Verbal Multiword Expressions for Identification of Metaphor

Omid Rohanian[†], **Marek Rei**[§ ‡], **Shiva Taslimipoor**[‡], **Le An Ha**[†]
[†]RGCL, University of Wolverhampton, United Kingdom
[‡]ALTA Institute, University of Cambridge, United Kingdom
[§]Department of Computing, Imperial College London, United Kingdom
{omid.rohanian, l.a.ha}@wlv.ac.uk
marek.rei@imperial.ac.uk, st797@cl.cam.ac.uk

Abstract

Metaphor is a linguistic device in which a concept is expressed by mentioning another. Identifying metaphorical expressions, therefore, requires a non-compositional understanding of semantics. Multiword Expressions (MWEs), on the other hand, are linguistic phenomena with varying degrees of semantic opacity and their identification poses a challenge to computational models. This work is the first attempt at analysing the interplay of metaphor and MWEs processing through the design of a neural architecture whereby classification of metaphors is enhanced by informing the model of the presence of MWEs. To the best of our knowledge, this is the first "MWE-aware" metaphor identification system paving the way for further experiments on the complex interactions of these phenomena. The results and analyses show that this proposed architecture reach state-of-the-art on two different established metaphor datasets.

1 Introduction

Human language is rife with a wide range of techniques that facilitate communication and expand the capacities of thinking and argumentation. One phenomenon of such kind is metaphor. Metaphor is defined as a figure of speech in which the speaker makes an implicit comparison between seemingly unrelated things which nonetheless have certain common characteristics (Shutova, 2010). This is done to convey an idea which is otherwise difficult to express succinctly or simply for rhetorical effect.

As an example, in the sentence *she devoured his novels*, the verb *devour* is used in a metaphorical sense that implies reading quickly and eagerly. The literal and metaphorical senses share the element of intense desire which in turn helps to decode the meaning of the word in its context.

It is clear that a mere literal understanding of semantics would not result in proper understanding of a metaphorical expression and a non-compositional approach would be required (Shutova et al., 2013; Vulchanova et al., 2019). The human brain is equipped with the necessary machinery to decode the intended message behind a metaphorical utterance. This involves mentally linking the seemingly unrelated concepts based on their similarities (Rapp et al., 2004).

Verbal MWEs (VMWEs) are another example of non-literal language in which multiple words form a single unit of meaning. These two phenomena share some common ground. Expressions like *take the bull by the horns*, *go places*, *kick the bucket*, or *break someone's heart* can be categorised as metaphorical VMWEs. Based on this observation we hypothesise that a metaphor classification model can be bolstered by knowledge of VMWEs.

In this work we focus on how identification of verbal metaphors can be helped by verbal MWEs. We devise a deep learning model based on attention-guided graph convolutional neural networks (GCNs) that encode syntactic dependencies alongside information about the existence of VMWEs and we test the model on two established metaphor datasets.

2 Related Works

The tasks of MWE and metaphor identification share some similarities. Many idiomatic MWEs can be considered as lexicalised metaphors.

Idioms are where the overlap becomes clear (Kordoni, 2018). It is important to note, however, that not all verbal metaphors are VMWEs. Metaphors that are less conventionalised and appear in creative context (e.g. within a poem or a literary piece) and are not established enough to make it as entries into dictionaries are examples of such cases. However, the distinction between these categories is not always clear, and few precise tests exist for the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2890–2895
July 5 - 10, 2020.

annotators to tell them apart (Gross, 1982). [1]

Most state-of-the-art MWE identification models are based on neural architectures (Ramisch et al., 2018; Taslimipoor and Rohanian, 2018) with some employing graph-based methods to make use of structured information such as dependency parse trees (Waszczuk et al., 2019; Rohanian et al., 2019). Top-performing metaphor detection models also use neural methods (Rei et al., 2017; Gao et al., 2018), with some utilising additional data such as sentiment and linguistic information to further improve performance (Mao et al., 2019; Dankers et al., 2019).

3 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016) are a variation of the classic CNNs that perform the convolution operation on nodes of a graph, making them suitable for capturing non-sequential inter-dependencies in the input.

Using the per-sentence formalism (Marcheggiani and Titov, 2017; Rohanian et al., 2019), GCN can be defined as:

$$GCN = f(WX^TA + b) \quad (1)$$

where W, X, A, b, and GCN refer to the weight matrix, representation of the input sentence, adjacency matrix, bias term, and the output of the convolution respectively. f is a nonlinearity which is often the relu function.

3.1 Multi-head Self-attention

Attention is a mechanism inspired by human visual attention which aims to encode sequences by emphasising their most informative parts through weighting. Self-attention (Cheng et al., 2016), also referred to as intra-attention, is a special case of the attention mechanism which relates different parts of the same sequence and relies only on information from the same sequence. When the sequence is a series of words, this means encoding the sentence by learning correlations between words in the sentence. Self-attention is a powerful method to learn long-range dependencies in a sequence.

In this work, we use a particular form of self-attention introduced by Vaswani et al. (2017) in which the weighting is determined by scaled dot product. Given the input representation X, three smaller sized vectors are created. These are Query, Key, and Value which are represented with Q, K, and V respectively. The output of self-attention is computed with:

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \quad (2)$$

N different self-attention mechanisms are activated in parallel. This approach is known as N-headed self-attention, where each head $H_i = Att(QW_i^Q, KW_i^K, V)$ and the projections W_i^Q and W_i^K are parameter matrices. The outputs from these individual heads are later used in GCN layers (Guo et al., 2019).

3.2 Attention Guided Adjacency

Central to GCN is the adjacency matrix where the relations between nodes are defined. Converting the graph of relations to an adjacency matrix involves a rule-based hard pruning strategy and potentially results in discarding valuable information due to the sparsity of the matrix. Influenced by Guo et al. (2019), in this work we consider dependency parse information as an undirected graph with adjacency A. To obtain $\tilde{A}$, we combine matrix A with matrices H_0, H_1,..., H_{N-1} induced by the N-headed self-attention mechanism defined in Section 3.1.

Given an N-headed attention, each A is converted to several $\tilde{A}_i$s where $i \in \{1, 2, ..N\}$ and each $\tilde{A}_i$ is a linear combination of A and H_i.

$$\tilde{A}_i = \alpha \times H_i + (1 - \alpha) \times A \quad (3)$$

Each $\tilde{A}_i$ can be interpreted as a fully connected graph where the relation strength between every two nodes is determined by a weight value. In this case, a higher weight signifies a stronger relation and a value close to zero would signal a lack of connection. These edge-weighted graphs are then fed to separate GCNs. A consolidated representation is finally achieved by a linear combination of the outputs from these N different GCNs.

The use of attention within the GCN network is motivated by the assumption that multi-hop paths between distantly related nodes could potentially be captured this way. We stack n layers of attention-guided GCNs using residual connections with n being a hyper-parameter that is tuned independently in each dataset.

Graph Attention (GAT) (Veličković et al., 2017) is a closely related work where the scope of attention is the neighbourhood of each node, whereas we make use of the entire sentence.

[1] See PARSEME annotation guidelines at https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/

3.3 MWE-Aware GCN

In order to inform the model of the structural hierarchy within the sentence and encode information about MWEs, our attention-guided GCN component integrates information from two separate sources; namely, the dependency parse information and token-level relations between components of existing MWEs in the sentence. These correspond to adjacencies $\tilde{A}_{DEP}$ and $\tilde{A}_{MWE}$ which are fed each into separate GCNs and the output is a concatenation of the outputs from both components:

$$GCN = concat[GCN_{s_{MWE}}; GCN_{s_{DEP}}] \quad (4)$$

4 Experiments

We describe the datasets used in the experiments and then provide details of the overall system.

4.1 Datasets

We apply the systems on two different metaphor datasets: MOH-X, and TroFi, which contain annotations for verb classification. Both of these datasets contain a set of sentences in which a single verb token is labelled as metaphorical or not. There is also an index provided that specifies the location of the target token in the sentence.

MOH-X. MOH-X is based on earlier work by Mohammad et al. (2016). It consists of short 'example' sentences from WordNet (Fellbaum, 1998)[2] with labels for metaphorical verbs along with associated confidence scores. Shutova et al. (2016) created a subset of this dataset, referred to as MOH-X, and added annotations for each verb and its argument. This dataset has 214 unique verbs.

TroFi. Similar to MOH-X, TroFi (Birke and Sarkar, 2006) has annotations for target verbs in each sentence. It has a comparatively longer average sentence length with 28.3 words per sentence compared to MOH-X's 8.0. The sentences in TroFi are constructed from the Wall Street Journal Corpus (Charniak et al., 2000). There are only 50 unique target verbs in this dataset.

4.2 MWE Identification

We extract MWEs using the GCN-based system proposed by Rohanian et al. (2019). Since we are focusing on verbal metaphors in this study, we train the system on the PARSEME English dataset (Ramisch et al., 2018), which is annotated for verbal MWEs. As a result, predicted MWE labels in our target datasets are IOB formatted, where B and I denote the *beginning* and *inside* tokens of an MWE and O signifies tokens not belonging to MWEs.

	TroFi	MOH-X
verbal metaphor	1627	315
MWE	257	77

Table 1: Number of predicted MWEs among target verbs.

We encode the relations between components of MWEs in each sentence using an adjacency matrix. Tokens of a sentence are nodes of the adjacency matrix; edges exist between tokens of an MWE. Relation matrices are then fed to the attention guided system as explained in Section 4.3.

The numbers of verbal MWEs in correlation with target verbs in metaphor datasets are shown in Table 1. As can be seen, almost 16% of metaphors in TroFi and 24% of metaphors in MOH-X are automatically labelled as VMWEs. This provides a strong motivation for incorporating this information into the metaphor identification system.

4.3 System Description

For our experiments, we devise two strong baselines and compare them against our proposed model. All three systems are built on top of a pre-trained BERT architecture (Devlin et al., 2019).

The starting baseline (BERTBaseline) is vanilla pre-trained BERT with a classification layer added on top. The other two models (BERT+GCN and BERT+MWE-Aware GCN) are created by adding extra layers with trainable parameters on top of the BERT model, augmenting its original structure. [3]

BERT+GCN is BERT plus an attention-guided GCN that uses dependency parse information. Finally, BERT+MWE-Aware GCN refers to the system that uses BERT along with the added MWE-aware GCN component that utilises both dependency and VMWE information as detailed in Section 3.3.

Adam (Kingma and Ba, 2014) is used for optimising the network; the learning rate is controlled with a linear warmup scheduler in which the rate

[2]Examples are sentences after the gloss that show in-context usage

[3]For all the experiments we use the pre-trained BERT model, `bert-base-uncased`, from the transformers library (Wolf et al., 2019).

Models	MOH-X				TroFi			
	Acc	P	R	F1	Acc	P	R	F1
Gao et al. (2018)	78.5	75.3	**84.3**	79.1	73.7	68.7	74.6	72.0
RNN-HG (Mao et al., 2019)	79.7	79.7	79.8	79.8	74.9	67.4	**77.8**	72.2
RNN-MHCA (Mao et al., 2019)	79.8	77.5	83.1	80.0	**75.2**	68.6	76.8	72.4
BERTBaseline	78.04	78.38	77.87	77.82	70.38	70.54	68.89	68.84
BERT+GCN	79.44	79.79	79.36	79.31	72.01	72.32	70.45	70.65
BERT+MWE-Aware GCN	**80.47**	**79.98**	80.40	**80.19**	73.45	**73.78**	71.81	**72.78**

Table 2: Performance of MWE-Aware GCN against baselines and state-of-the-art on MOH-X and TroFi

decreases linearly after increasing during a warmup period. In all the models, given the verb index in the dataset[4], and before passing the token-level output of the GCN to the softmax layer, we slice the output tensor based on the provided index and only select for the representation of the token of interest and subsequently pass this sliced tensor to the classification layer.

5 Results

We report the results in terms of accuracy, precision, recall and F_1-score, macro averaged over the measures obtained from 10 fold cross-validation. As can be seen in Table 2, our proposed model outperforms the baselines and also surpasses state-of-the-art in terms of F_1-score and precision in both datasets. As a whole, the results obtained for the two datasets are more homogeneous across the four metrics compared to previous state-of-the-art.

In order to have a fair comparison with the previous state-of-the-art, it is important to consider their architectures. Gao et al. (2018), which our model outperforms in most criteria across the two datasets, is a BiLSTM-based system that uses a combination of ELMo and GLoVe vectors for input representation. The two models by Mao et al. (2019) are more competitive, especially in accuracy and precision for the TroFi dataset. RNN-HG and RNN-MHCA are BiLSTM-based systems grounded in linguistic theories of Selectional Preference Violation (SPV) (Wilks, 1978) and Metaphor Identification Procedure (MIP) (Steen et al., 2007) which are based on the semantic contrast between the metaphorical word and its context or between the literal and contextualised meanings of a target token. These two models also make use of contextualised embeddings.

6 Discussion

The larger portion of annotated VMWEs in both datasets are figurative and thus provide a valuable signal to metaphoricity. TroFi proved to be more challenging as sentences can be as long as 118 tokens with several different VMWEs and only a single token of interest which could be labelled as literal. On the other hand, MOH-X is more focused and VMWEs, for the most part, coincide with the target verb.

A notable pattern in the results is when the baselines miss a metaphor and the proposed model correctly identifies it due to the presence of a non-compositional VMWE. A typical example is given below where *tack together*, identified initially as an MWE, signals metaphoricity:[5]

(1) He **tacked** together some verses.

There are examples of sentences falsely classified by BERT+GCN as metaphorical which are correctly identified as not by BERT+MWE-Aware GCN. This shows the model has picked up informative cues and general patterns. There are also metaphors missed by BERT+GCN that do not have explicitly tagged VMWEs, but the proposed model is still able to capture them. Example 2 is an instance of such case:

(2) The residents of this village **adhered** to Catholicism.

Due to their correlation with metaphoricity, VMWE information equips the model with the ability to identify metaphorical usage, which is reflected in the superior precision scores. However, this correlation is not always definitive, and in certain cases where a VMWE is realised in its literal meaning, the model might incorrectly associate its

[4] An index specifies the location of the target token.

[5] Target tokens are boldfaced

presence with metaphor. The following two sentences from MOH-X are examples of false positives influenced by VMWEs. Here, *jam the brake* and *land in* are VMWEs with literal meanings which can be idiomatic in other contexts:

(3) The driver **jammed** the brake pedal to the floor.

(4) The ship **landed** in Pearl Harbor

There are only a few such cases in MOH-X, however in TroFi, the problem is exacerbated by longer sentences with multiple target tokens. One possible remedy could be to not attend to all the tokens in each sentence but instead look at a certain window around the target token. We did not explore this idea in this work as it would defeat the purpose of attention-guided GCNs, but are open to considering it in future in such a way that accuracy is improved without hurting the precision scores which are higher in both datasets than previous state-of-the-art.

7 Conclusions and Future Work

In this work, we presented a neural model to classify metaphorical verbs in their sentential context using information from the dependency parse tree and annotations for verbal multiword expressions. To the best of our knowledge, this is the first MWE-aware metaphor identification system, that demonstrates how the knowledge of MWEs can enhance the performance of a metaphor classification model. Experiments showed that the resulting system sets a new state-of-the-art in several criteria across two benchmark metaphor datasets. The code used in the experiments will be made publicly available [6].

For future work, we plan to add VMWE annotations to the VU Amsterdam Corpus (Steen, 2010) which is the largest metaphor dataset and extend our experiments using that resource. Directionality of edges did not result in improvement in our models in this work, however for future, we plan to develop GCNs that incorporate edge typing, which would enable us to differentiate between different MWE types and dependency relations while comparing them against the current models.

[6] https://github.com/omidrohanian/metaphor_mwe

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. BLLIP 1987-89 WSJ corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561.

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.

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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

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.

Maurice Gross. 1982. Une classification des phrases figées du français. *Revue québécoise de linguistique*, 11(2):151–185.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Valia Kordoni. 2018. Beyond multiword expressions: Processing idioms and metaphors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 15–16, Melbourne, Australia. 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.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

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.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, et al. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions.

Alexander M Rapp, Dirk T Leube, Michael Erb, Wolfgang Grodd, and Tilo TJ Kircher. 2004. Neural correlates of metaphor processing. *Cognitive brain research*, 20(3):395–402.

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.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. *arXiv preprint arXiv:1902.10667*.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697. Association for Computational Linguistics.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

GJ Steen, LJ Cameron, AJ Cienki, P Crisp, A Deignan, W Raymond jr, J Grady, Z Kövecses, GD Low, and E Semino. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Shiva Taslimipoor and Omid Rohanian. 2018. SHOMA at parseme shared task on automatic identification of VMWEs: Neural multiword expression tagging with high generalisation. *arXiv preprint arXiv:1809.03056*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Mila Vulchanova, Evelyn Milburn, Valentin Vulchanov, and Giosuè Baggio. 2019. Boon or burden? the role of compositional meaning in figurative language processing and acquisition. *Journal of Logic, Language and Information*, 28(2):359–387.

Jakub Waszczuk, Rafael Ehren, Regina Stodden, and Laura Kallmeyer. 2019. A neural graph-based approach to verbal MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 114–124, Florence, Italy. Association for Computational Linguistics.

Yorick Wilks. 1978. Making preferences more active. *Artificial intelligence*, 11(3):197–223.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.