Enhanced Metaphor Detection via Incorporation of External Knowledge Based on Linguistic Theories

Chang Su, Kechun Wu, Yijiang Chen*
School of Informatics, Xiamen University, Xiamen, China
{suchang,cyj}@xmu.edu.cn, wukechun@stu.xmu.edu.cn

Abstract

Use of external knowledge is an important and effective method applied widely in metaphor detection. Although existing knowledge-based methods perform well, when leveraging external knowledge, they take little consideration on linguistic theories of metaphor detection. Based on Metaphor Identification Procedure (MIP) and Select Preference Violation (SPV), directly using examples and definitions of words from the Oxford Dictionary[1], we propose two BERT-based models for metaphor detection: ExampleBERT and DefinitionBERT. Experimental results show that our methods achieve state-of-the-art performance on two established metaphor datasets. Furthermore, we show that our DefinitionBERT is highly interpretable.

1 Introduction

Metaphor Detection (MD) is a high-level natural language processing (NLP) task, which aims to identify the metaphorical expressions/words in the text. Identifying metaphors, a cognitive activity in which humans use their experience in one field to explain or understand another field (Shutova et al., 2016), is a challenging task that requires rich prior knowledge and a high level of semantic understanding.

In earlier studies, many resources were exploited to develop rule-based and machine learning systems, such as domain types,word abstractness/concreteness(Turney et al., 2011; Tsvetkov et al., 2014). Recently, many deep learning based methods have been applied to metaphor detection (Kehat and Pustejovsky, 2020; Le et al., 2020; Rohanian et al., 2020), which achieve the current state-of-the-art performance. They also make use of external knowledge. Hence, we can infer that incorporating external knowledge is indeed important. In this paper, we show that some level of lexical semantic information, even if its just dictionary entries, can improve performance in identifying verbal metaphor.

A recent study (Mao et al., 2019) shows the effectiveness of taking advantage of linguistic theories when identifying metaphors. According to one of the linguistic theories, Metaphor Identification Procedure(**MIP**) (Semino et al., 2007; Steen et al., 2010), a metaphor is identified if the literal meaning of a word contrasts with the means that word takes in this context. For example, in the metaphorical sentence, *the deep learning model is flying during training*, the context meaning of 'flying' is 'the loss of the model is getting bigger and even become indefinite', which contrasts with its literal meaning of 'move through the air using wings' according to **Oxford** Dictionary. An alternative approach is Select preference Violation(**SPV**) (Wilks, 1975, 1978), wherein a metaphor is identified by noticing a semantic contrast between a target word and its context. For example, in *the deep learning model is flying during training*, 'fly' is unusual in the context of 'model' and 'training': a model cannot fly.

To incorporate external knowledge, we take advantage of the linguistic theories of metaphor detection. Following SPV, we use examples of the word from Oxford Dictionary, where the literal meanings of the word are expressed in the contextual examples for the most of time. Hence, some common contextual information of the word can be inferred from examples. In accordance with MIP, we use the definitions of a word from the Oxford dictionary, which directly express the literal meanings of the word. To better use this knowledge and conform the idea of linguistic theories, we propose (1) ExampleBERT, which, before it identifies metaphor, learns the common contextual information of the

[1] https://www.lexico.com/
*Corresponding author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1280–1287

target word and (2) DefinitionBERT, which, while identifying a metaphor, directly takes advantage of the literal meanings of the target word. In particular, our contribution is two-fold as follows:

1. We directly use the examples and definitions of the word from the Oxford Dictionary. To the best of our knowledge, it is the first time this knowledge is incorporated into metaphor detection.

2. We propose ExampleBERT and DefinitionBERT. Experimental results show that both of our models can outperform the state-of-the-art models on two verb metaphor detection datasets. Also, experimental analysis proves that our DefinitionBERT is indeed effective and has a strong interpretability.

2 Related Work

Metaphor identification is a linguistic metaphor processing task that identifies metaphors in textual data. Most of the earlier works on metaphor identification were based on feature-engineering. Unigrams, imageability, concreteness, abstractness, word embedding and semantic classes are features commonly employed by supervised machine learning (Turney et al., 2011; Assaf et al., 2013; Tsvetkov et al., 2014; Klebanov et al., 2016). Recently, many deep learning based methods have been proposed, which treat metaphor identification as a sequence tagging task. Considering whether to use external knowledge directly, we divide these methods into the following two categories:

Use of pre-trained word embeddings. The first methods use only pre-trained word embeddings, which are commonly used in NLP tasks. (Wu et al., 2018) proposed a model based on word2vec (Mikolov et al., 2013) and PoS tags and word clusters, which are encoded by a Convolutional Neural Network (CNN) and Bi-LSTM. The encoded information is directly fed into a softmax classifier. (Gao et al., 2018) and (Mao et al., 2019) concatenated Glove (Le et al., 2020) and ELMO (Peters et al., 2018) as the inputs of Bi-LSTM, the difference is (Mao et al., 2019), inspired by linguistic theories, uses attention mechanism to improve performance.

External knowledge. The second methods use different kinds of external knowledge to boost performance. (Kehat and Pustejovsky, 2020) use Vision-Language datasets to derive the concreteness scores of words and then convert them to Visibility Embeddings, which, like with (Gao et al., 2018) , finally feed to Bi-LSTM. (Le et al., 2020) propose a multi-mask learning method, which transfer knowledge from Word Sense Disambiguation (WSD); to improve performance, they also employe Graph Convolution Neural networks (GCN) with dependency trees. Like (Le et al., 2020), (Rohanian et al., 2020) also use GCN, but they incorporate annotations for verbal multiword expressions. Obviously, our methods belong to this second category.

3 Method

3.1 BERT

BERT (Devlin et al., 2019) is a powerful language representation model, whose architecture is a multi-layer bidirectional transformer encoder. The BERT model is pre-trained on a large corpus and two novel unsupervised prediction tasks, i.e., masked language model and next sentence prediction tasks are used in pre-training. Here, it must note that BERT is chosen as our base model, not only because of its excellent performance on many other NLP tasks, but also BERT is a bidirectional language model. More specifically, during training, BERT randomly mask some words in the sentence and then use all the unmasked words to predict them based on a self-attention mechanism. Hence, this procedure allow BERT to learn the common context of the target word, which is very useful for our task because if a target word appears in uncommon contexts, then BERT is more likely to predict it to be a metaphorical word.

3.2 BERT(Token-CLS)

To incorporate BERT to our metaphor detection task, we take the final hidden state of the token corresponding to the target word, and then add a classification layer to predict whether or not the target word is metaphorical. We compare this model as our baseline with ExampleBERT and DefinitionBERT mentioned below.

3.3 ExampleBERT

The intuition behind SPV is that metaphoricity is identified by detecting the incongruity between a target word and its context. Hence, we assume that, if a model has learned the common context information of a target word, then the model works more effectively. As described in Section 3.1 above, a bidirectional pre-training model satisfies our requirement. Therefore, our proposed ExampleBERT

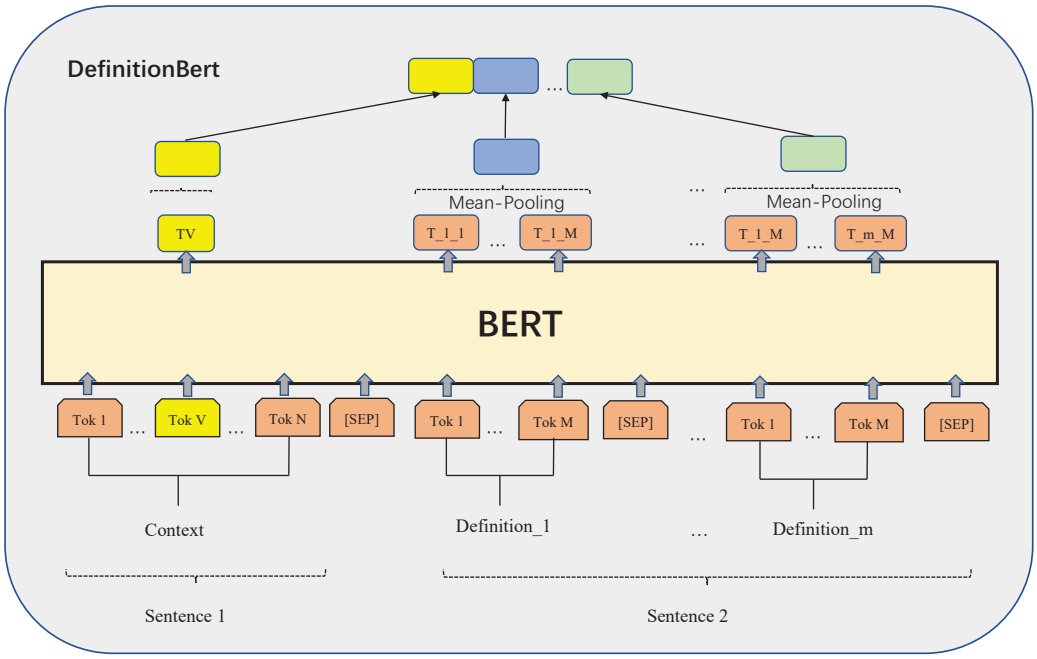

Figure 1: The overall of DefinitionBERT architecture and its context-definitions input pair.

model is built based on the standard BERT architecture (Devlin et al., 2019) which is based on the two-stage '*Pre-training*'-then-'*Fine-tuning*' pre-training language model approach, that recently become enormously popular in NLP. During the pre-training phase, we collect examples of the target word under its definitions from the Oxford dictionary and use only MaskLM as our pre-training objective. Here, we continue pre-training based on the pre-trained uncased BERT$_{BASE}$ model from (Wolf et al., 2020). The train strategy is the same as (Devlin et al., 2019). The training data generator chooses fifteen percent of the token positions at random for prediction. If the i-th token is chosen, we replace the i-th token with (1) the [MASK] token eighty percent of the time, (2) a random token ten percent of the time, and (3) the unchanged i-th token ten percent of the time. Here, our hypothesis is that most of the examples of a target word are expressing its literal meanings. Thus, whether or not a target word is selected, the model can also learn some common context information of a target word. During fine-tuning phase, we directly use the pre-trained ExampleBERT to fine-tune on the metaphor detection datasets as described in Section 3.2 above.

3.4 DefinitionBERT

Based on MIP, we assume that if we tell the model directly the literal meanings of the target word, then the model will work more effectively. Fortunately, BERT can explicitly model the relationship of a pair of texts, and this has been proved to be beneficial to many pair-wise natural language understanding tasks. Therefore, to fully leverage the definitions of words, we construct *context-definition* pair based on all possible definition of the target word from the Oxford dictionary, thereby treating MD task as a sentence pair classification problem seemingly. But, different from (Huang et al., 2019), here we cannot and don't need to match multiple definition and sentence directly one by one, because the contextual meaning of a metaphorical word is different from all its definitions. Also, we don't know which definition the contextual meaning of a non-metaphorical word corresponds to. Moreover, although there are word definition collections in WordNet (Miller, 1995), we find they cannot express accurately the literal meaning of words, and some of them are exactly the metaphorical meanings. For example, in WordNet, one of the definition for 'drink' is 'take in liquids'. On one hand, in the sentence, *car drinks gasoline*, that definition does not help us, or a model, identify that 'drink' is metaphorical. On the other hand, the Oxford Dictionary definition – 'take (a liquid) into the mouth and swallow' – can be of help. A car, which has no mouth and cannot swallow, is obviously unsuitable here. Hence, the latter is helpful to us.

As shown in Figure 1, we directly concatenate

the multiple definitions of the target word, and use "[SEP]" to separate them. Finally we use the *context-definitions* pair as the inputs for BERT. After encoding by BERT, we take the final hidden state of the target word as its context meaning. To obtain the literal meaning of its definition, we also take the final hidden states of the tokens of each definition, and use Mean-Pooling to average the hidden states of each definition, which represents literal meaning expressed by the definition. This is formulated as follows :

$$h_{di} = f_m(f_b(\boldsymbol{x_{di}})) \quad (1)$$

where f_b represents the BERT encoder, and $f_m : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ is a mean pooling function that maps from output vectors of n tokens to the definition vector. Then, we concatenate the vectors of the target word and definitions into one vector and apply a Feed-Forward Neural Network (FFNN) over the concatenated representations. This is formulated as:

$$h_f = FFNN([h_{tt}; h_{di}; ...h_{dn}]) \quad (2)$$

where h_{tt} indicates the hidden state of the target word from BERT. Then h_f is taken as input for a logistic regression classifier to make the prediction.

4 Experiments

4.1 Dataset

To be compatible with previous work (Gao et al., 2018; Mao et al., 2019; Le et al., 2020; Rohanian et al., 2020), we evaluate the proposed models using three widely used datasets for metaphor detection.

VUA (Steen and Gerard) It represents the largest public evaluation dataset for metaphor detection that is used by the NAACL-2018 Metaphor Shared Task(Klebanov et al., 2016). It contains 10,567 sentences and the average length of sentences is 19.4. Annotation for this dataset is based on MIP, for which every word in a sentence is labeled for metaphor identification. There are two versions of this dataset: (1) VUA ALL POS, where words of all types (e.g., nouns, verbs, adjectives) are labeled, and (2) VUA VERB, which focuses only on the verbs for metaphor detection. In this paper, we consider only the VUA VERB version.

MOH-X (Mohammad et al., 2016) Here, the sentences are shorter and simpler than those in the other datasets, as they are sampled from WordNet. It contains 647 sentences and the average length of sentences is 8.0. Only one single verb is labeled in each sentence in this dataset.

TroFi (Birke and Sarkar, 2006) This dataset consists of sentences from the 1987-89 Wall Street Journal Corpus Release 1 (Charniak et al., 2000). It contains 3737 sentences and the average length of sentences is 28.3. Each sentence has a single annotated target verb. There are only fifty unique target verbs in this dataset, which means, that for one target verb, there are many training samples.

4.2 Baselines

RNN-ELMo (Gao et al., 2018) This very representative model uses Glove and ELMo as features for sequential metaphor identification. The ELMO word vectors they trained has been adopted in many subsequent works.

RNN-HG & RNN-MHCA (Mao et al., 2019) These are BiLSTM-based systems grounded in linguistic theories of SPV and MIP, which are the first to explore using linguistic theories to directly inform the design of Deep Neural Networks (DNN) for metaphor identification. They use the Glove and ELMO word embeddings as the literal meaning of a word.

MUL-GCN (Le et al., 2020) This is a multi-task learning model for metaphor detection that, to improve performance, features graph convolutional neural networks to appropriately capture the following; important context words, the control mechanism to emphasize the target words, and the transference of knowledge from WSD.

BERT+MWE-Aware GCN (Rohanian et al., 2020) This is a neural model to classify metaphorical verbs in their sentential context using information from the dependency parse tree and annotations for verbal multiword expressions. It evaluates on the MOH-X and TroFi datasets.

4.3 Setup

For pre-training ExampleBERT, we collect about 40,000 examples of the verb words in all three datasets (See Section 4.1). The batch size is 128; the learning rate is 5e-5, and we train over ten epochs. For DefinitionBERT, because different words have a different number of definitions and to achieve batch computing, we choose the most common three definitions[2] for each word. If a word don't have three definitions, we simply use *"no*

[2]Although the dictionary doesn't state that definitions listed in the page are sorted by frequency, it is basically indeed this case according to our observation.

Models	VUA VERB				MOH-X				TroFi			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
RNN-ELMo (Gao et al., 2018)	69.1	53.4	65.6	58.9	78.5	75.3	84.3	79.1	73.7	68.7	74.6	72.0
RNN-HG (Mao et al., 2019)	82.1	69.3	72.3	70.8	79.7	79.7	79.8	79.8	74.9	67.4	**77.8**	72.2
RNN-MHCA (Mao et al., 2019)	81.8	66.3	75.2	70.5	79.8	77.5	83.1	80.0	75.2	68.6	76.8	72.4
MUL-GCN (Le et al., 2020)	83.2	72.5	70.9	71.7	79.9	79.7	80.5	79.6	**76.4**	73.1	73.6	**73.2**
BERTBaseline (Rohanian et al., 2020)	-	-	-	-	78.04	78.38	77.87	77.82	70.38	70.54	68.89	68.84
BERT+MWE-Aware GCN (Rohanian et al., 2020)	-	-	-	-	80.47	79.98	80.40	80.19	73.45	**73.78**	71.81	72.78
BERTbaseline(OURS)	85.48	**77.35**	72.19	75.07	80.05	80.65	76.70	78.43	75.05	73.02	67.86	70.30
ExampleBERT(OURS)	85.29	75.56	75.30	75.43	82.22	**83.21**	77.81	80.25	75.38	73.21	68.67	70.84
DefinitionBERT(OURS)	**85.65**	76.02	**76.15**	**76.09**	**84.24**	82.90	**84.09**	**83.38**	75.70	73.32	69.64	71.40

Table 1: Performance on three metaphor detection datasets

definition ." to replace the remaining insufficient definitions. We believe that three definitions cover most of the context of words; adding more uncommon definitions could become noise for the model. Following the settings in the prior work, we perform 10-fold cross validation on MOH-X and TroFi and use the same splits of training, validation and test sets for VUA VERB datasets. For VUA VERB datasets, we select the best checkpoint on validation data as the final model to evaluate test data performance. For the MOH-X and TroFi datasets, we train 10 epoch and select the last epoch model to evaluate the test data for every fold. Finally we take the 10 fold average for the performance of our final model.

To pre-train ExampleBERT and fine-tune DefinitionBERT, we all use the pre-trained uncased BERT_{base} model from (Wolf et al., 2020). The number of its transformer blocks is 12, the number of self-attention heads is 12, and the number of the hidden layer is 768. For the FFNN in Eq. 2 of DefinitionBERT, we simply use a 256 hidden units of fully connected layer, followed by a classification layer. The two models are all fine-tuned with shuffled minibatches of size 32. The Adam optimizer is used to update the parameters, and the initial learning rate is set at 5e-5.

4.4 Results

Results in terms of accuracy (Acc), precision (P), recall (R) and F1-score are given in Table 1. Scores with the best performances across all models are indicated in bold. Results not reported are indicated by (-). As shown in Table 1, our ExampleBERT and DefinitionBERT achieve state-of-the-art performance on VUA VERB and MOH-X datasets.

VUA VERB dataset. For the VUA VERB datasets, even our proposed BERT-Baseline model achieves excellent performance, gaining improvement over the best of the other methods (MUL-GCN) by a large margin: 2.28% and 3.37% on accuracy (Acc) and F1, respectively. Compared with our BERTBaseline, regarding F1, our ExampleBERT and DefinitionBERT show improvement of 0.36% and 1.02%, respectively.

MOH-X dataset. For the MOH-X dataset, our DefinitionBERT, compared with BERTBaseline and the best of the other models, achieves significant improvement across all results.

TroFi dataset. However,for the TroFi dataset, the performance of our ExampleBERT and DefinitionBERT is somewhat bad than other state-of-the-art results. Compared with our BERTBaseline, for F1, ExampleBERT and DefinitionBERT still show a gain of 0.54% and 1.10%, respectively, indicating that our method is effective also. The TroFi dataset, contains fewer samples than the VUA dataset, but with longer average sequence length (28.3). Thus, on one hand, it is more difficult for DefinitionBERT to capture the relationship between the target and its definitions. On the other hand, because the dataset contains only fifty unique verbs, there are many samples for a target verb, and most express the literal meanings of the word, e.g., the dataset contains 71 literal sentences and 25 metaphor sentences of the target word 'absorb'. Thus, the models can learn sufficient common contextual information and literal meanings of a target word from the dataset. That is to say, the prior knowledge we add provides only limited help. However, taking a step back, compared with the performances of other well-designed models, the performance of our model does not lag too far behind; therefore, we believe our method is still acceptable. Moreover, the results of MOH-X and TroFi dataset suggest that our two models are more useful when there exists only a small amount of training corpus.

We note that the DefinitionBERT always perform expect on precision, the possible reason is , when the model cannot obtain the context meaning

of the target world accurately(i.e.,when the sentence is complex) , or the definitions we get from the dictionary are highly summarized, there would exist a gap between the context meaning and literal meaning, although they express the same meaning actually. So the model could predict the literal one to metaphorical more likely, then the precision could be lower(precision=TP/(TP+FP), FP increased). Finally, this could be a inspiration that how to improve our methods in the future work.

4.5 Analysis

As described in Section 3.4, if DefinitionBERT works correctly, it should learn the differences and relationships between the contextual and literal meanings expressed by the definitions of the target word during identifying. Therefore, to understand how DefinitionBERT uses the definitions we provide, we compute the cosine similarity between h_{tt} and each h_{di} described in Eq. 2. If a word is predicted as metaphorical, then the cosine similarity between its definitions will be very small, and the definition which expresses the literal sense of its contextual meaning would always have the smallest cosine similarity. Inversely, if a word is predicted as non-metaphorical, the value will be larger, and the meaning with the greatest cosine similarity will always be the definition that expresses its contextual meaning.

A specific example is given in Table 2. For example, in the sentence, *Her husband often abuses alcohol*, 'abuses' is a metaphorical word; its context meaning is 'a man drinks too much resulting in a bad effect'. Thus, we infer this metaphorical meaning is based on the first definition that has the smallest similarity. In the sentence, *This boss abuses his workers*, 'abuse' is a non-metaphorical word; its context meaning is 'speaking in a insulting and offensive way', which obviously is the third definition that has the greatest similarity. That is to say, our DefinitionBERT takes advantage of the definitions during training. The definitions directly help the model distinguish the contextual and literal meanings of the target word, which exactly is our purpose.

4.6 Disscussion

The main reason for the improvements in our experimental results is that we use external knowledge based on linguistic theories, which is very suitable and effective for detecting metaphors. The ways we incorporate the examples and definitions of a word correspond exactly to the two pre-training objectives of BERT, which are also its advantages. However, it seems possible to combine ExampleBERT and DefinitionBERT to attain better performance, we can use pre-trained ExampleBERT to fine-tune DefinitionBERT. But, our experimental results show that, although its performance can surpass that of ExampleBERT, but cannot surpass DefinitionBERT. The possible reason is may due to the only MaskLM pre-training objective, the ability of pre-trained ExampleBERT to model the relationship of a pair-wise is weakened.

RNN-HG and RNN-MHCA proposed by (Mao et al., 2019), which are inspired also by linguistic theories, focus more on the model architecture suitable for SPV or MIP; whereas, we focus on external knowledge suited for SPV or MIP. Moreover, we believe our ExampleBERT and DefinitionBERT are just base models that can be further improved by other technology, such as GCN applied in (Rohanian et al., 2020).

Moreover, compared to previous state-of-the-art models, especially knowledge-based methods like (Le et al., 2020; Rohanian et al., 2020), our DefinitionBERT is highly interpretable while achieving excellent performance. As described in Section 4.5, because our DefinitionBERT locates the intended meaning of the metaphor in context, it helps us further interpret metaphors. One approache for metaphor interpretation is *Definition Generation* proposed in (Zayed et al., 2020), which aims to find the most probable definition/interpretation (if exists) of the highlighted expression among the given definitions. Obviously, our DefinitionBERT is very suitable for this task(dataset). Another approach is *Lexical Substitution* explored in (Mao et al., 2018), where the metaphoric word/phrase is replaced with its literal counterpart to clarify its semantic meaning. We also believe our DefinitionBERT can be an alternative method for (Mao et al., 2018).

5 Conclusion

We proposed two simple, but effective, methods for metaphor detection, which achieve state-of-the-art performance on two verb metaphor detection datasets. More importantly, we showed that our DefinitionBERT is highly interpretable and can be further applied to metaphor interpretation. For future work, we will explore how to use the external knowledge of words for a sequential task, such as the VUA ALL POS dataset, which is not evaluated

Abuse (definitions according to the Oxford Dictionary)					
d-1 : use (something) to bad effect or for a bad purpose;misuse.					
d-2 : treat with cruelty or violence, especially regularly or repeatedly.					
d-3 : speak to (someone) in an insulting and offensive way.					
Abuse (samples and cosine similarity between the three definitions)					
samples	**label**	**predict**	**d-1-s**	**d-2-s**	**d-3-s**
1.Her husband often abuses alcohol .	1	1	0.1632	0.1948	0.1800
2.This boss abuses his workers .	0	0	0.6448	0.6457	0.6478
3.The actress abused the policeman who gave her a parking ticket .	0	0	0.7344	0.7312	0.7403
4. Do n't abuse the system .	1	1	0.0274	0.0267	0.0330

Table 2: Examples for the word 'abuse' from the MOH-X dataset. **'d-1-s'** indicates the cosine similarity between the feature vector of the first definition and the feature vector of the target word extracted from DefinitionBERT.

in this paper. A simple, crude way is to collect all the examples of words in the datasets and then continue to use ExampleBERT according to SPV. If based on MIP, combining the definitions of all words into one sentence like this paper do seems to be a terrible implementation. Moreover, there are several dictionaries (Zayed et al., 2020) giving examples and definitions of the word, and the examples or definitions from different dictionaries are somewhat different in types and contents, which may cause a different result when combined with our methods. Therefore, to obtain better performance, we will try resources from different dictionaries, where the premise is the definitions of the words must be non-metaphorical.

Acknowledgments

We thank Mr.Michael McAllister for his valuable assistance in proofreading this paper. We also appreciate the anonymous reviewers for providing valuable suggestions.

References

Dan Assaf, Yair Neuman, Yohai Cohen, Shlomo Argamon, Newton Howard, Mark Last, Ophir Frieder, and Moshe Koppel. 2013. Why "dark thoughts" aren't really dark: A novel algorithm for metaphor identification. In *CCMB*, pages 60–65. IEEE.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, volume 1, pages 4171–4186.

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

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. In *EMNLP/IJCNLP*, volume 1, pages 3507–3512.

Gitit Kehat and James Pustejovsky. 2020. Improving neural metaphor detection with visual datasets. In *LREC*, pages 5928–5933. European Language Resources Association.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutiérrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *ACL*, volume 2.

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

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *ACL*, volume 1, pages 1222–1231.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *ACL*, volume 1, pages 3888–3898.

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

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

Saif Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In **SEM@ACL*.

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

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *ACL*, pages 2890–2895. Association for Computational Linguistics.

Elena Semino et al. 2007. Mip: a method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

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

Steen and Gerard. *A method for linguistic metaphor identification : from MIP to MIPVU*. John Benjamins Pub. Co.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.* Number 14 in Converging Evidence in Language and Communication Research. John Benjamins.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *ACL*, volume 1, pages 248–258.

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

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Fig-Lang@NAACL-HLT*, pages 110–114.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *LREC*, pages 5810–5819. European Language Resources Association.