# Label-Enhanced Hierarchical Contextualized Representation for Sequential Metaphor Identification

**Shuqun Li**[*], **Liang Yang**[†*], **Weidong He, Shiqi Zhang, Jingjie Zeng** and **Hongfei LIN**
Department of Computer Science and Technology, Dalian University of Technology, China
shuqunli@mail.dlut.edu.cn, liang@dlut.edu.cn
{heweidong,22009191,jjwind}@mail.dlut.edu.cn
hflin@dlut.edu.cn

## Abstract

Recent metaphor identification approaches mainly consider the contextual text features within a sentence or introduce external linguistic features to the model. But they usually ignore the extra information that the data can provide, such as the contextual metaphor information and broader discourse information. In this paper, we propose a model augmented with hierarchical contextualized representation to extract more information from both sentence-level and discourse-level. At the sentence level, we leverage the metaphor information of words that except the target word in the sentence to strengthen the reasoning ability of our model via a novel label-enhanced contextualized representation. At the discourse level, the position-aware global memory network is adopted to learn long-range dependency among the same words within a discourse. Finally, our model combines the representations obtained from these two parts. The experiment results on two tasks of the VUA dataset show that our model outperforms every other state-of-the-art method that also does not use any external knowledge except what the pre-trained language model contains.

## 1 Introduction

Metaphor is a type of figurative language and its essence is understanding and experiencing one kind of thing in terms of another(Lakoff and Johnson, 1980). As a common language expression, we often use metaphors to express our thoughts vividly and concisely in daily communication. For example, in the sentence *It is one of the keys for success of a commercial product*, *keys* is used to help understand the importance of *It*. However, this characteristic of metaphor makes it challenging to identify metaphors in texts. But the identification of metaphors is meaningful and can help us to understand the meaning of the texts, from which many downstream applications such as machine translation(Koglin, 2015) and opinion mining(Shutova et al., 2013) can benefit.

Recent metaphor researches(Gao et al., 2018; Mao et al., 2019), and ACL 2020 Metaphor Shared Task(Leong et al., 2020) regard it as a sequence labeling task. Although many previous works have explored ways to enhance the contextualized representation within a sentence(Gao et al., 2018; Mao et al., 2019), or to introduce some external knowledge(Rohanian et al., 2020; Chen et al., 2020; Wan and Xing, 2020), most of them do not make full use of the information in the dataset, from which the metaphor identification process may benefit.

Firstly, when considering the metaphoricity of the target word, the metaphor information of other words in the sentence can also be helpful. E.g., in the sentence *He find himself in the position of the gambler who gambled all and lost*, *gambler* and *gambled* are metaphors. The word *gambled* is the action of the *gambler*, and it's reasonable for a gambler to gamble. Thus, the model might prefer to classify *gambled* as literal. However, if we know that *gambler* is a metaphoric word, then it is obvious that *gambled* is also a metaphoric word that refers to the risky thing he did. Based on this observation, we propose a novel label-enhanced contextualized representation method to introduce the contextual metaphor information. It embeds the label of each word(i.e. metaphoric or literal) in the same space as the output of the encoder first and then takes both the output of the encoder and the label embedding as the input of a transformer(Vaswani et al., 2017). We believe it could enhance the reasoning ability of the model by attending to metaphor information of other words in the sentence. Besides, marking the metaphoric words in context could also help the target word understand the context better, especially in the com-

[*]Both authors contributed equally to this research
[†]Corresponding author



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3543
November 7–11, 2021. 

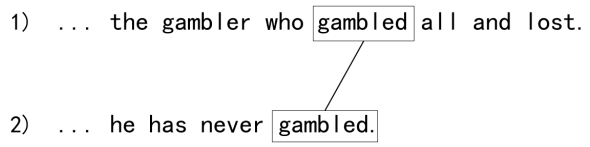


Figure 1: An example that shows the two occurrences of word *gambled* within a discourse

plicated sentence, because metaphorical words are not used as their literal meaning, increasing the difficulty of understanding the context. To the best of our knowledge, we are the first that proposes this method.

Secondly, Some existing benchmark metaphor datasets, such as VUAMC, contain sentences from long articles, and the contextual information in the articles will be very useful for metaphor identification. Some previous work used paragraph embedding(Mu et al., 2019) or neighbouring sentence representation(Dankers et al., 2020). Based on this, we use a discourse-level attention architecture that could capture both global and local features in the whole discourse for the target word. First, we introduce the work of Dankers et al. (2020) to extract local information. Then, we propose an improved method of Global Attention(Zhang et al., 2018), which is called position-aware global memory network, to represent global information of the target word. It is based on the observation that a metaphor brings another domain/frame into the discourse, so it is likely that metaphors mapping to the same domain/frame reoccur throughout the discourse, especially among the same words. Specifically, our model uses an attention mechanism between the target word and its other occurrences in the discourse. Figure 1 shows the two occurrences of word *gambled* in two sentences within a discourse.

Based on the above sentence-level and discourse-level methods, we propose a novel hierarchical contextualized representation model for metaphor identification, as shown in Figure 2. To verify the effectiveness of our model, we conduct experiments on the ALL POS and Verbs tasks of the VU Amsterdam Metaphor Corpus (VUA)(Steen, 2010). Our model outperforms several baseline models with 1.1% (VUA ALL POS) and 1.0% (VUA Verbs) improvement in F1 score. In addition, the results of our model surpass DeepMet(Su et al., 2020), which is the state-of-the-art model in metaphor identification, with the same experiment setup.

Our contributions in this paper can be summarized as follows.

- We propose a novel label-enhanced contextualized representation method to enhance the model's ability to reason about contextual metaphoric relationships and better understand the meaning of context.
- At the discourse level, we use an improved position-aware global memory network to introduce the long-range discourse information.
- Experiment results on the two tasks of the VUA dataset show that our model outperforms the state-of-the-art methods that also do not use external knowledge.

## 2 Related work

### 2.1 Metaphor Identification

Most of the early metaphor identification works employed machine learning approaches using linguistic features(Turney et al., 2011; Tsvetkov et al., 2013; Mohler et al., 2013; Klebanov et al., 2016; Bulat et al., 2017a). In recent years, neural metaphor identification has become highly popular for its end-to-end fashion and better performance. Wu et al. (2018) combined CNN and LSTM to obtain local and long-range information and achieved the best performance in the NAACL 2018 VUA Shared Task(Leong et al., 2018). Gao et al. (2018) applied the combined embedding of GloVe(Pennington et al., 2014) and ELMo(Peters et al., 2018) as the input of a Bi-LSTM, which introduced the contextualized word embedding. Based on the model of Gao et al. (2018), Mao et al. (2019) proposed RNN_HG and RNN_MHCA inspired by MIP(Group, 2007) and SPV(Wilks, 1978) theory respectively and gained certain improvements. Multi-task learning(Dankers et al., 2019; Le et al., 2020)and linguistic features(Rohanian et al., 2020; Wan and Xing, 2020) have also been explored to applied to the deep learning model. Su et al. (2020) achieved the best performance in ACL 2020 Metaphor Shared Task(Leong et al., 2020) by taking global text context, local text context, query word, General POS, and fine-grained POS as the input of a RoBERTa model(Liu et al., 2019).

There are also some works done on the relation-level metaphor identification. The early works employed machine learning models using linguistic features as well, including conceptual semantic features(Tsvetkov et al., 2014), visual fea-

tures(Shutova et al., 2016), and attribute-based semantics(Bulat et al., 2017b). The recent works mainly used deep learning model. Rei et al. (2017) proposed a supervised similarity network for relation-level metaphor identification. Zayed et al. (2020) introduced a novel architecture for identifying relation-level metaphoric expressions of certain grammatical relations based on contextual modulation, which achieved state-of-the-art results.

In this paper, we consider the token-level metaphor identification task for long discourse, and different from these previous works, we start from the known information that the data can provide, and use the label-enhanced contextualized representation to strengthen the model's reasoning ability by introducing contextual metaphor information.

### 2.2 Discourse-level Representation

Considering the datasets contain discourse information, some researchers enhanced word contextualized representation by introducing discourse features. Jang et al. (2015) used hand-crafted discourse-level features such as topical information and semantic relatedness. Mu et al. (2019) obtained the discourse contextual information by embedding the surrounding paragraph. Dankers et al. (2020) applied general attention and hierarchical attention on both the target sentence and its neighbouring sentences to get discourse representation. However, Mu et al. (2019) and Dankers et al. (2020) only considered the context close to the target word and used the same method as processing a sentence, which is not suitable for long context. To adapt to longer context of discourse, we consider the occurrences of words to avoid processing texts that are too long and capture the consistency in the use of metaphors in discourse. The work of Jang et al. (2017) had a similar idea with us, which paid attention to the similar words that appear globally in the discourse. However, their method must predefine frame and know what frame the target word belongs to. This limitation of their method in scalability makes it inapplicable to the general metaphor datasets, such as VUA.

In the field of Named Entity Recognition research, where document-level tasks are more common, there are some document representation methods that we can use for reference. Zhang et al. (2018) proposed Global Attention that establishes the relationship among the occurrences of the word within a document. Luo et al. (2020) adopted a key-value memory network to record the history hidden states. Based on their works, we propose an improved position-aware global memory network.

## 3 Methodology

### 3.1 Baseline Model

Given a sentence with a sequence of words $\{x_1, x_2, ..., x_n\}$ , our goal is to predict its metaphor label $\{y_1, y_2, ..., y_n\}$ as accurately as possible. Since many previous works(Dankers et al., 2020; Chen et al., 2020; Neidlein et al., 2020) have proven the effectiveness of the pre-trained language model in metaphor identification, we use BERT(Devlin et al., 2019) as our baseline model. Specifically, we follow the work of Dankers et al. (2020). That is, a word is considered metaphoric if any of its sub-word units tokenized by the Byte Pair Encoding(BPE) algorithm used in BERT is predicted as metaphoric. Thus, we can get the output hidden states of BERT:

$$(h_1, ..., h_n) = BERT(x_1, ..., x_n)$$

### 3.2 Discourse-level Representation

Sentences in some metaphor datasets, such as VUA, come from long texts. The semantic meaning of the sentences needs to be accurately obtained by considering the context at the discourse level. Therefore, we use hierarchical attention to extract the neighbouring sentence representation and position-aware discourse-level attention for capturing long-range dependency.

**Neighbouring sentence representation** Here we follow the work of Dankers et al. (2020). We use a context window of size $2k + 1$ sentences, which comprises $k$ preceding sentences, the target sentence, and $k$ succeeding sentences. Then they are fed into a hierarchical attention architecture(Yang et al., 2016), where the first encoder is BERT, and the second encoder is a transformer(Vaswani et al., 2017). At last, we concatenate the neighbouring representation $N$ obtained by the hierarchical attention with the output hidden states $h_i$ of BERT.

**Position-aware global memory network** To utilize the information of the whole discourse, we borrow the strategy of Global Attention(Zhang et al., 2018) to capture long-range dependency among the same words within a discourse. The main idea is to employ a global attention mechanism between the target word and other occurrences within the

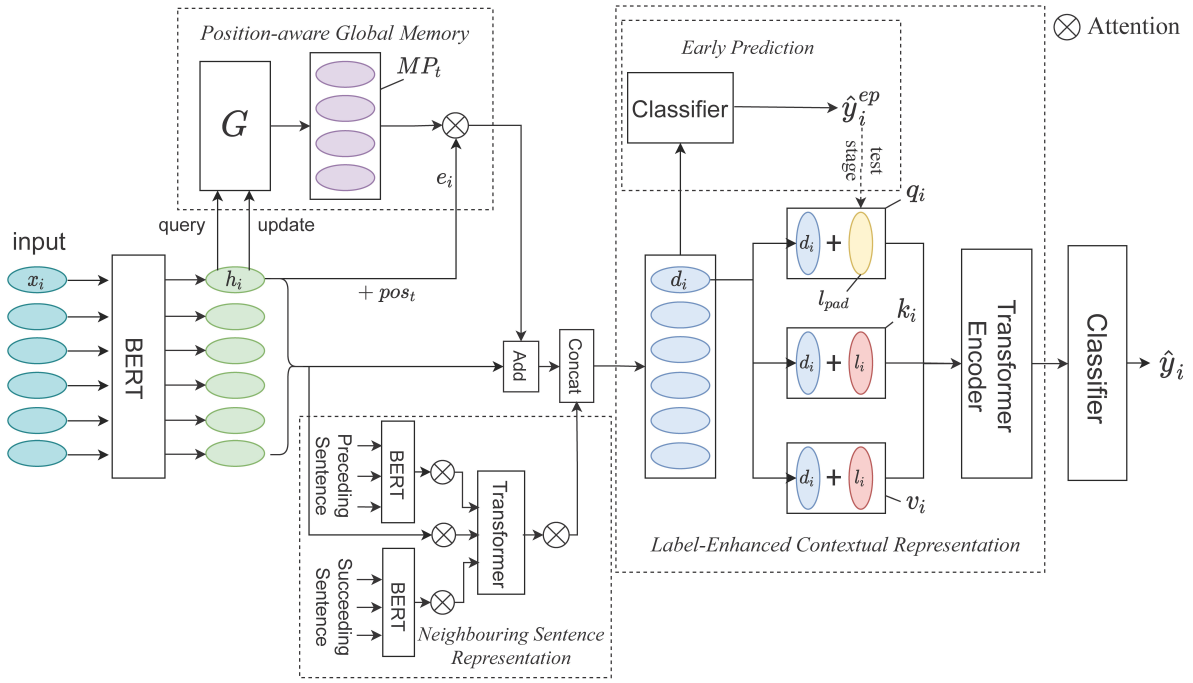


Figure 2: The overview of our model. The BERT encoder generates representations $h_i$. Then the Position-aware Global Memory Network and Neighbouring Sentence Hierarchical Attention generates discourse-level representations. The Label-Enhanced Contextualized Representation module introduced the contextual metaphor information, where Early Prediction is trained in the training phase and used in the testing phase. For clarity, only the operation of one word is shown.

discourse. Considering the time cost, we adopt the method of Luo et al. (2020), which records the history hidden states of other occurrences for each word instead of recalculating them. Thus, we call it global memory network.

Specifically, we record hidden states $h_i$ produced by the baseline model BERT for each word $x_i$ in sentences. Then, we put the hidden states of the $x_i$'s occurrences in the discourse into one group. The group containing word $x_i$ could be represent as follows:

$$G = \{h_1^m, h_2^m, ..., h_V^m\}$$

where $V$ is the number of the occurrences of $x_i$, and $h_j^m (j \in [1, V])$ is the hidden states of $x_i$'s occurrences. For each token $x_i$ and its output hidden states $h_i$ in the given sentence, we can get the corresponding group $G$.

Although there is no explicit sequence relation inside $G$, the position of words in $G$ still affects their contribution to the target word. For example, the words close to the target word may influence it more. Therefore, based on the global memory network, we add position embedding to $G$. Assuming that $x_i$ is located at the $t^{th}$ place in $G$, we remove the record of $x_i$ in $G$ and then get a matrix:

$$M_t = [h_1^m, ..., h_{(t-1)}^m, h_{(t+1)}^m, ..., h_V^m]$$

The position embedding is denoted as $pos = [pos_1, pos_2, ..., pos_V]$, then:

$$MP_t = [h_1^m + pos_1, ..., h_V^m + pos_V]$$

$$e_i = h_i + pos_t$$

We use $h_j^p = h_j^m + pos_j$, so the $MP_t$ can be represent as:

$$MP_t = [h_1^p, ..., h_{(t-1)}^p, h_{(t+1)}^p, ..., h_V^p]$$

A dot-product attention is applied on $e_i$ and $h_j^p \in MP_t$ to get the response of the global memory network:

$$\alpha_j = \frac{exp(e_i h_j^p)}{\sum_{j=1, j \neq t}^{j=V} exp(e_i h_j^p)}$$

$$r_i = \sum_{j=1, j \neq t}^{j=V} \alpha_j h_j^m$$

Finally, $h_i$ is used to update $G$ by replacing $h_t^m$. Then, we can get the final representation by fusing $h_i$, $N$ and $r_i$:

$$d_i = Concat(\lambda h_i + (1 - \lambda) r_i, N)$$

where $N$ is the neighbouring sentence representation.

### 3.3 Sentence-level Representation

In this section, we propose a novel label-enhanced contextualized representation that explicitly introduces contextual metaphor information, which is useful for understanding because the Specifically, the label embedding is adopted to represent each label, and then the early prediction is used to provide reference metaphor labels for the label embedding module.

**Label embedding** To fuse contextualized representation of words with label information, we use label embedding to map labels to the same space as the contextualized representation's. That is, every type of label(i.e. metaphoric or literal) corresponds to a vector via the label embedding. Therefore, we can obtain the label embedding $l_i$ of the word $x_i$ according to its label $y_i$. Then we take the sum of $d_i$ and $l_i$ as the input of a transformer encoder layer. Considering the particularity of label embedding, we modified the $Q$, $K$, and $V$ in the standard transformer architecture(Vaswani et al., 2017):

$$Q = [q_1, ..., q_n] = [d_1 + l_{pad}, ..., d_n + l_{pad}]$$

$$K = V = [d_1 + l_1, ..., d_n + l_n]$$

where the $l_{pad}$ is a padding embedding which has the same dimension as $l_i$. This is because the $l_i$ in the training steps comes from the golden label $y_i$, which will lead to leakage of the label if $Q$ contains the label information of the word itself. That is, the output of target word would contain its own golden label information. Similarly, $K_i$ and $V_i$ will introduce the label information of word $x_i$ when we calculate the attention of $q_i$ to $K$ and $V$. So we add a mask matrix to the self-attention mechanism:

$$AttentionMask = \begin{Bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{Bmatrix}$$

where the diagonal elements are all 0. It means each word ignores itself when calculating attention. Then we take $Q$, $K$, $V$ and $AttentionMask$ as the input of the transformer encoder:

$$s = Transformer(Q, K, V, Attention\ Mask)$$

$$\hat{y}_i = Classifier_{le}(s)$$

| | | #text | #sents | #tokens | % M |
|---|---|---|---|---|---|
| ALL POS | train | 90 | 12122 | 72611 | 15.2 |
| | test | 27 | 4080 | 22196 | 17.9 |
| Verbs | train | - | - | 17240 | 27.8 |
| | test | - | - | 5873 | 30.0 |

Table 1: Statistics of the VUA dataset.

We use $\hat{y}_i$ as the final prediction in the testing stage.

**Early prediction** The strategy of introducing contextual labels we adopt above uses contextual golden labels but not the labels predicted by the model in the training phase, which is similar to the teacher forcing strategy that is widely used in text generation tasks. However, it will be invalid in the testing stage since the golden label of the test set cannot be used as known information. To address this deficiency, we add an early prediction module:

$$\hat{y}_i^{ep} = Classifier_{ep}(d_i)$$

In this way, the model can predict the label of $x_i$ in advance. In the testing stage, the predicted metaphor label $\hat{y}_i^{ep}$ is provided to the label embedding phase as a substitute for the golden label.

### 3.4 Training Details

The final training objective of our model consists of two parts: (1) the early prediction $\hat{y}_i^{ep}$ and (2) the final prediction $\hat{y}_i$, both of which use cross-entropy loss function:

$$L_{EP} = -\sum_{D}\sum_{i=1}^{i=n} w_{y_i} log(\hat{y}_{ic}^{ep})$$

$$L_{LE} = -\sum_{D}\sum_{i=1}^{i=n} w_{y_i} log(\hat{y}_{ic})$$

where the $\hat{y}_{ic}^{ep}$ and $\hat{y}_{ic}$ are the predicted probabilities for the true label $y_i$, and the $w_{y_i}$ is the loss weight of $y_i$. The $D$ represents the whole dataset. The final loss is defined as the weighted summation of $L_{EP}$ and $L_{LE}$:

$$L = L_{LE} + \gamma L_{EP}$$

where $\gamma$ denotes the weighting parameter.

## 4 Experiment

### 4.1 Dataset

VU Amsterdam Metaphor Corpus (VUA)(Steen, 2010) consists of 117 fragments sampled across

| Method | VUA ALL POS | | | VUA Verbs | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| (Wu et al., 2018) | 60.8 | 70.0 | 65.1 | 60.0 | 76.3 | 67.1 |
| (Gao et al., 2018) | 68.4 | 59.7 | 63.8 | - | - | - |
| (Mao et al., 2019) | 71.7 | 60.2 | 65.5 | - | - | - |
| (Dankers et al., 2020) | 73.5 | 69.6 | 71.5 | - | - | 75.7 |
| (Dankers et al., 2020)† | 75.1 | 69.1 | 71.9 | 76.6 | 75.2 | 75.8 |
| BERT | 77.2 | 66.6 | 71.4 | 78.8 | 71.9 | 75.2 |
| Ours | 75.9 | 70.4 | **73.0*** | 77.5 | 76.1 | **76.8*** |

Table 2: The Precision, Recall and F1 score on the VUA ALL POS and VUA Verbs tasks. † denotes the model we implement according to their paper. * denotes $p < 0.05$ on a two-tailed t-test against the best competing model.

| Method | VUA ALL POS | | | VUA Verbs | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| DeepMet | 73.4 | 73.2 | 73.3 | 75.7 | 78.2 | 76.9 |
| Ours$_{cv}$ | 75.4 | 73.3 | **74.3*** | 77.4 | 79.1 | **78.3*** |

Table 3: The comparison between DeepMet and our model. Ours$_{cv}$ is obtained by training our model according to the settings of DeepMet. * denotes $p < 0.05$ on a two-tailed t-test against the best competing model.

four genres from the British National Corpus: Academic, News, Conversation, and Fiction. Every word in the corpus is labeled, guided by MIP. The corpus was used by the ACL 2020 Metaphor Shared Task(Leong et al., 2020). Similar to the shared task, we conduct experiments on the VUA ALL POS and VUA Verbs tasks. We do not choose TroFi(Birke and Sarkar, 2006) and MOH-X(Mohammad et al., 2016) datasets which are commonly used in the previous works. This is because neither of these two datasets contains discourse information, and words other than the target word within a sentence are all annotated as literal, which is useless for our model. Nonetheless, we believe that the results on the two tasks of the VUA dataset can well demonstrate the superiority of our model in both ALL POS and Verbs metaphor identification.

Table 1 shows the descriptive characteristics of the VUA dataset: the number of texts, sentences, tokens, and class distribution information for All POS and Verbs tasks.

### 4.2 Setup

We try to keep the hyper-parameters consistent with previous works which used BERT in metaphor identification. Our model is trained with a batch size of 16 for 4 epochs using the AdamW optimizer with a linear learning rate scheduler and a warm-up period of 10%. The maximum learning rate is 5e-5. We apply dropout to our model with a rate of 0.1. The weight in the loss function $w_{y_i} = 2$ if $y_i = 1$ (metaphor), otherwise $w_{y_i} = 1$. The $\lambda$ used in discourse-level representation is set as 0.8 empirically. The $k$ in neighbouring sentence representation is 2 as same as Dankers et al. (2020). The $\gamma$ used in early prediction is set as 0.2.

### 4.3 Results and Discussion

We compare our model with existing approaches which do not use external knowledge. We do not compare with the works that divided the dataset into the train set and test set by themselves, such as Wan and Xing (2020). Since Gao et al. (2018) and Mao et al. (2019) used a different subset of VUA, we use the results reported by Neidlein et al. (2020) on VUA ALL POS and VUA Verbs for comparison. Since the F1 score(71.4) of our BERT baseline is higher than that(70.3) in Dankers et al. (2020) even though the two models are basically the same, we re-implement their method. Our experimental results are obtained by averaging the results of five random runs. Table 2 shows that our model surpasses the highest results by 1.1% and 1.0% on VUA ALL POS and VUA Verbs tasks, respectively.

The current state-of-the-art model is DeepMet(Su et al., 2020), which takes global text context, local text context, query word, general POS, and fine-grained POS as the input. To make the comparison fairer, firstly, we removed their ensemble module, because simply modifying the hyper-parameters to vote is of little research significance, though it is helpful for the performance. Secondly, the DeepMet after removed the ensemble part is a 10-fold voting model, so we also adopt this strategy and remove our discourse-level module because DeepMet divides the training and validating sets at the sentence level, which will cause the sentences

| Genre | Model | P | R | F1 |
|---|---|---|---|---|
| Academic | BERT | 84.7 | 68.7 | 75.6 |
| | D-BERT | 82.7 | 72.2 | 77.1 |
| | Ours | 82.3 | 74.9 | **78.4** |
| News | BERT | 79.0 | 64.7 | 71.1 |
| | D-BERT | 77.3 | 66.5 | 71.5 |
| | Ours | 77.5 | 68.9 | **72.9** |
| Fiction | BERT | 70.1 | 67.6 | **68.8** |
| | D-BERT | 67.0 | 69.8 | 68.4 |
| | Ours | 68.7 | 68.7 | 68.7 |
| Conversation | BERT | 63.8 | 63.3 | 63.5 |
| | D-BERT | 62.2 | 65.6 | 63.8 |
| | Ours | 62.7 | 66.1 | **64.3** |

Table 4: Model performance on four different genres of VUA. D-BERT denotes the model we implement according to Dankers et al. (2020), which is the same as the model in Table 2.

| | ALL POS | Verbs |
|---|---|---|
| Ours | 73.0 | 76.8 |
| w/o label-enhance | 72.2 | 76.1 |
| w/o neighbour sentence | 72.8 | 76.2 |
| w/o global memory | 72.5 | 76.4 |
| w/o position | 72.8 | 76.6 |

Table 5: Ablation study on VUA ALL POS and VUA Verbs.

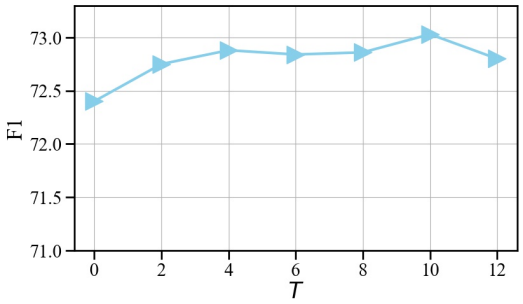


Figure 3: The performance of the model with different memory size $T$ on VUA ALL POS task.

in the same discourse to be scattered in the training set and validating sets. This will lead to incomplete discourse information in the training and validating sets. Finally, we use RoBERTa(Liu et al., 2019) as the baseline model same as DeepMet instead of BERT. This type of our model is marked as Ours$_{cv}$. We rerun the code of DeepMet and compare the results which are shown in Table 3. The F1 score of our model are 1% and 1.4% higher than DeepMet on ALL POS and Verbs, respectively.

As is shown in Table 2 and Table 3, both DeepMet and the proposed model show more gain for recall rather than for precision compared with BERT. In general, advanced pre-trained models, such as RoBERTa(Gong et al., 2020), or more semantic information(Dankers et al., 2020) will improve recall and worsen precision. Because the metaphor is a special(or high-level) way to use, it is difficult to identify complicated metaphorical expressions when the model cannot fully understand the meaning. Our model introduces contextual metaphorical information to enhance the model's ability to understanding complicated contexts. Meanwhile, by using the global memory network, the model might benefit from another well-understood context that contains the target word when processing the same word in a context that is difficult to understand. DeepMet used RoBERTa and reduced the threshold of classifying a word as a metaphor, making the model inclined to predict words as metaphors.

Table 4 reports the performance on the four genres of VUA dataset. Our model achieves better or comparable results against the baselines. It can be seen that our model performs well on news and academic genres. This is because each discourse in the two genres mainly describes one single event or stuff, which has strong logic internally. Thus, it is likely that the same metaphor appears in the discourse, which has a certain metaphorical consistency. Meanwhile, the label-enhanced representation module can enhance the ability to identifying the metaphorical expressions in long sentences which are common in these two genres. The improvement obtained on conversation is mainly because our model introduces more discourse information, which is important for understanding sentences in conversations.

### 4.4 Ablation Study

In this experiment, we remove the label-enhanced contextualized representation, neighbouring sentence representation, and position-aware global memory network modules from our model separately, and the experiment results are shown in Table 5. The last row in the table *w/o position information* refers to remove the position embedding from the position-aware global memory network. It turns out that each module of our model is useful, and removing any part of our model will cause the result to drop.

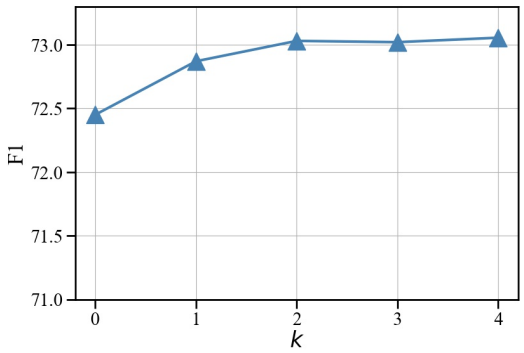


Figure 4: The performance of the model with different neighboring sentences number $k$ on VUA ALL POS task.

| | Word Num | | Metaphor Num | |
|---|---|---|---|---|
| | MO | OO | M>1 | M≤1 |
| BERT | 71.4 | 71.4 | 75.2 | **57.9** |
| Ours$_{-nei}$ | **73.3** | **72.4** | **77.1** | 57.4 |
| DeepMet | - | - | 77.8 | 56.8 |
| Ours$_{cv}$ | - | - | **79.3** | **57.5** |

Table 6: The F1 score of BERT, DeepMet, and our model. The first experiment calculates F1 scores for words that appear only once/more than once in the discourse. The second one is for words in the sentences where there are multiple metaphoric words (M>1), or only one or fewer metaphoric words (M≤1).

### 4.5 Influence of Hyper-parameters

**Memory size** In the position-aware global memory network module, if a word occurs more than $T$ times, we only record its first $T$ occurrences. Figure 3 shows the effect of $T$ on the performance of our model. when $T$ is 10, the result of our model is the best. Since the meaningful words are hard to appear many times, the performance of our model declines when $T$ is greater than 10, which may record more meaningless stop words.

**The effect of** $k$ We use $k$ to control the number of neighboring sentences. Figure 4 shows that the performance grows with $k$, and becomes stable when $k \geq 2$. Considering the time and memory cost, we choose $k = 2$ in the model.

### 4.6 Effectiveness Discussion

Table 6 shows the effectiveness experiment results of our model. To make the results more convincing, we remove the neighbouring sentence representation module from our model and denote it as Ours$_{-nei}$. We compare our model with BERT and DeepMet. Firstly, we compare the F1 scores of words in the test set that appear only once (OO) and more than once (MO) in their discourse. The results show that our model is 1.0% and 1.9% higher than BERT on OO and MO, respectively. The higher improvement obtained on MO suggests that the position-aware global memory network is indeed effective, which attends to other occurrences of the target word in the discourse to assist the identification process. Since the discourse representation module is removed from Ours$_{cv}$, we do not compare it with DeepMet. Secondly, we compare the F1 score of words in the sentences where there are multiple metaphoric words (M>1), or only one or fewer metaphoric words (M≤1). The results show that our model is 1.9% and 1.5% higher in F1 score than BERT and DeepMet respectively when M>1. This shows the effectiveness of our label-enhanced contextualized representation, because when a sentence contains multiple metaphoric words, it may be able to provide richer contextual metaphor information for the reasoning process. Moreover, we notice that the F1 scores of all models are very low when M≤1. This may be because there are many short sentences, which makes it difficult to understand the meaning of the words in the sentences. This needs further attention for metaphor identification research.

### 4.7 Error Analysis

Although introducing wider discourse information and label information, our model has limitations as well. If a word only appears once in the discourse, the global memory module will be invalid. In some cases, it is also difficult to judge the metaphoricity of some words even if they appear several times in the discourse. E.g., in the sentences *Tyson is not a gambling man*(VUA ID: aa3-fragment08-215) and *If you were a gambling man it would not affect you*(VUA ID: aa3-fragment08-232) where our model fails, the two *gambling* have similar context and usage, so it is difficult for our model to make the word benefit from another occurrence. Moreover, short sentences are also challenging because there are little contextual label information and semantic information, e.g., *No, but getting*(VUA ID: kb7-fragment48-13446), where there is not enough information for inference.

## 5 Conclusion

In this paper, we propose a hierarchical contextualized representation model to strengthen the model's

ability to leverage contextual information. Our model makes use of the contextual metaphor information in the sentence level and the long-range relation of the words in the discourse level. We improve the ability of the model to reason the contextual metaphoric relationships and understand the meaning of context by introducing contextual label representation for the target word. To obtain broader discourse information, we adopt a position-aware global memory network to extract relations among the occurrences of words in discourse. The results of our model on the two tasks of VUA dataset surpass the state-of-the-art models which also do not use external knowledge.

In future work, we will explore changing the golden label used in the label embedding stage to the iterative prediction result, which may avoid the deviation caused by the absence of golden labels during testing. Meanwhile, albeit limited, the work of Jang et al. (2017) could provide further direction for this research, such as using words belonging to the same topic/frame/domain instead of only the same words.

## Acknowledgements

We thank anonymous reviewers for their comments, which provided some insights on this research that will further influence our future work. This work is partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), and the Natural Science Foundation of China (No. 61702080, 61632011, 61806038, 61976036, 62076046, 62076051).

## References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017a. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 523–528. Association for Computational Linguistics.

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

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, FigLang@ACL 2020, Online, July 9, 2020*, pages 235–243. Association for Computational Linguistics.

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

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2218–2229. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

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

Hyeju Jang, Keith Maki, Eduard Hovy, and Carolyn Rosé. 2017. Finding structure in figurative language: Metaphor detection with topic-based frames. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330, Saarbrücken,

Germany. Association for Computational Linguistics.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Penstein Rosé. 2015. Metaphor detection in discourse. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 384–392. The Association for Computer Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutiérrez, 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Arlene Koglin. 2015. An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. *Translation & Interpreting*, 7(1):126.

George Lakoff and Mark Johnson. 1980. Metaphors We Live By. University of Chicago Press.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8139–8146. AAAI Press.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing, Fig-Lang@ACL 2020, Online, July 9, 2020*, pages 18–29. Association for Computational Linguistics.

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, Fig-Lang@NAACL-HLT 2018, New Orleans, Louisiana, 6 June 2018*, pages 56–66. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020. Hierarchical contextualized representation for named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8441–8448. AAAI Press.

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

Saif Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, *SEM@ACL 2016, Berlin, Germany, 11-12 August 2016*. The *SEM 2016 Organizing Committee.

Michael Mohler, D. Bracewell, Marc T. Tomlinson, and David R Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Learning outside the box: Discourse-level features improve metaphor identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 596–601, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546, Copenhagen, Denmark. Association for Computational Linguistics.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2890–2895. Association for Computational Linguistics.

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

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

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, Fig-Lang@ACL 2020, Online, July 9, 2020*, pages 30–39. Association for Computational Linguistics.

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

Yulia Tsvetkov, E. Mukomel, and A. Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51.

Peter D. 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, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 680–690. ACL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Mingyu Wan and Baixi Xing. 2020. Modality enriched neural network for metaphor detection. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3036–3042. International Committee on Computational Linguistics.

Yorick Wilks. 1978. Making preferences more active. *Artif. Intell.*, 11(3):197–223.

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, Fig-Lang@NAACL-HLT 2018, New Orleans, Louisiana, 6 June 2018*, pages 110–114. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.

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

Boliang Zhang, Spencer Whitehead, Lifu Huang, and Heng Ji. 2018. Global attention for name tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 86–96. Association for Computational Linguistics.