Metaphor Detection with Effective Context Denoising

Shun Wang[1]*, Yucheng Li[2]*, Chenghua Lin[1]†, Loïc Barrault[1,3] , Frank Guerin[2]
[1] Department of Computer Science, University of Sheffield, UK
[2] Department of Computer Science, University of Surrey, UK
[3] Meta AI
{swang209, c.lin}@sheffield.ac.uk
{yucheng.li, f.guerin}@surrey.ac.uk
loicbarrault@meta.com

Abstract

We propose a novel RoBERTa-based model, RoPPT, which introduces a target-oriented parse tree structure in metaphor detection. Compared to existing models, RoPPT focuses on semantically relevant information and achieves the state-of-the-art on several main metaphor datasets. We also compare our approach against several popular denoising and pruning methods, demonstrating the effectiveness of our approach in context denoising. Our code and dataset can be found at https://github.com/MajiBear000/RoPPT.

1 Introduction

Metaphor is a pervasive linguistic device, which attracts attention from both the fields of psycholinguistics and computational linguistics due to the key role it plays in the cognitive and communicative functions of language (Wilks, 1978; Lakoff and Johnson, 1980; Lakoff, 1993). Linguistically, metaphor is defined as a figurative expression that uses one or several words to represent another concept given the context, rather than taking the literal meaning of the expression (Fass, 1991). For instance, in the sentence "*This project is such a headache!*", the contextual meaning of *headache* is "a thing or person that causes worry or trouble", different from its literal meaning, "a continuous pain in the head".[1]

Metaphor detection is challenging, as it requires understanding the nuanced relationships between abstract concepts embodied by the metaphoric expression and its surrounding context. Recent studies on this direction show its potential in benefiting a wide range of NLP applications, including sentiment analysis (Li et al., 2022a), metaphor generation (Li et al., 2022b,c) and mental healthcare (Abd Yusof et al., 2017; Gutiérrez et al., 2017).

When modelling relevant context for metaphor detection, various strategies have been proposed. These range from using highly restricted forms of linguistic context such as subject-verb and verb-direct object word pairs (Gutiérrez et al., 2016), to a wider context accounting for a fixed window surrounding the target word (Do Dinh and Gurevych, 2016; Mao et al., 2018), and modelling the full sentential context (Gao et al., 2018; Choi et al., 2021). While it has been argued that modelling a wider context is beneficial (Cheng et al., 2021), it has also been noted that a wider context is likely to introduce noise into the representations, and hence hinder model's performance in metaphor detection (Le et al., 2020).

Some recent efforts (Le et al., 2020; Song et al., 2021a) attempt to improve context modelling by explicitly leveraging the syntactic structure (e.g., dependency parse tree) of a sentence in order to capture important context words, where the parse trees are typically encoded with graph convolutional neural networks. MelBERT (Choi et al., 2021) employs a simple chunking method which separates sub-sentences by commas. The sub-sentence that contains a target word is then marked with a special token type, signalling its contextual importance to the target. However, these strategies are either difficult to apply to batch optimisation due to their tree-dependent encoding process, or have limited effectiveness for context denosing. For instance, the simple chunking mechanism misses the syntactic structure, and thus can neither determine the degree of importance of context words, nor connect information across different subsentences.

In this paper, we propose a novel metaphor detection model RoPPT: RoBERTa with Pruning on target-oriented Parse Tree. RoPPT introduces a *flat*, *target-oriented* tree structure by reshaping and pruning the ordinary parse trees to extract semantically relevant neighbours of a target word. The resulting tree representation allows the model to

* The two authors contributed equally to this work.
† Corresponding author
[1] https://www.oxfordlearnersdictionaries.com

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1404–1409
May 2-6, 2023

focus on syntactically relevant information of a target word, and ignore irrelevant parts despite their position. It thus retains more relevant context for metaphor detection.

Extensive experiments conducted on three public benchmark datasets (i.e., VUA, MOH-X, TroFi) show that RoPPT can significantly improve metaphor detection on all datasets against several popular denoising and pruning methods. Our model also yields better or comparable performance to the state-of-the-art models (Choi et al., 2021; Song et al., 2021a) in Micro F1 measure. To further validate our approach, we conducted an additional investigation to assess the effect of sentence length on the performance of our model. Experimental results demonstrate a positive correlation between the increase in the performance of RoPPT and the length of the input sentence.

In summary, our paper makes three contributions: (1) we propose a flat, target word-oriented tree structure by reshaping and pruning the ordinary parse trees to retain the most relevant context for a target word; (2) we propose RoPPT, a RoBERTa-based model which can effectively encode the target-oriented parse tree for metaphor detection, achieving state-of-the-art results on three bench mark datasets; (3) we compare and evaluate a range of context denoising methods for metaphor detection, demonstrating the effectiveness of our proposed tree structure in context denoising.

2 Method

The overall architecture of RoPPT is shown in Figure 1, which can be divided into two parts: a target-oriented parse tree pruning module and a RoBERTa (Liu et al., 2019) contextual encoder.

2.1 Target-oriented Dependency Parse Tree

Connecting target words with their most relevant context words is crucial for metaphor detection and comprehension. While there have been attempts to employ dependency parse trees in graph convolutional neural networks to improve context modelling (Wang et al., 2020), it raises challenges of how to effectively encode and leverage such syntactic structure information for transformer-based mask language models for metaphor detection.

We tackle this challenge by introducing a target-oriented parse tree generated by three steps: **1**) reshape the original parse tree from existing parsers such as spaCy (Honnibal and Montani, 2017) and Biaffine (Dozat and Manning, 2016); **2**) root the tree at the target word; **3**) prune the tree according to the distance between leaves and root, coined as *neighbor range*. The rationale behind is that the target word is the focus of the task rather than the original root. So the re-rooting allows us to focus on the connections between target words and their relevant context. The resulting flat, target-oriented tree structure also enables simple encoding process into the model. Figure 1 shows an example of our reshaped tree, which retrains words with neighbor range $con = 1$ to the root 'bogged'.

2.2 RoBERTa-based Context Encoder

We employ two metaphor identification theories in our model, i.e., Metaphor Identification Procedure (Steen, 2010, MIP) and Selectional Preference Violation (Wilks, 1978, SPV). In MIP, a metaphor is detected when there is a contrast between target word's contextual and literal meanings, whereas in SPV a metaphorical word is identified by the semantic difference from its surrounding words. Therefore, we model three types of semantic representations for implementing MIP and SPV, i.e., the literal meaning and the contextual meaning of a target word, and the context meaning.

Formally, given a sentence $S = (w_0, ..., w_n)$, we first employ the RoBERTa network to produce representations for each word.

$$H = \text{RoBERTa\_Enc}(\text{emb}_{\text{cls}}, ..., \text{emb}_n) \quad (1)$$

Here CLS is a special token indicating the start of an input, $H = (h_{\text{cls}}, h_0, ..., h_n)$ the output hidden states, and emb_i the input embedding for word w_i. Specifically, $\text{emb}_i = \text{emb}_w + \text{emb}_{pos}$, where emb_w is the word embedding, and emb_{pos} the position encoding.

Context denoising with the target-oriented parse tree. When modelling sentence representation, existing works directly employed the CLS embedding as a common practice (Choi et al., 2021; Song et al., 2021b). In contrast, RoPPT employs the target-oriented parse tree to retain the most relevant context for a target word when computing the sentence embedding. Specifically, our sentence embedding is computed as follows.

$$v_S = \frac{1}{n} \sum h_i, \ i \in \mathcal{C}_n \quad (2)$$

Here v_S is the sentence representation; $\mathcal{C}_n$ represents the n neighbour words within the neighbour

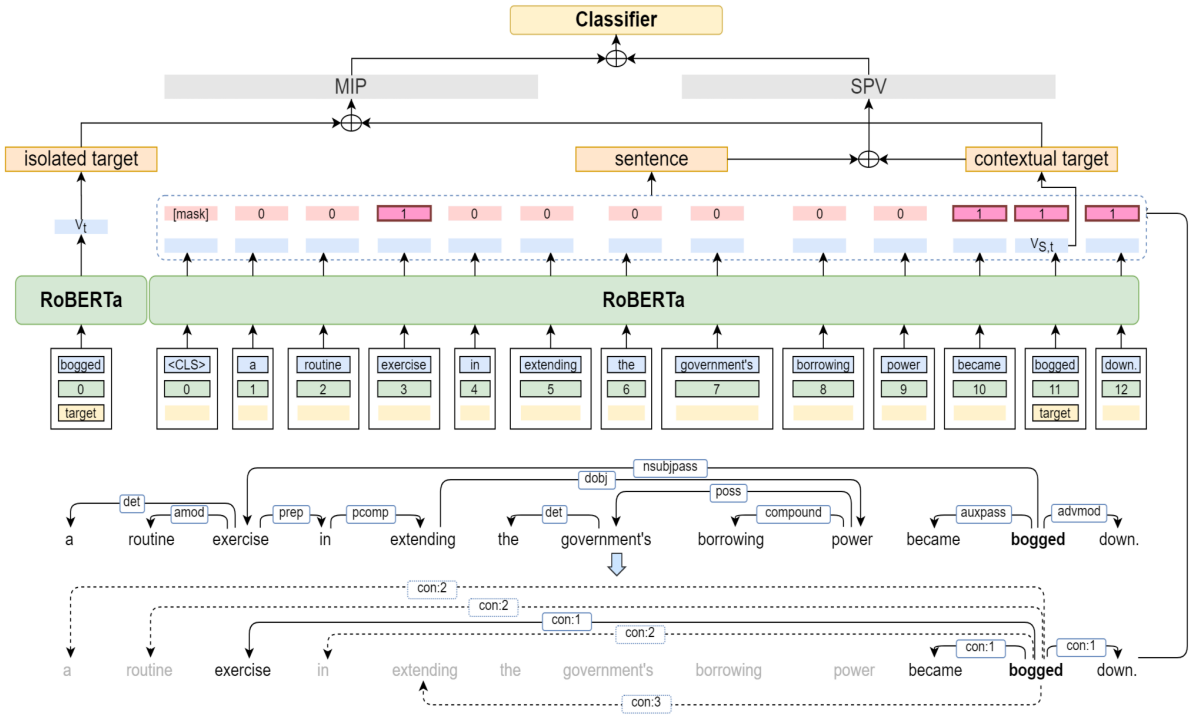

Figure 1: The overall framework of RoPPT. The parse tree of a sentence is reshaped to a target-oriented tree, and the context is pruned with a pre-set threshold. The sentence embedding is the average pooling result of hidden states for pruned context from RoBERTa. $\bigoplus$ denotes concatenation.

range of the target-oriented parse tree, and h_i is the hidden state of w_i. In other words, we do average pooling on the most relevant context words as the sentence representation and ignore other words in the sentence. We also design an alternative strategy by directly masking the original input sentence to the encoder according to the pruned parse tree. We denote this intuitive solution as RoPPT with Input Mask (**RoPPT_IM**) and discuss the performance difference between these two variants in §4.

Similar to Choi et al. (2021), we use the hidden state of target word w_t as the contextual target word embedding (i.e. $v_{S,t} = h_t$), and the literal target word embedding v_t is obtained by feeding the single target word w_t to the RoBERTa network.

$$v_t = \text{RoBERTa\_Enc}(\text{emb}_t) \quad (3)$$

We then model SPV (h_{SPV}) by concatenating the sentence embedding v_S and contextual target embedding $v_{S,t}$, and MIP (h_{MIP}) by concatenating the contextual and literal target embeddings v_t, followed by a MLP layer (i.e. $f_1(\cdot)$ and $f_2(\cdot)$).

$$h_{\text{SPV}} = f_1([v_S, v_{S,t}]) \quad (4)$$

$$h_{\text{MIP}} = f_2([v_{S,t}, v_t]) \quad (5)$$

Finally, we combine two hidden vectors h_{MIP} and h_{SPV} to compute a prediction score $\hat{y}$, and use binary cross entropy loss to train the overall framework for metaphor prediction.

$$\hat{y} = \sigma(W^\top [h_{\text{MIP}}; h_{\text{SPV}}] + b) \quad (6)$$

$$\mathcal{L} = -\sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \quad (7)$$

3 Experimental Setup

Dataset. We conduct experiments on four public benchmark datasets. **VUA-18** (Leong et al., 2018) and **VUA-20** (Leong et al., 2020) are the largest available datasets, released in the metaphor detection shared tasks in 2018 and 2020. VUA-20 extends VUA-18 with about 12K sentences for training set and 3.6K sentences for test and validation sets. The **MOH-X** dataset is constructed by sampling sentences from WordNet (Miller, 1998). Only a single target verb in each sentence is annotated. The average sentence length is 8 tokens, the shortest of our three datasets. **TroFi** (Birke and Sarkar, 2006) consists of sentences from the 1987-89 Wall Street Journal Corpus (Charniak et al., 2000), with an average length of 28.3 tokens per sentence.

Baselines. **RoBERTa_SEQ** (Leong et al., 2020) is a fine-tuned RoBERTa sequence labeling model for metaphor detection. **MelBERT** (Choi et al., 2021)

Model	VUA18			VUA20		
	Prec	Rec	F1	Prec	Rec	F1
RNN_ELMo	71.6	73.6	72.6	-	-	-
RoBERTa_SEQ	80.1	74.4	77.1	75.1	67.1	70.9
MrBERT	82.7	72.5	77.2	-	-	-
MelBERT*	79.6	76.4	77.9	76.3	68.6	72.2
MelBERT	80.1	76.9	78.5	75.9	69.0	72.3
RoBERTa_tree	78.9	76.1	77.4	74.8	68.6	71.6
RoChunk	76.6	80.0	78.2	73.9	70.0	71.9
RoWindow	78.0	78.1	78.0	75.0	68.8	71.8
RoPPT_IM	73.4	74.3	73.9	67.7	66.8	67.2
RoPPT	80.0	78.2	**79.1**	75.9	70.0	**72.8**

Table 1: Performance comparison on **VUA** dataset (best is in **bold**). NB: * indicates the reproduced results of MelBERT using the original source code and setting of (Choi et al., 2021). RNN_ELMo and MrBERT have no results on VUA20 in their original paper. Popular denoising methods are also compared. **RoChunk** means chunk sentence by comma on RoBERTa input, **RoWindow** means denoising by a context window (size=4). **RoPPT_IM** represent masking sentence before input to transformer encoder.

Models	TroFi			MOH-X		
	Prec	Rec	F1	Prec	Rec	F1
RoBERTa_SEQ	53.6	70.1	60.7	80.6	77.7	78.7
DeepMet	53.7	72.9	61.7	79.9	76.5	77.9
MrBERT	53.8	75.0	62.7	75.9	84.1	79.8
MelBERT*	53.1	73.2	61.6	78.0	79.5	78.8
MelBERT	53.4	74.1	62.0	79.3	79.7	79.2
RoBERTa_tree	50.3	77.8	61.1	76.9	83.5	79.3
RoPPT	54.2	76.2	**63.3**	77.0	83.5	**80.1**

Table 2: Performance comparison on TroFi and MOH-X datasets (NB: **bold** denotes the best result).

realises MIP and SPV theories via a RoBERTa based model. **MrBERT** (Song et al., 2021b) is the recent SOTA on verb metaphor detection based on BERT with verb relations encoded.

Hyperparameter. We set the hyperparameter neighbour range $con = 4$ based on the validation set results. All the parser results are based on spaCy as it performs better than Biaffine empirically (see §4 for more discussion).

4 Experimental Results

Overall results. Table 1 shows a comparison of the performance of our models against the baseline models on VUA18 and VUA20, respectively. It is clear that our RoPPT outperforms all baselines on VUA18 and VUA20, including the state-of-the-art model MelBERT. A two-tailed t-test was conducted based on 10 paired results from RoPPT and the strongest baseline MelBERT* on both VUA-18 ($p = 0.014$) and VUA-20 ($p = 0.019$).

We also compared our method against several common denoising strategies. The results show that our tree-based denoising method is more effective than other popular denoising approaches such as RoChunk and RoWindow, which are sequence-based methods. We also apply our target-oriented tree to RoBERTa_SEQ, denoted as the RoBERTa_tree model. The improvement of RoBERTa_tree over RoBERTa_SEQ on two VUA datasets (i.e. 0.3% and 0.7%) further demonstrates the utility of our tree-based denoising method.

Following the setup of Choi et al. (2021), we also conducted a zero-shot transfer learning experiment shown in Table 2. Specifically, our model is trained on the training set of VUA20 and directly tested on the entire Trofi and MOH-X datasets. This is intended to test the generalisation power of trained models. RoPPT shows the best performance on both datasets (significant test on RoPPT against MelBERT*: TroFi $p = 0.0001$; MOH-X $p = 0.021$; we cannot compare with MrBERT as the code is unavailable). It can be observed that our model gives a larger margin of improvement over the baselines on TroFi (i.e., 1.3% gain over MelBERT and 0.6% over MrBERT) than MoH-X (i.e., 0.9% gain over MelBRTT and 0.3% over MrBERT).

Model performance vs. Sentence length. As the averaged sentences length of TroFi (28.3 tokens) is significantly longer than that of MoH-X (8 tokens), it is worth investigating whether our model gives more performance boost on data with longer context as it is likely to be noisier. To verify this hypothesis, we evaluated the performance boost of our RoPPT against the SOTA baseline MelBERT. Table 3 shows the results of VUA18 with the testset splitted into 3 different groups based on sentence length. The results demonstrate a clear positive correlation between performance boost and sentence length.

Impact of Parsers. We also investigated how the choice of parsers impacts the metaphor detection performance of our model. Specifically, we tested two parsers for constructing the target-oriented dependency parse trees, namely, the CNN-based parser Biaffine and the RoBERTa-based parser spaCy. When tested on the validation set, our model achieves 78.0% with spaCy and 77.7% with Biaffine in F1 for metaphor detection, respectively. This shows that the impact of the parse choice is relatively small for our model.

Case Studies. RoPPT shows its strength in the following example with the target word far away from

Sent. len.	RoPPT			MelBERT*			F1 diff.	Pruning comp.	# of Sent.
	Prec	Rec	F1	Prec	Rec	F1			
<20	76.4	74.8	**75.6**	75.0	75.2	75.1	**0.5**	10.7 / 12.3	18,515
20-40	81.8	79.9	**80.8**	79.2	79.1	79.2	**1.6**	16.4 / 29.4	17,729
>40	82.3	80.0	**81.1**	78.5	76.8	77.6	**3.5**	19.5 / 53.6	7,703

Table 3: RoPPT and MelBERT* performance comparison on sentences with different length range from VUA18. 'Pruning comp.' is the comparison of the average length of (pruned) / (original) sentences.

its subject, which is correctly labeled by RoPPT but incorrectly by baseline models. For the instance with metaphorical target word *bogged*, "*a routine exercise in extending the government's borrowing power to $3.1 thousand billion became bogged down.*", the target word *bogged* is separated from its subject by a long phrase, which causes baselines (including MelBERT) to fail to detect the metaphor. Thanks to the parse tree, RoPPT links *exercise* directly to the target and produces the right label.

5 Conclusion

In this paper, we proposed, RoPPT, an effective approach to extract contextual information for target words for metaphor detection based on a target-oriented parse tree structure. Extensive experiments show that our model can yield better performance compared to the state-of-the-art. In addition, our method is particularly effective in denoising long sentences, despite its simplicity.

6 Limitations

Empirical experiments show that our method is more effective in denoising long sentences with the proposed target-oriented parse tree. While this is somewhat expected as shorter sentences tend to have cleaner context, it raises a question or limitation of how can we improve the proposed method to better deal with short sentences and improve its performance in these cases. One possibility is to exploit external knowledge (e.g. ConceptNet) to support the detection of the most important contextual words.

References

Noor Fazilla Abd Yusof, Chenghua Lin, and Frank Guerin. 2017. Analysing the causes of depressed mood from depression vulnerable individuals. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 9–17.

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*.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978, Online. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.

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

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

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

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930, Copenhagen, Denmark. Association for Computational Linguistics.

E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic

models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

George Lakoff. 1993. The contemporary theory of metaphor.

George Lakoff and Mark Johnson. 1980. Metaphors we live by: University of chicago press. *Chicago, IL*.

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

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29.

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.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022a. The secret of metaphor on expressing stronger emotion. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022b. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In *International Conference on Computational Linguistics*.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022c. Nominal metaphor generation with multitask learning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 225–235.

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. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021a. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251, Online. Association for Computational Linguistics.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021b. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251.

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

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

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