Neural Metaphor Detection in Context

Ge Gao,[◇] Eunsol Choi,[◇] Yejin Choi,[◇†] and Luke Zettlemoyer[◇]
University of Washington[◇]
Allen Institute for Artificial Intelligence[†]
{ggao,eunsol,yejin,lsz}@cs.washington.edu

Abstract

We present end-to-end neural models for detecting metaphorical word use in context. We show that relatively standard BiLSTM models which operate on complete sentences work well in this setting, in comparison to previous work that used more restricted forms of linguistic context. These models establish a new state-of-the-art on existing verb metaphor detection benchmarks, and show strong performance on jointly predicting the metaphoricity of all words in a running text.

1 Introduction

Metaphors are pervasive in natural language, and detecting them requires challenging contextual reasoning about whether specific situations can actually happen. (Lakoff and Johnson, 1980). For example, in Table 1, "examining" is metaphorical because it is impossible to literally use a "microscope" to examine an entire country. In this paper, we present end-to-end neural models for metaphor detection, which can learn rich contextual word representations that are crucial for accurate interpretation of figurative language.

In contrast, most previous approaches focused on limited forms of linguistic context, for example by only providing SVO triples such as (car, **drink**, gasoline) to the model (Shutova et al., 2016; Tsvetkov et al., 2013; Rei et al., 2017; Bulat et al., 2017). While the verbal arguments provide strong cues, providing the full sentential context supports more accurate prediction, as seen in Table 1. Even in the few cases when the full sentence is used (Köper and im Walde, 2017; Turney et al., 2011; Jang et al., 2016) existing models have used unigram-based features with limited expressivity.

We investigate two common task formulations: (1) given a target verb in a sentence, classifying whether it is metaphorical or not, and (2) given a sentence, detecting all of the metaphorical words (independent of their POS tags). We find that relatively standard architectures based on bi-directional LSTMs (Hochreiter and Schmidhuber, 1997) augmented with contextualized word embeddings (Peters et al., 2018) perform surprisingly well on both tasks, even with modest amount of training data. We improve the previous state-of-the-art by 7.5 F1 on the VU Amsterdam Metaphor Corpus (VUA) for the sequence labeling task (Steen et al., 2010), by 2.5 F1 on the VUA verb classification dataset, and by 4.9 F1 on the MOH-X dataset (Mohammad et al., 2016). Our code is publicly available at `https://github.com/gao-g/metaphor-in-context`.

The experts started **examining** the Soviet Union with a microscope to study perceived changes.
Rockford teachers are honored for saving a *drowning* student.
You're **drowning** in student loan debt.

Table 1: Metaphorical usages of the target word are bold faced, and literal usages are italicized. Full sentence context is crucial for metaphor detection.

2 Task

We study two task formulations.

Sequence Labeling: Given a sentence $x_1, \ldots, x_n$, predict a sequence of binary labels $l_1, \ldots, l_n$ to indicate the metaphoricity of each word.

Classification: Given a sentence $x_1, \ldots, x_n$ and a target verb index i, predict a binary label l to indicate the metaphoricity of the target x_i.

While both formulations have been studied in previous work, it is worth noting that the sequence labeling task generalizes the classification task in

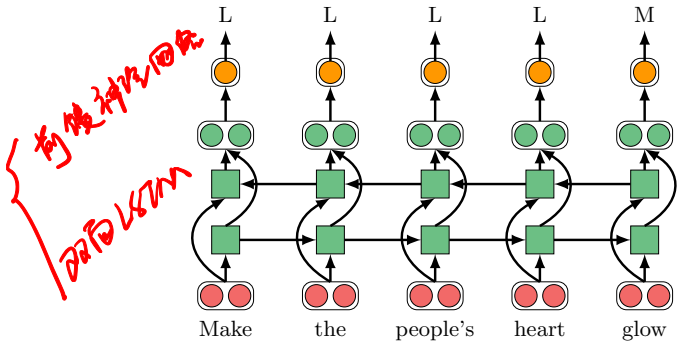

Figure 1: A sequence labeling model for metaphor detection. Every word in a sentence is classified.

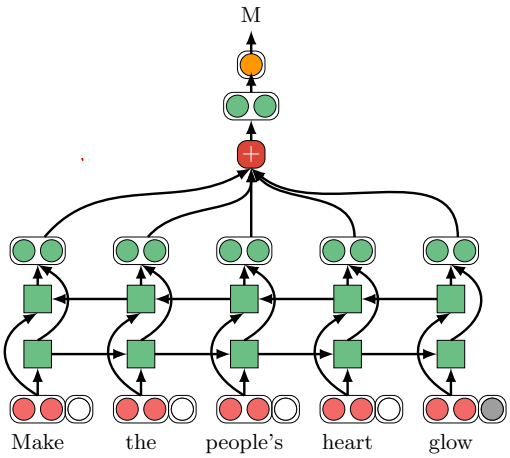

Figure 2: A classification model for metaphor detection. Only a single word per sentence is labeled as metaphorical or literal.

that the prediction for the target verb can be extracted from the full sentence predictions. In addition, as will be shown in Section 5, we find that given accurate annotations for all words in a sentence, the sequence labeling model outperforms the classification model even when the evaluation is set up as a classification task.

3 Model

Our models use a bidirectional LSTM to encode a sentence, and a feedforward neural network for classification, optimized for the log-likelihood of gold labels.

Sentence encoding For both sequence labeling and classification, we represent each token x_i in the input sentence with a pre-trained word embedding w_i. To further encode contextual information, we also concatenate ELMo (Embeddings from Language Models) vectors e_i from Peters et al. (2018). These vectors have been shown to be useful for word sense disambiguation, a task closely related to metaphor detection (Birke and Sarkar, 2006).

3.1 Sequence Labeling Model

Figure 1 shows the model architecture. We input the word representation $[w_i; e_i]$ to a bidirectional LSTM, producing a contextualized representation h_i for each token. Then we use a feedforward neural network that takes h_i to predict a label l_i for each word x_i.

When the dataset does not contain annotations for every word, we make the simplifying assumption that every unannotated word is used literally.

3.2 Classification Model

Figure 2 shows the model architecture. We concatenate an index embedding n_i, which indicates whether x_i is the target verb. We use $[w_i; e_i; n_i]$ as an input to a bidirectional LSTM, producing a contextualized representation h_i.

We add an attention layer by computing the attention weight a_i for token x_i, and compute the representation c as a weighted sum of LSTM output states where W_a and b_a are learned parameters.

$$a_i = \text{SoftMax}_i(W_a h_i + b_a)$$

$$c = \sum_{i=1}^{n} a_i h_i$$

Finally, we feed c to a feedforward network to compute the label scores for target verb.

4 Dataset

We evaluate performance on a number of benchmark datasets, including two for classification (TroFi and MOH\MOH-X) and one for tagging (VUA).[1] Table 2 shows statistics for the verb classification datasets. Despite being two times larger than the MOH dataset, the TroFi dataset contains only 50 unique verbs, and the larger VUA dataset contains over 2K unique verbs. The MOH dataset contains shorter and simpler sentences (example sentences in WordNet), compared to sentences in other datasets which come from resources such as

[1]For detailed information about each dataset, please refer to original papers: TroFi (Birke and Sarkar, 2006), MOH (Mohammad et al., 2016), VUA (Steen et al., 2010). MOH-X refers to a subset of MOH dataset used in previous work (Shutova et al., 2016) where verb and its argument are extracted from each sentence.

	# Expl.	% Metaphor	# Uniq. Verb	Avg # Sent. Len
MOH-X	647	49%	214	8.0
MOH	1,639	25%	440	7.4
TroFi	3,737	43%	50	28.3
VUA	23,113	28%	2047	24.5

Table 2: Verb classification dataset statistics. % Metaphor refers to sentence-level percentage.

	Train	Dev	Test
# Unique tokens	13,843	7,458	7,200
# Tokens	116,622	38,628	50,175
# Unique sent.	6,323	1,550	2,694
% Metaphor	11.2	11.6	12.4

Table 3: VUA sequence labeling dataset statistics. % Metaphor refers to token-level percentage.

Model	P	R	F1	Acc.
Lexical Baseline	68.6	45.2	54.5	90.6
Wu (2018) ensemble	60.8	70.0	65.1	-
Ours (SEQ)	**71.6**	**73.6**	**72.6**	**93.1**

Table 4: Performance on the VUA sequence labeling test set for all POS tags.

POS	#	% metaphor	P	R	F1.
VERB	**20K**	18.1	68.1	71.9	69.9
NOUN	20K	13.6	59.9	60.8	60.4
ADP	13K	**28.0**	**86.8**	**89.0**	**87.9**
ADJ	9K	11.5	56.1	60.6	58.3
PART	3K	10.1	57.1	59.1	58.1

Table 5: The breakdown of performance on the VUA sequence labeling test set by POS tags. We show data statistics (count, % metaphor) on the training set. We only show POS tags whose % metaphor > 10.

news articles. The TroFi and MOH-X datasets are constructed to have higher percentages of metaphor, compared to the natural likelihood of metaphor in a running text, as seen in the VUA dataset.

Classification Experiment Setup We perform 10 fold cross-validation on the MOH-X and TroFi datasets, following prior work. For the VUA dataset, we use the original training and test split (Klebanov et al., 2016), and set aside 10% of the training set as a development set.

Sequence Labeling Experiment Setup The VUA dataset contains annotations for all words in each sentence. We divide the data into training, development, and test set following the same split for the VUA verb classification task. While the label classes are less balanced (only 11% metaphors at the token level), this dataset is much bigger. Table 3 shows the data statistics.

5 Experiments

Evaluation Metric We report precision, recall and F1 measure for the metaphor class as well as the overall accuracy. For the VUA dataset, we also report macro-averaged F1 score across four genres (conversation, academic writing, fiction and news).

Comparison Systems We propose a simple yet effective lexical baseline. It assigns the metaphor label if the word is annotated metaphorically more frequently than as literally in the training set, and the literal label otherwise. We also compare our models to previously published work, including: (1) a logistic regression classifier with features that indicate verb lemmas and the verbs' semantic class from WordNet (Klebanov et al., 2016), (2) a neural similarity network with skip-gram word embeddings (Rei et al., 2017), (3) a balanced logistic regression classifier on target verb lemma that uses a set of features based on multi-sense abstractness rating (Köper and im Walde, 2017), and (4) a CNN-LSTM ensemble model with weighted-softmax classifier which incorporates pre-trained word2vec, POS tags, and word cluster features (Wu et al., 2018).[2]

We experiment with both sequence labeling model (SEQ) and classification model (CLS) for the verb classification task, and the sequence labeling model (SEQ) for the sequence labeling task.

Implementation Details We used 300d GloVe vectors (Pennington et al., 2014) and 1024d ELMo vectors. We used additional 50d index embedding for the classification task. The LSTM module has a 300d hidden state. We applied dropout on the input to LSTM and on the input to the feedforward layer. We fine-tuned learning rate and dropout rate for each model on each dataset. We used SGD to optimize the CLS model and Adam (Kingma and Ba, 2013) for the SEQ model. We used spaCy (Honnibal and Montani, 2017) for lemmatization, tokenization, and part-of-speech tagging.

[2]The best performing model on the VUA Metaphor Detection Shared Task at the NAACL 2018 workshop on Figurative Language Processing.

Model	MOH-X (10 fold)				TroFi (10 fold)				VUA - Test				
	P	R	F1	Acc.	P	R	F1	Acc.	P	R	F1	Acc.	MaF1
Lexical Baseline	39.1	26.7	31.3	43.6	**72.4**	55.7	62.9	71.4	67.9	40.7	50.9	76.4	48.9
Klebanov (2016)	-	-	-	-	-	-	-	-	-	-	-	-	60.0
Rei (2017)	73.6	76.1	74.2	74.8	-	-	-	-	-	-	-	-	-
Köper (2017)	-	-	-	-	-		**75.0**	-	-	-	62.0	-	-
Wu (2018) ensemble	-	-	-	-	-	-	-	-	60.0	**76.3**	67.2	-	-
CLS	75.3	84.3	**79.1**	**78.5**	68.7	**74.6**	72.0	73.7	53.4	65.6	58.9	69.1	53.4
SEQ	**79.1**	73.5	75.6	77.2	70.7	71.6	71.1	**74.6**	**68.2**	71.3	**69.7**	**81.4**	**66.4**

Table 6: Model performances for the verb classification task. Our models achieve strong performance on all datasets. The CLS model performs better than the SEQ model when only one word per sentence is annotated by human (TroFi and MOH-X). When all words in the sentence are accurately annotated (VUA), the SEQ model outperforms the CLS model.

Model	P	R	F1.	Acc.
SEQ	**68.3**	**72.0**	**70.4**	**83.5**
-ELMo	59.4	64.3	61.7	78.2
CLS	52.4	63.0	57.3	74.3
-ELMo	52.0	48.7	50.8	74.1

Table 7: Ablation study on VUA development set for the verb classification task.

Sequence Labeling Results Performance on the sequence labeling task is reported in Table 4. While prior work (Klebanov et al., 2014; Özbal et al., 2016) reported on the same dataset, the experiment setting is not comparable (they did cross validation on a smaller training set).[3] Our lexical baseline performs strongly in terms of precision, as some words and POS tags are almost exclusively annotated as literal. Our sequence labeling model mainly improves recall.

Table 5 reports the breakdown of performance by POS tags. Not surprisingly, tags with more data are easier to classify. Adposition is the easiest to identify as metaphorical and is also the most frequently metaphorical class (28%). On the other hand, particles are challenging to identify, since they are often associated with multi-word expressions, such as "put **down** the disturbances".

Verb Classification Results Table 6 shows performance on the verb classification task for three datasets (MOH-X , TroFi and VUA).[4]

Our models achieve strong performance on all datasets, outperforming existing models on the MOH-X and VUA datasets. On the MOH-X dataset, the CLS model outperforms the SEQ model, likely due to the simpler overall sentence structure and the fact that the target verbs are the only words annotated for metaphoricity. For the VUA dataset, where we have annotations for all words in a sentence, the SEQ model significantly outperforms the CLS model. This result shows that predicting metaphor labels of context words helps to predict the target verb. We hypothesize that Köper et al. (2017) outperforms our models on the TroFi dataset for a similar reason: their work uses concreteness labels, which highly correlate to metaphor labels of neighboring words in the sentence. Also, their best model uses the verb lemma as a feature, which itself provides a strong clue in the dataset of 50 verbs (see lexical baseline).

Table 7 shows an ablation study on input representations (with or without ELMo vectors). Contextualized word vectors improve the performance of both models by a large margin.

Error Analysis We sampled 100 errors of our best model from the VUA verb classification development set for analysis. Table 8 shows examples. Following the original annotation guideline,[5] we classify metaphors into five categories: direct metaphor, indirect metaphor, implicit metaphor, personification, and borderline case. Indirect metaphor, the most common type for verbs, means that the basic meaning of a word is different from its contextual meaning. Implicit metaphor occurs due to an underlying link which points to a recoverable metaphorical concept.

About half of the errors were false positives, and the other half were false negatives. Among the false negatives, 33% are indirect metaphors, 18% are personifications, and 2% are direct metaphors. Among 55 false positives, 31% of verbs have im-

[3] As a point of reference, their macro-averaged F1 scores were 33.25 / 50.6 respectively.

[4] We did not compare to Shutova et al. (2016) as their experiment setting is not comparable.

[5] http://www.vismet.org/metcor/documentation/home.html

CLS	SEQ	Sentence	Metaphor Type
✗	✗	To *throw* up an impenetrable Berlin Wall between you and them could be tactless.	-
✗	✗	In reality you just invent a tale, as if you were **sitting** round a fire in a cave.	direct metaphor
✗	✗	So they **bought** immunity.	indirect metaphor
✗	✗	During the early states of the phased evacuation the logistical problem **facing** the police was the street-by-street warning of the population to make ready for evacuation.	indirect metaphor
✗	✔	There are few things worse than being **bludgeoned** into reading a book you hate.	indirect metaphor
✗	✔	He thought of thick, fat, hot motorways **carving** up that land.	personification
✗	✔	One might *ask* whether motorists are ever justified in knowingly taking risks with other people's lives.	-
✗	✔	The abstract talk of *commuting* by rail or road being replaced by information technology finds a concrete expression in the idea of telecottages.	-
✗	✔	A fly landed on the empty, staring vizor, and *crawled* across it.	-

Table 8: Some examples from the VUA verb classification development set. Metaphorical usages of the target word are bold faced, and literal usages are italicized. Leftmost columns show the correctness of prediction.

plicit arguments that are not explicitly mentioned in the context, 15% have long range dependencies (at least five words away) from core arguments, 10% have arguments with rare word senses, and 5% have anthropomorphic arguments. Finally, we found about half of false negatives and 20% of false positives to be borderline cases, showing the subjective nature of the task.

We sampled 257 dev examples that the CLS model gets wrong but the SEQ model gets correct. We found that the SEQ model outperforms the CLS model on detecting personifications, indirect metaphors, and direct metaphors involving uncommon verbs.

6 Related Work

There has been significant work on studying different features for metaphor detection, including concretenesss and abstractness (Turney et al., 2011; Tsvetkov et al., 2014; Köper and im Walde, 2017), imaginability (Broadwell et al., 2013; Strzalkowski et al., 2013), feature norms (Bulat et al., 2017), sensory features (Tekiroglu et al., 2015; Shutova et al., 2016), bag-of-words features (Köper and im Walde, 2016), and semantic class using WordNet (Hovy et al., 2013; Tsvetkov et al., 2014). More recently, embedding-based approaches (Köper and im Walde, 2017; Rei et al., 2017) showed gains on various benchmarks.

Many neural models with various features and architectures were introduced in the 2018 VUA Metaphor Detection Shared Task. They include LSTM-based models and CRFs augmented by linguistic features, such as WordNet, POS tags, concreteness score, unigrams, lemmas, verb clusters, and sentence-length manipulation (Swarnkar and Singh, 2018; Pramanick et al., 2018; Mosolova et al., 2018; Bizzoni and Ghanimifard, 2018; Wu et al., 2018). Researchers also studied different word embeddings, such as embeddings trained from corpora representing different levels of language mastery (Stemle and Onysko, 2018) and binarized vectors that reflect the General Inquirer dictionary category of a word (Mykowiecka et al., 2018). We show that contextualized word embedding significantly improves metaphor detection. We also study both sequence labeling and classification approaches, suggesting that sequence labeling approach enhances performance when used to jointly predict the metaphoricity of all words in a sentence.

7 Conclusion

In this paper, we present simple BiLSTM models augmented with contextualized word representation for metaphor detection. Our models establish new state-of-the-arts across multiple existing benchmarks, and our error analysis shows remaining challenges for metaphor detection.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was supported in part by the NSF (IIS-1714566 and IIS-1252835), the ARO (W911NF-16-1-0121), the DARPA CwC program through ARO (W911NF-15-1-0543), and gifts from Google and Facebook.

References

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

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, NAACL*, pages 91–101.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah M. Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *SBP*.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *EACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard H. Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP, ACL*, pages 52–57.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Penstein Rosé. 2016. Metaphor detection with topic transition, emotion and cognition in context. In *ACL*.

Diederik P Kingma and Jimmy Ba. 2013. Adam: A method for stochastic optimization.

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

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP, ACL*.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of german particle verbs. In *HLT-NAACL*.

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 First Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. The University of Chicago Press.

Saif Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*.

Anna Mosolova, Ivan Bondarenko, and Vadim Fomin. 2018. Conditional random fields for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing, NAACL*, pages 121–123.

Agnieszka Mykowiecka, Aleksander Wawer, and Magorzata Aneta Marciniak. 2018. Detecting figurative word occurrences using recurrent neural networks. In *Proceedings of the Workshop on Figurative Language Processing, NAACL*, pages 124–127.

Gözde Özbal, Carlo Strapparava, Serra Sinem Tekiroglu, and Daniele Pighin. 2016. Learning to identify metaphors from a corpus of proverbs. In *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matthew Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. abs/1802.05365.

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, NAACL*, pages 67–75.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *EMNLP*.

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

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. 21(4):765796.

Egon W. Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing, NAACL*, pages 133–138.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphor from novel data. In *Proceedings of the First Workshop on Metaphor in NLP, ACL*.

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, NAACL*, pages 115–120.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP, ACL*.

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

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP, ACL*.

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

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, NAACL*, pages 110–114.