formulations Most neural networks based approaches cast the task as a classification or sequence labeling problem (Do Dinh and Gurevych, 2016; Gao et al., 2018). As shown in Figure 1, the classification paradigm predicts a single binary label to indicate the metaphoricity of the target verb, while the sequence labeling paradigm predicts a sequence of binary labels to all tokens in a sentence.

Based on the basic formulations, various approaches have tried to enhance feature representations by using globally trained contextual word embeddings (Gao et al., 2018) or incorporating wider context with powerful encoders such as BiLSTM (Gao et al., 2018; Mao et al., 2019) and Transformers (Dankers et al., 2019; Su et al., 2020).

Limitations and recent trends However, the above two paradigms have some limitations.

First, contextual information is mostly used to enhance the representation of the target word, but the interactions between the target word and its contexts are not explicitly modeled (Zayed et al., 2020; Su et al., 2020). To alleviate this, Su et al. (2020) proposed a new paradigm by viewing metaphor detection as a reading comprehension problem, which uses the target word as a query and captures its interactions with the sentence and clause. A concurrent work to this work (Choi et al., 2021) adopted a pre-trained contextualized model based late interaction mechanism to compare the basic meaning and the contextual meaning of a word.

Second, exploiting wider context will bring in more noise and may lose the focus. Fully depending on data-driven models to discover useful contexts is difficult, given the scale of available datasets for metaphor detection is still limited. The grammar structures, such as verb arguments, are important for metaphor processing (Wilks, 1978), but is not well incorporated into neural models. Stowe et al. (2019) showed that data augmentation based on syntactic patterns can enhance a standard model. Le et al. (2020) adopted graph convolutional networks to incorporate dependency graphs, but did

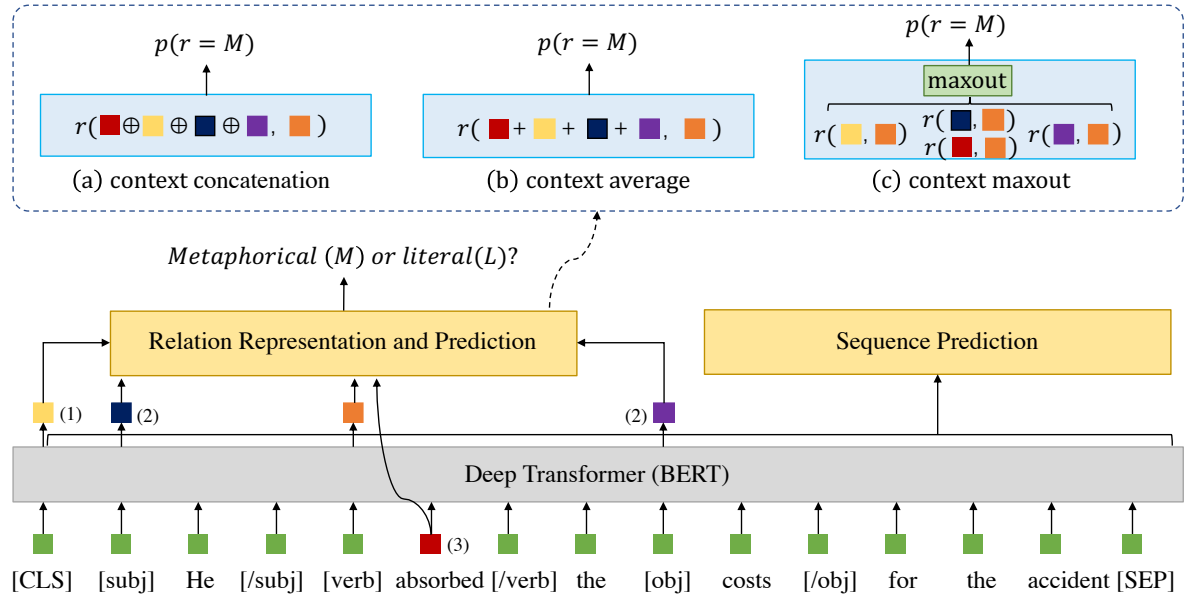

Figure 2: An example shows MrBERT's main architecture. MrBERT considers the representations of (1) the sentential global context, (2) the grammatical local context, and (3) the basic meaning of the verb as a distant context. Three context integration strategies for modeling contextual relations are adopted: (a) context concatenation, (b) context average, and (c) context maxout. Contextual relation r is modeled to indicate the probability of being metaphorical, where *linear, bilinear* and *neural tensor models* can be applied to capture interactions between the verb and its contexts. The *relation-level* and *sequence-level* predictions are jointly optimized.

not consider specific grammatical relations. It is interesting to further explore how to integrate explicit linguistic structures for contextual modeling.

This paper presents a new paradigm for verb metaphor detection to overcome these limitations, by viewing the task as a relation extraction task. We assume a target verb and its multiple contexts are entities, and metaphor detection is to determine whether a metaphorical relation holds between the verb and its contexts.

We will introduce the proposed model in Section 3. Before diving into details, we argue that viewing metaphor as a relation is reasonable and consistent with existing metaphor theories. According to Wilks (1978), metaphors show a violation of selectional preferences in a given context. The conceptual metaphor theory views metaphors as transferring knowledge from a familiar, or concrete domain to an unfamiliar, or more abstract domain (Lakoff and Johnson, 1980; Turney et al., 2011). The metaphor identification procedure (MIP) theory (Group, 2007) aims to identify metaphorically used words in discourse based on comparing their use in particular context and their basic meanings. All the theories care about a kind of relations between a target word and its contexts, which may help identify metaphors.

3 Metaphor-Relation BERT (MrBERT)

We propose the Metaphor-relation BERT (MrBERT) model to realize verb metaphor detection as a relation classification task.

Figure 2 shows the architecture of MrBERT. We use the pre-trained language model BERT as the backbone model. There are three main procedures: (1) extract and represent contexts; (2) model the contextual relations between the target verb and its contexts; (3) manipulate the contextual relations for predicting the verb's metaphoricity.

3.1 Contexts and their Representations

3.1.1 Types of Contexts

A metaphor can result when a target word interacts with a certain part in a sentence. Previous work often explored individual context types, such as verb arguments through grammatical relations or the whole sentence/clause. Little work has attempted to summarize and combine different contexts.

We summarize the following contexts, which would help determine verbs' metaphoricity:

- **Global context**: We view the whole sentence as the global context. A metaphorically used word may seem divergent to the meaning or topic of the sentence.

- **Local context**: We view the words that have a close grammatical relation to the target words as the local context, which is widely studied to capture selectional preference violations.

- **Distant context**: Motivated by the MIP theory, the difference between the contextual usage of a word and its basic meaning may indicate a metaphor so that we view the basic meaning of the target verb as a distant context.

Then, we have to extract and represent these contexts.

3.1.2 Context Extraction and Representation

We call the target verb's contexts as *context components*. To get the contextual or basic meanings of these components. we use the deep transformer models, such as BERT.

We first use Stanford dependency parser (Chen and Manning, 2014) to parse each sentence and extract verb-subject and verb-direct object relations with VB head and NN dependent. The nominal subjects and objects are used as the local context components.

Motivated by (Baldini Soares et al., 2019), we introduce 6 component marker tokens, $[subj]$, $[/subj]$, $[verb]$, $[/verb]$, $[obj]$ and $[/obj]$, to explicitly label the boundaries of the target verb, its subject and object in each sentence. We also use $[CLS]$ and $[SEP]$ to mark the whole sentence. For example, the marker inserted token sequence for the sentence *He absorbed the costs for the accident* is shown in Figure 2. The whole token sequence is fed into BERT's tokenizer, and then the transformer layers.

To get the contextual representations, we use the hidden states of the final transformer layer. For each marked component, we use the start marker (e.g., $[subj]$) or the averaged embedding between the start and the end markers (e.g., $[subj]$ and $[/subj]$) as the component representation.

The contextual representation of the whole sentence is read from the final hidden state of $[CLS]$.

To represent the basic meaning of the verb, we use the output from the BERT tokenizer to get the context independent verb representation. If word pieces exist, their averaged embedding is used.

3.2 Modeling the Contextual Relation

The relation between the target verb and one of its contexts is called a *contextual relation*. Our purpose is to utilize the contextual relation(s) to determine the metaphoricity of the verb.

The representations of the verb and a context component are denoted as $v \in \mathbb{R}^d$ and $c \in \mathbb{R}^k$, respectively. We adopt three ways to explicitly define the form of the relation r for capturing the interactions between v and c.

- **Linear model** We use a parameter vector $V_r \in \mathbb{R}^{d+k}$ and a bias b_r to represent the relation r, and the probability of the relation being metaphorical is computed according to

$$p(r|v,c) = \sigma(V_r^\top \begin{pmatrix} v \\ c \end{pmatrix} + b_r), \quad (1)$$

  where σ is the `sigmoid` function.

- **Bilinear model** We use a parameter matrix $A_r \in \mathbb{R}^{d \times k}$ and a bias b_r to represent the relation r:

$$p(r|v,c) = \sigma(v^\top A_r c + b_r). \quad (2)$$

  The components and the relation can interact more sufficiently with each other in this way.

- **Neural tensor model** We also exploit a simplified neural tensor model for relation representation:

$$p(r|v,c) = \sigma(v^\top A_r c + V_r^\top \begin{pmatrix} v \\ c \end{pmatrix} + b_r). \quad (3)$$

3.3 Integrating Contextual Relations for Prediction

We focus on 3 types of contextual relations:

- **Verb-global relation** The relation between the contextual representations of the verb v and the whole sentence c_{CLS}.

- **Verb-local relation** The relation between the contextual representations of the verb v and its subject c_{subj} or object c_{obj}.

- **Verb-distant relation** The relation between the verb v and its basic meaning v_{bsc}.

The representations of $c_{subj}, c_{obj}, c_{CLS}$ and v_{bsc} can be obtained as described in Section 3.1.2. We try three ways to integrate the contextual relations. The first two ways build a combined context c first:

- **Context concatenation** We can concatenate the representations of context components together as the combined context, i.e., $c = c_{subj} \oplus c_{obj} \oplus c_{CLS} \oplus v_{bsc}$.

- **Context average** Similarly, we can use the averaged representation of all context components as the combined context, i.e., $c = \texttt{average}(c_{subj}, c_{obj}, c_{CLS}, v_{bsc})$.

Then we compute the probability that the relation is metaphorical, i.e., $p(r|v, c)$, where either linear, bilinear or neutral tensor model can be applied.

The other way is to choose the most confident single prediction, i.e.,

- **Context maxout** The prediction is based on $\texttt{max}\{p(r|v, c)\}$, where c belongs to $\{c_{CLS}, c_{subj}, c_{obj}, v_{bsc}\}$.

To train the relation-level prediction model, we use binary cross-entropy as the loss function,

$$\mathcal{L}_0 = -\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i y_i + (1-\hat{y}_i)(1-y_i)), \quad (4)$$

where N is the number of training samples; $\hat{y}_i$ is the golden label of a verb with $\hat{y}_i = 1$ indicating a metaphorical usage and $\hat{y}_i = 0$ indicating a literal usage; y_i is the probability of being metaphorical predicted by our model.

We further combine relation-level and sequence-level metaphor detection via multi-task learning. The sequence metaphor detection uses the hidden states of the final layer and a `softmax` layer for predicting the metaphoricity of each token. We use cross-entropy as the loss function and denote the average loss over tokens in training samples as $\mathcal{L}_1$. The final loss of MrBERT is $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$.

4 Evaluation

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics

VUA dataset We mainly conduct experiments on the VUA (Steen, 2010) dataset. It is the largest publicly available metaphor detection dataset and has been used in metaphor detection shared tasks (Leong et al., 2018, 2020). This dataset has a training set and a test set. Previous work utilized the training set in different ways (Neidlein et al., 2020). We use the preprocessed version of the VUA dataset provided by Gao et al. (2018). The first reason is that this dataset has a fixed development set so that different methods can adopt the same model selection strategy. The second reason is that several recent important methods used the same dataset (Mao et al., 2018; Dankers et al., 2019; Stowe et al., 2019; Le et al., 2020). Therefore it is convenient for us to compare the proposed method with previous work.

	Train	Dev	Test
# tokens	116,622	38,628	50,175 (5,873)
# unique sent.	6,323	1,550	2,694
% metaphor	11.2	11.6	12.4

Table 1: Basic statistics of the preprocessed VUA dataset provided by (Gao et al., 2018). 50,175 and 5,873 tokens are used for evaluating All-POS and Verb tracks, respectively.

There are two tracks: Verb and All-POS metaphor detection. Some basic statistics of the dataset are shown in Table 1. We focus on the Verb track since we mainly model metaphorical relations for verbs. We use MrBERT's relation-level predictions for the verb track and use its sequence labeling module to deal with the All-POS track.

MOH-X and TroFi datasets MOH-X (Mohammad et al., 2016) and TroFi (Birke and Sarkar, 2006) are two relatively smaller datasets compared with VUA. Only a single target verb is annotated in each sentence. We will report the results on MOH-X and TroFi in three settings: zero-shot transfer, re-training and fine-tuning.

Metrics The evaluation metrics are accuracy (Acc), precision (P), recall (R) and F1-score (F1), which are most commonly used in previous work.

4.1.2 Baselines

We compare with the following approaches.

- Gao et al. (2018) use contextual embeddings ELMo to enhance word representations and use BiLSTM as the encoder. It has two settings: classification (CLS) and sequence labeling (SEQ).

- Mao et al. (2019) exploit two linguistic theory motivated intuitions based on the basis of (Gao et al., 2018). This work motivates us to further explore contextual relation modeling with pre-trained language models.

- Stowe et al. (2019) exploit grammatical relations for data augmentation to enhance (Gao et al., 2018).

- Le et al. (2020) propose a multi-task learning approach with graph convolutional neural networks and use word sense disambiguation as an auxiliary task.

Parameter	Value
Learning Rate	5e-5
Optimizer	Adam
Batch-size	16
Dropout	0.1
Weight decay	0.01
Linear warmup	used

Table 2: Hyper-parameters for BERT based systems.

- Neidlein et al. (2020) (BERT-SEQ) provide a detail setting for a BERT based sequence labeling model. This method is used as a main pre-trained language model based baseline.

The above methods all used Gao et al. (2018)'s dataset for evaluation so that their results can be directly read from their papers for comparison.

- Su et al. (2020) (DeepMet) view metaphor detection as a reading comprehension problem with RoBERTa as the backbone model. It obtained the best performance on 2020 metaphor detection shared task.

- Choi et al. (2021) (MelBERT) present a concurrent work to ours. The method shares similar ideas and architecture with us, but it does not consider the grammatical relations.

Notice that the systems participating in the VUA metaphor detection shared tasks (Leong et al., 2018, 2020) can use any way to manipulate the training set for model selection and ensemble learning so that the reported results in the task report are not directly comparable to us. The results of DeepMet and MelBERT are based on the single model evaluation in (Choi et al., 2021).

The first four baselines do not utilize pre-trained language models, while the last three baselines use BERT or RoBERTa. These baselines support comprehensive comparisons from multiple aspects.

4.1.3 Parameter Configuration

During context component extraction, if the target verb does not have a subject or an object, we use a fixed zero vector instead. We use the *bert-base-uncased* model and the standard tokenizer. The values of hyper-parameters are shown in Table 2.

For MrBERT, we view the ways of component representation (*start marker* or *averaged embedding*, see Section 3.1.2), relation modeling (*linear, bilinear*, and *neural tensor (NT)*) models, see Section 3.2) and context integration (*context concatenation, average* and *maxout*, see Section 3.3) strategies as hyper-parameters as well. We run each model for 10 epoches, and choose the best combination according to the performance on the development set. The best combination uses the averaged embeddings, the bilinear model and the context average strategy, and it will represent MrBERT for performance report in Section 4.2.

4.2 Main Results on VUA Dataset

Table 3 shows the results of the baselines and MrBERT. Except for (Gao et al., 2018)-CLS, all methods use the annotation information of all tokens. For the All-POS track, we report the performance on either all POS tags or 4 main POS tags for comparison with previous work.

We can see that MrBERT achieves superior or competitive performance compared with previous work on verb metaphor detection. The use of pre-trained language models improves the performance in general, compared with several LSTM based methods. Recent proposed models, such as DeepMet, MelBERT and MrBERT, gain further improvements compared with BERT-SEQ.

MrBERT outperforms (Stowe et al., 2019) and (Le et al., 2020) largely. The two baselines attempt to make use of grammar information, through data augmentation or graph neural networks. In contrast, MrBERT provides a simple yet effective way to incorporate verb arguments and new contexts into a pre-trained language model.

MrBERT also has competitive performance compared with DeepMet and MelBERT. We share the similar idea to enhance interactions between the target verb and its contexts, but implement in different ways. DeepMet and MelBERT base on the pre-trained model RoBERTa and use additional POS or FGPOS information. Moreover, these two models are trained for every token so that the training might be more sufficient. In contrast, we mainly model metaphorical relation for verbs. This is perhaps also the reason that on the All-POS metaphor detection track, MrBERT has slightly worse results compared with MelBERT. However, our model is flexible and can be applied to tokens with other POS tags as well. We leave this as future work.

4.3 Analysis

We further analyze the effects of modeling contextual relations from several aspects.

Relation modeling and context integration strategies Table 4 shows the results of different

	VUA Verb				VUA All-POS				VUA All-POS (4 POS)			
Model	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Gao et al. (2018)-CLS	69.1	53.4	65.6	58.9	–	–	–	–	–	–	–	–
Gao et al. (2018)-SEQ	81.4	68.2	71.3	69.7	93.1	71.6	73.6	72.6	–	–	–	–
Mao et al. (2019)	81.8	66.3	**75.2**	70.5	93.8	73.0	75.7	74.3	–	–	–	–
Stowe et al. (2019)	–	–	–	69.5	–	–	–	73.5	–	–	–	–
Le et al. (2020)	83.2	72.5	70.9	71.7	93.8	74.8	75.5	75.1	–	–	–	–
Neidlein et al. (2020)	84.9	78.0	69.0	73.2	94.5	**83.0**	71.9	77.0	**91.8**	77.9	**64.6**	70.7
DeepMet (Su et al., 2020)	–	79.5	70.9	74.9	–	82.0	71.3	76.3	–	–	–	–
MelBERT (Choi et al., 2021)	–	78.7	72.9	75.7	–	80.1	**76.9**	**78.5**	–	–	–	
MrBERT	**86.4**	**80.8**	71.5	**75.9**	**94.7**	82.7	72.5	77.2	**91.8**	**78.4**	**64.6**	**70.9**

Table 3: Results on the VUA dataset. MrBERT uses the bilinear model for relation modeling and the context-average integration strategy. VUA All-POS (4 POS) indicates the performance on 4 main POS tags.

	VUA-verb			
Model	Acc	P	R	F1
BERT-SEQ	85.1	77.5	70.8	74.0
Average-Linear	85.7	79.8	70.2	74.7
Average-Bilinear	**86.4**	**80.8**	71.5	**75.9**
Average-NT	85.7	77.4	**73.8**	75.6
Maxout-Linear	85.2	78.1	70.2	73.9
Maxout-Bilinear	85.3	75.7	**74.8**	**75.3**
Maxout-NT	**85.6**	**78.8**	70.9	74.7
Concat-Linear	**85.5**	**80.3**	68.6	74.0
Concat-Bilinear	85.2	77.6	71.2	**74.3**
Concat-NT	85.0	76.4	**72.3**	**74.3**

Table 4: The effects of the ways for modeling contextual relations and integrating multiple contexts.

combinations of relation modeling and context integration strategies.

BERT-SEQ here refers to the re-trained baseline with model selection based on the performance on the development set, and surpasses the reported results in (Neidlein et al., 2020). We can see that most combinations outperform BERT-SEQ, and have consistent performance. The bilinear and neural tensor models perform better than the linear model. This means that sophisticated relation modelling techniques can benefit the performance.

Context average and context maxout strategies perform better than context concatenation. The reason may be that context concatenation is more difficult to be trained due to more parameters.

Effects of different contexts Table 5 shows the performance of MrBERT when it considers the global context (MrBERT-G), the global and the local contexts (MrBERT-GL), and the full model with the distant context (MrBERT-GLD). Each model is trained separately, with the same model selection procedure. We can see that integrating multiple contexts leads to better performance.

	VUA-verb			
Model	Acc	P	R	F1
MrBERT-G	85.2	77.3	71.9	74.5
MrBERT-GL	85.5	76.8	**73.9**	75.3
MrBERT-GLD	**86.4**	**80.8**	71.5	**75.9**

Table 5: The performance of MrBERT when considering different types of contexts: G, L and D indicate global, local and distant contexts, respectively.

MrBERT explicitly incorporates verb arguments through grammatical relations as the local context, which differs from other methods. We are interested in the effect of such information.

We analyze MrBERT-G and MrBERT-GL. Table 6 shows the distribution of auto-extracted verb-subject and verb-direct object relations in the VUA test dataset. ΔF_1 values indicate the improvements of MrBERT-G compared with BERT-SEQ in F1. We can see that MrBERT-G outperforms BERT-SEQ mainly when verb's arguments are incomplete. For verbs with complete verb-subject and verb-direct object structures, little improvement is gained.

Table 7 shows the corresponding performance of MrBERT-GL. Better performance is obtained for verbs with all status of grammatical relations. The improvement on verbs in the lower right corner is obvious. In these cases, the verbs are usually intransitive verbs or used as a noun or an adjective. The benefit of involving grammatical relations may be that it helps keep a dynamic and balanced focus between the global and local contexts according to the signals expressed by the grammatical structure.

Intuitively, the effect of incorporating grammatical relations should be more obvious for metaphor detection in long sentences, since the local and global contexts are quite different. To verify this, we divide sentences in the test dataset into bins

		Verb-direct object: Yes	Verb-direct object: No	total
Verb-subject	Yes	1,324 (36%) ΔF_1=0.0	2,035 (23%) ΔF_1= +0.57	3,359
	No	1,201 (38%) ΔF_1=+0.05	1,313 (27%) ΔF_1= +1.51	2,514
total		2,525	3,348	

Table 6: The distribution of available syntactic patterns in VUA-verb test dataset and the improved F1 score of MrBERT-G compared with BERT-SEQ. The figures in brackets are the percentage of metaphors.

		Verb-direct object: Yes	Verb-direct object: No	total
Verb-subject	Yes	1,324 (36%) ΔF_1=0.47	2,035 (23%) ΔF_1= +0.65	3,359
	No	1,201 (38%) ΔF_1=0.93	1,313 (27%) ΔF_1= +4.29	2,514
total		2,525	3,348	

Table 7: Similar to Table 6, this table shows the improved F1 score of MrBERT-GL, instead of MrBERT-G, compared with BERT-SEQ.

according to the number of clauses. Figure 3 confirms our hypothesis that MrBERT obtains larger improvements on sentences with more clauses, indicating that incorporating grammatical relations can help filter noisy information.

Finally, the use of distant context obtains a further improvement. This observation is consistent with the conclusion of (Choi et al., 2021). It also indicates that the BERT tokenizer's embedding can be used to approximate the representation of the target verb's basic meaning.

4.4 Results on MOH-X and TroFi Datasets

Table 8 shows the results on the MOH-X and TroFi datasets.

In the zero-shot transfer setting, MrBERT obtains better performance compared with DeepMet and MelBERT on both datasets. The performance of DeepMet and MelBERT is read from (Choi et al., 2021). The results means MrBERT has good zero-shot transferability, although these datasets have quite different characteristics.

In the 10-fold cross-validation setting, the re-trained MrBERT can also obtain superior or competitive results compared with previous work. If we continue to fine-tune the pre-trained MrBERT on the target datasets, better performance can be obtained, especially on the MOH-X dataset.

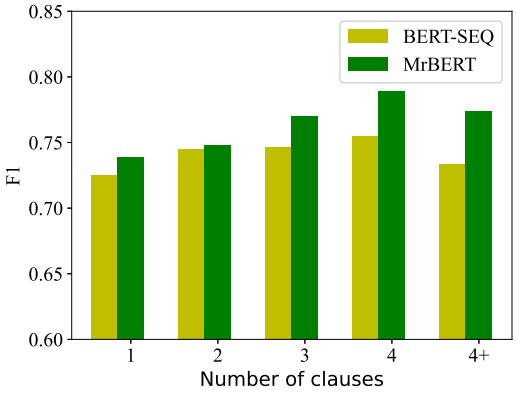

Figure 3: The F1 scores of MrBERT and BERT-SEQ for sentences with different number of clauses.

	Model	Acc	P	R	F1
	MOH-X				
CV	Gao et al. (2018)	78.5	75.3	84.3	79.1
	Mao et al. (2019)	79.8	77.5	83.1	80.0
	Le et al. (2020)	79.9	79.7	80.5	79.6
	MrBERT	81.9	80.0	85.1	82.1
	MrBERT-finetune	**84.9**	**84.1**	**85.6**	**84.2**
Trans.	DeepMet	-	**79.9**	76.5	77.9
	MelBERT	-	79.3	79.7	79.2
	MrBERT	79.3	75.9	**84.1**	**79.8**
	TroFi				
	Model	Acc	P	R	F1
CV	Gao et al. (2018)	74.6	70.7	71.6	71.1
	Mao et al. (2019)	75.2	68.6	**76.8**	72.4
	Le et al. (2020)	76.4	73.1	73.6	**73.2**
	MrBERT	75.1	70.4	74.3	72.2
	MrBERT-finetune	**76.7**	**73.9**	72.1	72.9
Trans.	DeepMet	-	53.7	72.9	61.7
	MelBERT	-	53.4	74.1	62.0
	MrBERT	61.1	**53.8**	**75.0**	**62.7**

Table 8: The experimental results on MOH-X and TroFi, where CV indicates 10-fold cross-validation and Trans. indicates transferring the trained MrBERT on VUA to the target datasets.

5 Related Work

Metaphor detection is a key task in metaphor processing (Veale et al., 2016). It is typically viewed as a classification problem. The early methods were based on rules (Fass, 1991; Narayanan, 1997),

while most recent methods are data-driven. Next, we summarize data-driven methods from the perspective of context types that have been explored.

Grammatical relation-level detection This line of work is to determine the metaphoricity of a given grammatical relation, such as verb-subject, verb-direct object or adjective-noun relations (Shutova et al., 2016). The key to this category of work is to represent semantics and capture the relation between the arguments.

Feature-based methods are based on handcrafted linguistic features. Shutova and Teufel (2010b) proposed to cluster nouns and verbs to construct semantic domains. Turney et al. (2011) and Shutova and Sun (2013) considered the abstractness of concepts and context. Mohler et al. (2013) exploited Wikipedia and WordNet to build domain signatures. Tsvetkov et al. (2014) combined abstractness, imageability, supersenses, and cross-lingual features. Bulat et al. (2017) exploited attribute-based concept representations.

The above handcrafted features heavily rely on linguistic resources and expertise. Recently, distributed representations are exploited for grammatical relation-level metaphor detection. Distributed word embeddings were used as features (Tsvetkov et al., 2014) or to measure semantic relatedness (Gutiérrez et al., 2016; Mao et al., 2018). Visual distributed representations were also proven to be useful (Shutova et al., 2016). Rei et al. (2017) designed a supervised similarity network to capture interactions between words. Song et al. (2020) modeled metaphors as attribute-dependent domain mappings and presented a knowledge graph embedding approach for modeling nominal metaphors. Zayed et al. (2020) identified verb-noun and adjective-noun phrasal metaphoric expressions by modeling phrase representations as a context.

Token-level detection Another line of work formulates metaphor detection as a single token classification or sequence labeling problem (Do Dinh and Gurevych, 2016; Gao et al., 2018; Mao et al., 2019). These approaches are mostly based on neural network architectures and learn representations in an end-to-end fashion. These approaches depend on token-level human annotated datasets, such as the widely used VUA dataset (Steen, 2010).

BiLSTM plus pre-trained word embeddings is one of the popular architectures for this task (Gao et al., 2018; Mao et al., 2019). Recently, Transformer based pre-trained language models become the most popular architecture in the metaphor detection shared task (Leong et al., 2020). Multi-task learning (Dankers et al., 2019; Rohanian et al., 2020; Le et al., 2020; Chen et al., 2020) and discourse context (Dankers et al., 2020) have been exploited as well.

Discussion The grammatical relation-level and token-level metaphor detection consider different aspects of information. Grammatical relations incorporate syntactic structures, which are well studied in selectional preferences (Wilks, 1975, 1978) and provide important clues for metaphor detection. However, sentential context is also useful but is ignored. In contrast, token-level metaphor detection explores wider context and gains improvements, but syntactic information is neglected and as discussed in (Zayed et al., 2020), interactions between metaphor components are not explicitly modeled.

This paper aims to combine the grammatical relation-level, token-level and semantic-level information through pre-trained language model based contextual relation modeling.

6 Conclusion

This paper presented the Metaphor-relation BERT (MrBERT) model for verb metaphor detection. We propose a new view to formulate the task as modeling the metaphorical relation between the target verb and its multiple context components, i.e., contextual relations. We propose and evaluate various ways to extract, model and integrate contextual relations for metaphoricity prediction. We conduct comprehensive experiments on the VUA dataset. The evaluation shows that MrBERT achieves superior or competitive performance compared with previous methods. We also observe that incorporating grammatical relations can help balance local and global contexts, and the basic meaning of the verb as a distant context is effective. Further experiments on small datasets MOH-X and TroFi also show good model transferability of MrBERT.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61876113, 61876112), Beijing Natural Science Foundation (No. 4192017), Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan (CIT&TCD20170322). Lizhen Liu is the corresponding author.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. Association for Computational Linguistics.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In Proceedings of the Second Workshop on Figurative Language Processing, pages 235–243, Online. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1763–1773, Online. Association for Computational Linguistics.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In Proceedings of the Second Workshop on Figurative Language Processing, pages 227–234, Online. Association for Computational Linguistics.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

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

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In Proceedings of the Fourth Workshop on Metaphor in NLP, pages 28–33, San Diego, California. Association for Computational Linguistics.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. Computational linguistics, 17(1):49–90.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. Metaphor and symbol, 22(1):1–39.

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. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. University of Chicago press.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In AAAI, pages 8139–8146.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In Proceedings of the Second Workshop on Figurative Language Processing, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In Proceedings of the Workshop on Figurative Language Processing, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, pages 23–33.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In Proceedings of the First Workshop on Metaphor in NLP, pages 27–35, Atlanta, Georgia. Association for Computational Linguistics.

Srini Narayanan. 1997. Knowledge-based action representations for metaphor and aspect (KARMA). Ph.D. thesis, Ph. D. thesis, University of California at Berkeley.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3722–3736.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1537–1546, Copenhagen, Denmark. Association for Computational Linguistics.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2890–2895, Online. Association for Computational Linguistics.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 160–170, San Diego, California. Association for Computational Linguistics.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 978–988, Atlanta, Georgia. Association for Computational Linguistics.

Ekaterina Shutova and Simone Teufel. 2010a. Metaphor corpus annotated for source - target domain mappings. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Ekaterina Shutova and Simone Teufel. 2010b. Metaphor corpus annotated for source-target domain mappings. In LREC, volume 2, pages 2–2. Citeseer.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2020. A knowledge graph embedding approach for metaphor processing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:406–420.

Gerard Steen. 2010. A method for linguistic metaphor identification: From MIP to MIPVU, volume 14. John Benjamins Publishing.

Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic analysis improves neural metaphor detection. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 362–371, Hong Kong, China. Association for Computational Linguistics.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In Proceedings of the Second Workshop on Figurative Language Processing, pages 30–39, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In Proceedings of the First Workshop on Metaphor in NLP, pages 45–51, Atlanta, Georgia. Association for Computational Linguistics.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In Proceedings of the 2011 Conference on Empirical

Methods in Natural Language Processing, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. Synthesis Lectures on Human Language Technologies, 9(1):1–160.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. Artificial intelligence, 6(1):53–74.

Yorick Wilks. 1978. Making preferences more active. Artificial intelligence, 11(3):197–223.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2020. Contextual modulation for relation-level metaphor identification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 388–406, Online. Association for Computational Linguistics.