Enhancing Metaphor Detection by Gloss-based Interpretations

Hai Wan[1], **Jinxia Lin**[1], **Jianfeng Du**[2,3]*, **Dawei Shen**[1], **Manrong Zhang**[1]
[1] School of Computer Science and Engineering,
Sun Yat-sen University, Guangzhou 510006, P.R.China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing,
Guangdong University of Foreign Studies, Guangzhou 510006, P.R.China
[3] Pazhou Lab, Guangzhou 510330, P.R.China
wanhai@mail.sysu.edu.cn, jfdu@gdufs.edu.cn,
{linjx36, shendw5, zhangmr7}@mail2.sysu.edu.cn

Abstract

This paper focuses on utilizing *metaphor interpretation* to enhance *metaphor detection*. Considering that existing approaches to metaphor interpretation are limited by ambiguous meanings of the metaphorical substitute words, this paper proposes a novel interpretation mechanism that utilizes *glosses* to interpret metaphorical words. Since there is no dataset annotated for both metaphor detection and metaphor interpretation, we enhance three datasets TroFi, VUA, and PSUCMC from the field of metaphor detection with gloss annotations. Accordingly, we develop a model for jointly conducting metaphor detection and gloss-based interpretation (named MDGI-Joint for short). Experimental results demonstrate that MDGI-Joint outperforms state-of-the-art models on all the three enhanced datasets and that gloss-based metaphor interpretation benefits metaphor detection.[1]

1 Introduction

Metaphor has been defined as words or other linguistic expressions representing another concept with the language from a more concrete conceptual domain (Kövecses and Zoltán, 2002; Lagerwerf and Meijers, 2008). According to existing studies on metaphor (Kövecses and Zoltán, 2002; Steen, 2010), metaphor has been used so frequently in daily language that it almost occurs in one out of three natural language sentences.

Metaphor detection aims to identify all *metaphorical words* in given texts and has been demonstrated to be of great value in many Natural Language Processing (NLP) tasks, such as machine translation (Mao et al., 2018), machine reading comprehension (Shutova et al., 2013), *etc.* Therefore, metaphor detection has drawn increasing interests in recent years. There have emerged a number of methods for metaphor detection, including SEQ (Gao et al., 2018), RNN_HG (Mao et al., 2019), RNN_MHCA (Mao et al., 2019), MUL_GCN (Le et al., 2020), DeepMet (Su et al., 2020), *etc.*

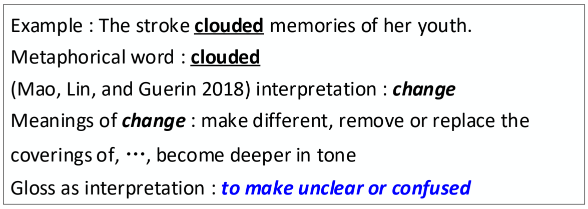
Example : The stroke **clouded** memories of her youth.
Metaphorical word : **clouded**
(Mao, Lin, and Guerin 2018) interpretation : ***change***
Meanings of ***change*** : make different, remove or replace the coverings of, ···, become deeper in tone
Gloss as interpretation : ***to make unclear or confused***

Figure 1: An example for metaphor detection, where the word **clouded** is a metaphorical word (highlighted in bold and underline). Interpretation by a substitute word (highlighted in **bold**) is computed by (Mao et al., 2018). Gloss as interpretation (highlighted in ***blue bold italics***) is picked from the Merriam Webster dictionary.

To capture the meanings of metaphorical words, *metaphor interpretation* has also been studied, which paraphrases metaphorical expressions into literal expressions that maintain the intended meanings of given texts (Mao et al., 2018). Existing approaches have treated metaphor interpretation as extraction of transferred properties, identification of the underlying conceptual mapping, or generation of a literal substitute paraphrase (Rai and Chakraverty, 2020). All of them are yet limited by ambiguous meanings of the metaphorical substitute words. Consider the example shown in Figure 1. The metaphorical word **clouded** is interpreted as **change** in (Mao et al., 2018), where **change** has 10 diverse meanings according to WordNet (Miller, 1995). Due to multiple meanings of **change**, it is difficult to precisely capture the intended meaning of the metaphorical word **clouded**.

It has been pointed out by Group (2007) that, metaphors have a clear distinction between their

*Corresponding author

[1] Source code and datasets are available at `https://github.com/sysulic/MDGI`.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1971–1981
August 1–6, 2021. ©2021 Association for Computational Linguistics

basic meanings and contextual meanings. Consider the example shown in Figure 1 again. The gloss-based interpretation for **clouded** gives the meaning "*to make unclear or confused*", which is evidently different from the basic meaning "*to grow cloudy*" of **clouded**. In other words, different from the substitute word, the *gloss* extracted from dictionaries can provide an unambiguous interpretation of **clouded** which exhibits a clear distinction to its basic meaning. Therefore, we can utilize gloss-based metaphor interpretation to enhance the metaphor detection task.

Based on the above observations, we propose to use glosses to interpret metaphorical words. Specifically, we formulate metaphor interpretation as a task of predicting the best gloss among a set of candidate glosses. Considering that *Word Sense Disambiguation* (WSD) (Kilgarriff, 2004) is a popular technique for identifying the correct meaning of a target word, modern WSD methods can be adapted to metaphor interpretation.

By now there is a lack of dataset annotated for both metaphor detection and metaphor interpretation. In order to study whether gloss-based interpretation benefits metaphor detection, we enhance three benchmark datasets in the field of metaphor detection, including two English datasets (TroFi and VUA) and one Chinese dataset (PSUCMC). For each dataset, we construct a set of candidate words and annotate these words with glosses extracted from a dictionary.

Accordingly we develop a joint model for **M**etaphor **D**etection and **G**loss-based **I**nterpretation (named MDGI-Joint for short). To be specific, our joint model encodes contexts and glosses independently for every given word. Based on both the contextual word embedding and gloss embeddings, a probability distribution for all glosses of the given word is computed and then used to predict the best gloss, in a similar way as the state-of-the-art WSD method (Blevins and Zettlemoyer, 2020), Afterwards, an attention mechanism is employed to compute an integrated representation of all glosses. This integrated representation is then concatenated with the contextual word embedding to determine whether the given word is metaphorical through a classical prediction layer. The joint model is trained by minimizing a combined loss from both the metaphor detection task and the metaphor interpretation task.

We conduct experiments on the aforementioned three enhanced datasets. Experimental results demonstrate that gloss-based metaphor interpretation does benefit metaphor detection. On one hand, the proposed model achieves state-of-the-art performance on all three enhanced datasets in the metaphor detection task. On the other hand, it also achieves comparable performance with the outstanding WSD method in the metaphor interpretation task.

The main contributions of this paper can be summarized as follows.

- We provide a novel interpretation mechanism that utilizes glosses to interpret metaphorical words.
- We enhance three metaphor detection datasets (TroFi, VUA, and PSUCMC) with annotations of glosses for metaphorical words.
- We develop a joint model for metaphor detection and gloss-based interpretation and empirically show that metaphor interpretation with glosses benefits metaphor detection.

2 Related Work

2.1 Metaphor Detection

Early studies on metaphor detection (Shutova et al., 2010; Rei et al., 2017) usually follow the linguistics theory (Lakoff and Johnson, 1980) and construct mappings from the source domain to the target domain.

Subsequent studies focus on metaphor detection over subject-verb-objects, adjective-noun tuples, or metaphorical phrases. Among these studies, a lot of semantic features including the degree of abstractness, the degree of concreteness, the degree of imageability, semantic super-senses (namely coarse semantic categories originating in WordNet), lemma unigrams, and grammatical dependencies are added to improve performance of metaphor detection (Turney et al., 2011; Tsvetkov et al., 2014; Klebanov et al., 2016; Özbal et al., 2016; Jang et al., 2015). Jang et al. (2015) took topic distribution into consideration. Jang et al. (2016) explored topic transition between a metaphor and its context. For handing multi-modal information, Shutova et al. (2016) considered both word embeddings and visual embeddings whereas Bulat et al. (2017) introduced a cross-modal method to integrate linguistic representations and property-based representations. A broader context of discourse was considered by (Jang et al., 2015) and (Mu et al., 2019).

Regarding word-level metaphor detection, metaphor detection can be treated as a sequence tagging task. Wu et al. (2018) proposed a neural model, which uses Word2Vec (Mikolov et al., 2013) as text representation and encodes part-of-speech (POS) tags, word clusters with Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) network. Gao et al. (2018) and Mao et al. (2019) respectively utilized GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) embeddings. Le et al. (2020) proposed to construct a graph CNN guided by dependency trees of sentences for metaphor detection and to construct a multi-task learning framework for the WSD task, utilizing the knowledge from WSD to improve metaphor detection. Instead of treating metaphor detection as a sequence tagging task, Su et al. (2020) proposed a novel reading comprehension paradigm based on a pre-trained language model, using features from POS tags and local texts.

2.2 Metaphor Interpretation

Metaphor interpretation is an intricate task, having a challenge in deciphering the meaning conveyed by a metaphorical expression (Rai and Chakraverty, 2020). Existing approaches to metaphor interpretation can be grouped into three categories. For the first category, the problem of metaphor interpretation is treated as a problem of extraction of transferred properties. It is often assumed that a metaphor is essentially a projection of a specific set of salient concept properties from the source domain, known as *property matching* (Su et al., 2016; Ortony, 1980). Su et al. (2016) extracted perceptual properties from the source domain and the target domain, and then searched for metaphor interpretation by expanding the extracted properties with synonymy relationships from WordNet. In contrast to property matching, the second category defines the problem of metaphor interpretation as a problem of identifying the underlying conceptual mapping, usually focusing on re-conceptualization of the target domain (Marmolejo-Ramos et al., 2013; Semino, 2010). Martin (2006) empirically found that metaphor interpretation has certain contextual clues (*e.g.*, the appearance of a concept from the target domain) and related metaphorical expressions. For the last category, the problem of metaphor interpretation is treated as the generation of a literal substitute paraphrase (Mao et al., 2018; Shutova et al., 2012). Mao et al. (2018) used hypernyms and synonyms as candidate substitutes and computed the best substitute word by the cosine similarity between the embedding of the given word and the embedding of a candidate substitute.

All the above methods for metaphor interpretation fail to capture the contextual meaning of a metaphorical word due to ambiguous interpretation of the metaphor. To tackle this issue, in this work we provide a novel interpretation mechanism that utilizes glosses to interpret metaphorical words.

2.3 Word Sense Disambiguation

Word sense disambiguation (WSD) aims to predict a specific meaning of a word that occurs in a particular context (Navigli, 2009). Understanding the meaning of a word in context is critical to many NLP tasks, such as machine translation (Vickrey et al., 2005; Neale et al., 2016; Gonzales et al., 2017) and information extraction (Ciaramita and Altun, 2006; Bovi et al., 2015). One category of WSD is class-based, which provides coarse-grained labels that are shared among different words. The other category of WSD is word-based, aiming to disambiguate every word in texts (Palmer et al., 2001; Moro and Navigli, 2015; Blevins and Zettlemoyer, 2020). Our proposed metaphor interpretation scheme belongs to this category. Some neural models for word-based WSD exploit encoders for better feature extraction. Based on an encoder, they either train classifiers on top of extracted features (Kågebäck and Salomonsson, 2016) or introduce a shared output space to label words (Raganato et al., 2017). Other neural models augment word representations with additional data by semi-supervised learning (Melamud et al., 2016; Yuan et al., 2016). BEM (Blevins and Zettlemoyer, 2020), which inspires our model, is a state-of-the-art method for word-based WSD. It introduces a bi-encoder model to embed the target word with its surrounding context and its glosses.

3 The Proposed MDGI-Joint Model

Given a sentence s consisting of n words $\{w_0, w_1, \ldots, w_i, \ldots, w_{n-1}\}$ and a list of all m_i candidate glosses $G_i = \{g_i^0, g_i^1, \cdots, g_i^j, \ldots, g_i^{m_i-1}\}$ for the target word w_i, the task of metaphor interpretation aims to select a gloss from G_i to interpret the intended meaning of w_i, whereas the task of

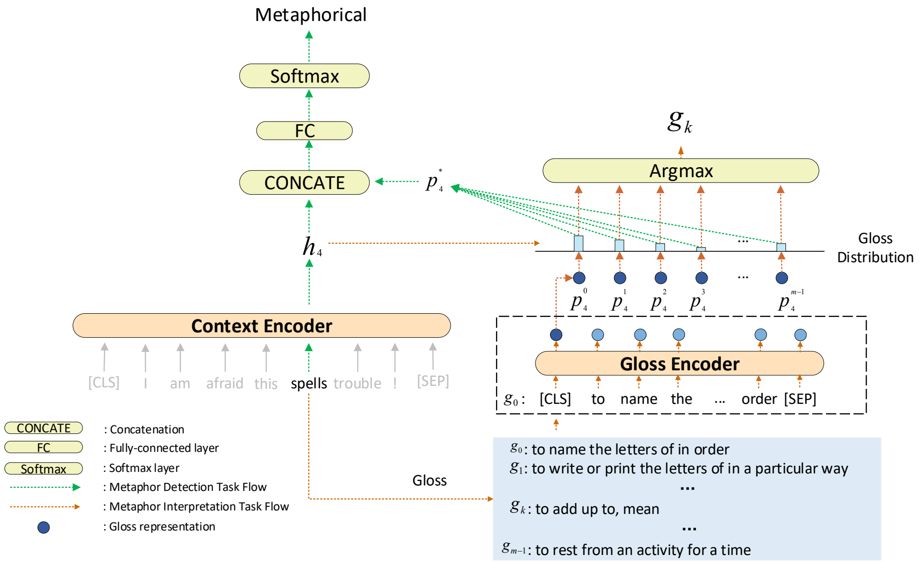
Metaphorical
Softmax
FC
CONCATE
p_4^*
g_k
Argmax
h_4
Gloss Distribution
...
p_4^0
p_4^1
p_4^2
p_4^3
p_4^{m-1}
Context Encoder
[CLS]
I
am
afraid
this
spells
trouble
!
[SEP]
Gloss Encoder
g_0: [CLS] to name the ... order [SEP]
CONCATE : Concatenation
FC : Fully-connected layer
Softmax : Softmax layer
: Metaphor Detection Task Flow
: Metaphor Interpretation Task Flow
● : Gloss representation
Gloss
g_0: to name the letters of in order
g_1: to write or print the letters of in a particular way
...
g_k: to add up to, mean
...
g_{m-1}: to rest from an activity for a time

Figure 2: The architecture of our proposed joint model for **M**etaphor **D**etection and **G**loss-based **I**nterpretation.

metaphor detection focuses on predicting whether the word w_i is metaphorical or literal. Since gloss-based metaphor interpretation provides a rich representation for the contextual meaning of the target word, it is appropriate to construct a joint model to incorporate gloss-based metaphor interpretation and metaphor detection. Therefore, we propose a model named MDGI-Joint to incorporate metaphor detection and metaphor interpretation.

The architecture of MDGI-Joint is given by Figure 2. MDGI-Joint employs two encoders to generate respectively the contextual representation and the gloss representation for a target word. The probability distribution over all candidate glosses is computed by an attention mechanism. The gloss with the highest probability is predicted as the interpretation of the target word. Afterwards, an integrated representation of all candidate glosses is computed and then concatenated with the contextual representation. The concatenated representation is finally used to determine whether the target word is metaphorical, through a fully-connected layer followed by the softmax classifier.

3.1 Encoding Module for Sentences

Given a sentence s consisting of n words $\{w_0, w_1, \ldots, w_i, \ldots, w_{n-1}\}$ as well as a target word w_i, the contextual representation of w_i will be encoded into a vector. Considering that the pre-trained language model BERT (Devlin et al., 2019) has been proved to be effective in transfer learning, contributing to state-of-the-art performance in many NLP tasks, we fine-tune a BERT model to be the context encoder. Initially, we construct a token sequence "[CLS], $w_0, w_1, \cdots, w_{n-1}$,[SEP]", where [CLS] and [SEP] are special tokens introduced in BERT. Then this token sequence is taken as input to BERT, yielding a sequence of 768-dimensional vectors "$h_{[CLS]}, h_0, h_1, \ldots, h_{n-1}, h_{[SEP]}$" as the output of BERT. The corresponding vector h_i for the target word w_i is treated as the contextual representation of w_i.

3.2 Encoding Module for Glosses

For the target word w_i, we collect the set of candidate glosses $G_i = \{g_i^0, g_i^1, ..., g_i^{m_i-1}\}$ from an existing dictionary. We fine-tune another BERT model to be the gloss encoder. Similar to the encoding module for sentences, we also construct a token sequence for each gloss and feed it into BERT. For each gloss $g_i^j \in G_i$ where $0 \leq j \leq m_i - 1$, we use the output 768-dimensional vector corresponding to the first token "[CLS]" as the gloss representation, which is written as p_i^j.

3.3 Prediction for Metaphor Interpretation

Based on the contextual representation h_i of w_i and the gloss representation p_i^j of each gloss $g_i^j \in G_i$ where $0 \leq j \leq m_i - 1$, the probability α_i^j for a gloss g_i^j to represent the intended meaning of the word w_i is computed by an attention mechanism, formally defined as follow:

$$\alpha_i^j = \frac{\exp(h_i^T p_i^j)}{\sum_{k=0}^{m_i-1} \exp(h_i^T p_i^k)} \qquad (1)$$

3.4 Prediction for Metaphor Detection

We define the joint representation q_i of the target word w_i as the concatenation of the contextual representation h_i and the weighted sum p_i^* of gloss representations $p_i^0, \ldots, p_i^{m_i-1}$, which is formally defined below:

$$\begin{aligned} p_i^* &= \sum_{j=0}^{m_i-1} \alpha_i^j p_i^j, && (2) \\ q_i &= [h_i; p_i^*] && (3) \end{aligned}$$

where $[;]$ denotes the concatenation of two vectors.

By transforming the joint representation q_i through a fully connected layer followed by the softmax classifier, the probability distribution l_i that the target word w_i is metaphorical or literal is formally defined below:

$$l_i = \mathsf{softmax}(W_1 q_i + b_1) \qquad (4)$$

where $W_1 \in \mathbb{R}^{2\times1536}$ and $b_1 \in \mathbb{R}^2$ are learnable parameters. The probability distribution l_i is of the form $[l_i^0, l_i^1]$, where l_i^0 is the predicted probability that w_i is literal, and l_i^1 is the predicted probability that w_i is metaphorical.

3.5 Training Objective Function

Based on predicted probabilities $\alpha_i^0, \ldots, \alpha_i^{m_i-1}$ over glosses, the loss value for metaphor interpretation about the target word w_i is defined as:

$$\text{loss}_i^{\text{MI}} = -\sum_{j=0}^{m_i-1} I(f_i = g_i^j) \log(\alpha_i^j)) \qquad (5)$$

where f_i is the correct gloss for w_i in sentence s, $I(X) = 1$ if X is true and $I(X) = 0$ otherwise.

As for the task of metaphor detection, we employ the binary cross entropy loss for predicting the target word w_i to be literal or metaphorical, defined as follow:

$$\text{loss}_i^{\text{MD}} = -(1-y_i)\log(l_i^0) - y_i \log(l_i^1) \qquad (6)$$

where y_i is the correct label, $y_i = 0$ if the label is literal or $y_i = 1$ if the label is metaphorical.

The proposed joint model is trained by minimizing the following combined loss for every training sentence.

$$\text{loss} = \sum_{i=0}^{n-1} \text{loss}_i^{\text{MD}} + I(w_i \in \mathcal{C}) * \text{loss}_i^{\text{MI}} \qquad (7)$$

where n is the number of words in the considering sentence, and $\mathcal{C}$ is a predefined set of candidate words required to be interpreted, which is collected from all metaphorical words in our experiments.

Table 1: Kappa-score for annotations in every dataset.

	TroFi	VUA	PSUCMC
kappa-score	0.82	0.86	0.89

Table 2: Statistics of the enhanced datasets. #sentences is the number of sentences; #tokens is the total number of words needed to be detected, %M is the metaphor percentage over the detected words, and #glosses is the number of samples with gloss annotations.

Dataset	#sentences	#tokens	%M	#glosses
TroFi_train	2,989	2,989	58.3	2,989
TroFi_val	374	374	52.4	374
TroFi_test	374	374	54.8	374
VUA_train	6,323	116,622	11.2	2,710
VUA_val	1,550	38,628	11.6	635
VUA_test	2,694	50,175	12.4	905
VERB_test	2,694	5,873	30.0	905
PSUCMC_train	1,381	28,572	8.3	5,486
PSUCMC_val	173	3,520	8.0	674
PSUCMC_test	173	3,727	7.4	730
VERB_test	165	736	16.3	120

4 Experiments

4.1 Three Enhanced Datasets

We enhance three datasets TroFi, VUA and PSUCMC from the field of metaphor detection, where TroFi and VUA are in English and PSUCMC is in Chinese.

- TroFi (Birke and Sarkar, 2006). This is a benchmark metaphor dataset with verb metaphors annotated. Following the work (Mao et al., 2019), we treat unlabeled words as literal in the training phase.

- VUA (Steen, 2010). The VU Amsterdam Metaphor Corpus (VUA) samples fragments from the British National Corpus. All words in the corpus are labeled. We evaluate on both the VUA ALL POS track and the VUA-Verb track.

- PSUCMC (Lu and Wang, 2017; Nacey et al., 2019). The PSU Chinese Metaphor Corpus

Table 3: Data splits of PSUCMC for different POS tags, including NOWN, VERT, ADJ and ADV.

Dataset	NOUN	VERB	ADJ	ADV
all	929	2,547	390	19
train	760	1,988	334	13
val	105	244	31	4
test	64	315	25	2

Table 4: Accuracy on the metaphor interpretation task. All numbers are in %.

Model	VUA	PSUCMC	TroFi
BEM	45.5	**82.3**	71.9
MDGI-Joint-S	**46.4**	**82.3**	**72.2**
MDGI-Joint	44.3	80.8	70.9

consists of text samples from the Lancaster Corpus of Mandarin Chinese where are annotated for metaphor-related words following MIPVU (Steen, 2010). All words in PSUCMC are labeled. Following the experimental setting of VUA, we evaluate on both the PSUCMC ALL POS track and the PSUCMC-Verb track.

All these datasets are enhanced to support the gloss-based metaphor interpretation task. We refer to the set of words that need to be interpreted as the *candidate set*. For TroFi, the candidate set includes all labeled verbs. For VUA, the words in the candidate set are randomly selected from verbs. As for PSUCMC, we randomly choose a set of words and filter the meaningless words to construct a candidate set. Words in the candidate set are annotated by human annotators.

For English datasets, when given a word in the candidate set, the annotators are asked to look up the Merriam-Webster dictionary[2] to fetch its glosses. For the Chinese dataset, word glosses are extracted from the Baidu Dictionary[3].

For every dataset, four annotators are recruited to annotate the dataset independently. All annotators need to select the most appropriate gloss that expresses the contextual meaning of the target word, by comparing all glosses with the given context of the target word. They also need to discuss and determine the final labels after generating their own annotations.

To verify the reliability of the annotations, we use kappa-score to measure inter-annotator agreements for the annotations. Kappa-score (Siegel, 1956) has been widely used in computational linguistics to measure the reliability of an annotation scheme. Table 1 shows the kappa-score for annotations in every dataset, which demonstrates that the annotations have a high degree of reliability.

For PSUCMC and TroFi, we randomly split the samples into a training set, a validation set, and a test set according to the proportion of 8 : 1 : 1. For VUA, the data splits provided by (Mao et al., 2019) are reused. Table 2 reports the statistics of all experimental datasets, whereas Table 3 gives more details about PSUCMC.

4.2 Experimental Setup

In our experiments, BERT[4] is used as both the context encoder and the gloss encoder, where the uncased BERT base model is used for TroFi and VUA, and the Chinese BERT base model is used for PSUCMC.

There are two variants for our proposed model. The first variant, named MDGI-Joint-S, shares parameters between the context encoder and the gloss encoder. The second one, named MDGI-Joint, implements two independent encoders for context and gloss, respectively.

To train these two models, we set the learning rate as 2e-5. The maximum number of epochs is set to 20. We set the dropout probability to 0.2 for the fully connected layer. The max sequence length is set to 128 for TroFi and VUA, and 256 for PSUCMC. The batch sizes for the two tasks are all set to 8 for VUA and TroFi, and 16 for PSUCMC. We keep the best model that maximizes the F1 score on the validation set for the metaphor detection task. This model is then used to evaluate the test set.

4.3 Compared Methods

We compare our method with the following methods.

- SEQ (Gao et al., 2018). SEQ is a neural model taking ELMo embeddings and GloVe embeddings as input. It uses a Bi-LSTM encoder to capture the contextual information of the target word, and then employs a fully connected layer followed by the softmax classifier to predict whether a word is metaphorical or not.

- RNN_HG (Mao et al., 2019). RNN_HG also takes ELMo embeddings and GloVe embeddings as input. Different from SEQ (Gao et al., 2018), RNN_HG concatenates the encoded representation with the GloVe embedding to capture the contextual information of the target word.

[2] https://www.merriam-webster.com/dictionary/

[3] https://dict.baidu.com

[4] https://github.com/google-research/bert

Table 5: Accuracy on the metaphor detection task. All numbers are in %. '-' denotes no evaluation on the corresponding dataset.

	VUA ALL POS				VUA-Verb				PSUCMC ALL POS				PSUCMC-Verb				TroFi			
Method	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
SEQ	71.6	73.6	72.6	93.1	68.2	71.3	69.7	81.4	74.9	57.5	65.0	95.4	71.6	40.2	51.5	87.5	91.2	85.9	88.4	87.7
RNN_HG	71.8	76.3	74.0	93.6	69.3	72.3	70.8	82.1	70.6	69.8	70.2	95.6	67.0	56.8	61.5	88.2	90.4	82.4	86.2	85.6
RNN_MHCA	73.0	75.7	74.3	93.8	66.3	75.2	70.5	81.8	73.5	68.7	71.1	95.9	63.1	58.3	60.6	87.5	86.6	84.9	85.7	84.5
MUL_GCN	74.8	75.5	75.1	93.8	72.5	70.9	71.7	83.2	-	-	-	-	-	-	-	-	-	-	-	-
DeepMet	73.8	73.2	73.5	90.5	76.2	78.3	**77.2**	86.2	66.7	73.9	70.1	96.4	61.4	67.3	64.2	93.1	92.2	81.0	86.2	85.8
MDGI-Joint-S	81.3	73.2	77.0	94.6	78.8	71.5	75.0	85.6	89.7	69.8	78.5	97.2	85.9	60.8	**71.2**	92.0	89.2	84.9	87.0	86.1
MDGI-Joint	82.5	72.5	**77.2**	94.7	78.9	70.9	74.7	85.4	89.0	70.6	**78.7**	97.2	85.9	60.8	**71.2**	92.0	89.3	89.3	**89.3**	88.2

- RNN_MHCA (Mao et al., 2019). RNN_MHCA has a similar architecture as RNN_HG. But it adopts a multi-head contextual attention mechanism to capture the contextual information of the target word.

- MUL_GCN (Le et al., 2020). MUL_GCN is a multi-task learning framework. It exploits the similarity between word sense disambiguation and metaphor detection, by employing a graph convolutional neural network (GCN) to connect the words of interest with context words for metaphor detection.

- DeepMet (Su et al., 2020). DeepMet is a RoBERTa (Ott et al., 2019) based model with an ensemble strategy. It also takes features including POS tags and local texts as input. For fairness, we only compare our models with the single model of DeepMet.

- BEM (Blevins and Zettlemoyer, 2020). BEM is a model for the WSD task. It consists of two independent encoders. One is the context encoder, which embeds the target word and its surrounding context. The other encoder is the gloss encoder, which embeds the gloss for each word sense. Both encoders are deep transformer networks initialized from BERT.

For the evaluation on PSUCMC and TroFi, we use the publicly released code for all the compared methods except for MUL_GCN whose code is unavailable. Thus we cannot obtain results for MUL_GCN on both PSUCMC and TroFi. For VUA, we present results that are reported in the published papers of the compared methods.

It should be noted that all compared methods target the English domain only in their published papers. Since PSUCMC is a dataset in the Chinese domain, to evaluate the compared methods on PSUCMC, we use ELMo embeddings trained on the Xinhua proportion of Chinese gigawords[5], which is a Chinese corpus released by Che et al. (2018) and Fares et al. (2017); moreover, we place the GloVe embeddings with word embeddings trained on the zh-wiki corpus[6] based on Word2Vec (Mikolov et al., 2013).

4.4 Experimental Results

We use P (precision), R (recall), F1 (F1 score) and Acc (accuracy) as the evaluation metrics for the metaphor detection task, and Acc (accuracy) for the metaphor interpretation task. All the reported numbers are in percent. For the metaphor detection task, all words in the test set are evaluated. For the metaphor interpretation task, although corresponding glosses can be computed for all words, only the words with annotations in the test set are evaluated.

4.4.1 Results for Metaphor Interpretation

It can be seen from Table 4 that, no matter whether the two encoders share parameters or not, the proposed model achieves similar results comparable with BEM. MDGI-Joint-S is even slightly superior to BEM. The implemented BEM model does not share parameters between the two encoders. Therefore, it can be confirmed that the improved performance of MDGI-Joint-S comes from capturing the interaction between the two tasks, rather than from inheriting parameters from BEM.

4.4.2 Results for Metaphor Detection

From Table 5, it is evident that MDGI-Joint outperforms other methods over all three datasets except for the VUA-Verb track. Although DeepMet achieves the best performance on the VUA-Verb track, it exhibits lower performance than our proposed model on other datasets and the VUA ALL POS track. There results indicate that joint training for metaphor detection and metaphor interpretation does improve the performance of metaphor

[5] https://github.com/HIT-SCIR/ELMoForManyLangs

[6] https://dumps.wikimedia.org/zhwiki/latest/

Table 6: Case studies to analyze the beneficial effect of gloss-based interpretations.

Text	Detected word	Gold label	Gold gloss	Prediction	Method
Iranian guns pummeled Basra on the war's first day, and in the following eight years, about 65, 000 shells rained down.	rained	metaphorical	to fall like rain	literal	DeepMet
				metaphorical	MDGI-Joint
Stand up man.	stand	literal	to support oneself on the feet in an erect position	metaphorical	DeepMet
				literal	MDGI-Joint

Table 7: Ablation study for the metaphor detection task. All numbers are given in %.

	VUA ALL POS				VUA-Verb				PSUCMC ALL POS				PSUCMC-Verb				TroFi			
Method	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
MDGI-Joint-S	81.3	73.2	77.0	94.6	78.8	71.5	**75.0**	**85.6**	89.7	69.8	78.5	**97.2**	85.9	60.8	**71.2**	**92.0**	89.2	84.9	87.0	86.1
MDGI-Joint	82.5	72.5	**77.2**	**94.7**	78.9	70.9	74.7	85.4	89.0	70.6	**78.7**	**97.2**	85.9	60.8	**71.2**	**92.0**	89.3	89.3	**89.3**	**88.2**
MD-MGI	83.3	71.5	76.9	**94.7**	78.4	69.1	73.4	85.5	76.2	75.6	75.9	96.5	72.3	67.5	69.8	90.5	91.2	85.4	88.2	87.4
MD (only)	80.9	72.2	76.3	94.4	77.5	70.1	73.6	84.8	80.3	69.8	74.7	96.5	77.9	61.7	68.8	90.9	92.5	83.9	88.0	87.4

Table 8: Ablation study for the gloss-based metaphor interpretation task. The measure is accuracy in %.

Model	VUA	PSUCMC	TroFi
MDGI-Joint-S	**46.4**	**82.3**	**72.2**
MDGI-Joint	44.3	80.8	70.9
MD-MGI	44.8	80.8	70.3

detection. It can also be seen that all the compared methods perform poorly on the Chinese dataset PSUCMC. This is probably due to that these methods are strongly language sensitive. The reason why DeepMet outperforms our model on the VUA-Verb track is due to extra information such as subclasses of POS tags being taken as input in DeepMet. It is interesting to see whether our proposed model can be further improved by exploiting extra information. But this question is out of the scope of this work. It will be explored in our future work.

4.5 Case Study

As shown in Table 6, we use two cases to exemplify the superiority of our model in the metaphor detection task, demonstrating that gloss-based interpretation improves performance of metaphor detection.

The first example is picked from the TroFi dataset (see Row 2 in Table 6). In this example the detected word **rained** is a metaphorical word, but DeepMet predicts it as literal. The reason could be that the word **rained** commonly occurs in the given context. In contrast, MDGI-Joint correctly predicts it as metaphorical according to its gloss-based interpretation ***to fall like rain*** since a gloss of the form *do like some behavior* is commonly used as the gloss of a metaphorical word.

The second example is selected from the VUA dataset (see Row 3 in Table 6). In this example the detected word **stand** has a literal meaning in its context. DeepMet wrongly predicts it as metaphorical, while MDGI-Joint gives a correct prediction based on its predicted gloss-based interpretation which has an expression style as that of general glosses of literal words.

4.6 Ablation Study

To investigate the effect of joint training for the two tasks, we further conduct experiments on the following weakened models.

- MD-MGI. In this model we only use the context representation to predict whether a target word is metaphorical. In other words, the weighted sum of gloss representations is not considered in the metaphor detection task, although the two tasks are still jointly trained without sharing parameters between the context encoder and the gloss encoder.

- MD (only). This model addresses the metaphor detection task only and neglects the metaphor interpretation task. It detects metaphors based on the contextual representation of words only.

From Table 7 we observe that the performance of MD-MGI drops slightly on the metaphor detection task compared to the proposed model. The reason is probable that, in MD-MGI the two tasks only interact through the context encoder, leading to limited benefits for metaphor detection. From Table 8 we can see that all the three compared variants are comparable in the metaphor interpretation task, where MDGI-Joint-S is slightly better than others. These results show that whether the metaphor detection task uses the weighted sum of gloss representations or not has few impacts on the metaphor interpretation task. It can be seen

Table 9: Performance on the VUA ALL POS track for metaphor detection, separated by different POS tags. In-vocabulary words are annotated words whereas out-of-vocabulary words have no annotations. Only a portion of verbs in VUA are annotated.

POS	P	R	F1	Acc
ADJ	78.6	57.1	66.2	92
ADV	81.5	65.2	72.4	96.4
NOUN	78.3	60.8	68.8	91.6
VERB(in-vocabulary)	79.2	75.5	77.3	82.0
VERB(out-of-vocabulary)	78.1	70.2	73.9	91.6
Other POS	88.3	84.6	86.4	97.5

from Table 7 that the weakest variant namely MD (only) performs worse than all the other variants, indicating that joint training for metaphor detection and metaphor interpretation leads to performance improvement in metaphor detection.

4.7 Error Analysis on VUA

Taking the VUA dataset as example, there are more false negatives than false positives generated by MDGI-Joint; *i.e.*, the recall is lower than the precision, as shown in Table 9. Especially, adjectives, adverbs and nouns have a significantly lower recall than verbs. The reason for this phenomenon is two-fold. On one hand, only a partial set of verbs has annotations in the VUA dataset, so the detection of metaphorical adjectives, adverbs or nouns can only gain limited benefits from the metaphor interpretation task. On the other hand, in VUA a number of metaphors appear at the phrase level, but MDGI-Joint is only able to detect metaphors at the word level, thus it is more likely to predict false negatives. Consider the following example picked from VUA id:a7w-fragment01#41: *But only two million out of the 20 million journeys which ambulance crews carry out each year are emergency calls.* MDGI-Joint only detects the word *carry* as metaphorical. It treats words separately and cannot predict the preposition *out* in the phrase *carry out* as metaphorical. It also wrongly predicts *crews* as literal possibly due to that the word is a noun.

5 Discussion

Out-of-vocabulary words are treated differently in the training phase and in the test phase. In the training phase, the loss of metaphor interpretation over out-of-vocabulary words is not computed according to Equation (7). In the test phase, the evaluation on metaphor detection involves all words, but the evaluation on metaphor interpretation only targets in-vocabulary words. For the metaphor detection task, in-vocabulary words and out-of-vocabulary words have no different treatments. We have conducted extra experiments for metaphor detection on the VUA ALL POS track. The results reported in Table 9 show that MDGI-Joint achieves significantly higher performance on in-vocabulary verbs (77.3%) than on out-of-vocabulary verbs (73.9%), but on all out-of-vocabulary words, MDGI-Joint achieves similar performance (77.2%).

Due to the limited glosses extracted from existing dictionaries, there will be correct glosses for novel metaphors that do not appear in the training set. It has been shown (Rai and Chakraverty, 2020) that interpreting novel metaphors in general situations is difficult. Hence our proposed gloss-based interpretations are definitely more suitable for conventional or lexical metaphors. Nevertheless, by considering that we can expand glosses with more external resources, the effectiveness of our proposed approach is not limited to a fixed set of metaphors.

6 Conclusion and Future Work

Metaphor detection is of great value in many natural language processing tasks. In this paper we have utilized gloss-based metaphor interpretation to enhance metaphor detection. The novelty mainly lies in the interpretation mechanism, *i.e.*, utilizing glosses to interpret metaphorical words. Accordingly, we propose a joint model for metaphor detection and gloss-based interpretation. We enhance three datasets in the field of metaphor detection to evaluate the joint model. Experimental results confirm that metaphor interpretation in gloss improves the performance of metaphor detection. Future work will extend our approach to other rhetoric identification tasks.

Acknowledgments

This work is supported by the Guangdong Province Science and Technology Plan projects (2017B010110011), the National Natural Science Foundation of China (No. 61876204, 61976232, and 51978675), the National Key R&D Program of China (No.2018YFC0830600), Guangdong Province Natural Science Foundation (No. 2018A030313086), All-China Federation of Returned Over-seas Chinese Research Project (No. 17BZQK216).

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*, pages 329–336.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *ACL*, pages 1006–1017.

Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015. Knowledge base unification via sense embeddings and disambiguation. In *EMNLP*, pages 726–736.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *EACL*, pages 523–528.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *CoNLL*, pages 55–64.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *EMNLP*, pages 594–602.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *NoDaLiDa*, pages 271–276.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *EMNLP*, pages 607–613.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *WMT*, pages 11–19.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Penstein Rosé. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *ACL*, pages 216–225.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Penstein Rosé. 2015. Metaphor detection in discourse. In *SIGDIAL*, pages 384–392.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. *arXiv preprint arXiv:1606.03568*.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *TSD*, volume 3206, pages 103–112.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutiérrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *ACL*, pages 101–106.

Kövecses and Zoltán. 2002. *Metaphor: A Practical Introduction*. Oxford University Press.

Luuk Lagerwerf and Anoe Meijers. 2008. Openness in metaphorical and straightforward advertisements: Appreciation effects. *Journal of Advertising*, 37(2):19–30.

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

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

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of mandarin chinese. *Lang. Resour. Evaluation*, 51(3):663–694.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *ACL*, pages 1222–1231.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *ACL*, pages 3888–3898.

Fernando Marmolejo-Ramos, María Rosa Elosúa, Yuki Yamada, Nicholas Francis Hamm, and Kimihiro Noguchi. 2013. Appraisal of space words and allocation of emotion words in bodily space. *PLoS One*, 8(12):e81688.

James H Martin. 2006. A corpus-based analysis of context effects on metaphor comprehension. *Trends in Linguistics Studies and Monographs*, 171:214.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *CoNLL*, pages 51–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *SemEval@NAACL-HLT*, pages 288–297.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Learning outside the box: Discourse-level features improve metaphor identification. In *NAACL-HLT*, pages 596–601.

S. Nacey, A.G. Dorst, T. Krennmayr, and W.G. Reijnierse. 2019. *Metaphor Identification in Multiple Languages: MIPVU around the world*, volume 22. John Benjamins Publishing Company.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Steven Neale, Luís Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *LREC*, pages 2777–2783.

Andrew Ortony. 1980. Understanding metaphors. *Center for the Study of Reading Technical Report; no. 154*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT, Demonstrations*, pages 48–53.

Gözde Özbal, Carlo Strapparava, Serra Sinem Tekiroglu, and Daniele Pighin. 2016. Learning to identify metaphors from a corpus of proverbs. In *EMNLP*, pages 2060–2065.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *SENSEVAL@ACL*, pages 21–24.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *EMNLP*, pages 1156–1167.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *EMNLP*, pages 1537–1546.

Elena Semino. 2010. Descriptions of pain, metaphor, and embodied simulation. *Metaphor and Symbol*, 25(4):205–226.

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *COLING*, pages 1121–1130.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *NAACL-HLT*, pages 160–170.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *COLING*, pages 1002–1010.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Sidney Siegel. 1956. Nonparametric statistics for the behavioral sciences.

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

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Eng. Appl. Artif. Intell.*, 48:188–203.

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 *Fig-Lang*, pages 30–39.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *ACL*, pages 248–258.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *EMNLP*, pages 680–690.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *HLT/EMNLP*, pages 771–778.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, and Yongfeng Huang. 2018. Neural metaphor detecting with cnn-lstm model. In *Fig-Lang*, pages 110–114.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. pages 1374–1385.