Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5928–5933
Marseille, 11–16 May 2020

Improving Neural Metaphor Detection with Visual Datasets

Gitit Kehat and James Pustejovsky
Brandeis University
Waltham, MA, USA
{gititkeh, jamesp}@brandeis.edu

Abstract

We present new results on Metaphor Detection by using text from visual datasets. Using a straightforward technique for sampling text from Vision-Language datasets, we create a data structure we term a *visibility word embedding*. We then combine these embeddings in a relatively simple BiLSTM module augmented with contextualized word representations (ELMo), and show improvement over previous state-of-the-art approaches that use more complex neural network architectures and richer linguistic features, for the task of verb classification.

Keywords: Metaphor Detection, Visual datasets, Visibility word embeddings, neural models

1. Introduction

Metaphors play a special role in human language and thought, as they evoke a complex array of hidden connotations, past experiences, feelings, and humor, in the service of helping the speaker convey their message in a way that is easier to relate to. However, by their very nature, metaphors continue to pose a challenge to Natural Language Processing (NLP) systems, and their identification is crucial for many tasks, such as Machine Translation, Information Retrieval, and others.

In most cases, metaphor identification is done at the sentence level, where the input consists of some or all of the words in the sentence, and the output refers to the metaphoricity of the word(s) in the specific context. Often, Metaphor Identification takes the form of one of two tasks: (1) Sequence Labeling, in which each token in the sentence is classified as either "metaphorical" or "literal" (multiple outputs per sentence), or (2) Classification of a specific target word, usually the main verb (one output per sentence). In this paper, we deal with the second task, which more formally takes a sentence $w_1, ...w_n$ and a verb index i as input, and outputs a label for the target verb w_i of either "metaphorical" or "literal", in relation to its role in the sentence (See Figure 1 for examples for non-metaphorical (literal) and metaphorical usages of the same verb in different contexts).

In our approach to improve metaphor detection, we follow Black (1979)'s observation that a metaphor is essentially an interaction between two terms, creating an "implication-complex" to resolve two incompatible meanings. Operationally, we follow Turney et al. (2011) and their adoption of Lakoff and Johnson (1980)'s notion that metaphor is a way to move knowledge from a concrete domain to an abstract one. Hence, there should be a correlation between the "degree of abstractness in a word's context [...] with the likelihood that the word is used metaphorically" (Turney et al., 2011). Recent studies have suggested that there is a strong correlation between the concreteness scores of words, as annotated by humans, and the visibility of words, as calculated as a function of their occurrences in a visual corpus (Kehat and Pustejovsky, 2017). In the present paper, we take this notion one step further and use visibility of words directly as a feature of the system.

More specifically, we further improve on the recently presented results by Gao et al. (2018) on the task of verb classification for metaphor detection. In their work, Gao et al. (2018) used contextual information, in the form of contextualized word embeddings (ELMo) (Peters et al., 2018), as well as the GloVe embeddings (Pennington et al., 2014), both concatenated and fed as an input to a simple BiLSTM. We use a number of popular Vision-Language Datasets to create what we call *Visibility Embeddings*. These embeddings are created by a simple sampling technique from visual corpora (the textual part of vision-language datasets, usually in the form of a list of image-caption sentences). We show that these Visibility Embeddings are useful when combined in a simple concatenation manner with the previously presented architecture by Gao et al. (2018). Our code is available at `https://github.com/gititkeh/visibility_embeddings`.

2. Background and Related Work

2.1. Metaphor Detection

Currently, neural methods are dominating the task of Metaphor Detection, with recent state-of-the-art results by Gao et al. (2018) and Mao et al. (2019), using BiLSTMs and contextualized word embeddings (ELMo) (Peters et al., 2018), demonstrated on a number of popular annotated Metaphor Detection datasets by Mohammad et al. (2016) (MOH-X), Steen et al. (2010) (the VU Amsterdam Metaphor Corpus (VUA)) and Birke and Sarkar (2006) (TroFi). In the recent 2018 VUA Metaphor Detection Shared Task, several neural models with different architectures were introduced. Most of the teams in the task used LSTM's combined with other linguistic features, such as part-of-speech tags, WordNet data, concreteness scores and more (Wu et al., 2018; Swarnkar and Singh, 2018; Pramanick et al., 2018; Bizzoni and Ghanimifard, 2018).

Previous work by Turney et al. (2011), Tsvetkov et al. (2014) and Köper and im Walde (2017) showed concreteness scores to be effective for Metaphor Detection. Embedding-based approaches such as in Köper and im Walde (2017) and Rei et al. (2017) also proved to work effectively on several annotated datasets. Different types of word embeddings were studied by researchers, including

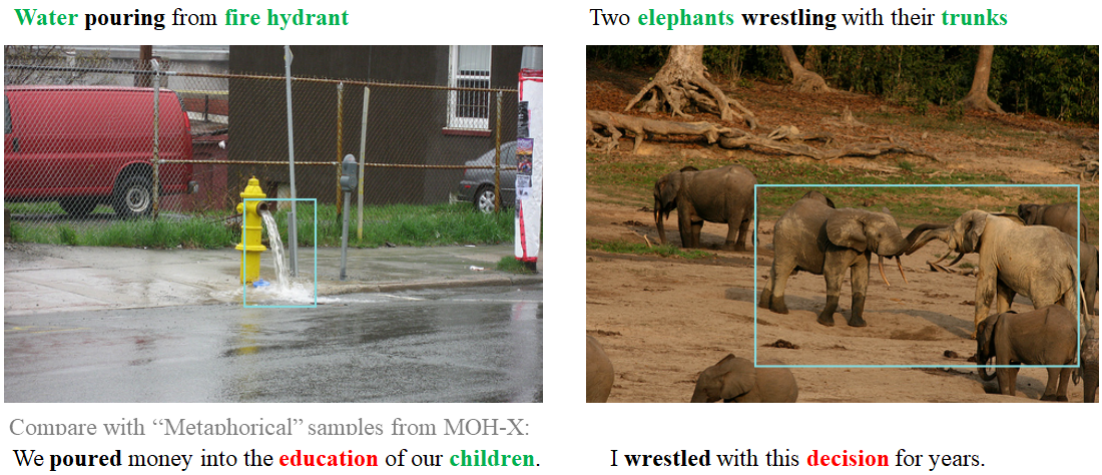

Figure 1: Example sentences with non-metaphorical (literal) and metaphorical usages of the verbs "pour" and "wrestle". The literal sentences (as well as the images) are taken from the Visual Genome dataset (Krishna et al., 2016), and are the captions of the regions highlighted in squares in the respective images. The metaphorical sentences are taken from the MOH-X datasets (Mohammad et al., 2016). Words with significantly higher concreteness scores are highlighted in green, and words that are considered abstract are highlighted in red.

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).

In our work, we study the effect of using embeddings created from visual datasets, which were shown to be useful in Metaphor Detection (Shutova et al., 2016), as well as in the task of estimating concreteness scores (Kehat and Pustejovsky, 2017).

2.2. Vision-Language datasets

The field of Vision and Language has become extremely popular in the last several years. New tasks involving both images and texts were introduced to both the Computer Vision and Natural Language Processing communities, such as Visual Question Answering (Antol et al., 2015) and visual entailment (Krishnamurthy, 2015).

This growing interest has led to an explosion of datasets combining visual and textual information, mostly in the form of an image (or segmented regions of an image) and its corresponding or associated textual caption. Many of the most popular vision-language datasets are based on extensive crowdsourcing. The most famous ones to date are the Visual Genome (Krishna et al., 2016) (See examples in Figure 1), Microsoft COCO (Lin et al., 2014), Imagenet (Deng et al., 2009), which is a visual version of WordNet (Miller, 1995), and Flickr30K (Young et al., 2014). Other vision-language datasets, like the SBU dataset (Ordonez et al., 2011) were created automatically by simply querying the web.

In our work we use what we call "visual corpora", which are the text-only parts of vision and language datasets. These texts tend to represent words and ideas of higher concreteness on average, helping us to solve concreteness-related tasks such as metaphor detection (Kehat and Pustejovsky, 2017).

2.3. Word Concreteness

The concreteness of a word commonly refers to what extent the word represents things that can be perceived directly through the five senses (Brysbaert et al., 2014; Turney et al., 2011), such as *water* and *blue*. Accordingly, an abstract word represents a concept that is far from immediate perception, or alternatively, could be explained only by other words (as opposed to being demonstrated through image, taste, etc.), like *decision* and *fun*.

The most common resources for concreteness ratings of English words are the list of 40K scores by Brysbaert et al. (2014), with assigned concreteness scores between 1.0-5.0, and the MRC psycholinguistic database (Coltheart, 1981) that contains over 4K words and their concreteness scores (range from 158 to 670), given by human subjects through psychological experiments.

3. Improving Metaphor Detection

As presented in previous work, certain lexical features, like concreteness scores, have been shown to improve metaphor detection models (Mykowiecka et al., 2018; Turney et al., 2011). Nevertheless, these models were based on hand-annotated resources, such as the MRC Psycholinguistic Corpus (Coltheart, 1981). One of the major disadvantages of using these lists is the fact that they contain a limited number of words and are usually available and evaluated for English only and are hard to reproduce for other languages, as noted by Mykowiecka et al. (2018).

In order to introduce information about the concreteness of word to the models without having to use an annotated dataset or a dictionary, we take a similar approach to Kehat and Pustejovsky (2017), and use vision-language datasets as a reference. Many of the available vision-language datasets were created by crawling image-sharing social networks like Flickr (Ordonez et al., 2011), which are already popular among users throughout the web.

In the following sections, we show our results on two commonly used annotated datasets for metaphor detection:

The dataset by Mohammad et al. (MOH) (Mohammad et al., 2016), was created as part of a bigger dataset that also contains annotations about the emotional level and emotional polarity of words. In this dataset, about 1,600 sentences were annotated in a binary fashion, as either "metaphorical" or "literal", in relation to a certain verb occurrence. The MOH dataset is commonly cut into a smaller dataset, called the MOH-X dataset, which contains only about 650 sentences, and is more balanced in terms of the number of labels for each class (the original MOH dataset contains many more "literal" annotations than "metaphorical" ones).

the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010) is the largest available metaphor dataset to date. In this dataset, every word (not just a target verb) is labeled through an exhaustive annotating scheme. We use the Verbs subset of the VUA metaphor dataset, as used in the 2018 shared task (See section 2.1). This subset concsists of more than 17K training samples and over 5K test samples, taken from the British National Corpus (BNC).

3.1. Visibility Embeddings

In their work, Kehat and Pustejovsky (2017) showed that visual corpora (text derived from vision-language datasets) tend to have higher "concreteness level", and used this fact to automatically estimate concreteness scores of words, by checking if the given word and its nearest neighbors (in a semantic vector space) are contained in the visual corpus. We aim to improve upon the suggested model by (Gao et al., 2018), which already use embeddings such as the GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018). These inherently carry information about the semantic vector space structure and neighbors. Therefore, our approach is even simpler, and checks only if the specific given word is in the visual corpus.

We base our sampling method on the shown relatively high differences in the "concretness level" of different visual and non-visual corpora. The concreteness level of a corpus is calculated as follows: given a concreteness score list (usually the 40K or MRC), we divide the words in the list into two non-overlapping sets (words contained in the corpus and words not contained in the corpus), and calculate the average concreteness score of each set, as well as the difference of the two averages normalized by the score range of the list ('Diff/Range%').

Table 1 contains the Diff/Range percentages of several visual and non-visual corpora and their subsets (as sets of words). Like Kehat and Pustejovsky (2017), we refer to the BVC as the Big Visual Corpus, a unified corpus consists of several common visual corpora, which showed to have the higher Diff/Range ratio. As a balanced non-visual corpus, we take the Brown corpus (Francis and Kucera, 1964), which showed to have the smaller, almost zero, Diff/Range ratio (means, it is balanced in terms of concreteness).

Corpus	D/R% 40K	D/R% MRC
BVC	25.49%	24.53%
$Brown$	2.74%	-0.28%
$Brown - BVC$	-17.30%	-24.44%
$Brown\&BVC$	14.84%	13.34%

Table 1: The Diff/Range% of the Big Visual Corpus (BVC), the Brown corpus, and their subsets. Higher Diff/Range ratio indicates the corpus is more concrete on average.

3.2. The Construction of the Visibility Embeddings

In this section, we show how to build word embeddings out of the visual and non-visual corpora discussed above. In the next section, we show how to plug these vectors in a BiLSTM model, improving existing results.

For each seen word in a sentence, we build a vector of length l, consisting of l values sampled from a normal distribution around mean m with variance v. We choose m such that it can have one of three values, -1.0, 0.0 or 1.0, where -1.0 aims to represent abstractness and 1.0 aims to represent concreteness.

In order to determine m, we use several of the corpora in Table 1 as reference. Based on the Diff/Range ratios, we determine m as follows:

```
For each word in a sentence:
  If the word is a stopword or punctuation:
    assign m = 0.0.
  Else, if the word is in Brown − BVC:
    assign m = -1.0.
  Else, if the word is in BVC:
    assign m = 1.0.
  Else:
    assign m = 0.0.
```

First we check if a word is in $Brown - BVC$ since this sub-corpus is small with a very low Diff/Range ratio. We continue checking if the word is in the BVC (we don't check for $BVC - Brown$ since, according to our calculations, it is less concrete on average than the BVC). If the word is in neither corpora or if it is a stopword, we choose m to be the neutral o.o.

Following Kehat and Pustejovsky (2017) and Gao et al. (2018), we do not normalize the tokens before building the visibility embeddings (or generally inputting them into the system). Our experiments show that without special handling of contextual ambiguity, too much information is lost, due to the derivative nature of the English language. For example, for the lemma "woman", we can construct both "women" and "womanize", which are highly different in terms of concreteness scores.

3.3. Experiment Setting and Results

We further build on the model proposed by Gao et al. (2018) by adding our own Visibility Embeddings to the set of embeddings mapped to each word in a given sentence. Originally, Gao et al. (2018) concatenated three types of vectors: embeddings created with ELMo (of dimension

1024), GloVe embeddings (Pennington et al., 2014) (of dimension 300), and binary verb embeddings (of dimension 50) which indicated the verb index in the sentence. We kept the same structure and dimensions of the vectors and also added the new Visibility Embeddings of dimension 50 (See Figure 2).

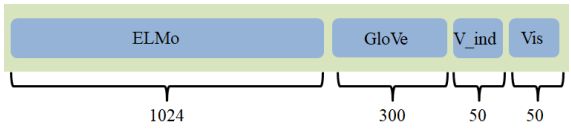

Figure 2: The embeddings used in the model consist of the ELMo output, GloVe, Verb Index binary embeddings, and Trinary Visibility Embeddings.

The model consists of three main layers (See Figure 3): (1) A Bidirectional-LSTM layer; (2) An attention layer, in which we apply linear softmax on the result and then calculate the similarity of the created vector and the matrix created from the Bi-LSTM output; (3) A classification layer, a feed-forward layer with softmax log to get the classification label of each sentence.

We implemented the model in Python using the AllenNLP package for deep semantic NLP (Gardner et al., 2017). The input for each learning iteration of the model is a batch of embedded sentences. We also apply three dropout factors: before the Bi-LSTM layer, inside the Bi-LSTM layer, and before the classifier layer. To accommodate the new embeddings, we also changed a few constants, such as learning rates, dropout, and number of epochs, but kept the structure of the model and all the other parameters as in Gao et al. (2018).

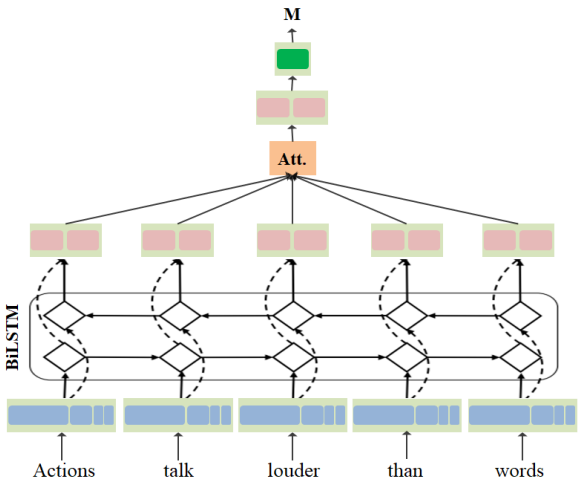

Figure 3: The classification model architecture.

The results of our model, compared with other recent top results, are shown in Tables 2 and 3. We compare our work with the best results gained so far, and with the recent work by Gao et al. (2018), to show more clearly the effect of adding our Visibility Embeddings to their system. We fine-tuned the hyperparameters of the models for each of the discussed metaphor detection datasets. We can notice that by just adding our simply constructed visibility vectors to the already existed model by Gao et al. (2018), we can achieve significant improvement over their previous results on both the MOH-X and VUA datasets.

Model	P	R	F1
Lexical Baseline	39.1	26.7	31.3
Mao et al. (2019)	77.5	83.1	80.0
Gao et al. (2018)	75.3	84.3	79.1
Gao et al. (2018)+Vis	79.5	81.84	80.46
Gao et al. (2018)+Vis (rand)	80	80.62	80.02
Elmo+verb+Vis	79.35	84.6	81.57
Elmo+verb+Vis (rand)	81.16	81.03	80.85

Table 2: Results on the MOH-X dataset. Our model improves upon the previous state of the art by Mao et al. (2019).

Model	P	R	F1
Lexical Baseline	67.9	40.7	50.9
Mao et al. (2019)	69.3	72.3	70.8
Wu et al. (2018)	60	76.3	67.2
Gao et al. (2018)	53.4	65.6	58.9
Gao et al. (2018)+Vis	70.11	64.33	67.1
40K scores	71.65	60.87	65.82

Table 3: Comparison of recent algorithms on the VUA verb classification task. Our model, which is a variation on the one by Gao et al. (2018), gets very close to the state of the art achieved by Wu et al. (2018)

For the MOH-X dataset shown in Table 2, we can see that by simply adding our visibility vectors, we can gain +1.36 to the F1-score. We experimented also with variations of the models that do not include the GloVe embeddings (i.e., of dimension 1024+50+50), and found the system to perform better in this settings for the MOH-X dataset (though not for the VUA dataset). These results are shown in the last rows of Table2.

We note the difficulty in the evaluations of the results reported by Gao et al. (2018). Though not mentioned in their paper, the code that was made available online suggested that the 10-fold cross-validation was performed without shuffling. Also, the reported maximal score was computed by sampling within a given number of iterations (rather than in the end of every epoch). When running their code, we discovered a steady difference between running on the same pre-chosen sets over unshuffled samples (like they apparently did), and randomly choosing the validation set (as traditionally done by researchers), with the right sampling in the end of each epoch. Therefore, to maintain consistency with future results, we also bring our models' performances when tested on randomly chosen 10-fold cross-validation sets, which are, in fact, the ones we should report.

In general, we can observe that the higher results are on the MOH-X dataset. this is due to the fact that for this dataset, only the metaphoricity of the target verb is known, and the sentences are relatively short. Other methods, such as labeling each token of a sentence, give better results on datasets like the VUA.

Specifically for the VUA dataset, we also experiment with

actual concreteness scores annotated by humans, from the list of 40K concreteness ratings by Brysbaert et al. (2014). For each word, we build a similar normalized vector using the concreteness score from the list as the mean m. To set up the variance, we tried to use both the inter-annotators standard deviation as appears in the list, and a constant standard deviation (as in the Visibility Vectors case), and found the last one to give better results. All the means and variances were normalized to have the same range as the visibility embeddings, and the results are shown in the last row of Table 3.

We found that using the concreteness scores directly showed less improvement than using the Visibility Embeddings. The overall F1-score is lower because of a lower recall, yet the precision is higher. We hypothesize that the high variance of the concrete and non-concrete terms in our construction of the Visibility Embeddings is more significant than the finer differences naturally occurring in the human annotation, hence their effect as part of the vectorized input is more noticeable.

4. Summary

In this paper, we have presented a simple and direct way to use visual corpora as a reference to certain visibility properties of words. We showed that by adding Visibility Embeddings, built in the same way, to existing deep learning models for metaphor detection, we can compare with or improve upon most classification scores for the task of verb classification. Furthermore, our approach is much simpler than previous models, and is not limited to English.

Acknowledgements

We would like to thank the reviewers for their helpful comments. 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.

5. Bibliographical References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*.

Bizzoni, Y. and Ghanimifard, M. (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, June. Association for Computational Linguistics.

Black, M. (1979). More about metaphor.[in] a. ortony (ed.), metaphor and thought.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

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

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. (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.

Francis, W. N. and Kucera, H. (1964). Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1.

Gao, G., Choi, E., Choi, Y., and Zettlemoyer, L. (2018). Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform.

Kehat, G. and Pustejovsky, J. (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)*, volume 2, pages 103–108.

Köper, M. and im Walde, S. S. (2017). Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. *SENSE 2017*, page 24.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Li, F. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Krishnamurthy, J. (2015). Visually-verifiable textual entailment: A challenge task for combining language and vision. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 1–3, Lisbon, Portugal, September. Association for Computational Linguistics.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago press.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (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.

Mao, R., Lin, C., and Guerin, F. (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, July. Association for Computational Linguistics.

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

Mohammad, S., Shutova, E., and Turney, P. (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.

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

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 1143–1151.

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

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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)*, volume 1, pages 2227–2237.

Pramanick, M., Gupta, A., and Mitra, P. (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, June. Association for Computational Linguistics.

Rei, M., Bulat, L., Kiela, D., and Shutova, E. (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, September. Association for Computational Linguistics.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 160–170.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., and Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

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

Swarnkar, K. and Singh, A. K. (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, June. Association for Computational Linguistics.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (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.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (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.

Wu, C., Wu, F., Chen, Y., Wu, S., Yuan, Z., and Huang, Y. (2018). Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana, June. Association for Computational Linguistics.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.