# CATE: A Contrastive Pre-trained Model for Metaphor Detection with Semi-supervised Learning

**Zhenxi Lin**[1,2], **Qianli Ma**[1*], **Jiangyue Yan**[1], **Jieyu Chen**[3]
[1]School of Computer Science and Engineering,
South China University of Technology, Guangzhou, China
[2]Tencent Jarvis Lab, Shenzhen, China
[3]Department of English and Communication, The Hong Kong Polytechnic University
zhenxi_lin@foxmail.com, qianlima@scut.edu.cn
jiangyue9606@gmail.com, 18043507r@connect.polyu.hk

## Abstract

Metaphors are ubiquitous in natural language, and detecting them requires contextual reasoning about whether a semantic incongruence actually exists. Most existing work addresses this problem using pre-trained contextualized models. Despite their success, these models require a large amount of labeled data and are not linguistically-based. In this paper, we proposed a **C**ontr**A**stive pre-**T**rained mod**E**l (CATE) for metaphor detection with semi-supervised learning. Our model first uses a pre-trained model to obtain a contextual representation of target words and employs a contrastive objective to promote an increased distance between target words' literal and metaphorical senses based on linguistic theories. Furthermore, we propose a simple strategy to collect large-scale candidate instances from the general corpus and generalize the model via self-training. Extensive experiments show that CATE achieves better performance against state-of-the-art baselines on several benchmark datasets.

## 1 Introduction

Conceptual metaphors are figurative languages widely used in our daily communication, implying a mapping between two conceptual domains (Lakoff and Johnson, 2008). At a linguistic level, metaphor is defined as a linguistic expression representing other concepts rather than taking literal meanings of words in context (Lagerwerf and Meijers, 2008). For instance, in the sentence "I have *digested* all this information," the word *digested* does not literally mean converting food into absorbable substances. Instead, this word means "*arrange and integrate in the mind*" in the context.[1] This metaphor conceptualizes the concept of **ideas** in terms of the properties of **food**. Metaphorical associations as such are broad generalizations that allow us to project knowledge and inferences across domains and are beneficial for various downstream NLP applications, such as machine translation (Shi et al., 2014), sentiment analysis (Cambria et al., 2017; Dankers et al., 2019), and dialogue systems (Dybala and Sayama, 2012).

Given the prevalence of metaphors in human communication, the effective detection of metaphors plays an essential role in natural language understanding. Hence, many efforts have been devoted to metaphor detection (MD), which aims to identify metaphorical expressions in a text automatically. Most previous methods (Mason, 2004; Turney et al., 2011; Tsvetkov et al., 2014; Shutova et al., 2016) for MD are based on various hand-crafted linguistic features and rely on manually annotated resources to extract them. Recently, significant progress has been made in applying deep learning techniques for MD (Wu et al., 2018; Gao et al., 2018; Mao et al., 2019; Rohanian et al., 2020; Le et al., 2020). These methods directly embed textual semantic information into a low-dimensional space by deep neural networks. Nevertheless, these methods are unable to model the multiple meanings of polysemous words in context (Choi et al., 2021). With the rapid development of contextualized representations, a number of methods (Su et al., 2020; Chen et al., 2020; Choi et al., 2021) adopt pre-trained language models to effectively capture context-dependent information with respect to the target words and fine-tune them to obtain state-of-the-art performances for MD.

Although these pre-trained models have achieved promising results, several problems remain unsolved. First, the current models lack the discrimination between the literal meaning and non-literal meaning of the target words, which can be enhanced by analogical comparison in the specific context based on Metaphor Identification Procedure (MIP) (Praglejaz Group, 2007). Second, one challenge for fine-tuned language

*Corresponding author

[1]http://wordnetweb.princeton.edu/perl/webwn?s=digest

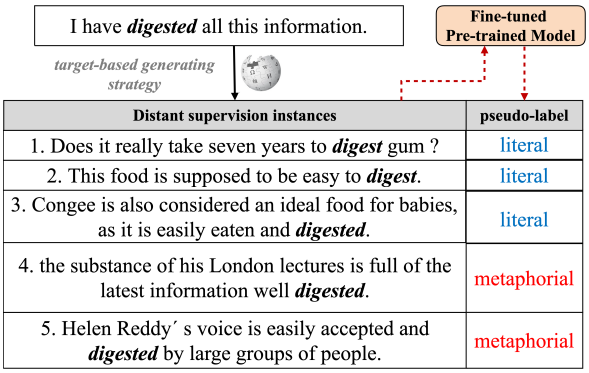


*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898
November 7–11, 2021. 

models is they still require large amounts of labeled data for obtaining state-of-the-art performances on downstream tasks (Du et al., 2020; Yu et al., 2020; Karamanolakis et al., 2021). However, due to the expensive and labor-intensive labeling, existing public MD datasets are relatively small. In addition, labeling metaphorical words can be influenced by subjective input and may need expert knowledge (Tsvetkov et al., 2014), which poses a significant challenge for metaphor detection.

The above challenges motivate us to propose a **C**ontr**A**stive Pre-**T**rained Mod**E**l (**CATE**) for metaphor detection, using a contrastive objective to model the distance between the target word's literal and metaphorical senses, enhancing the model generalization performance via self-training with unlabeled data generated by a simple strategy. Firstly, we utilize pre-trained models (i.e., BERT and RoBERTa) to capture contextual information about a target word in the sentence. If the target word is a metaphor, its semantic meaning is context-specific and different from its literal meaning. The word's literal meaning can be described through non-metaphorical instances. Therefore, we incorporate a contrastive objective to enhance contextual representations between the literal and metaphorical meaning of a target word to make it more distinguishable, in which way the classifier can make a more informed decision. To address the label scarcity issue, we propose a simple target-based generating strategy to automatically generate training data inspired by a distantly supervised paradigm (Mintz et al., 2009; Hoffmann et al., 2011). Concretely, if a given word serves as the detection target in a sentence, all sentences containing this word in a specific corpora are retrieved and regarded as candidate instances. To expand the training data, we use the pre-trained model to generate pseudo-labels for these candidate instances and incorporate them into training data, where the pre-trained model is first fine-tuned on the original training set, as shown in Figure 1. We update the pseudo-labels and the model iteratively by self-training for improving the generalization power.

In summary, the contributions of this paper are as follows: (1) We propose a novel pre-trained model with a contrastive objective for capturing the semantic incongruence in metaphors based on MIP linguistic theories. (2) To our best knowledge, this is the first attempt to combine semi-supervised learning with self-training to alleviate the label scarcity issue for MD. (3) Empirically, we perform experiments on widely used datasets to verify the effectiveness of our approach. Experimental results show that our approach obtains state-of-the-art performance over several benchmark datasets.




| Distant supervision instances | pseudo-label |
|---|---|
| 1. Does it really take seven years to ***digest*** gum ? | literal |
| 2. This food is supposed to be easy to ***digest***. | literal |
| 3. Congee is also considered an ideal food for babies, as it is easily eaten and ***digested***. | literal |
| 4. the substance of his London lectures is full of the latest information well ***digested***. | metaphorial |
| 5. Helen Reddy´ s voice is easily accepted and ***digested*** by large groups of people. | metaphorial |

Figure 1: Some candidate instances are obtained by target-based generating strategy.

## 2 Related Work

Early approaches mainly use a variety of linguistic features to detect metaphors, such as Part of Speech, unigrams (Klebanov et al., 2014), concreteness/abstractness (Turney et al., 2011; Tsvetkov et al., 2014), WordNet supersenses (Klebanov et al., 2016), and sensory features (Tekiroğlu et al., 2015; Shutova et al., 2016), etc. They rely heavily on numerous carefully designed feature engineering.

In recent years, various models have been widely used in MD based on end-to-end neural architectures. Wu et al. (2018) reformat the MD task as a sequence labeling problem and combine CNN and LSTM layers with ensemble learning to generate the best performance in the NAACL-2018 metaphor shared task (Leong et al., 2018). Subsequently, Gao et al. (2018) presented simple BiLSTM augmented with contextualized word representation, which achieved better results. Mao et al. (2019) further adopted two linguistic theories on top of the structure of (Gao et al., 2018). In addition, some approaches employed multi-task learning to transfer knowledge from the related tasks and resources to improve the performance of MD (Do Dinh et al., 2018; Dankers et al., 2019; Rohanian et al., 2020; Le et al., 2020). These neural models are capable of properly capturing the relations between metaphors and their contexts without linguistic analyses. However, the superficial structures make them difficult to represent different aspects of words in context.

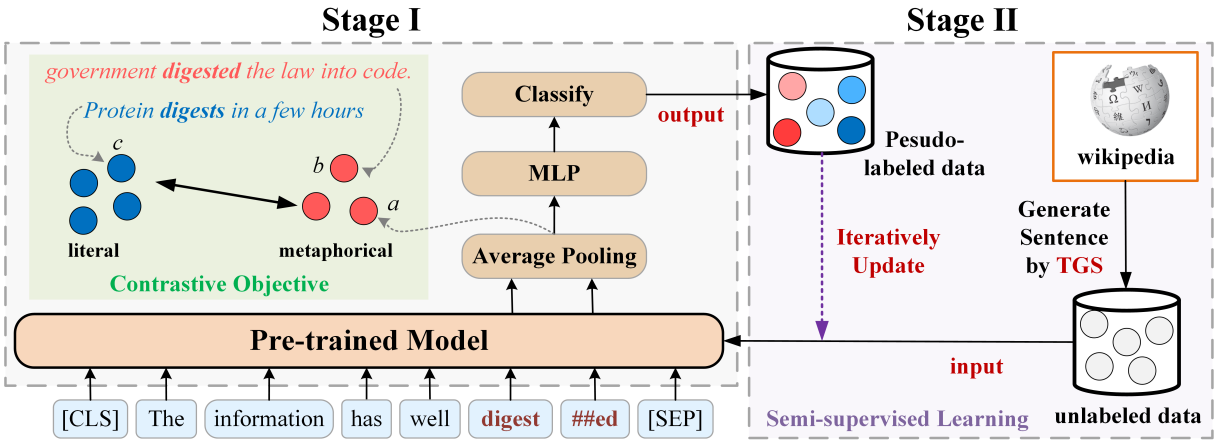



Figure 2: The diagram of CATE model with two stages. In stage I, the proposed pre-trained model is fine-tuned with labeled data using a contrastive objective. In stage II, we design a target-based generating strategy (TGS) to collect unlabeled data and adopt self training to iteratively augment the training data by generating pesudo-labels.

Recently proposed pre-trained language models (Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019) have shown dramatic improvements on several NLP tasks with appropriate fine-tuning. Therefore, some efforts (Maudslay et al., 2020; Gong et al., 2020; Su et al., 2020; Choi et al., 2021) are made to leverage the strong expressive power of pre-trained models, such as BERT, RoBERTa, to effectively capture general semantics and context-dependent information of target words for improving the performance of metaphor detection. Despite their success, one bottleneck for fine-tuning pre-trained models is the requirement of labeled data. When labeled data are scarce, the fine-tuned models often suffer from degraded performance, and the large number of parameters can lead to severe overfitting (Xie et al., 2019; Du et al., 2020; Yu et al., 2020). However, it is time-consuming and human-intensive to manually annotate large-scale training data for MD.

## 3 Proposed Method

The MD task is to predict whether a target word in a given sentence is metaphorical or literal. Some previous work (Wu et al., 2018; Gao et al., 2018; Mao et al., 2019) regards metaphor detection as a sequence labeling task that predicts the metaphoricity of each word in a given sentence. Nevertheless, this format introduces the noise of treating all non-target words as literal, which negatively impacts the model learning the difference between literal and metaphorical words (Mao et al., 2019). In this paper, we convert the MD task as a classification task based on the target word, like (Le et al., 2020; Choi et al., 2021). Formally, given a sentence $S = \{w_1, w_2, ..., w_n\}$ with $n$ words and a target word $w_t \in S$, the task involves predicting a binary label $l_t \in \{0, 1\}$ to indicate the metaphoricity (i.e., metaphorical or literal) of the target word $w_t$. Figure 2 gives an overview of CATE.

### 3.1 Pre-trained Model for MD

Given a sentence $S$ with target word $w_t$, our model leverages the power of BERT as a sentence encoder, which is particularly attractive to this task due to its strong expressive power to capture general semantics and contextual information effectively. Following (Devlin et al., 2018), we insert two special tokens '[CLS]' and '[SEP]', at the beginning and end of the input sentence, respectively. We feed the sentence $S$ with two special tokens into the BERT backbone to obtain the final hidden states $\mathbf{H}$:

$$\mathbf{H} = \text{BERT}([\text{CLS}], w_1, w_2, ..., w_n, [\text{SEP}]). \quad (1)$$

Our goal is to identify whether the semantic meaning of the target word $w_t$ within the sentence $S$ is metaphorical or not. We should calculate the context-specific representation of $w_t$ to classify. The pre-training models (e.g., BERT) usually employ the WordPiece techniques (Wu et al., 2016; Radford et al., 2019) to tokenize the word to reduce the size of the vocabulary so that a word may be divided into multiple word pieces. For example, the word *digested* is segmented into two word pieces "digest" and "##ed". Hence, we use the average operation to obtain a fixed-sized feature vector. Assuming that the hidden states corresponding to the subwords of the target word $w_t$ are from $\boldsymbol{h}_i$ to $\boldsymbol{h}_j$,

we average these hidden states:

$$\boldsymbol{c} = \frac{1}{j-i+1}\sum_{k=i}^{j}\boldsymbol{h}_k, \quad (2)$$

where $\boldsymbol{c}$ is the contextualized feature of target word $w_t$. Then we feed $\boldsymbol{c}$ into an MLP layer with tanh activation function and a softmax layer to predict the metaphoricity of the target word $w_t$. This process can be mathematically formalized as follows:

$$\boldsymbol{p} = \text{Softmax}(\mathbf{W}_2(\tanh(\mathbf{W}_1\boldsymbol{c} + \mathbf{b}_1) + \mathbf{b}_2), \quad (3)$$

where $\mathbf{W_1} \in \mathbb{R}^{d\times d}, \mathbf{b}_1 \in \mathbb{R}^d, \mathbf{W}_2 \in \mathbb{R}^{2\times d}$, and $\mathbf{b}_2 \in \mathbb{R}^d$ ($d$ is the hidden state size from BERT). The parameters are updated by minimizing the cross-entropy loss between the true label $\boldsymbol{y}$ and the metaphoricity distribution $\boldsymbol{p}$:

$$\mathcal{L}_{cls} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{y}_m \log(\boldsymbol{p}_m), \quad (4)$$

where $M$ is the number of instances in the dataset.

### 3.2 Contrastive Objective

Metaphor Identification Procedure (MIP) dictates that a word is identified as a metaphor if the literal meaning of a word contrasts with the meaning that word adopts in this context (Pragglejaz Group, 2007). According to MIP, the contrast between the contextual and literal meaning of a word serves as an important criterion for detecting its metaphoricity. Although some work (Mao et al., 2019; Choi et al., 2021) has attempted to explore the contrastive relationship between literal and contextual meaning corresponding to target word by simply concatenating the semantic features extracted from different branches of models, it remains to be unclear whether this contrastive relationship is effectively modeled.

This section explicitly incorporates a contrastive objective to capture this contrastive relationship, making the classifier more distinguishable. The objective enables the metaphorical instances of a target word to have closer semantic representations and keep literal instances separated. As shown in the shaded green part in Figure 2, the target word "digest" in both instances *a* and *b* is metaphorical and means "arrange and integrate some information in the mind", rather than its literal meaning "converting food into absorbable substances" in instance *c*. Therefore, we expect the contextual representation of the target word "digest" in sentences *a* and *b* to be more similar, and be far away from the representation in sentence *c*.

Formally, given a sentence $S_a$ with target word $w_t$ as an anchor, $S_p$ is a *positive* example with target word $w_t$ belonging to the same class as $S_a$ in batch $\mathcal{B}$, while $S_n$ is a *negative* example with target word $w_t$ belonging to another class in batch $\mathcal{B}$. We calculate their contextualized features $\boldsymbol{c}_a$, $\boldsymbol{c}_p$ and $\boldsymbol{c}_n$ by Eq. (2), respectively. The contrastive objective is defined:

$$\mathcal{L}_{co} = \sum_{(a,p,n)\in\mathcal{B}} d(\boldsymbol{c}_a, \boldsymbol{c}_p) + [\gamma - d(\boldsymbol{c}_a, \boldsymbol{c}_n)]_+, \quad (5)$$

where $[\cdot]_+$ denotes the function $f(x) = \max(0, x)$; $d(\cdot,\cdot)$ denotes the L2-normalized euclidean distance; $\gamma$ controls the margin.

This loss means capturing similarities between examples of the same class and contrasting them with examples from other classes. When the samples are from different classes (that is, one is metaphorical and the other is literal), the contrastive loss increases the distance between them and keep them apart by at least a margin $\gamma$. Modelling the distance in embedding space between the target word's literal and metaphorical semantics is an important characteristic for metaphor detection.

### 3.3 Semi-supervised Learning

The scarcity of labeled data is another challenge for MD. Currently, only relatively small training sets are available for MD, and labeling metaphorical words requires manual efforts from metaphor experts, which is time-consuming and labor-intensive. Although recent advances on pre-trained models reduce the annotation workload, they still require large amounts of labeled data to avoid overfitting (Du et al., 2020). In this section, we propose a simple strategy called Target-based Generating Strategy (TGS) to construct a large-scale training dataset with no need of metaphor experts or sophisticated pre-defined rules.

**Target-based Generating Strategy (TGS)** The TGS is based on a heuristic process that if a word serves as the detection target in a sentence, all other sentences containing this word in a specific corpus serve as potential candidate instances. This strategy effectively obtains a large-scale candidate set $\mathcal{U}$ based on the target words in the labeled data as heuristic seeds, which can cover more topics without any special manual design. It is natural

to use the fine-tuned model to predict the labels of candidate instances and then select high-confidence samples as the expanded data, but this way relies on the performance of the pre-trained model, which may lead to prediction bias and introduce noise.

**Self-Training (ST)** To alleviate the noise in $\mathcal{U}$, we adopt self-training (Rosenberg et al., 2005; Lee et al., 2013) to generate pseudo-labels for the candidate instances by the fine-tuned model and incorporate them into the training set, with which the pseudo-labels and the model are updated in an iterative manner. There are two alternatives for generating the pseudo-labels for candidate instances, namely hard labeling (Lee et al., 2013) and soft labeling (Xie et al., 2016). Hard labeling selects the highest-confidence prediction as the class label for each instance, which is prone to cause error propagation when having the wrong prediction (Yu et al., 2020). Alternatively, we choose to generate soft pseudo-labels $\hat{\boldsymbol{y}}_i \in \mathbb{R}^K$ for each instance $u_i \in \mathcal{U}$:

$$\hat{\boldsymbol{y}}_{ij} = \frac{\boldsymbol{p}_{ij}^2 / f_j}{\sum_{j'} \boldsymbol{p}_{ij'}^2 / f_{j'}}, \tag{6}$$

where $f_j = \sum_i \boldsymbol{p}_{ij}$ is the sum over soft frequencies of class $j$, $\boldsymbol{p}_{ij}$ is $j$-th class prediction of $u_i$. Eq. (6) derives $\hat{\boldsymbol{y}}_i$ by strengthening high-confidence predictions while reducing low-confidence ones via squaring and normalizing the current predictions, and it retains more information than hard labels. We define the ST objective as a KL-divergence loss between the pseudo-labels distributions $\hat{Y}$ and the model's current prediction $P$:

$$\mathcal{L}_{st} = \mathrm{KL}(\hat{Y}||P) = \sum_{i=1}^{|\mathcal{U}|} \sum_{j=1}^{K} \hat{\boldsymbol{y}}_{ij} \log \frac{\hat{\boldsymbol{y}}_{ij}}{\boldsymbol{p}_{ij}}. \tag{7}$$

### 3.4 Traning Procedure of CATE

The overall objective function of CATE includes contrastive loss $\mathcal{L}_{co}$, classification loss $\mathcal{L}_{cls}$ for labeled data and KL loss for unlabeled data $\mathcal{U}$:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{co} + \beta \mathcal{L}_{st}, \tag{8}$$

where $\alpha$ and $\beta$ are hyperparameters for balancing the strength of the contrastive loss and KL loss, respectively. CATE includes a two-stage training procedure: In the first stage, we fine-tune the pre-trained model with the *first two terms* of Eq. (8) using the labeled data, which can significantly learn contrastive relationship in metaphors and improve the quality of prediction for MD. Then we use the fine-tuned model to predict the *soft* pseudo-labels for all unlabeled data collected by TGS. In the second stage, we apply a self-training strategy to augment the training data with pseudo-labeled data and update the pre-trained model in an iterative manner. During self-training, we iteratively compute soft pseudo-labels based on current predictions and refine model parameters with Eq. (8). The procedures are summarized in Algorithm 1.

**Algorithm 1:** Training Procedure of **CATE**

**Input:** labeled instances $\mathcal{S}$; candidate instances $\mathcal{U}$ collected by GTS; Pre-trained Model $f(\cdot;\theta)$.

*// Stage I: fine-tune model with labeled data.*

Update $\theta$ on $\mathcal{S}$ by **first two terms** in Eq. (8).

*// Stage II: refine model with unlabeled data.*

**for** $t = 1, 2, ..., T$ **do**

  Generate pesudo-labels for $\mathcal{U}$ by Eq. (6).

  Update $\theta$ on $\mathcal{S}$ and $\mathcal{U}$ by Eq. (8).

**end**

**Output:** The final fine-tuned model $f(\cdot, \theta)$.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets** To evaluate the effectiveness of our model, we conduct experiments on three widely-studied datasets: (1) VUA (Steen, 2010) is currently the largest publicly available dataset used by NAACL-2018 Metaphor Shared Task. Follow previous work (Gao et al., 2018; Mao et al., 2019), we examine our model on two tracks, i.e., VUA ALL POS and VERB metaphor detection. (2) MOH-X (Mohammad et al., 2016) is a verb metaphor detection dataset that only a single target verb is labeled in each sentence. The sentences are sampled from WordNet. (3) TroFi (Birke and Sarkar, 2006) is also a verb metaphor detection dataset, and the sentences are extracted from the 1987-89 Wall Street Journal Corpus Release 1. Statistics of these datasets are listed in Table 1.

**Baselines** we compare CATE against state-of-the-art baselines in metaphor detection, including **RNN_CLS** (Gao et al., 2018): a classification model combining attention-based BiLSTM and ELMo embedding. **RNN_SEQ_ELMo** and **RNN_SEQ_BERT** (Gao et al., 2018): a sequence labeling model with attention-based BiLSTM combining the ELMo embedding and BERT embedding, respectively. **RNN_HG** (Mao et al., 2019):

| Dataset | #Tokens | %Meta. | #Sent. | Avg. Len |
|---|---|---|---|---|
| VUA-ALL$_{tr}$ | 116,622 | 11.2 | 6,323 | 18.4 |
| VUA-ALL$_{dev}$ | 38,628 | 11.6 | 1,550 | 24.9 |
| VUA-ALL$_{te}$ | 50,175 | 12.4 | 2,694 | 18.6 |
| VUA-VERB$_{tr}$ | 15,516 | 27.9 | 7,479 | 20.2 |
| VUA-VERB$_{dev}$ | 1,724 | 26.9 | 1,541 | 25.0 |
| VUA-VERB$_{te}$ | 5,873 | 30.0 | 2,694 | 18.6 |
| MOH-X | 647 | 48.7 | 647 | 8.0 |
| TroFi | 3,737 | 43.5 | 3,737 | 28.3 |

Table 1: Detailed dataset statistics. **#Tokens**: the number of target tokens whose metaphoricity is to be identified. **%Meta.**: the percentage of metaphoric tokens among target tokens. **#Sent.**: the number of sentences. **Avg. Len**: the average length of sentences.

a sequence labeling model that concatenates the GloVe embedding and hidden states from BiLSTM based on MIP principle. **RNN_MHCA** (Mao et al., 2019): a sequence labeling model that utilizes multi-head attention to capture the contextual representations based on SPV principle. **MUL_GCN** (Le et al., 2020): joint learning metaphor detection with word sense disambiguation, and utilize GCN to capture important context words. **BERT+MWE_GCN** (Rohanian et al., 2020): an attention-guided GCN that encodes syntactic dependencies alongside information about the existence of verb multiword expressions. **DeepMet** (Su et al., 2020): utilize pre-trained transformer to encode global and local context and incorporate with various linguistic features. **MelBERT** (Choi et al., 2021): utilize RoBERTa as backbone and model the contextual meaning and literal meaning based on siamese architecture.

**Implementation Details** In experiment, we first collect a target word set in all datasets as triggers and use TGS to recall large-scale target-related candidate instances from the common corpus for semi-supervised learning. We use Wikipedia as the knowledge base because it contains a wide variety of domains which makes it an ideal general-purpose corpus and is usually easily and cheaply accessible. We extract and filter text from the English Wikipedia dump[†] to construct a large-scale candidate set and apply the NLTK package (Bird et al., 2009) to turn documents into sentences and perform deduplication. Besides, we filter sentences longer than 150 words due to potential noise and memory limitations.

Following (Su et al., 2020; Choi et al., 2021), we use RoBERTa (Liu et al., 2019) as the realization of BERT. The number of transformer layers is 12, and the hidden size is 784. We use AdamW (Peters et al., 2019) optimizer with a learning rate of 3e-5 to update the parameters. The number of training epochs is 5, and the batch size is 32. The margin $\gamma$ in contrastive loss is set to 1.0. The hyper-parameters $\alpha$ and $\beta$ are set to 0.2 and 0.05, respectively. We perform 10-fold cross-validation on MOH-X and TroFi and split the VUA datasets into training, validation, and test sets the same as the previous work (Gao et al., 2018; Mao et al., 2019) for the fair comparison.

### 4.2 Overall Results

We report the results in Table 2 in terms of accuracy, precision, recall, and F1-score, where F1-score is the main measurement for metaphor detection (Mao et al., 2019). We can found that CATE achieves strong performance on all datasets, is superior to existing models on 3 out of 4 datasets in terms of F1-score (improved by 0.5%, 4.5% and 1.3% compared with the previous best model in VUA ALL POS, MOH-X and TroFi, respectively), and achieves similar performance on VUA VERB with MelBERT. Noteworthily, DeepMet and MelBERT additionally utilize linguistic features, such as POS features in their model, while CATE does not use any linguistic features. Meanwhile, it can be observed that the improvement of our model is more obvious on small-scale datasets (i.e., MOH-X and TroFi). The reason is that the massive parameters in the pre-trained model easily lead to overfitting of the model when only relatively small training sets are available. However, CATE can make full use of a large number of unlabeled data collected by the proposed target-based generating strategy and improve the model generalization by self-training. Compared with RNN_HG, which also considers the MIP principle, our model significantly outperforms it because ours explicitly captures the contrast between the literal and metaphorical meaning of target words by a contrastive objective. Not surprisingly, the approaches based on pre-trained language models (e.g., CATE, MelBERT, DeepMet) are consistently superior to the RNN-based models (e.g., RNN_CLS, RNN_HG, RNN_MHCA) due to the strong expressive power of pre-trained models to encode rich semantic and contextual information into the representations.

[†] https://dumps.wikimedia.org/enwiki/20210201/

| Model | VUA ALL POS | | | | VUA VERB | | | | MOH-X(10-fold) | | | | TroFi(10-fold) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc |
| RNN_CLS | - | - | - | - | 53.4 | 65.6 | 58.9 | 69.1 | 75.3 | 84.3 | 79.1 | 78.5 | 68.7 | 74.6 | 72.0 | 73.7 |
| RNN_SEQ_ELMo | 71.6 | 73.6 | 72.6 | 93.1 | 68.2 | 71.3 | 69.7 | 81.4 | 79.1 | 73.5 | 75.6 | 77.2 | 70.7 | 71.6 | 71.1 | 74.6 |
| RNN_SEQ_BERT | 71.5 | 71.9 | 71.7 | 92.9 | 66.7 | 71.5 | 69.0 | 80.7 | 75.1 | 81.8 | 78.2 | 78.1 | 70.3 | 67.1 | 68.7 | 73.4 |
| RNN_HG | 71.8 | 76.3 | 74.0 | 93.6 | 69.3 | 72.3 | 70.8 | 82.1 | 79.7 | 79.8 | 79.8 | 79.7 | 67.4 | **77.8** | 72.2 | 74.9 |
| RNN_MHCA | 73.0 | 75.7 | 74.3 | 93.8 | 66.3 | **75.2** | 70.5 | 81.8 | 77.5 | 83.1 | 80.0 | 79.8 | 68.6 | 76.8 | 72.4 | 75.2 |
| MUL_GCN | 74.8 | 75.5 | 75.1 | 93.8 | 72.5 | 70.9 | 71.7 | 83.2 | 79.7 | 80.5 | 79.6 | 79.9 | 73.1 | 73.6 | 73.2 | 76.4 |
| BERT+MWE-GCN | - | - | - | - | - | - | - | - | 80.0 | 80.4 | 80.2 | 80.5 | 73.8 | 71.8 | 72.8 | 73.5 |
| DeepMet | **82.0** | 71.3 | 76.3 | - | **79.5** | 70.8 | 74.9 | - | - | - | - | - | - | - | - | - |
| MelBERT | 80.1 | 76.9 | 78.5 | - | 78.7 | 72.9 | **75.7** | - | - | - | - | - | - | - | - | - |
| CATE w/o CO | 81.7 | 75.4 | 78.4 | **94.8** | 79.0 | 71.6 | 75.1 | **85.8** | 83.3 | 83.8 | 83.3 | 83.9 | 73.3 | 74.9 | 74.0 | 77.1 |
| CATE w/o ST | 78.8 | 78.7 | 78.7 | 94.7 | 77.0 | 73.8 | 75.4 | 85.5 | 84.1 | 82.0 | 82.7 | 83.5 | 72.9 | 74.9 | 73.6 | 76.7 |
| CATE | 79.3 | **78.8** | **79.0** | **94.8** | 78.1 | 73.2 | 75.6 | **85.8** | **85.7** | **84.6** | **84.7*** | **85.2** | **74.4** | 74.8 | **74.5*** | **77.7** |

Table 2: Experimental results of on three metaphor detection benchmarks. The best performance is in **bold** and the second best performance is underlined. * denotes $p < 0.01$ for a two-tailed t-test against the best baseline.

| Genre | Model | P | R | F1 | Acc |
|---|---|---|---|---|---|
| **Academic** | RNN_ELMo | 78.2 | 80.2 | 79.2 | 92.8 |
| | RNN_BERT | 76.7 | 76.0 | 76.4 | 91.9 |
| | RNN_HG | 76.5 | **83.0** | 79.6 | 92.7 |
| | RNN_MHCA | 79.6 | 80.0 | 79.8 | 93.0 |
| | DeepMet | 88.4 | 74.7 | 81.0 | - |
| | MelBERT | 85.3 | 82.5 | 83.9 | - |
| | CATE | **88.5** | 81.0 | **84.2** | **94.1** |
| **Conversation** | RNN_ELMo | 64.9 | 63.1 | 64.0 | 94.6 |
| | RNN_BERT | 64.7 | 64.2 | 64.4 | 94.6 |
| | RNN_HG | 63.6 | **72.5** | 67.8 | 94.8 |
| | RNN_MHCA | 64.0 | 71.1 | 67.4 | 94.8 |
| | DeepMet | 71.6 | 71.1 | 71.4 | - |
| | MelBERT | 70.1 | 71.7 | 70.9 | - |
| | CATE | **72.2** | 72.2 | **72.2** | **95.8** |
| **Fiction** | RNN_ELMo | 61.4 | 69.1 | 65.1 | 93.1 |
| | RNN_BERT | 66.5 | 68.6 | 67.5 | 93.9 |
| | RNN_HG | 61.8 | 74.5 | 67.5 | 93.4 |
| | RNN_MHCA | 64.8 | 70.9 | 67.7 | 93.8 |
| | DeepMet | 76.1 | 70.1 | 73.0 | - |
| | MelBERT | 74.0 | **76.8** | 75.4 | - |
| | CATE | **77.8** | 74.1 | **75.9** | **95.7** |
| **News** | RNN_ELMo | 72.7 | 71.2 | 71.9 | 91.6 |
| | RNN_BERT | 71.2 | 72.5 | 71.8 | 91.4 |
| | RNN_HG | 71.6 | **76.8** | 74.1 | 91.9 |
| | RNN_MHCA | 74.8 | 75.3 | 75.0 | 92.4 |
| | DeepMet | 84.1 | 67.6 | 75.0 | - |
| | MelBERT | 81.0 | 73.7 | **77.2** | - |
| | CATE | **84.3** | 71.0 | 77.1 | **96.6** |

Table 3: Model performance on different genres of texts in VUA ALL POS. The best performance is in **bold** and the second best performance is underlined.

| POS | Model | P | R | F1 | Acc |
|---|---|---|---|---|---|
| **Verb** | RNN_ELMo | 68.1 | 71.9 | 69.9 | - |
| | RNN_BERT | 67.1 | 72.1 | 69.5 | 87.9 |
| | RNN_HG | 66.4 | 75.5 | 70.7 | 88.0 |
| | RNN_MHCA | 66.0 | 76.0 | 70.7 | 87.9 |
| | DeepMet | **78.8** | 68.5 | 73.3 | - |
| | MelBERT | 74.2 | **75.9** | 75.1 | - |
| | CATE | 77.1 | 74.4 | **75.7** | **90.9** |
| **Adjective** | RNN_ELMo | 56.1 | 60.6 | 58.3 | - |
| | RNN_BERT | 58.1 | 51.6 | 54.7 | 88.3 |
| | RNN_HG | 59.2 | **65.6** | 62.2 | 89.1 |
| | RNN_MHCA | 61.4 | 61.7 | 61.6 | 89.5 |
| | DeepMet | **79.0** | 52.9 | 63.3 | - |
| | MelBERT | 69.4 | 60.1 | 64.4 | - |
| | CATE | 74.4 | 59.0 | **65.8** | **91.6** |
| **Adverb** | RNN_ELMo | 67.2 | 53.7 | 59.7 | 94.8 |
| | RNN_BERT | 64.8 | 61.1 | 62.9 | 94.8 |
| | RNN_HG | 61.0 | 66.8 | 63.8 | 94.5 |
| | RNN_MHCA | 66.1 | 60.7 | 63.2 | 94.9 |
| | DeepMet | 79.4 | 66.4 | 72.3 | - |
| | MelBERT | **80.2** | 69.7 | 74.6 | - |
| | CATE | 76.9 | **74.2** | **75.5** | **95.5** |
| **Noun** | RNN_ELMo | 59.9 | 60.8 | 60.4 | - |
| | RNN_BERT | 63.3 | 56.8 | 59.9 | 88.6 |
| | RNN_HG | 60.3 | **66.8** | 63.4 | 88.4 |
| | RNN_MHCA | 69.1 | 58.2 | 63.2 | 89.8 |
| | DeepMet | 76.5 | 57.1 | 65.4 | - |
| | MelBERT | 75.4 | 66.5 | **70.7** | - |
| | CATE | **77.9** | 60.0 | 68.0 | **91.5** |

Table 4: Model performance on different open-class words in VUA ALL POS. The best performance is in **bold** and the second best performance is underlined.

### 4.3 Ablation Study

To investigate different components in CATE, we compare CATE variants without the contrastive objective (w/o CO) and without the self-training (w/o ST). As we can see from the last three lines of Table 2, each component is important for the proposed model as excluding any of them would hurt the performance significantly. When the self-training is removed, the F1-score respectively drops

by 2.0% and 0.9% on small-scale MOH-X and TroFi datasets, which demonstrates the necessity of integrating semi-supervised learning to improve the generalization performance. The contrastive objective learns the difference between the target word's literal and metaphorical semantics and is also beneficial to our model.

### 4.4 Model Analysis

**VUA Breakdown Analysis** Table 3 and 4 respectively report the breakdown of performance by different genres and open-class words based on the VUA ALL POS test dataset, which in line with Leong et al. (2018) and Mao et al. (2019). CATE shows very promising overall results against other competitive baselines in both breakdown datasets. In Table 3, all models achieve better results on Academic due to the expressions used in academic articles are formal and normative with abundant context. Particularly, CATE presents a substantial improvement in terms of F1-score against the second best with a gain of 1.3% and 0.5% on Conversation and Fiction, respectively. This is meaningful because Conversation and Fiction are more challenging and have lower F1-score than other genres due to their fragmented or rare expressions, such as *er*, *yeah*, *na*. We speculate that the reason for the improvement of CATE is that the target-based generating strategy has the ability to automatically construct diverse training data from Wikipedia containing different topics and avoid the model tend to be biased toward a specific domain. In Table 4, all models perform better results on Verb as it has the largest training instances. Properly, CATE provides strong performance on almost all open-class words and achieves large improvements against MelBERT in Verb (0.6%), Adjective (1.4%), Adverb (0.9%) in terms of F1-score.

**Embedding Visualization** In Figure 3, we visualize TroFi sample contexual embeddings in Eq. (2) for specific target words *attack* and *cool*. As shown in Figure 3 (a)(c), when the contrastive objective is removed, the literal and metaphorical representations are mixed together and indistinguishable. Based on the MIP principle, a metaphor is identified if the literal meaning of the target word contrasts with its contextual meaning. As excepted, the proposed contrastive objective explicitly extends the distance between the target word's literal and metaphorical meanings in embedding space and learns more compact representations for data

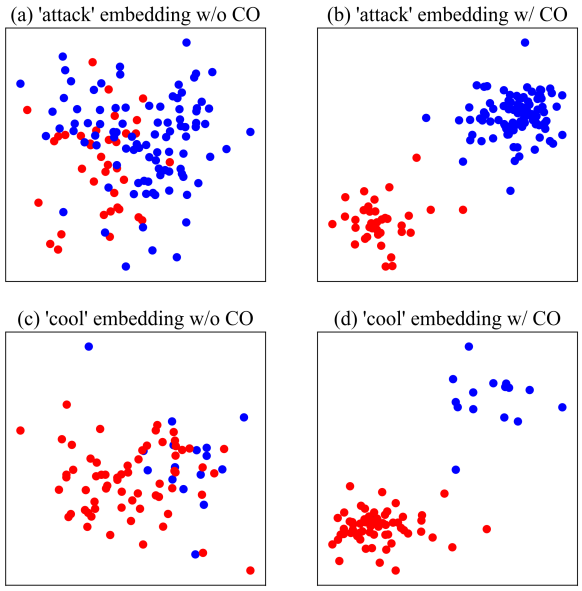



Figure 3: t-SNE visualization on TroFi for target words *attack* and *cool*. The red color denotes the metaphorical instances and blue color denotes the literal instances.

from the same class, as shown in Figure 3 (b)(d).

**Impact of available labeled data** We further investigate the effectiveness of self-training when using different ratios of supervised data. The results on MOH-X and TroFi are reported in Figure 4. As the labeled data size continues to increase, the gain of self-training gradually decreases. When little supervised data is available, self-training can be regarded as a regularizer to effectively improve the prediction ability and generalization of the model.

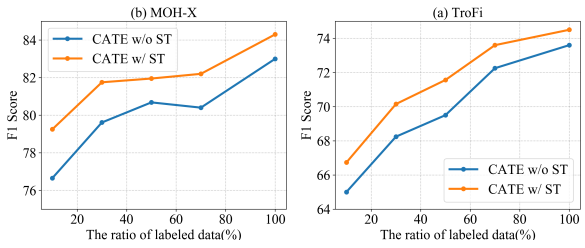



Figure 4: Effect of self-training on different proportions of supervised data on MOH-X and TroFi.

**Hyperparameters Discussion** We examine the effects of hyper-parameters $\alpha$ and $\beta$ on the MOH-X dataset, as shown in Figure 5. With the increase of parameter $\alpha$ or $\beta$, the model performance increases first and then decreases. When $\alpha$ is too large, it easily leads to overly penalize the distance and overlooks the metaphorical associations between different senses, whereas when $\beta$ is too large, it also deteriorates performance due to injecting too much noise from unlabeled data.

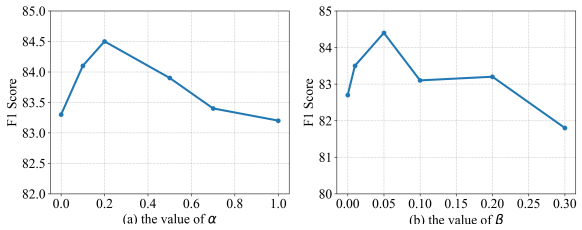



Figure 5: Impact of different $\alpha$ and $\beta$ on MOH-X.

## 5 Conclusion

This paper takes advantage of self-training and designs a simple but effective metaphor detection model based on the pre-trained backbone to capture the contextualized features. To be specific, we incorporate a contrastive objective into the model to capture the semantic incongruence in metaphors and use a simple strategy to automatically construct substantial training data ready for self-training. The evaluation on multiple benchmarks has shown that our model can achieve state-of-the-art performance. In future work, we plan to explore how to use unlabeled data more effectively and discover potentially valuable metaphor examples to reduce efforts of manual annotation.

## Acknowledgments

We thank the anonymous reviewers for their helpful feedbacks. The work described in this paper was partially funded by the National Natural Science Foundation of China (Grant No. 61502174, and 61872148), the Natural Science Foundation of Guangdong Province (Grant No. 2017A030313355, 2019A1515010768 and 2021A1515011496), the Guangzhou Science and Technology Planning Project (Grant No. 201704030051, and 201902010020), the Key R&D Program of Guangdong Province (No. 2018B010107002) and the Fundamental Research Funds for the Central Universities.

## References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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*.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Xianyang Chen, Chee Wee 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.

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. *arXiv preprint arXiv:2104.13615*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint arXiv:2010.02194*.

Pawel Dybala and Kohichi Sayama. 2012. Humor, emotions and communication: Human-like issues of human-computer interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

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.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational*

*linguistics: human language technologies*, pages 541–550.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. *arXiv preprint arXiv:2104.05514*.

Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17.

Beata Beigman Klebanov, Chee Wee Leong, E Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106.

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. 2008. *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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8139–8146.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3.

Chee Wee 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. *FIGURATIVE LANGUAGE PROCESSING*, pages 221–226.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

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.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, et al. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2890–2895.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models.

Chunqi Shi, Toru Ishida, and Donghui Lin. 2014. Translation agent: a new metaphor for machine translation. *New Generation Computing*, 32(2):163–186.

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.

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

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.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39.

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.

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.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with cnn-lstm model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2020. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. *arXiv preprint arXiv:2010.07835*.