Neural Metaphor Detection with Visibility Embeddings

Gitit Kehat and **James Pustejovsky**
Department of Computer Science
Brandeis University
Waltham, MA 02453 USA
{gititkeh,jamesp}@brandeis.edu

Abstract

We present new results for the problem of sequence metaphor labeling, using the recently developed *Visibility Embeddings*. We show that concatenating such embeddings to the input of a BiLSTM obtains consistent and significant improvements at almost no cost, and we present further improved results when visibility embeddings are combined with BERT.

1 Introduction

When browsing through vision-language datasets, one can make the intuitive observation that their textual parts ("visual corpora") contain more physical language, mostly descriptive, which tends to be non-metaphorical by nature (See, for example, typical images from the Visual Genome dataset in Figure 1). Recently, this property was used to build visibility embeddings, which aim to provide a good estimation of a word's concreteness, a feature that has been long related to metaphoricity (Lakoff and Johnson, 1980; Turney et al., 2011).

Many metaphors indeed involve noticeable differences between the abstractness of words constructing them, like "clean conscience" (vs. "clean air"). Metaphors are not created in isolation, commonly do not stand alone as non-literal expressions, and are highly context-dependent in nature. Even the most concrete and physical text can be considered as metaphorical when mentioned in a different context than its original one, or in proximity to another text from the target domain. For example, a single use of a verb like "push" or "leak" can have both literal and metaphorical meanings, in relation to its context (see Figure 1).

Technically, the task of metaphor detection at the sentence level is commonly approached as one of the following two tasks:
(1) **Sequence Labeling**, in which each token in the sentence is classified as either "metaphorical" or "literal" (multiple outputs per sentence).
(2) **Classification** of a specific target word, usually the main verb (one output per sentence). This task is sometimes called "verb classification".

L: **Water leaking** on the **road**.
M: The **news leaked** out despite his **secrecy**.

L: A **woman pushing** a **cart**.
M: The **liberal party pushed** for **reforms**.

Figure 1: Images from the Visual Genome (Krishna et al., 2016) along with their literal ("L") description, and a metaphorical ("M") sentence with a similar verb from the MOH-X dataset (Mohammad et al., 2016) (concrete words in green and abstract words in red).

Recently, Kehat and Pustejovsky (2020) presented the simply constructed Visibility Embeddings (VE), which use references to visual/non-visual corpora to estimate word concreteness, and applied it to the task of verb-classification. In this paper we apply VE also to the sequence labeling task, and show how they consistently improve the result of a BiLSTM model with BERT. We also discuss possible problems when reporting results

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 222–228
August 5–6, 2021, Bangkok, Thailand (online)

on very small annotated datasets, and the effect on adding GloVe to the model input.

2 Background and Previous Work

2.1 Visibility Embeddings

Visibility embeddings (VE) were shown by (Kehat and Pustejovsky, 2020) to be useful for metaphor detection when concatenated to the input of BiLSTM models for the verb classification task. These simple and no-cost embeddings, are created by checking the occurrence of each word in a set of different visual and non-visual corpora, as a way to estimate its concreteness. They developed the *big visual corpus* (BVC), which contains the textual parts of multiple vision-language datasets, such as Visual Genome (Krishna et al., 2016), ImageNet (Deng et al., 2009), MSCOCO (Lin et al., 2014) and Flick r 30K (Young et al., 2014), as well as a "non-visual" corpus, $Brown - BVC$, which is the subtraction of the BVC from the Brown corpus (Francis and Kucera, 1964). These two corpora were previously shown (Kehat and Pustejovsky, 2017) to be highly concrete and highly abstract on average, respectively.

2.2 Metaphor Detection

The current state-of-the-art in metaphor detection is achieved by neural methods, enriched with contextual word embeddings (such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019)), and commonly combined with varied linguistic features and metrics. Some notable results are the ones by Gao et al. (2018) who used ELMo and BiLSTM, Mao et al. (2019) who also experimented with BERT and features that rely on human metaphor processing, Dankers et al. (2019) who performed joint learning with emotion prediction, and Le et al. (2020) who used graph convolutional neural networks with dependency parse trees.

Impressive results[1] were presented in the 2018 Metaphor Detection Shared Task (Leong et al., 2018), with most of the groups using neural models with other linguistic elements like POS tags, WordNet features, concreteness scores and more (Wu et al., 2018; Swarnkar and Singh, 2018; Pramanick et al., 2018; Bizzoni and Ghanimifard, 2018), as well as in the more recent 2020 Shared Task (Leong et al., 2020), with the majority of groups using some variation of BERT in addition to the other features (Su et al., 2020; Gao and Zhang, 2002; Kuo and Carpuat, 2020; Torres Rivera et al., 2020; Kumar and Sharma, 2020; Hall Maudslay et al., 2020; Stemle and Onysko, 2020; Liu et al., 2020; Brooks and Youssef, 2020; Chen et al., 2020; Alnafesah et al., 2020; Li et al., 2020; Wan et al., 2020; Dankers et al., 2020).

Embedding-based approaches such as in Köper and Schulte im Walde (2017) and Rei et al. (2017) proved to work effectively on several annotated datasets. Different types of word embeddings were studied, including embeddings trained on corpora representing different levels of language mastery (Stemle and Onysko, 2018), and embeddings representing different dictionary categories in the form of binary vectors for each word (Mykowiecka et al., 2018). Previous work by Turney et al. (2011), Tsvetkov et al. (2014) and Köper and Schulte im Walde (2017) showed concreteness scores to be effective for Metaphor Detection, however, they all used fix concreteness score lists, such as the MRC (Coltheart, 1981) and the 40K list by Brysbaert et al. (2014), either as a reference or for training.

3 Model Details

As a base structure we use the simple BiLSTM architectures presented by Gao et al. (2018). The sequence labeling model (see Figure 2) consists of two layers, a BiLSTM and a feedforward layer, to get a label for each word in the sentence. We implemented the model in Python using the AllenNLP package (Gardner et al., 2017).

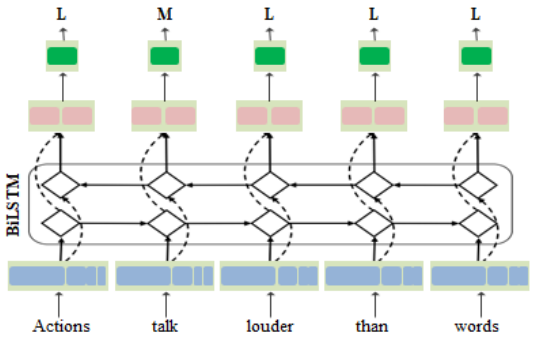

Figure 2: Simple sequence model with multiple outputs, one per word in the sentence.

We use a pretrained BERT model provided by the AllenNLP package, with 24 layers and 1024 hidden states, trained on cased English text. The

[1] yet not directly comparable to ours, since they used different train-test separations and evaluation, see Dankers et al. (2020)

input vector for the model consists of the concatenation of the 1024-dimensions BERT vector (using all the layers of the BERT model), the GloVe embeddings (Pennington et al., 2014) (not in all cases, see discussion in Section 4.3), and the VE of varied length (we experimented with vectors from length 50 and 300). Hyperparameters are fine-tuned on each dataset.

4 Experiment Setting and Results

We present results and comparison for two of the most common datasets for metaphor detection: **VUA** (Steen et al., 2010) and **MOH-X** (Mohammad et al., 2016). Annotated datasets for the validation and training of metaphor detection systems are not easily created, and require a level of expertise. The available datasets are therefore relatively small, hand crafted sets of several hundreds to a few thousands sentences, mostly only partially annotated for the metaphoricity of their main verb. As a result, the F1-scores vary highly, even with the slight change in parameters. In order to provide a consistent evidence to our algorithm's performance, we chose to compare not only the maximal F1-scores gained by each model, but also present a "parameterized" F1-score, over different learning rates. This would allow us to analyze the results while ignoring very highly-frequent fluctuations in the performance of the models.

4.1 VUA

We used the labels assigned to each token by the original VUA annotators. The verbs used for verb-testing are the ones used by Gao et al. (2018) (a large subset of all the verbs). Adding VE to the simple BiLSTM-BERT model achieves very high results (See Table 1). In order to provide more detailed comparison with previous models, results per POS are shown in Table 2.

Figure 3 demonstrates the consistent improvement gained by using VE by comparing four types of input vectors with different BERT - VE - GloVe combinations. Very similar learning rates (+-0.0001) can vary in up to +2 F1-Score, demonstrating the high variance those models have given the relatively small dataset. The random vector is of the same length and value range as the VE, with each value chosen randomly, to demonstrates that the length of the input vector has some effect on the results in terms of when the model reaches its maximum F1-score, as seen by the shifted gray line (BERT only). In this specific case, adding the GloVe vector improves the results (see discussion in Section 4.3).

	Model	P	R	F1
Verb Testing	Wu et al.*	60.0	76.3	67.2
	Gao et al.	68.2	71.3	69.7
	Mao et al.	69.3	72.3	70.8
	Su et al.*	78.9	81.9	80.4
	BERT+VE	72.2	75.0	73.6
All POS Testing	Wu et al.*	60.8	70.0	65.1
	Gao et al.	71.6	73.6	72.6
	Mao et al.	73.0	75.7	74.3
	Dankers et al.	—	—	76.8
	Su et al.*	75.6	78.3	76.9
	BERT+VE	77.1	77.8	**77.4**

Table 1: Sequence metaphor labeling on the VUA. Results denoted by * are not directly comparable.

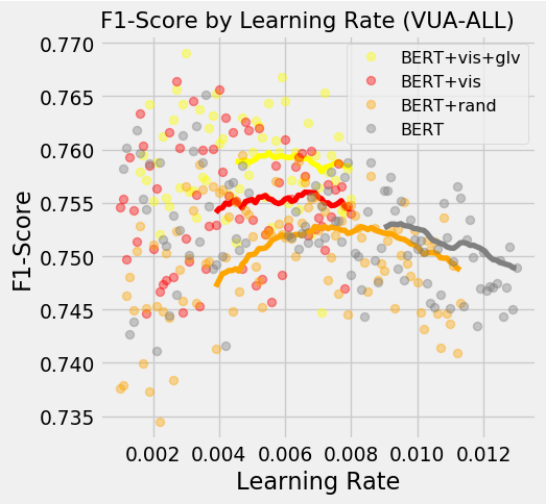

Figure 3: F1-scores as a function of the learning rate for VUA-ALL. The lines are moving averages of the corresponding points. **Yellow** - BERT + Visibility Embeddings + GloVe, **Red** - BERT + Visibility Embeddings, **Orange** - BERT + random vector, **Gray** - BERT only.

Figure 4 shows a similar comparison but in this case, the model is maximized on just the verbs of the classification task (as opposed to all words above). In all cases, adding visibility embeddings to the BERT embeddings achieves a no-cost improvement in the F1-score, both on average and as the maximal result gained for the model (over the given learning-rates gaps).

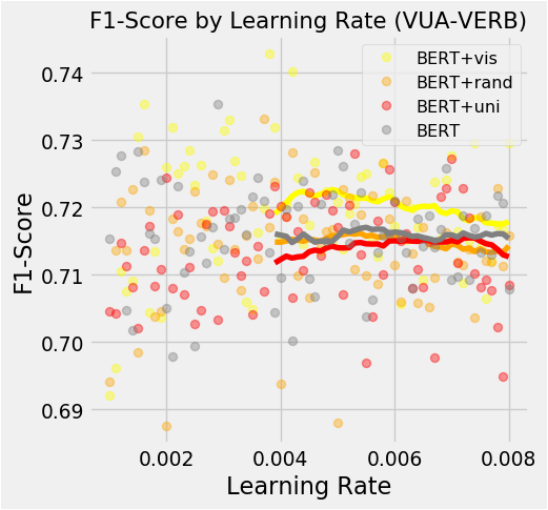

Figure 4: VUA target verbs testing. **Yellow** - BERT + Visibility Embeddings, **Orange** - BERT + random vector, **Red** - BERT with embeddings based on vocabulary (uniform positive value for all valid words), **Gray** - BERT only.

POS	P	R	F1
VERB	71.92	75.62	73.72
NOUN	71.33	67.85	69.55
ADP	89.36	91.55	90.44
ADJ	69.10	62.84	65.82

Table 2: Results by POS tags of the best model for VUA-All (BERT + GloVe + VE).

4.2 MOH-X

The MOH-X dataset (as a subset of the largest MOH dataset) was originally annotated for the main verbs only. It is small, and contains around 650 sentences. For the sequence labeling task, we use the default base case of assigning the rest of the tokens a "literal" label (as demonstrated in previous work). The results are presented in Table 3.

As a direct result from its size, testing on the MOH-X using ten-fold-CV with random splits yields fluctuating results. After conducting 50 random ten-fold-CVs (500 splits over all), we got an average F1-score of 82.3, with a maximum of 84.0 and a minimum of 81.0. Even though these two vary significantly, the minimum F1-score obtained is still higher in 1.0 F1-score point than the one recently reported by Mao et al. (2019) .

The above observation makes it hard to optimize and fine-tune the parameters of the model. We noticed that in general, higher F1-scores are gained for splits where the training set and evaluation set contain instances of the same verbs. Previously reported results did not explicitly mention this issue. To maintain consistency with the results by Gao et al. (2018) and Le et al. (2020), we present our results both on their prechosen sets, as well as on randomly chosen splits (rand-CV).

Model	P	R	F1
Gao et al. (2018)	79.1	73.5	75.6
Le et al. (2020)	79.7	80.5	79.6
Mao et al. (2019)	77.5	83.1	80.0
BERT+VE	83.8	85.8	84.6
BERT+VE (rand-CV)	80.8	84.7	82.3

Table 3: Results on the MOH-X dataset using sequence labeling. Our model improves upon the previous state of the art by Mao et al. (2019).

4.3 Further Discussion

In some cases, adding the GloVe to the input vector does not help to improve the results, and even worsens them. This is true for both the sequence and classification tasks on the MOH-X dataset, and varies in the VUA (as can be seen in Figures 3, 4), though the differences are relatively small.

Concatenating GloVe to the input vector provides additional generalized non-domain-specific (the pre-trained GloVe was trained on Wikipedia) context for each word in a sentence. The MOH-X dataset contains shorter sentences, so on average, every word in the sentence has more weight when determining the metaphoricity of the target verb. In particular, when the verb is used metaphorically, the few other words in the sentence play a special role in giving us clues about it, say, when they belong to different domains. Adding the information from GloVe might smooth this effect.

When applied to the VUA, the Glove's effect is minimized, since it contains longer sentences and we have more words that are not directly related to the main metaphor presented by the target verb. In general, the VUA gets much lower results than the MOH-X on all performed tasks, since it was created from real sentences, while the MOH-X was handcrafted from WordNet sample sentences for the specific task of detecting non-direct language. in real world texts, we should expect similar lower performances.

5 Summary

We have presented new and improved results for sequence metaphor labeling for the VUA and MOH-X datasets using visibility embeddings and BERT

as inputs for a simply constructed BiLSTM. We provided detailed comparison for the effect of adding VE to the model, and showed it to be a useful no-cost component to a metaphor detection system.

Acknowledgements

This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract #W911NF-15-C-0238 at Brandeis University. The points of view expressed herein are solely those of the authors and do not represent the views of the Department of Defense or the United States Government. Any errors or omissions are, of course, the responsibility of the authors.

References

Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210, Online. Association for Computational Linguistics.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Brooks and Abdou Youssef. 2020. Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 244–249, Online. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

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

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

W. Nelson Francis and Henry Kucera. 1964. Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1.

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

Jianfeng Gao and Min Zhang. 2002. Improving language model size reduction using better pruning criteria. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 176–182, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226, Online. Association for Computational Linguistics.

Gitit Kehat and James Pustejovsky. 2017. Integrating vision and language datasets to measure word concreteness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 103–108, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Gitit Kehat and James Pustejovsky. 2020. Improving neural metaphor detection with visual datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5930–5935, Marseille, France. European Language Resources Association.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Tarun Kumar and Yashvardhan Sharma. 2020. Character aware models with similarity learning for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 116–125, Online. Association for Computational Linguistics.

Kevin Kuo and Marine Carpuat. 2020. Evaluating a Bi-LSTM model for metaphor detection in TOEFL essays. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 192–196, Online. Association for Computational Linguistics.

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

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

Chee Wee (Ben) 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, Online. Association for Computational Linguistics.

Chee Wee (Ben) 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, New Orleans, Louisiana. Association for Computational Linguistics.

Shuqun Li, Jingjie Zeng, Jinhui Zhang, Tao Peng, Liang Yang, and Hongfei Lin. 2020. ALBERT-BiLSTM for sequential metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 110–115, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255, Online. Association for Computational Linguistics.

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

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Agnieszka Mykowiecka, Aleksander Wawer, and Malgorzata Marciniak. 2018. Detecting figurative word occurrences using recurrent neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 124–127, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana. Association for Computational Linguistics.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546, Copenhagen, Denmark. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

Egon Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 133–138, New Orleans, Louisiana. Association for Computational Linguistics.

Egon Stemle and Alexander Onysko. 2020. Testing the role of metadata in metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 256–263, Online. Association for Computational Linguistics.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Krishnkant Swarnkar and Anil Kumar Singh. 2018. Di-LSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, New Orleans, Louisiana. Association for Computational Linguistics.

Andrés Torres Rivera, Antoni Oliver, Salvador Climent, and Marta Coll-Florit. 2020. Neural metaphor detection with a residual biLSTM-CRF model. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 197–203, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, 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 *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 680–690.

Mingyu Wan, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109, 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 *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.