Testing the role of metadata in metaphor identification

Egon W. Stemle
Eurac Research / Bolzano-Bozen (IT)
Masaryk University / Brno (CZ)
egon.stemle@eurac.edu

Alexander Onysko
University of Klagenfurt / Klagenfurt (AT)
alexander.onysko@aau.at

Abstract

This paper describes the adaptation and application of a neural network system for the automatic detection of metaphors. The LSTM BiRNN system participated in the shared task of metaphor identification that was part of the Second Workshop of Figurative Language Processing (FigLang2020) held at the Annual Conference of the Association for Computational Linguistics (ACL2020). The particular focus of our approach is on the potential influence that the metadata given in the ETS Corpus of Non-Native Written English might have on the automatic detection of metaphors in this dataset. The article first discusses the annotated ETS learner data, highlighting some of its peculiarities and inherent biases of metaphor use. A series of evaluations follow in order to test whether specific metadata influence the system performance in the task of automatic metaphor identification. The system is available under the APLv2 open-source license.

1 Introduction

Research on metaphors, particularly in the framework of conceptual metaphor theory, continues to grow in all genres of language use and across diverse disciplines of linguistics (cf., among others, Littlemore, 2019; Gibbs Jr, 2017; Charteris-Black, 2016; Kövecses, 2020; Callies and Degani for recent and forthcoming overviews and extensions, and Veale et al., 2016 for a book-length discussion of computational linguistic perspectives). While the importance of metaphor in thought and everyday language use has long been acknowledged (Lakoff and Johnson, 1980), the practice of metaphor research still faces two methodological and analytical challenges: first of all, the identification of metaphors and, secondly, their actual description through source and target domains.

In computational linguistics, a great amount of recent work has been concerned with addressing the challenge of identifying metaphors in texts. This is evident in the series of four Workshops on Metaphor in NLP from 2013 to 2016 and in the two Workshops on Figurative Language Processing in 2018 (Beigman Klebanov et al., 2018b) and 2020 (Leong et al., 2020), each of which involved a shared task (ST) in automatic metaphor detection. Identification systems that achieved the best results in the first shared task relied on neural networks incorporating long-term short-term memory (LSTM) architectures (see Mu et al., 2019 for a discussion). Further advances in the field using deep learning approaches have been reported in Dankers et al. (2019), Gao et al. (2018), and Rei et al. (2017).

This paper extends a system proposed in Stemle and Onysko (2018), which combines word embeddings (WEs) of corpora like the BNC (British National Corpus Consortium, 2007) and the TOEFL11 language learner corpus (see Blanchard et al., 2013). With the modified system, we participated in *The Second Shared Task on Metaphor Detection*. The difference to the 2018 edition of the ST is a new set of data. As in the first task, one part of the dataset is based on the VU Amsterdam (VUA) Metaphor Corpus manually annotated according to the MIPVU procedure (Steen et al., 2010). Additionally, the second task includes a sample of 240 argumentative learner texts. These texts are taken from the ETS Corpus of Non-Native Written English (synonymous to the TOEFL11 corpus) and have been manually annotated (Beigman Klebanov et al., 2018a).

Since the learner essays are a very specific kind of data, the aim of this study is to build upon observations from Stemle and Onysko (2018), who found that a combination of word embeddings from the BNC and the TOEFL11 learner corpus yielded the best results of metaphor identification in a bi-

Proceedings of the Second Workshop on Figurative Language Processing, pages 256–263
July 9, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

directional recursive neural network (BiRNN) with LSTM. These results triggered the hypothesis that learner language can lead to an information gain for neural network based metaphor identification. To explore this hypothesis further, the current study puts an explicit focus on the metadata provided for the ETS Corpus of Non-Native Written English and specifically tests the potential influence of proficiency ratings, essay prompt, and the first language (L1) of the author. In addition, we also test whether a combined training on the diverse datasets and the sequence of this training will have an impact on our system of neural network based metaphor identification.

To address these aims, our paper is structured as follows: Section 2 provides observations on the annotated learner corpus dataset. Section 3 describes the system of metaphor identification. This is followed in Section 4 by the results of the experiments, which are briefly discussed in light of the observations on the annotated learner corpus data.

2 Observations on the data

The VUA Metaphor Corpus and its application in the first shared task has been concisely described in Leong et al. (2018). The authors have reported the relatively high inter-annotator agreement ($\kappa > 0.8$), which is in part due to the MIPVU protocol (Steen et al., 2010) and the close training of annotators in the Amsterdam Metaphor Group. Interestingly, the results of the first task across all submitted systems showed a clear genre bias with academic texts consistently displaying the highest correct identification rates and conversation data (i.e. spoken texts) the lowest Leong et al. (2018, p.60). This might be related to the fact that academic discourse is more schematic and formulaic (e.g. in the use of sentential adverbials and verbal constructions) and might rely to a greater extent on recurrent metaphorical expressions than spoken conversations, which are less schematic and can thus display a higher degree of syntactic and lexical variation. In other words, similarities in the data between training and test sets might be higher in the academic than in the conversation genre, leading to different genre-specific training effects in neural networks.

Apart from the VUA metaphor corpus, the second shared task introduces a novel dataset culled from the ETS Corpus of Non-Native Written English. In their description, Beigman Klebanov et al. (2018a) report an average inter-annotator agreement of $\kappa = 0.62$ on marking argumentation-relevant metaphors. Disagreement in the annotations was positively compensated in that all metaphor annotations were included even if only given by one of the two raters. While a focus on argumentation-relevant metaphors coheres with the genre of short argumentative learner essays written in response to one of eight prompts during TOEFL examinations in 2006-2007 (Blanchard et al., 2013), the scope of metaphor annotation is more restricted in the ETS sample than in the VUA corpus, which follows the more stringent MIPVU protocol. This explains to some extent why the overall amount of metaphor-related words in the training sets is considerably lower in the ETS sample (an average of 7% in All-POS and 14% among verbs; see Beigman Klebanov et al., 2018a, p.88) than in the VUA Metaphor Corpus (15% in All-POS and 28.3% among verbs; see Leong et al., 2018, p.58).

The relatively small size of the ETS sample inspired us to look into the structure of the data more closely to check whether any potential biases exist that might play a role for the automatic detection of metaphors. Beigman Klebanov et al. (2018a, p.89) report a significant positive correlation of the number of metaphors and the proficiency ratings of the texts as medium or high in the data. This relationship is confirmed by an independent samples t-test in the training partition of the data (180 texts). The group of highly proficient learners ($N = 95$) uses more metaphors ($M = 13.98, SD = 8.23$) compared to the group of medium proficient learners ($N = 85, M = 9.55, SD = 5.96$) at a significantly higher rate ($t = 4.07, p = 0.000071$). The L1 background of the learners (i.e. Arabic, Italian, Japanese) did not influence the mean number of metaphors in the texts as confirmed by a one-way ANOVA ($F = 1.619, p = 0.201$; L1 Arabic: $N = 63, M = 12.48, SD = 8.37$; L1 Italian: $N = 59, M = 12.71, SD = 7.79$; L1 Japanese: $N = 59, M = 10.44, SD = 6.38$).

Since the Corpus of Non-Native Written English consists of argumentative learner essays that were written in response to one of eight different prompts, another factor to consider in the annotated ETS sample is the role the prompt might have on metaphor use. Table 1 summarizes the descriptive statistics on the number of metaphors per prompt.

From left to right, the columns in Table 1 re-

Prompt	# words	# metaph. types	# metaph. tokens	Metaph. type/toks	% metaph. per words	mean # of metaph.
P1	8059	205	361	0.57	4.5	15.696
P2	7493	199	330	0.60	4.4	15.714
P3	7947	222	397	0.59	4.8	17.227
P4	8076	146	173	0.84	2.1	7.522
P5	8455	172	206	0.83	2.4	8.957
P6	8446	134	188	0.71	2.2	7.833
P7	7516	170	243	0.70	3.2	11.045
P8	7923	197	260	0.76	3.3	11.818

Table 1: Number of metaphors (types and tokens) per prompt in the annotated ETS training set.

port, per prompt, the total number of words, the number of metaphor types, the overall number of metaphor tokens (i.e. all words annotated for metaphor), the type token ratio of metaphors, the relative amount of metaphor tokens among all words, and the mean values of metaphors. The two rightmost columns illustrate an uneven occurrence of metaphors across the diverse prompts. Three groups emerge from the data according to their similarly high (or low) values: P1, P2, P3 as the highest scoring group, P4, P5, P6 as the lowest scoring group, and P7, P8 whose values are in-between the other two groups. A one-way ANOVA for independent samples ($F = 7.0919, p = .00001$) confirms a significant difference between the groups. T-tests comparing the minimal and maximal numerical distances between the high, the medium, and the low clusters show that the high cluster is significantly different from the low cluster (P1: $N = 23, M = 15.70, SD = 8.138$ compared to P5: $N = 23, M = 8.96, SD = 4.117$) at $t(44) = 3.54, p = .000948$. The differences between the low and the medium groups as well as the high and the medium group do not reach a significance threshold of $p < .01$.

When looking for an explanation of these biased metaphor occurrences, some interesting patterns emerge among the high frequency group (P1, P2, and P3). In all these instances, the prompts trigger certain metaphor-related words (MRW) that occur at a high rate. Table 2 provides an overview of the metaphorical expressions triggered by the prompts P1, P2, and P3. The 30 most frequent MRW were closely analyzed for each of the prompts.

In Table 2, MRW that cannot be related to the prompt are preceded by an asterisk. All the other terms are triggered by the prompts. For P1, the expression "broad knowledge" from the prompt that instantiates an objectification metaphor of knowledge (KNOWLEDGE IS AN OBJECT) is frequently reiterated in the test takers' essays and is by far the most frequent metaphorical expression among all annotated MRW in P1. The metaphorical uses of the lexeme *focus* as in "focus on a particular subject/field" is triggered by the prompt as a synonymous phrase for "...specialize in one specific subject". Similarly, the term *wide/-er* is used by some learners as a synonym of "broad knowledge". In P2, the metaphorical phrase "have time" is prevalent. It is thematically triggered by the phrase "enjoy life", which stimulates people to write about time as a (precious) possession that allows you to enjoy life. The metaphorical expression of "spending time" is evoked by the same conceptual metaphor. The LIFE IS A JOURNEY metaphor triggered by P2 is instantiated in the recurrent expression of stages in life. The mention of "time" in P3 evokes the same TIME IS A PRECIOUS POSSESSION conceptual metaphor as in P2. Again, the by far most recurrent MRW are the verbs *give, have*, and *spend* that objectify time in that metaphor. In addition, the use of the verb *support* as in "support communities" is directly related to the prompt ("...to helping their communities") as are the metaphorical collocations "free time" and "dedicate time".

In all the other prompts, trigger effects do not occur or are not as quantitatively relevant as in P1 to P3. P4, for example, does not show any spikes in metaphor frequencies with the most frequent MRW (*image*) merely occurring 5 times. The same is true in P5 with the terms *ruining, reach, comfortable*, and *advancement* being mentioned 4 times each as the most frequent MRW. A weak effect can be observed in P6 where the prompt "The best way to travel is in a group led by a tour guide" triggers the metaphorical collocation to "take a trip" that recurs 14 times across the learner texts. In P7 and P8, the most frequent MRW are not stimulated by the prompt, and there are similarly low token frequencies leading to a flat frequency distribution among the MRW. Incidentally, P8 ("Successful people try new things and take risks rather than only

P1: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.

P2: Young people enjoy life more than older people do.

P3: Young people nowadays do not give enough time to helping their communities.

metaphorical expression	**# of occur.**	**metaphorical expression**	**# of occur.**	**metaphorical expression**	**# of occur.**
P1		**P2**		**P3**	
“broad(er) knowledge”	60	have/-ing/has “have time”	55	give/-s/-ing/-en “give time”, “give help”	29
focus/-ed “focus on a particular subject/field”	12	spend/-s/-ing “spend time/hours/years”	14	have/-ing/has “have time”	24
*give/-s/-ing	10	*face “face problems / change / responsibilities”	6	spend/-ing/-t “spend time”	16
wide/-er	9	*get	5	support/-s/-ing/-ed “support communities”	14
*lead/-s	7	stage “stage of life”	4	*strong/-er/-ly	14
*spend	6			free “free time”	9
				dedicate/-s “dedicate time”	9

Table 2: Most frequent metaphorical expressions in P1, P2, and P3.
(*) MRW does not occur verbatim in the prompt

doing what they already know how to do well”) contains the metaphorical expression “take risks” that recurs in the learner essays 43 times. However, it has not been manually annotated as an MRW in the ETS Corpus.

If we take a look at how the MRW are distributed across the different parts of speech, it is interesting to note that verbs are by far more often marked as metaphors than nouns and adjectives/adverbs. Among the 30 most frequent MRW per prompt in the training set, 416 metaphorical verbs precede over 198 adjectives/adverbs and 141 nouns.

Finally, the learner data poses another peculiarity that is worth considering in the automatic metaphor identification. There is a total of 99 misspelled MRW across all prompts in the data (4.6% of all MRW) as in *messege, actractivity, strenght, knowled, isolte, dangerousness*, and *broadn* to randomly pick out a few. Finding a way to factor in the out-of-vocabulary words will increase the performance of automatic metaphor detection.

3 Design and implementation

Our approach extends the system from Stemle and Onysko (2018), which combines a small set of established machine learning (ML) methods and tools. In particular, the system uses fastText[1] word embeddings from different corpora in a bi-directional recursive neural network architecture with long-term short-term memory (LSTM BiRNN) and implements a flat sequence-to-sequence neural network with one hidden layer using TensorFlow+Keras (Abadi et al., 2015) in Python. With the goals we introduced in Section 1, it seemed sufficient to use a successful system from the last ST instead of improving the overall system by integrating the latest, highly successful[2] developments from the field of NLP (see, e.g., Wang

[1] https://fasttext.cc/

[2] The current development of increasing the complexity in neural network architectures by adding more processing layers in systems comes with the trade-off of loosing insights into the mechanisms of how the improvements are achieved. See also Mikolov (2020).

et al., 2020 for an overview).

In their experimental design Stemle and Onysko (2018) use word embeddings from corpora such as Wikipedia[3] and BNC[4] as well as from texts of language learners (TOEFL11). This is to follow the intuition that the use of metaphors can vary depending on the competence of the learners and that these differences can be helpful in training a metaphor detection system.

For the current ST, the system was slightly extended. We

- bumped the requirements of the used tools to current versions (in particular for TensorFlow+Keras, gensim[5], FastText, and scikit-learn[6]),
- adapted the system to the new format of the TOEFL data set,
- improved the use of sub-word information encoded in FastText word representations (we fixed a bug that prevented the system to use proper subword character n-gram representations in some cases), and
- added an option to integrate metadata into the input representations for the neural network.

For the last point, we adapted the way the input for the neural network is represented: The number of input layers corresponds to the number of features, i.e. for multiple features, e.g. multiple WE models or additional PoS tags, sequences are concatenated on the word level such that the number of features for an individual word grows. For the metadata, we added an input layer with the number of dimensions varying with the number of encoded metadata.

We maintain the implementation in a source code repository[7]. The system is available under the APLv2 open-source license.

4 Experiments

4.1 Combining VUA and TOEFL

In our first experiment, we tried to extend the training data and combine the two available datasets. Given the discussion in Section 2, we expected confounding effects due to the fact that the manual classification of the All-POS- and Verb-metaphors are different in these two sets.

First, we shuffled both datasets individually and then combined them in three ways: A re-shuffled combination of the two sets and two combinations where we put one set at the beginning and the other one at the end. For the evaluation we emulated a 10-fold CV, with training on combinations of the original datasets and testing on a held out part of one of the datasets: We trained on one of our combined sets and tested on one of 10 parts of the uncombined dataset, which had been held out from the training, and repeated this for all 10 parts. As word embeddings, we used BNC and the complete TOEFL11 data.

Table 3 shows that, most notably, the highest recall is achieved on the Verbs task when using first the VUA data and subsequently the TOEFL data, and testing on the TOEFL data. We interpret that in a way that the learning of the VUA data is mostly 'forgotten' by the neural network, but its focus on Verb-metaphors leaves a strong initialization bias towards verbs.

Overall, the results of the various runs show that the much larger VUA data dominates the learnt properties of the model, and that the matching focus on Verb-metaphors in both datasets improves recall.

4.2 Metadata

In this experiment, we added available metadata as additional information to learn from. The difference is that compared to other information, such as POS tags, this metadata applies to whole sentences and the entire texts. Also, this experiment only addresses the TOEFL dataset.

The available metadata are the following:

- Prompt: The prompt that triggered the production of the respective sentence and text (P1-P8)
- Proficiency: The proficiency of the language learner who produced the text (medium or high)
- L1: The language learner's L1 (ARA, JPN, ITA)
- Text ID: The text's unique ID (which represents the individual language learner)

[3] https://fasttext.cc/docs/en/english-vectors.html

[4] https://embeddings.sketchengine.co.uk/static/index.html

[5] https://radimrehurek.com/gensim/

[6] https://scikit-learn.org/

[7] https://github.com/bot-zen/

		Test on VUA		Test on TOEFL	
Training		Verbs	All-POS	Verbs	All-POS
Shuffled VUA + TOEFL	Pr	**0.58 (+/- 0.03)**	0.55 (+/- 0.04)	0.46 (+/- 0.09)	0.42 (+/- 0.04)
	Re	0.65 (+/- 0.03)	0.64 (+/- 0.05)	0.69 (+/- 0.07)	0.67 (+/- 0.07)
	F1	**0.61 (+/- 0.02)**	0.59 (+/- 0.02)	0.54 (+/- 0.06)	0.52 (+/- 0.03)
Sequential 1:VUA 2:TOEFL	Pr	0.56 (+/- 0.05)	0.55 (+/- 0.06)	0.43 (+/- 0.06)	0.44 (+/- 0.04)
	Re	0.67 (+/- 0.06)	0.64 (+/- 0.09)	**0.71 (+/- 0.07)**	0.68 (+/- 0.04)
	F1	0.60 (+/- 0.03)	0.58 (+/- 0.03)	0.53 (+/- 0.04)	0.53 (+/- 0.03)
Sequential 1:TOEFL 2:VUA	Pr	0.56 (+/- 0.04)	0.53 (+/- 0.06)	0.46 (+/- 0.06)	0.42 (+/- 0.05)
	Re	0.68 (+/- 0.04)	0.67 (+/- 0.08)	0.69 (+/- 0.08)	0.70 (+/- 0.06)
	F1	0.61 (+/- 0.03)	0.59 (+/- 0.02)	0.55 (+/- 0.05)	0.52 (+/- 0.04)
Baseline: VUA and TOEFL individually	Pr	0.55 (+/- 0.04)	0.55 (+/- 0.03)	0.57 (+/- 0.07)	0.53 (+/- 0.06)
	Re	0.69 (+/- 0.04)	0.68 (+/- 0.04)	0.63 (+/- 0.08)	0.68 (+/- 0.06)
	F1	0.61 (+/- 0.01)	0.61 (+/- 0.01)	0.59 (+/- 0.03)	0.59 (+/- 0.03)

Table 3: 10-fold CV comparison of training on (un)shuffled VUA and TOEFL data for the Verbs and All-POS Tasks.

- Text length: The length (in tokens) of the complete text the respective sentence belongs to

Given the discussion in Section 2, we expected confounding effects for some of the metadata, and we hoped to improve the results when factoring in the metadata that showed significant effects on the number and use of metaphors.

As word embeddings, we used only the BNC. The input data was constructed for both tasks (All-POS and Verbs) by adding the metadata information for every single word in the input sequence. Testing was done by 10-fold cross-validation.

Table 4 shows that, most notably, the overall metadata does *not* improve the results in a systematic, meaningful way. Also, some metadata, like the *Text ID* even considerably degrades performance. Overall, there is no clear tendency towards metadata being more – if at all - helpful.

The complete held-out test set was not available at the time of writing, but we had evaluated some combinations of metadata during the shared task (via CodaLab) and found that the Verbs task - contrary to our 10-fold CV - gained slightly from the use of metadata. An evaluation on the complete test set would have been preferable. Additionally, representing the metadata at the level of the entire sequence instead of for each word individually could also noticeably influence the results.

		Verbs	All-POS
Baseline (no metadt.)	Pr	0.53 (+/- 0.04)	0.53 (+/- 0.04)
	Re	0.64 (+/- 0.07)	0.63 (+/- 0.05)
	F1	0.57 (+/- 0.02)	0.57 (+/- 0.03)
Prompt	Pr	0.50 (+/- 0.06)	0.51 (+/- 0.05)
	Re	0.66 (+/- 0.10)	0.64 (+/- 0.05)
	F1	0.56 (+/- 0.02)	0.56 (+/- 0.03)
Proficiency	Pr	0.51 (+/- 0.07)	0.51 (+/- 0.05)
	Re	0.68 (+/- 0.09)	0.65 (+/- 0.07)
	F1	0.57 (+/- 0.02)	0.57 (+/- 0.04)
L1	Pr	0.49 (+/- 0.05)	0.53 (+/- 0.06)
	Re	0.68 (+/- 0.08)	0.60 (+/- 0.06)
	F1	0.57 (+/- 0.02)	0.56 (+/- 0.03)
Prom.+ Prof. + L1	Pr	0.54 (+/- 0.07)	0.52 (+/- 0.06)
	Re	0.62 (+/- 0.07)	0.64 (+/- 0.05)
	F1	0.57 (+/- 0.03)	0.57 (+/- 0.03)
Text ID	Pr	0.42 (+/- 0.08)	0.48 (+/- 0.08)
	Re	0.72 (+/- 0.08)	0.60 (+/- 0.10)
	F1	0.52 (+/- 0.05)	0.52 (+/- 0.03)
Text length	Pr	0.50 (+/- 0.06)	0.53 (+/- 0.05)
	Re	0.66 (+/- 0.11)	0.62 (+/- 0.05)
	F1	0.56 (+/- 0.03)	0.57 (+/- 0.03)
All	Pr	0.45 (+/- 0.07)	0.44 (+/- 0.08)
	Re	0.68 (+/- 0.08)	0.64 (+/- 0.08)
	F1	0.54 (+/- 0.04)	0.51 (+/- 0.04)

Table 4: 10-fold CV comparison of training with different metadata on the TOEFL dataset.

5 Conclusion

This paper has focused on the structure of the learner data used in the Second Shared Task of Metaphor Identification. We aimed at exploring possible factors that influence this kind of data and tested whether these play a role for the automated identification using word embeddings in an established LSTM BiRNN system from the first ST in 2018. A descriptive investigation of the manually annotated sample of the ETS Corpus of Non-Native Written English (TOEFL11 corpus) shows that the factors of proficiency and especially the essay prompt exhibit significant correlations to the amount and type of metaphors found in the annotated training set. The data also show a numerical bias towards the annotation of verbs as metaphors compared to other content words.

A sequential training of the bidirectional neural network using both the VUA and the TOEFL partitions of the shared task points to the different structure of the datasets, in particular towards an emerging bias of overidentifying verbal metaphors in the neural network based classification. The hypothesized influence of the metadata in the TOEFL set, in particular the observed dependencies on proficiency and the essay prompt, was not confirmed by the results of the automated identification. While the factors of L1, proficiency, prompt and essay length did not influence the baseline results, the essay ID (i.e. the individual learner) reduced the performance of the system as did a combination of all metadata. For the future, more tests with different ways of modelling the metadata in the neural network architecture and on the test set of the task will provide further insights.

Acknowledgments

The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018a. A corpus of non-native written english annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee. 2018b. *Proceedings of the Workshop on Figurative Language Processing*. Association for Computational Linguistics, New Orleans, Louisiana.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

British National Corpus Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition).

Marcus Callies and Marta Degani, editors. *Metaphor in Language and Culture across World Englishes*. Bloomsbury, London. Forthcoming.

Jonathan Charteris-Black. 2016. *Fire Metaphors: Discourses of Awe and Authority*. Bloomsbury Academic.

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.

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.

Raymond W. Gibbs Jr. 2017. *Metaphor Wars: Conceptual Metaphors in Human Life*. Cambridge University Press, Cambridge.

Zoltán Kövecses. 2020. *Extended Conceptual Metaphor Theory*, first edition. Cambridge University Press.

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

Chee Wee 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*, Seattle, WA.

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.

Jeannette Littlemore. 2019. *Metaphors in the Mind: Sources of Variation in Embodied Metaphor*, first edition. Cambridge University Press.

Tomáš Mikolov. 2020. Complexity and simplicity in machine learning. Keynote at the GeCKo symposium (Integrating Generic and Contextual Knowledge) 2020.

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, Volume 1 (Long and Short Papers)*, pages 596–601, Minneapolis, Minnesota. 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, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

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, LA. Association for Computational Linguistics.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*.