# Being neighbourly: Neural metaphor identification in discourse

**Verna Dankers[1], Karan Malhotra[1], Gaurav Kudva[1],**
**Volodymyr Medentsiy[1], Ekaterina Shutova[2]**
[1]University of Amsterdam, The Netherlands
[2]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
vernadankers@gmail.com
{karan.malhotra,gaurav.kudva,volodymyr.medentsiy}@student.uva.nl
e.shutova@uva.nl

## Abstract

Existing approaches to metaphor processing typically rely on local features, such as immediate lexico-syntactic contexts or information within a given sentence. However, a large body of corpus-linguistic research suggests that situational information and broader discourse properties influence metaphor production and comprehension. In this paper, we present the first neural metaphor processing architecture that models a broader discourse through the use of attention mechanisms. Our models advance the state of the art on the all POS track of the 2018 VU Amsterdam metaphor identification task. The inclusion of discourse-level information yields further significant improvements.

## 1 Introduction

Metaphor widely manifests itself in natural language. It is used to make an implicit comparison between two distinct domains that have certain common aspects (Lakoff and Johnson, 1980). For instance, in the sentence "The *price* of the commodity is *rising*", the target domain of quantity (*price*) can be understood through the source domain of directionality (*rising*).

The majority of computational approaches to metaphor focus on the task of its identification in text. Early approaches utilised hand-crafted features based on word classes (Beigman Klebanov et al., 2016), concreteness and imageability ratings (Turney et al., 2011; Broadwell et al., 2013) or selectional preferences (Wilks et al., 2013). Succeeding research has moved on to corpus-based techniques, such as the use of distributional and vector space models (Shutova, 2011; Gutierrez et al., 2016; Bulat et al., 2017), and more recently, deep learning methods (Rei et al., 2017). Current metaphor identification approaches cast the problem in the sequence labelling paradigm and apply convolutional (Wu et al., 2018), recurrent (Gao et al., 2018; Mao et al., 2019; Dankers et al., 2019) and transformer-based neural models (Dankers et al., 2019).

However, these approaches model only local linguistic context, i.e. information about the sentence in which the metaphor resides. Yet, a large body of corpus-linguistic research suggests that metaphor production and comprehension is influenced by situational information and wider discourse properties (Musolff, 2000; Semino, 2008; Jang et al., 2015b). Previously presented computational models of metaphor incorporating discourse use handcrafted features (Jang et al., 2015a) or neural sentence embeddings (Mu et al., 2019) within a simple classification paradigm. Since these methods employ shallow classification models, their task performance is subpar compared to deep neural architectures. Nonetheless, these studies established that discourse-level information is beneficial for metaphor detection.

Improving upon prior methods, we present a novel neural metaphor identification architecture that incorporates broader discourse. To model discourse, we investigate two types of attention mechanisms: a shallow general attention mechanism and a hierarchical one (Yang et al., 2016). The former builds a sentence representation by applying word-level attention. The latter combines attention at both word and sentence level. We apply our models to the 2018 VU Amsterdam (VUA) metaphor identification shared task (Leong et al., 2018), specifically to the all POS subtask. This task involves metaphor detection for all open-class words – i.e. verbs, adjectives, nouns and adverbs. Our results confirm that modelling discourse is beneficial for metaphor detection and our models improve upon the previous state-of-the-art task performance (Wu et al., 2018) by 5.2 and 6.4 F1-points, for our best-performing baseline and discourse models, respec-



*Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234

https://doi.org/10.18653/v1/P17

tively. To the best of our knowledge, this is the first end-to-end neural approach investigating the effect of broader discourse on metaphor identification.

## 2 Related Work

### 2.1 Deep Learning for Metaphor Identification

The approach of Wu et al. (2018) obtained the highest performance in the 2018 VUA metaphor identification task. Their model combined a convolutional neural network and a bidirectional LSTM (Bi-LSTM), thus utilising local and long-range contextual information in the immediate sentence. F1-scores of 65.1% and 67.2% were obtained in the task's all POS and verbs-only subtasks, respectively. See Leong et al. (2018) for an overview of other systems submitted to the task.

Afterwards, Gao et al. (2018) proposed a sequence labelling model for metaphor identification that employed GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) embeddings as input to a Bi-LSTM followed by a classification layer. The main difference compared to previously presented neural models was the inclusion of contextualised word embeddings, which significantly improved metaphor detection. Gao et al. (2018) reported results on the full sequence labelling task using the VUA metaphor corpus. However, their performance is not comparable to the all POS subtask of the 2018 shared task. The evaluation of Gao et al. (2018) included both closed- *and* open-class words and their models were trained on a different subset of the VUA metaphor corpus, causing incomparable main task performance measures.[1]

Mao et al. (2019) and Dankers et al. (2019) recently presented improved approaches to modelling metaphors by relying on (psycho)linguistically motivated theories of human metaphor processing. Mao et al. (2019) proposed two adaptations of the model of Gao et al. (2018): Firstly, concatenating the hidden states of the Bi-LSTM to a context representation capturing surrounding words within the current sentence, to model selectional preferences. Secondly, including word embeddings both at the input and classification layer, to explicitly model the discrepancy between a word's literal and its contextualised meaning. Dankers et al. (2019) improved metaphor identification through joint learning with emotion prediction, motivated by the finding that metaphorical phrases tend to be more emotionally evocative than their literal counterparts. Joint learning was applied to the model of Gao et al. (2018) as well as to BERT (Devlin et al., 2019). The latter setup is the current state-of-the-art approach in metaphor identification. Mao et al. (2019) and Dankers et al. (2019) used the data subset and evaluation setup of Gao et al. (2018), which complicates direct performance comparisons to the 2018 shared task. We only compare to these studies in our performance breakdown per POS tag, for the four open-class POS categories.

### 2.2 Metaphor Identification and Discourse

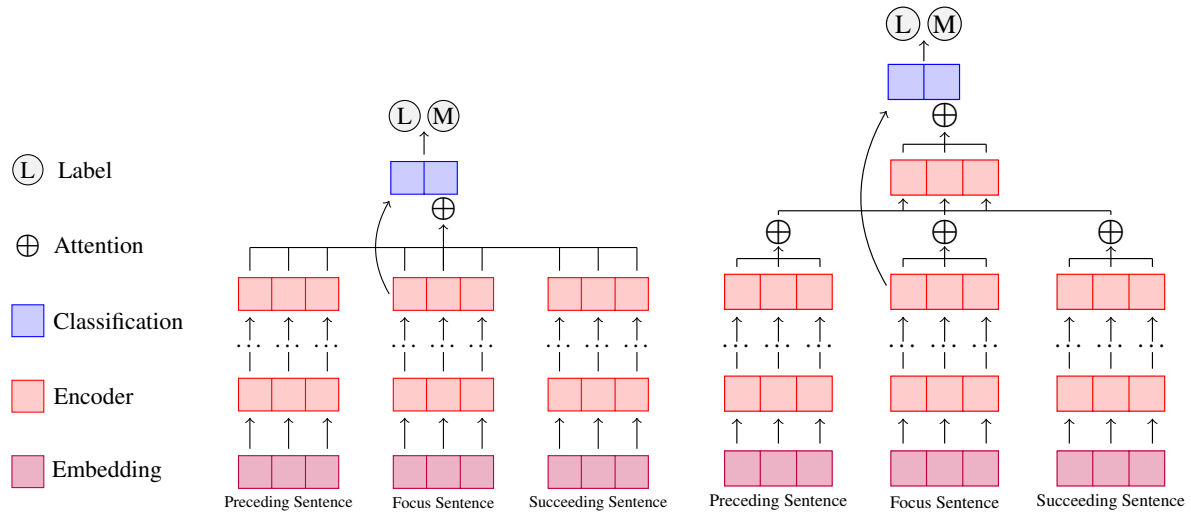
The work of Jang et al. (2015a) was the first to investigate the effects of broader discourse in a computational model of metaphor. Their approach used hand-crafted features and coarse-grained lexical information extracted from a broader discourse such as topical information, lexical chains and unigram features. However, they did not directly model the effect of including neighbouring sentences in metaphor identification.

Mu et al. (2019) considered metaphor identification for the verbs-only subtask of the 2018 shared task. They obtained the broader context of verbs by embedding the surrounding paragraph with a range of methods: GloVe, ELMo, skip-thought (Kiros et al., 2015) and doc2vec (Le and Mikolov, 2014). The context embedding, along with the verb lemma and syntactic arguments, was used to train a gradient boosting decision tree classifier. The authors have shown that metaphor identification is positively influenced by including paragraph-level context. Their best-performing model achieved an F1-score of 66.8% falling just shy of the 2018 verbs-only subtask's highest performance of Wu et al. (2018).

## 3 Data

The task revolves around performing binary classification – identifying whether a word is metaphorical or literal – on the all POS subtask of the 2018 VUA metaphor identification shared task. This task uses a dataset consisting of 117 text excerpts from the British National Corpus (Clear, 1993), labelled in the VUA metaphor corpus (Steen et al., 2010). Each excerpt has been retrieved from one of the following four genres: academic, news, conver-

[1] Closed-class function words such as prepositions are considerably easier to classify than open-class words. The systems evaluated in the 2018 shared task setup essentially addressed a more challenging task, which would make a task performance comparison unfair.




(a) General Attention (b) Hierarchical Attention

Figure 1: Visualisation of the sequence labelling architecture, for the two discourse computation mechanisms with context window size $k = 1$. To avoid visual clutter, only the classification of the first word in the focus sentence is shown.

sation, and fiction. Metaphorical expressions are annotated at word level. Following the shared task setup, we evaluate our models on open-class words only – i.e. verbs, adjectives, nouns and adverbs. The dataset contains 72,611 and 22,196 labelled words in the training and test set, respectively. 15% and 18% of these words are metaphorical for the two sets, respectively. We randomly sample 10% of the training data texts for validation purposes.

## 4 Methods

We construct a neural architecture that is optimised to predict binary metaphoricity at word level, by embedding the input words, applying encoder layers followed by a classification layer and softmax activation to yield per-word predictions. We present two variants of the model. The first method is *feature-based*, similar to the model of Gao et al. (2018). We embed input words through the concatenation of their non-contextualised (GloVe) and contextualised (ELMo) embeddings. The concatenation of GloVe and ELMo feeds into a one-layer Bi-LSTM encoder. During optimisation, we learn the parameters of the Bi-LSTM and classification layer.

As our second method, we use a *fine-tuning* architecture, similar to Dankers et al. (2019), in which the embeddings and recurrent encoder are replaced with the pretrained $\text{BERT}_{\text{base}}$ model (Devlin et al., 2019). BERT uses embeddings for subword units that are encoded with twelve transformer encoder layers. During optimisation we fine-tune BERT and learn the parameters of the classification layer. A word is considered metaphorical if any of its subword units is labelled as metaphorical. This choice is based on the assumption that it is more likely that a common prefix or suffix is not considered metaphorical while a word's main piece is than the other way around.

To include discourse information in both architectures, the output of the last encoder layer is concatenated to a discourse representation and fed to the classification layer, such that the classification layer contains dedicated parameters for both the discourse and (sub)word representations. We further detail the two mechanisms employed to compute the discourse representations below.

### 4.1 Modelling Discourse

**Discourse Definition** To represent discourse, we use a context window of size $2k + 1$ sentences. It comprises $k$ preceding sentences, the immediate sentential context of the word to be classified (the *focus* sentence), and $k$ succeeding sentences. However, based on the position of the focus sentence in the corresponding text, the number of preceding or succeeding sentences can be less than $k$. A value of 0 for $k$ implies a context window containing the focus sentence only.

**General Attention** The first attention mechanism constructs a discourse representation by applying general attention to all tokens within the

| Model | $k$ | P | R | F1 |
|---|---|---|---|---|
| Wu et al. (2018) | - | 60.8 | 70.0 | 65.1 |
| ELMo-LSTM | | | | |
| - General Att. | 0 | 66.3 | 64.8 | 65.5±.3 |
| | 1 | 66.3 | 66.6 | 66.4±.4 |
| | 2 | 66.8 | 67.8 | **67.3**±.2 |
| | 3 | 66.6 | 67.3 | 66.9±.4 |
| - Hierarchical Att. | 1 | 67.5 | 65.5 | 66.5±.4 |
| | 2 | 67.6 | 66.1 | 66.8±.3 |
| | 3 | 68.1 | 66.1 | 67.1±.6 |
| BERT | | | | |
| - General Att. | 0 | 73.5 | 67.4 | 70.3±.5 |
| | 1 | 72.6 | 69.8 | 71.1±.6 |
| | 2 | 73.7 | 68.9 | 71.1±.5 |
| | 3 | 72.8 | 70.0 | 71.3±.5 |
| - Hierarchical Att. | 1 | 73.1 | 69.1 | 71.0±.4 |
| | 2 | 73.5 | 69.6 | **71.5**±.5 |
| | 3 | 73.8 | 68.9 | 71.3±.5 |

Table 1: Main task performance for the all POS 2018 VUA metaphor identification task. The highest performance per model type is shown in bold font.

| Model | $k$ | VB | ADJ | NN | ADV |
|---|---|---|---|---|---|
| Mu et al. (2019)† | - | 66.8 | - | - | - |
| Wu et al. (2018) | - | 67.4 | 65.1 | 62.9 | 58.8 |
| Gao et al. (2018)† | - | 69.9 | 58.3 | 60.4 | - |
| Mao et al. (2019)† | - | 70.8 | 62.2 | 63.4 | 63.8 |
| ELMo-LSTM | | | | | |
| - General Att. | 0 | 69.7 | 62.0 | 62.5 | 57.3 |
| | 2 | 71.2 | 63.5 | 65.0 | 57.6 |
| - Hierarchical Att. | 3 | 71.5 | 63.2 | 64.2 | 57.1 |
| BERT | | | | | |
| - General Att. | 0 | 74.6 | 65.9 | 67.5 | 64.3 |
| | 3 | 75.6 | 66.4 | 68.9 | 64.0 |
| - Hierarchical Att. | 2 | 75.7 | 66.0 | 69.3 | 63.2 |

Table 2: Task performance (F1-score) breakdown per POS category, for the baseline systems and best-performing setup per model and attention module type. †Due to differences in the data subset used and evaluation setup these results are not directly comparable to ours.

context window. The encoder layers are applied to each of the sentences within the context window individually. The discourse representation is the weighted combination of the outputs, where the weights are computed by applying a linear layer followed by the softmax function. The architecture is shown in Figure 1a.

**Hierarchical Attention** Secondly, we replace the general attention with hierarchical attention inspired by the work of Yang et al. (2016) to combine word- and sentence-level attention. The former type provides fine-grained information needed for disambiguation and possibly co-reference resolution, whereas the latter is more suited to capture coarse-grained topical information. First, the individual sentences from the context window are encoded, and general attention is applied per sentence, yielding sentence representations. Second, the sentence representations are fed to a sentence-level encoder, and sentence-level attention is utilised to produce a discourse representation. For the recurrent architecture, the encoder is a Bi-LSTM, and for BERT, it is a transformer layer. The hierarchical attention module is visualised in Figure 1b.

## 5 Experiments and Results

### 5.1 Experimental Setup

The *feature-based* approach uses GloVe and ELMo embeddings of dimensionalities 300 and 1,024, respectively. The hidden state dimensionality of the Bi-LSTM is 128. Training lasts for 10 epochs, with a maximum learning rate of 0.005 and batches of size 64.

The *fine-tuning* method includes the $\text{BERT}_{\text{base}}$ model that has 12 pretrained transformer layers with a hidden dimensionality of 768. The BERT model is fine-tuned for 4 epochs with a batch size of 16, and a maximum learning rate of $5\text{e}{-5}$.

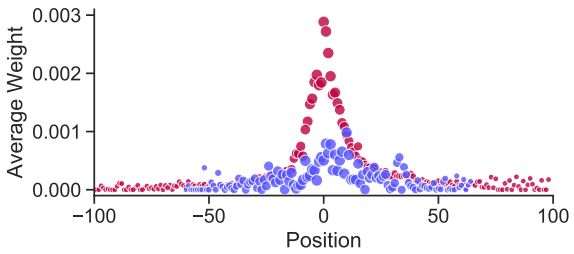
Both model types are trained using the AdamW optimiser with a cosine-based learning rate scheduler and a warm-up period of 10%. Tokens with a POS tag other than the four open-class categories are included in the sentence, albeit that their metaphoricity is not considered during optimisation. We use the negative log-likelihood loss along with class weights to account for the class imbalance. The weights are annealed during training from 0.9 and 0.1 to 0.7 and 0.3 for the metaphorical and literal classes, respectively. We report the precision (P), recall (R), and F1-scores for the metaphor class achieved by each model averaged over ten randomly initialised runs.

### 5.2 Results

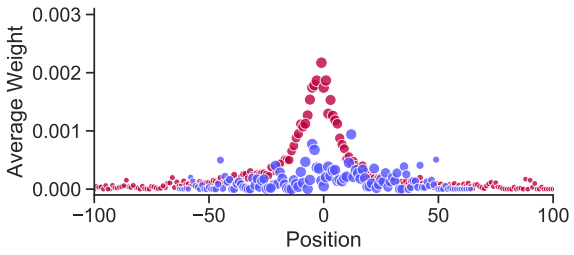
Table 1 presents our main task performance. We compare our models to the previous highest performing system designed by Wu et al. (2018). Our baseline systems using $k = 0$ only incorporate the sentential context of the focus sentence. The recurrent and BERT-based models with $k = 0$ already outperform the approach of Wu et al. (2018) by the margins of 0.4 and 5.2 F1-points, respectively.

The inclusion of discourse representations further improves the F1-scores over the baseline methods, for both model types and both attention modules. The performance differences are significant as per a $t$-test ($p < 5e{-}3$) for all experimental setups including wider discourse ($k > 0$). This finding is in accordance with the findings of Mu et al. (2019). Generally, the largest performance gain is achieved by increasing $k$ from 0 to 1. This indicates that the immediate neighbouring sentences are the most informative. This claim is supported by Bizzoni and Ghanimifard (2018) who mention that the metaphors in the VUA metaphor corpus generally do not require long-distance information for their resolution. The top-performing model is the BERT setup with hierarchical attention and $k = 2$ that achieves a state-of-the-art F1-score of 71.5%. Overall, we observe that using the hierarchical attention module is more effective at increasing the precision, while the general attention module is more likely to improve the recall.

Table 2 displays more fine-grained F1-scores per POS category for the best-performing experimental setups per model type and attention module. We notice that the increase in performance is mainly achieved by increments in the F1-scores for verbs and nouns rather than adjectives and adverbs.




(a) ELMo-LSTM, general attention, $k = 1$




(b) BERT, general attention, $k = 1$

Figure 2: Distribution of the attention weight per token position, for (1) all test sentences and (2) the test sentences whose metaphoricity labels were corrected by including discourse information.

## 6 Analysis and Discussion

In order to investigate the types of information that discourse provides, we have conducted a qualitative analysis of the sentences where our discourse-aware models improved the labelling of a word over the discourse-agnostic baselines. We found that discourse helps primarily in two ways: (1) it provides information about the topic of the text, which is often needed for disambiguation, particularly for shorter sentences; and (2) it allows the models to implicitly perform co-reference resolution through the use of word-level attention.

We observe that while the attention distributions of the hierarchical mechanisms are rather diffused, the general attention mechanism uses very sparse weights. We hypothesise that the former type is more suited for modelling coarse-grained topical information. The latter is able to highlight specific words and phrases in neighbouring sentences that may be needed for disambiguation and co-reference resolution. The attention distributions displayed in Table 3 are exemplary of distributions in the hierarchical and general attention modules with regard to sparseness. This observation holds for both the recurrent and BERT-based models.

To further investigate the internal functioning of the attention modules and the influence of discourse, we use majority voting across the randomly initialised runs to obtain samples that are consistently improved by all discourse setups compared to the baseline per model type. For the recurrent models, this subset contains 109 samples: 53 metaphorical tokens and 56 literal ones. For BERT, these statistics are 57 and 19, respectively. To assert that within these subsets, the surrounding sentences affect the discourse representation, we visualise the average weight per word position, as measured from the middle of the focus sentence. Figure 2 demonstrates that for sentences in this subset, the distribution is more diffused compared to all test sentences. Thus, on average, the broader discourse has a more substantial effect on the model's classification for these samples. This finding supports the hypothesis that the numerical performance gain observed is caused by the inclusion of discourse representations.

The examples listed in Table 3 are drawn from the consistently improved subset for all recurrent discourse setups. Specifically, the annotations of the words *mincemeat*, *spread* and *hurry* were corrected. In the first example, the context contains

| **Hierarchical Attention** *Sentence* | *Word* | **General Attention** |
|---|---|---|
| .297 | I 've got_L some cooking_L apples_L out there Oh is n't he , I could hit_L him ! | apples_L |
| .419 | Why does n't he make_L make_L your own bloody_M mincemeat_L then ! | bloody_M, mincemeat_L |
| .284 | Yeah that 's cos_L I make_L the the pastry_L and you can | - |
| .276 | As you always_L still_L continue_L to tell_L them yes_L you do . | - |
| .321 | If I want_L it spread_M around . | spread_M, around |
| .403 | That gives_M you your bit_M of character_M . | gives_M, bit_M, character_M |
| .344 | The accelerator_L is the one on the right_L sir_L ! | - |
| .268 | He is an old_L gentleman_L , dear_L but er_L he is n't in a hurry_L what so ever_L . | hurry_L |
| .388 | Look_L he 's slowing_L down there ! | Look_L, slowing_L, down |

Table 3: Visualisation of the attention distributions for three example sentences in which the underlined token was correctly labelled after including wider discourse information, for ELMo-LSTM with $k = 1$. The colour intensity represents the word-level attention weights. For hierarchical attention, the sentence weights (first column) influence the effect of the word-level weights (second column). Since the general attention distributions were rather sparse, we only include the key words (column three).

food references ("cooking apples" and "make the pastry") and hence provides the required topical information, emphasising that *mincemeat* is used in a text about cooking. If "making mincemeat of something" were used metaphorically, one would expect neighbouring sentences to discuss defeats, demolition or devastation. The second example illustrates a case where the wider context is needed for anaphora resolution: the meaning of *it* is unclear from the sentence itself, which hinders the metaphoricity resolution of *spread*. While the context does not specify the referent of *it*, it still incorporates clues through the phrases "continue to tell them" and "gives you character" that indicate something is not spread physically but socially. When looking into the even broader context (available to models with $k > 1$), *it* appears to be gossip about a wild night involving alcohol. This example indicates that the directly neighbouring sentences are not always sufficient for complete clarity – i.e. that increasing $k$ beyond 1 can occasionally be helpful. In the final example, it becomes apparent from the metaphor's discourse context that "being in a hurry" in fact concerns the actual speed of the accelerator, as opposed to the metaphorical use of hurrying: the obstacle that keeps us all from living life fully.

The results in Table 2 show that the use of discourse information primarily improves performance for verbs and nouns, and less so for adjectives and adverbs. We hypothesise that much of this improvement is due to pronominal co-reference resolution, which is most critical for verbs and nouns. Pronouns replace nouns and noun phrases in sentences and are themselves in direct grammatical relations with the verbs (as their subject or object). A verb may be used metaphorically or literally, but without knowing the identity of its subject or object, its metaphoricity may be difficult to determine. For adverbs and adjectives that would not be the case, as they never modify a pronoun.

## 7 Conclusion

In this work, we presented deep neural architectures that make use of attention mechanisms to investigate the impact of discourse in the task of word-level metaphor detection. Our models establish new state-of-the-art results in the all POS track of the 2018 VUA metaphor identification shared task (Leong et al., 2018). Two attention mechanisms were experimented with for modelling discourse, a general and hierarchical one. Both modules yield significant performance increases, but our qualitative analysis indicates that they serve a different purpose.

Considering the high variety in the corpus's sentence lengths, future work could include defining the context window in terms of words instead of sentences and the merging of techniques to reap the benefits of both co-reference resolution and capturing topical information.

## References

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

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of NAACL-HLT 2018 Workshop on Figurative Language Processing*, pages 91–101.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 102–110. Springer.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, volume 2, pages 523–528.

Jeremy H. Clear. 1993. The British National Corpus. In *The Digital Word: Text-Based Computing in the Humanities*, page 163–187. MIT Press, Cambridge, MA, USA.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, volume 1, pages 4171–4186.

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 (EMNLP)*, pages 607–613.

E Dario Gutierrez, 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 (ACL)*, volume 1, pages 183–193.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rosé. 2015a. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 384–392.

Hyeju Jang, Miaomiao Wen, and Carolyn Rose. 2015b. Effects of situational factors on metaphor detection in an online discussion forum. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 1–10.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 2, pages 3294–3302.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of NAACL-HLT 2018 Workshop on Figurative Language Processing*, pages 56–66.

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 (ACL)*, pages 3888–3898.

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 (NAACL-HLT)*, volume 1, pages 596–601.

Andreas Musolff. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. Iudicium, Muenchen.

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. 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 (NAACL-HLT)*, volume 1, 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 (EMNLP)*, pages 1537–1546.

Elena Semino. 2008. *Metaphor in Discourse*. Cambridge University Press, Cambridge.

Ekaterina Shutova. 2011. *Computational Approaches to Figurative Language*. Ph.D. thesis, University of Cambridge.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

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 Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.