FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning

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

Abstract

In this paper, we propose FrameBERT, a RoBERTa-based model that can explicitly learn and incorporate FrameNet Embeddings for concept-level metaphor detection. FrameBERT not only achieves better or comparable performance to the state-of-the-art, but also is more *explainable* and *interpretable* compared to existing models, attributing to its ability of accounting for external knowledge of FrameNet.

1 Introduction

Metaphor is a pervasive linguistic device, which attracts attention from both 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; a problem", different from its literal meaning, "a continuous pain in the head".

Metaphor Detection presents a significant challenge as it necessitates comprehending the intricate associations between the abstract concepts embodied by the metaphoric expression and the related context. Recent studies in this field have demonstrated its potential to positively impact various Natural Language Processing (NLP) applications, including sentiment analysis (Cambria et al., 2017; Li et al., 2022a), metaphor generation (Tang et al., 2022; Li et al., 2022b,c), and mental health care (Abd Yusof et al., 2017; Gutiérrez et al., 2017). Different strategies have been proposed for modeling relevant context, including employing limited linguistic context such as subject-verb and verb-direct object word pairs (Gutiérrez et al., 2016), incorporating a wider context encompassing a fixed window surrounding the target word (Do Dinh and Gurevych, 2016; Mao et al., 2018), and considering the complete sentential context (Gao et al., 2018; Choi et al., 2021). Some recent efforts attempt to improve context modelling by explicitly leveraging the syntactic structure (e.g., dependency tree) of a sentence in order to capture important context words, where the parse trees are typically encoded with graph convolutional neural networks (Le et al., 2020; Song et al., 2021).

Despite the progress, we also observe the inadequacy of existing models in semantic modelling, which plays a crucial role in metaphor detection. For instance, it has been noted that BERT's tendency to aggregate shallow semantics instead of precise meaning, as its objective, may limit the context modelling ability (Xu et al., 2020). External knowledge such as FrameNet has been widely used to provide extra semantic information and has been shown useful in a wide range of NLP tasks, such as question answering (Shen and Lapata, 2007), machine reading comprehension (Guo et al., 2020), and identifying software requirements (Alhoshan et al., 2019). Very recently, FrameNet has also been employed to the task of metaphor generation via learning mappings between domains, with promising results achieved (Stowe et al., 2021). However, such a valuable source of knowledge, surprisingly, has not been explored in the deep learning literature for metaphor detection. We hypothesise that incorporating external knowledge of concepts is essential for improving metaphor detection and model explainability.

In this paper, we propose FrameBERT, a BERT-based model for conceptual metaphor detection underpinned by learning and incorporating embedding representation of semantic frames in FrameNet. FrameBERT directly addresses the lim-

* The two authors contributed equally to this work.
† Corresponding author

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1558–1563
May 2-6, 2023 ©2023 Association for Computational Linguistics

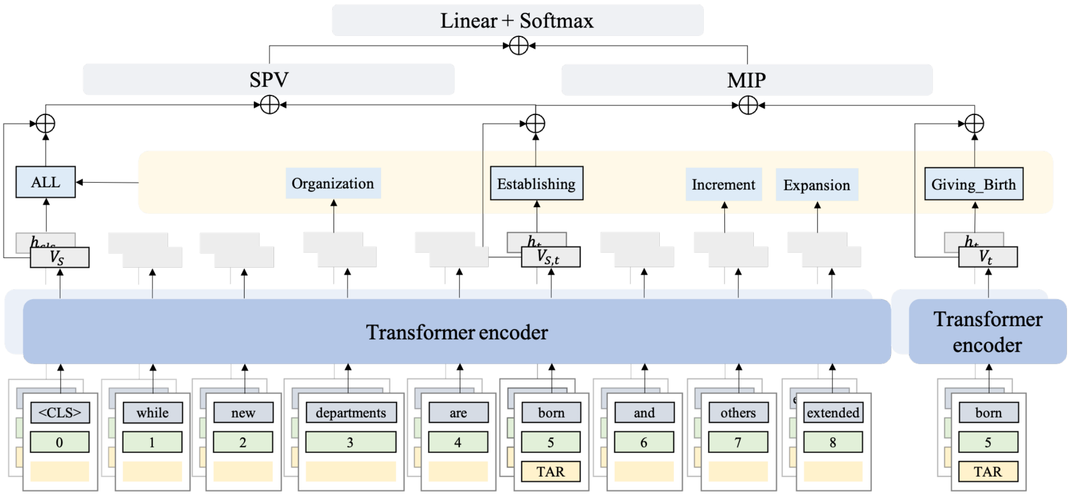

Figure 1: The overall framework. The surface encoder illustrates sentence encoder providing hidden-state representations and the insider one shows concept encoder producing concepts information. The frame embedding and hidden state embedding are concatenated to make final predictions.

itation of the existing works, which solely rely on the shallow semantics captured by hand-crafted psycholinguistics features or encoded by large pre-trained language models such as BERT. This is achieved by explicitly learning and incorporating FrameNet embeddings into the model training process. To our knowledge, this is the first attempt to apply FrameNet in deep learning models for metaphor detection. We also leverage Metaphor Identification Procedure (Group, 2007; Steen, 2010, MIP) and Selectional Preference Violation (SPV) (Wilks, 1975, 1978) to inform our model design.

Extensive experiments conducted on four public benchmark datasets (i.e., VUA MOH-X, TroFi) show that FrameBERT can significantly improve metaphor detection for all datasets compared to our base model without exploiting FrameNet embeddings. Our model also yields better or comparable performance to state-of-the-art models in Micro F1 measure. Furthermore, we show the *explainable* feature of FrameBERT, attributing to its ability of extracting semantic frames from text. The code and dataset can be found at https://github.com/liyucheng09/MetaphorFrame.

2 Model

We propose FrameBERT, a novel model that can explicitly learn and incorporate FrameNet embeddings for concept-level metaphor detection. Figure 1 illustrates the overall architecture of FrameBERT, which consists of two components: a sentence encoder (§ 2.1) and a concept encoder (§ 2.2).

2.1 Sentence Encoder

Similar to the prior work (Choi et al., 2021; Song et al., 2021), we develop the sentence encoder to produce the sentence encoding $\mathbf{v}_S$, the contextualised encoding for the target word $\mathbf{v}_{S,t}$, as well as isolated encoding for the target word $\mathbf{v}_t$. Formally, given an input sequence $S = (w_0, ..., w_n)$, RoBERTa (Liu et al., 2019) encodes each word into a set of contextualised embedding vectors $\mathbf{H} = (\mathbf{h}_{cls}, \mathbf{h}_0, ..., \mathbf{h}_n)$:

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

where CLS token is a special token used to indicate the the beginning of the input; emb_i is the input embedding for word w_i represented as

$$\mathrm{emb}_i = \mathrm{emb}_w + \mathrm{emb}_{pos} + \mathrm{emb}_{type} \quad (2)$$

Here emb_w represents the word embedding, emb_{pos} is the position encoding for w_i, emb_{type} token type encoding indicating whether a word is a target or non-target word. We employ the CLS hidden state $\mathbf{h}_{cls}$ as the sentence representation, i.e., $\mathbf{v}_S = \mathbf{h}_{cls}$, the hidden states $\mathbf{h}_t$ of target word w_t as the contextual target word embedding, i.e., $\mathbf{v}_{S,t} = \mathbf{h}_t$. For the isolated word embedding for w_t, we directly feed w_t to RoBERTa in order to obtain the literal representation of the target word, i.e., $\mathbf{v}_t = \mathrm{RoBERTa}(\mathrm{emb}_t)$.

FrameBERT using MIP and SPV. With MIP, a metaphorical word is identified by the gap between the contextual and literal meaning of a word. To incorporate MIP, we employ the contextualised $\mathbf{v}_{S,t}$

and isolated embedding $\mathbf{v}_t$ vectors for w_t. With SPV, a metaphorical word is identified by the semantic difference from its surrounding words, i.e., the contrast between $\mathbf{v}_S$ and $\mathbf{v}_{S,t}$. We formalise our incorporation of these two metaphor identification theories below. Note that $\oplus$ is an operation for vector concatenation.

$$\mathbf{h}_{MIP} = \mathbf{v}_t \oplus \mathbf{v}_{S,t} \tag{3}$$

$$\mathbf{h}_{SPV} = \mathbf{v}_{S,t} \oplus \mathbf{v}_S \tag{4}$$

2.2 Conceptual Encoder

One of the key contributions of our paper is that our model can explicitly learn and incorporate FrameNet Embeddings for concept-level metaphor detection. This is achieved via the conceptual encoder, where we first fine-tuning a RoBERTa model on the FrameNet (Fillmore et al., 2002) dataset with a objective to classify frame lables, and then join the conceptual encoder with the sentence encoder.

Given an input sentence $S = (w_0, ..., w_n)$, we add a special token CLS at the beginning of the sentence and apply a stack of Transformer encoder layers on the tokenised input to obtain the contextualised hidden states for each word $\mathbf{H} = (\mathbf{h}_{cls}, \mathbf{h}_0, ..., \mathbf{h}_n)$ and the CLS token, similar to § 2.1. We then leverage the contextual target word hidden states and CLS hidden states (as sentence representation) to predict the target word's frame and all frames detected in the sentence. Formally, given CLS hidden states $\mathbf{h}_{cls}$ and a list of contextualised target word hidden states $\mathbf{H} = (\mathbf{h}_0, ..., \mathbf{h}_k)$, we obtain the frame distribution for sentence and targets as follows:

$$\hat{\mathbf{y}}^{\mathbf{f}}_{cls} = \text{sigmoid}(\mathbf{W}_0\mathbf{h}_{cls} + \mathbf{b}_0) \tag{5}$$

$$\hat{\mathbf{y}}^{\mathbf{f}} = \text{softmax}(\mathbf{W}_1\mathbf{H} + \mathbf{b}_1) \tag{6}$$

where $\mathbf{W}_0$ and $\mathbf{W}_1$ are learnable parameters, $\mathbf{b}_0$ and $\mathbf{b}_1$ are bias. Note that $\hat{\mathbf{y}}^{\mathbf{f}}_{cls}$ should be applied on all frame classes, that is compute it on each possible frame classess. We then minimise the following loss functions:

$$\mathcal{L}_{target} = -\sum \mathbf{y} \log \hat{\mathbf{y}}^{\mathbf{f}} \tag{7}$$

$$\mathcal{L}_{cls} = -\sum^{N} \sum_{i=0}^{L} \mathbf{y}_i \log \hat{\mathbf{y}}^{\mathbf{f}}_{cls} \tag{8}$$

$$+ (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}^{\mathbf{f}}_{cls}) \tag{9}$$

where N is the number of training samples. L is number of frame labels, which means we are optimising the objective on all possible frame classes. We use λ as a hyperparameter controling weights between two losses: $\mathcal{L} = \lambda * \mathcal{L}_{cls} + \mathcal{L}_{target}$; and we set it to 2 in our experiments.

After the pre-training stage, the conceptual encoder will provide frame information for metaphor detection. As shown in Figure 1, in the MIP module, we concatenate the contextualised frame embedding $\mathbf{h}_{S,t}$ and isolated frame embedding $\mathbf{h}_t$ of target word to $\mathbf{h}_{MIP}$ (eq. 10). In the SPV module, we concatenate the CLS frame embedding $\mathbf{h}_{\mathbf{cls}}$ and contextualised target word frame embedding $\mathbf{h}_{S,t}$ to $\mathbf{h}_{SPV}$ (eq. 11).

$$\mathbf{h}_{MIP} = \mathbf{v}_t \oplus \mathbf{v}_{S,t} \oplus \mathbf{h}_t \oplus \mathbf{h}_{S,t} \tag{10}$$

$$\mathbf{h}_{SPV} = \mathbf{v}_S \oplus \mathbf{v}_{S,t} \oplus \mathbf{h}_{cls} \oplus \mathbf{h}_{S,t} \tag{11}$$

We then combine two hidden vectors $\mathbf{h}_{MIP}$ and $\mathbf{h}_{SPV}$ to compute a prediction score.

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}_T(\mathbf{h}_{MIP} \oplus \mathbf{h}_{SPV}) + \mathbf{b}) \tag{12}$$

Finally, we use binary cross entropy loss to train the overall framework for metaphor prediction.

$$\mathcal{L} = -\sum_{i=0}^{N} \mathbf{y}_i \log \hat{\mathbf{y}}_i - (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}_i) \tag{13}$$

3 Experiments

Dataset. We conduct experiments on four public bench datasets. **VUA-18** (Leong et al., 2018) and **VUA-20** (Leong et al., 2020), the extension of VUA-18, are the largest publicly available datasets. 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 length of sentences 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 Release 1 (Charniak et al., 2000). The average length of sentences is the longest of our datasets (i.e., 28.3 tokens per sentence). At last, the concept encoder was pre-trained on **FrameNet release 1.7** (Fillmore et al., 2002) with about 19k, 6k, 2k annotations for training, testing and evaluation respectively.
Baselines. RNN_ELMo (Gao et al., 2018) combined ELMo and BiLSTM as a backbone model. **RNN_MHCA** (Mao et al., 2019): they introduced MIP and SPV into RNN_ELMo and capture the contextual feature within window size by multi-head attention. **MUL_GCN** (Le et al., 2020) apply a GCN based multi-task framework by modelling

Models	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
MelBERT ⋆	79.6	76.4	77.9	76.4	68.6	72.3
MelBERT	80.1	76.9	78.5	75.9	69.0	72.3
MrBERT	82.7	72.5	77.2	-	-	-
FrameBERT	82.7	75.3	**78.8***	79.1	67.7	**73.0***

Table 1: Performance comparison on VUA datasets (best results in **bold**). NB: ⋆ indicates the reproduced results of MelBERT using the original source code and setting of (Choi et al., 2021). * denotes our model outperforms the competing model with $p < 0.05$ for a two-tailed t-test; except MrBERT whose code is not published.

	Models	Prec	Rec	F1
TroFi	RNN_MHCA	68.6	76.8	72.4
	MUL_GCN	73.1	73.6	73.2
	MrBERT	73.9	72.1	72.9
	FrameBERT	70.7	78.2	**74.2**
MOH-X	RNN_MHCA	77.5	83.1	80.0
	MUL_GCN	79.7	80.5	79.6
	MrBERT	84.1	85.6	**84.2**
	FrameBERT	83.2	84.4	83.8

Table 2: Performance comparison of our method with baselines on TroFi and MOH-X (best results in **bold**). We do not perform a significance test since the code of MrBERT is not published.

metaphor detection and word sense disambiguation. **RoBERTa_SEQ** (Leong et al., 2020) is a fine-tuned RoBERTa model in sequence labeling setting for metaphor detection. **MelBERT** (Choi et al., 2021) realize MIP and SPV theories via a RoBERTa based model. **MrBERT** (Song et al., 2021) is the recent sota on verb metaphor detection based on BERT with verb relation encoded.

4 Experimental Results

Overall results. Table 1 shows a comparison of the performance of our model against the baseline models on VUA18 and VUA20 respectively. Our model outperforms all the baseline models on VUA-20, including the state-of-the-art-model MelBERT (with $p < 0.05$ for a two-tailed t-test). For VUA-18, we outperformed all the baselines significantly including the *re-produced* results for MelBERT. Table 2 shows the results on the MOH-X and TroFi dataset. The results shows our method beats SOTA method on TroFi benchmark and gains comparable performance on MOH-X dataset.

Models	Prec	Rec	F1
FrameBert	82.7	75.3	**78.8**
rand_in_eval	81.8	58.7	68.3
rand_in_train_&_eval	79.3	72.6	75.8
w/o frame fine-tuning	79.1	76.3	77.6

Table 3: Results of ablation study, tested on VUA18. *rand_in_eval* represents the first experiment where the shuffle process is only performed in evaluation stage and *rand_in_train_&_eval* represents the second experiment where the shuffle process is performed in both training and evaluation stages. In *w/o frame fine-tuning* experiment, we remove the frame fine-tuning process.

Ablation Study. We performed three experiments to test the effectiveness of conceptual information. First, the system is fed with shuffled conceptual embeddings in the batch during evaluation. Second, in both training and evaluation processes, we shuffle the conceptual embeddings in the batch. Third, we remove the concept fine-tuning process. In all experiments, the overall framework remains the same as the original setting. The results are provided in Table 3. Based on the results, the performance in terms of F1 drops by 13% and 3.7% while feeding random conceptual information in only evaluation stage and both training and evaluation stages (likely collapse into the base model) respectively, which demonstrates the extent to which the conceptual information is incorporated in the overall framework (i.e. especially when we shuffle only for evaluation). The third experiment shows the performance declines 1.2% while removing the frame fine-tuning procedure, which proves the effectiveness of frame embedding learning.

Concept Analysis. In this section, we illustrate how the proposed approach detect metaphor in an interpretable way and how well the method using frame features. We performed an exploratory analysis on 200 examples where our system had a correct classification, but MelBERT failed. The following two examples show how frame information works in the metaphor detection procedure. The forst is a true positive example with the target word in bold: '*Local people mutter and march, make speeches and throw things; staff **face** sarcasm in nearby pubs* . . . .' Here our system had the following concepts as the literal meaning for '*face*': '*Body_parts, Facial_expression, Change_posture*', which are more basic meanings, relating to the face as a part of the body. In contrast, contextual concepts are extracted as follows: '*Confronting_problem, Resolve_problem, Surviving*'. These capture well the contextual meaning of 'face' in the sentence. The contextual meanings are more abstract, and the contrast between literal and contextual concepts helps the model to detect the metaphorical usage of *face* here. An example of a true negative is: '*...**hot** computers are slow, the warmth might damage...*'. 'Hot' is a word that can often be used metaphorically (e.g. hot topic, hot pants, hot properties), but in this sentence our model correctly identified it as literal and contextual concepts identified were identical: '*Temperature, Fire_Burning*'. In terms of how well our method using frame features, we measured the accuracy of the frame prediction module manually for these 200 examples, and found the correct frame label was identified in the top 3 frame label prediction for 178 of 200 examples (89%). This indicates our method is effective extracting frame features.

5 Conclusion

We proposed FrameBERT, the first conceptual model for metaphor detection by explicitly learning and incorporating FrameNet Embeddings for concept-level metaphor detection. Extensive experiments show that our model can yield better or comparable performance to the state-of-the-art.

6 Limitations

This paper mainly models frame information by representation learning on the frame classification task. However, other features such as Frame Elements (FEs) and Lexical Units (LUs) in FrameNet have not been explored in this paper, where our case analysis shows these features could provide useful signals for metaphor detection. It might also be promising to explore other types of knowledge such as context graphs (Cheng et al., 2021) for improving metaphor detection. We leave these directions to future works.

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.

Waad Alhoshan, Riza Batista-Navarro, and Liping Zhao. 2019. Using frame embeddings to identify semantically related software requirements. In *REFSQ Workshops*.

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.

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.

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.

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

Charles J Fillmore, Collin F Baker, and Hiroaki Sato. 2002. The framenet database and software tools. In *LREC*.

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

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

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896.

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.

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.

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.

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

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. 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.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Cheng Tang, Frank Guerin, Yucheng Li, and Chenghua Lin. 2022. Recent advances in neural text generation: A task-agnostic survey. *ArXiv*, abs/2203.03047.

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

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

Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. 2020. Symmetric regularization based bert for pair-wise semantic reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1901–1904.