Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions

Omid Rohanian[†*], **Shiva Taslimipoor**[†*], **Samaneh Kouchaki**[‡], **Le An Ha**[†], **Ruslan Mitkov**[†]
[†]Research Group in Computational Linguistics, University of Wolverhampton, UK
[‡]Institute of Biomedical Engineering, University of Oxford, UK
{omid.rohanian, shiva.taslimi, l.a.ha, r.mitkov}@wlv.ac.uk
samaneh.kouchaki@eng.ox.ac.uk

Abstract

We introduce a new method to tag Multiword Expressions (MWEs) using a linguistically interpretable language-independent deep learning architecture. We specifically target discontinuity, an under-explored aspect that poses a significant challenge to computational treatment of MWEs. Two neural architectures are explored: Graph Convolutional Network (GCN) and multi-head self-attention. GCN leverages dependency parse information, and self-attention attends to long-range relations. We finally propose a combined model that integrates complementary information from both, through a gating mechanism. The experiments on a standard multilingual dataset for verbal MWEs show that our model outperforms the baselines not only in the case of discontinuous MWEs but also in overall F-score.[1]

1 Introduction

Multiword expressions (MWEs) are linguistic units composed of more than one word whose meanings cannot be fully determined by the semantics of their components (Sag et al., 2002; Baldwin and Kim, 2010). As they are fraught with syntactic and semantic idiosyncrasies, their automatic identification remains a major challenge (Constant et al., 2017). Occurrences of discontinuous MWEs are particularly elusive as they involve relationships between non-adjacent tokens (e.g. ***put*** *one of the blue masks* ***on***).

While some previous studies disregard discontinuous MWEs (Legrand and Collobert, 2016), others stress the importance of factoring them in (Schneider et al., 2014). Using a CRF-based and a transition-based approach respectively, Moreau et al. (2018) and Al Saied et al. (2017) try to capture discontinuous occurrences with help from dependency parse information. Previously explored neural MWE identification models (Gharbieh et al., 2017) suffer from limitations in dealing with discontinuity, which can be attributed to their inherently sequential nature. More sophisticated architectures are yet to be investigated (Constant et al., 2017).

Graph convolutional neural networks (GCNs) (Kipf and Welling, 2017) and attention-based neural sequence labeling (Tan et al., 2018) are methodologies suited for modeling non-adjacent relations and are hence adapted to MWE identification in this study. Conventional GCN (Kipf and Welling, 2017) uses a global graph structure for the entire input. We modify it such that GCN filters convolve nodes of dependency parse tree on a per-sentence basis. Self-attention, on the other hand, learns representations by relating different parts of the same sequence. Each position in a sequence is linked to any other position with $O(1)$ operations, minimising maximum path (compared to RNN's $O(n)$) which facilitates gradient flow and makes it theoretically well-suited for learning long-range dependencies (Vaswani et al., 2017).

The difference in the two approaches motivates our attempt to incorporate them into a hybrid model with an eye to exploiting their individual strengths. Other studies that used related syntax-aware methods in sequence labeling include Marcheggiani and Titov (2017) and Strubell et al. (2018) where GCN and self-attention were separately applied to semantic role labelling.

Our contribution in this study, is to show for the first time, how GCNs can be successfully applied to MWE identification, especially to tackle discontinuous ones. Furthermore, we propose a novel architecture that integrates GCN with self-attention outperforming state-of-the-art. The resulting models not only prove superior to existing

*The first two authors contributed equally.

[1] The code is available on https://github.com/omidrohanian/gappy-mwes.

Proceedings of NAACL-HLT 2019, pages 2692–2698
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

methods in terms of overall performance but also are more robust in handling cases with gaps.

2 Methodology

To specifically target discontinuity, we explore two mechanisms both preceding a Bi-LSTM: 1) a GCN layer to act as a syntactic ngram detector, 2) an attention mechanism to learn long-range dependencies.

2.1 Graph Convolution as Feature Extraction

Standard convolutional filters act as sequential ngram detectors (Kim, 2014). Such filters might prove inadequate in modeling complex language units like discontinuous MWEs. One way to overcome this problem is to consider non-sequential relations by attending to syntactic information in parse trees through the application of GCNs.

GCN is defined as a directed multi-node graph $G(V, E)$ where $v_i \in V$ and $(v_i, r, v_j) \in E$ are entities (words) and edges (relations) respectively. By defining a vector x_v as the feature representation for the word v, the convolution equation in GCN can be defined as a non-linear activation function f and a filter W with a bias term b as:

$$c = f\Big(\sum_{i \in r(v)} W x_i + b\Big) \tag{1}$$

where $r(v)$ shows all words in relation with the given word v in a sentence, and c represents the output of the convolution.

Following Kipf and Welling (2017) and Schlichtkrull et al. (2017), we represent graph relations using adjacency matrices as mask filters for inputs. We derive associated words from the dependency parse tree of the target sentence. Since we are dealing with a sequence labelling task, there is an adjacency matrix representing relations among words (as nodes of the dependency graph) for each sentence. We define the sentence-level convolution operation with filter W_s and bias b_s as follows:

$$C_s = f(W_s X^T A + b_s) \tag{2}$$

where X, A, and C are representation of words, adjacency matrix, and the convolution output, all at the level of sentence. The above formalism considers only one relation type, while depending on the application, multiple relations can be defined.

Kipf and Welling (2017) construct separate adjacency matrices corresponding to each relation type and direction. Given the variety of dependency relations in a parse tree (e.g. obj, nsubj, advcl, conj, etc), and per-sentence adjacency matrices, we would end up with an over-parametrised model in a sequence labeling task. In this work, we simply treat all relations equally, but consider only three types of relations: 1) the head to the dependents, 2) the dependents to the head, and 3) each word to itself (self-loops). The final output is obtained by aggregating the outputs from the three relations.

2.2 Self-Attention

Attention (Bahdanau et al., 2014) helps a model address the most relevant parts of a sequence through weighting. As attention is designed to capture dependencies in a sequence regardless of distance, it is complementary to RNN or CNN models where longer distances pose a challenge. In this work we employ multi-head self-attention with a weighting function based on scaled dot product which makes it fast and computationally efficient.

Based on the formulation of Transformer by Vaswani et al. (2017), in the encoding module an input vector x is mapped to three equally sized matrices K, Q, and V (representing key, query and value) and the output weight matrix is then computed as follows:

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^T}{\sqrt{d}}\Big)V \tag{3}$$

The timing signal required for the self-attention to work is already contained in the preceding CNN layers alleviating the need for position encoding.

2.3 Model Architecture

The overall scheme of the proposed model, composed of two parallel branches, is depicted in Figure 1. We employ multi-channel CNNs as the step preceding self-attention. One channel is comprised of two stacked 1D CNNs and the other is a single 1D CNN. After concatenation and batch normalisation, a multi-head self attention mechanism is applied (Section 2.2).

Parallel to the self-attention branch, GCN learns a separate representation (Section 2.1). Since the GCN layer retains important structural information and is sensitive to positional data from the syntax tree, we consider it as a position-based approach. On the other hand, the self-attention layer

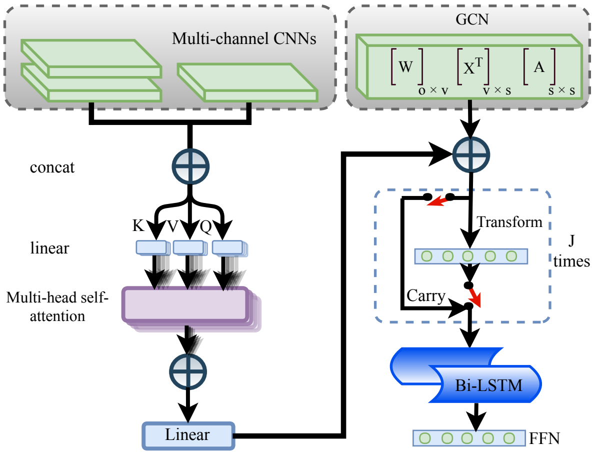

Figure 1: A hybrid sequence labeling approach integrating GCN (o: output dimension; v: word vectors dimension; s: sentence length) and Self-Attention.

is intended to capture long-range dependencies in a sentence. It relates elements of the same input through a similarity measure irrespective of their distance. We therefore regard it as a content-based approach. As these layers represent different methodologies, we seek to introduce a model that combines their complementary traits in our particular task.

Gating Mechanism. Due to the considerable overlap between the GCN and self-attention layers, a naive concatenation introduces redundancy which significantly lowers the learning power of the model. To effectively integrate the information, we design a simple gating mechanism using feed-forward highway layers (Srivastava et al., 2015) which learn to regulate information flow in consecutive training epochs. Each highway layer consists of a Carry (Cr) and a Transform (Tr) gate which decide how much information should pass or be modified. For simplicity Cr is defined as $1 - Tr$. We apply a block of J stacked highway layers (the section inside the blue dotted square in Figure 1). Each layer regulates its input x using the two gates and a feedforward layer H as follows:

$$y = Tr \odot H + (1 - Tr) \odot x \qquad (4)$$

where $\odot$ denotes the Hadamard product and Tr is defined as $\sigma(W_{Tr}x + b_{Tr})$. We set b_{Tr} to a negative number to reinforce carry behavior which helps the model learn temporal dependencies early in the training.

Our architecture bears some resemblance to Marcheggiani and Titov (2017) and Zhang et al. (2018) in its complementary view of GCN and BiLSTM. However there are some important differences. In these works, BiLSTM is applied prior to GCN in order to encode contextualised information and to enhance the teleportation capability of GCN. Marcheggiani and Titov (2017) stack a few BiLSTM layers with the idea that the resulting representation would enable GCN to consider nodes that are multiple hops away in the input graph. Zhang et al. (2018) use a similar encoder, however the model employs single BiLSTM and GCN layers, and the graph of relations is undirected.

In our work, we use pre-trained contextualised embeddings that already contain all the informative content about word order and disambiguation. We put BiLSTM on top of GCN, in line with how CNNs are traditionally applied as feature generating front-ends to RNNs. Furthermore, Marcheggiani and Titov (2017) use an edge-wise gating mechanism in order to down-weight uninformative syntactic dependencies. This method can mitigate noise when parsing information is deemed noisy, however in Zhang et al. (2018) it caused performance to drop. Given our low-resource setting, in this work we preferred not to potentially down-weight contribution of individual edges, therefore treating them equally. We rely on gating as the last step when we combine GCN and self-attention.

3 Experiments

Data. We experiment with datasets from the shared task on automatic identification of verbal Multiword Expressions (Ramisch et al., 2018). The datasets are tagged for different kinds of verbal MWEs including idioms, verb particle constructions, and light verb constructions among others. We focus on annotated corpora of four languages: French (FR), German (DE), English (EN), and Persian (FA) due to their variety in size and proportion of discontinuous MWEs. Tags in the datasets are converted to a variation of IOB which includes the tags `B` (beginning of MWEs), `I` (other components of MWEs), and `O` (tokens outside MWEs), with the addition of `G` for arbitrary tokens in between the MWE components e.g. $\textit{make}_{\texttt{[B]}}$ $\textit{important}_{\texttt{[G]}}$ $\textit{decisions}_{\texttt{[I]}}$.

ELMo. In our experiments, we make use of ELMo embeddings (Peters et al., 2018) which are contextualised and token-based as opposed to

		All		Discontinuous			
		Token-based	MWE-based				
					MWE-based		
L	model	F	F	%	P	R	F
EN	baseline	41.37	35.38	32	24.44	10.48	14.67
	GCN-based	39.78	39.11		39.53	16.19	22.97
	Att-based	33.33	31.79		46.88	14.29	21.90
	H-combined	41.63	**40.76**		63.33	18.10	**28.15**
DE	baseline	62.27	57.17	43	69.50	45.37	54.90
	GCN-based	65.48	**61.17**		65.19	47.69	55.08
	Att-based	61.20	58.19		67.86	43.98	53.37
	H-combined	63.80	60.71		68.59	49.54	**57.53**
FR	baseline	76.62	72.16	43	75.27	52.04	61.54
	GCN-based	79.59	75.15		79.58	56.51	66.09
	Att-based	78.21	74.23		71.49	60.59	65.59
	H-combined	80.25	**76.56**		77.94	59.11	**67.23**
FA	baseline	88.45	86.50	14	67.76	55.88	61.29
	GCN-based	87.78	86.42		78.72	54.41	64.35
	Att-based	87.55	84.20		62.32	63.24	62.77
	H-combined	88.76	**87.15**		75.44	63.24	**68.80**

Table 1: Model performance (P, R and F) for development sets for all MWE and only discontinuous ones (%: proportion of discontinuous MWES)

type-based word representations like `word2vec` or `GLoVe` where each word type is assigned a single vector. Token-based embeddings better reflect the syntax and semantics of each word in its context compared to traditional type-based ones. We use the implementation by Che et al. (2018) to train ELMo embeddings on our data.

Validation. In the validation phase, we start with a strong baseline which is a CNN + Bi-LSTM model based on the top performing system in the VMWE shared task (Taslimipoor and Rohanian, 2018). Our implemented baseline differs in that we employ ELMo rather than `word2vec` resulting in a significant improvement. We perform hyper-parameter optimisation and make comparisons among our systems, including GCN + Bi-LSTM (GCN-based), CNN + attention + Bi-LSTM (Att-based), and their combination using a highway layer (H-combined) in Table 1.

4 Evaluation and Results

Systems are evaluated using two types of precision, recall and F-score measures: strict MWE-based scores (every component of an MWE should be correctly tagged to be considered as true positive), and token-based scores (a partial match between a predicted and a gold MWE would be considered as true positive). We report results for all MWEs as well as discontinuous ones specifically.

According to Table 1, GCN-based outperforms Att-based and they both outperform the strong baseline in terms of MWE-based F-score in three out of four languages. Combining GCN with attention using highway networks results in further improvements for EN, FR and FA. The H-combined model consistently exceeds the baseline for all languages. As can be seen in Table 1, GCN and H-combined models each show significant improvement with regard to discontinuous MWEs, regardless of the proportion of such expressions.

In Table 2 we show the superior performance (in terms of MWE-based F-score) of our top systems on the test data compared to the baseline and state-of-the-art systems, namely, ATILF-LLF (Al Saied et al., 2017) and SHOMA (Taslimipoor and Rohanian, 2018). GCN works the best for discontinuous MWEs in EN and FA, while H-combined outperforms based on results for all MWEs except for FA. The findings are further discussed in Section 5.

5 Discussion and Analysis

The overall results confirm our assumption that a hybrid architecture can mitigate errors of individual models and bolster their strengths. To demonstrate the effectiveness of the models in detecting discontinuous MWEs, in Figure 2 we plot their performance for FR and EN given a range of different gap sizes. As an ablation study, we show the results for the baseline, GCN-based, Att-based only, as well as H-combined models. GCN and Att-based models each individually outperform the baseline, and the combined model clearly improves the results further.

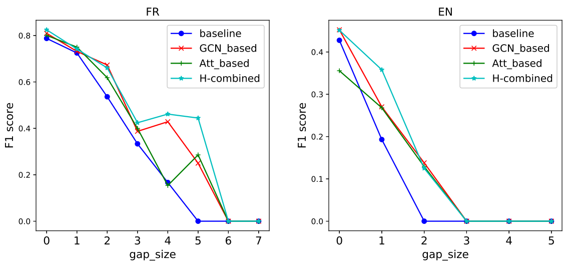

Figure 2: Model performance given different gap sizes

The example in Figure 3 taken from the English dataset demonstrates the way GCN considers relations between non-adjacent tokens in the sentence. Our baseline is prone to disregarding these links. Similar cases captured by both GCN and H-combined (but not the baseline) are ***take*** *a final* ***look***, ***picked*** *one* ***up***, and ***cut*** *yourself* ***off***.

	All | Discontinuous			
	EN	DE	FR	FA
baseline	33.01 | 16.53	54.12 | 53.94	67.66 | 58.70	**81.62** | 61.73
GCN-based	36.27 | **24.15**	56.96 | 54.87	70.79 | 59.95	81.00 | **62.35**
H-combined	**41.91** | 22.73	**59.29** | **55.00**	**70.97** | **63.90**	80.04 | 61.90
ATILF-LLF	31.58 | 09.91	54.43 | 40.34	58.60 | 51.96	77.48 | 53.85
SHOMA	26.42 | 01.90	48.71 | 40.12	62.00 | 51.43	78.35 | 56.10

Table 2: Comparing the performance of the systems on test data in terms of MWE-based F-score

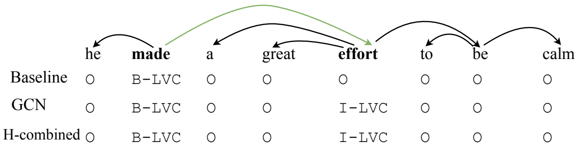
he made a great effort to be calm
Baseline O B-LVC O O O O O O
GCN O B-LVC O O I-LVC O O O
H-combined O B-LVC O O I-LVC O O O

Figure 3: Sample sentence with a discontinuous occurrence of an English MWE, *make an effort*.

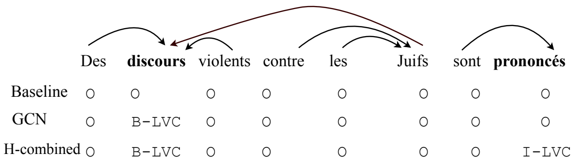
Des discours violents contre les Juifs sont prononcés
Baseline O O O O O O O O
GCN O B-LVC O O O O O O
H-combined O B-LVC O O O O O I-LVC

Figure 4: Example sentence with a discontinuous occurrence of a French MWE, *prononcer un discours* 'to make a speech'.

In more complicated constructs where syntactic dependencies might not directly link all constituents, GCN alone is not always conducive to optimal performance. In Figure 4, the French sentence is in the passive form and MWE parts are separated by 5 tokens. This is an MWE skipped by GCN but entirely identified by the H-combined model.

It is important to note that model performance is sensitive to factors such as percentage of seen expressions and variability of MWEs (Pasquer et al., 2018). In FA for instance, 67% of the MWEs in the test set are seen at training time, making them easy to be captured by the baseline (Taslimipoor et al., 2018). Furthermore, only 21% of MWEs in FA and 15% in EN are discontinuous as opposed to 44% in FR and 38% in DE. In this case, a sequential model can already learn the patterns with high accuracy and the potential of a GCN and self-attention is not fully exploited.

Also in DE, a sizable portion of MWEs are verbal idioms (VIDs) which are known for their lexico-syntactic fixedness and prevalence of tokens that lack a standalone meaning and occur only in a limited number of contexts (also known as cranberry words). Furthermore, MWEs in the Persian dataset are all Light Verb Constructions (LVCs), which can be modelled using lexical semantic templates (Megerdoomian, 2004). For such MWEs, our models compete with strong sequential baselines.

6 Conclusion and Future Work

In this paper, we introduced the application of GCN and attention mechanism to identification of verbal MWEs and finally proposed and tested a hybrid approach integrating both models. Our particular point of interest is discontinuity in MWEs which is an under-explored area. All the individual and combined models outperform state-of-the-art in all considered criteria. In future, we will further develop our system using structured attention (Kim et al., 2017) and try to improve the accuracy of parsers in multi-tasking scenarios.

References

Hazem Al Saied, Matthieu Constant, and Marie Candito. 2017. The ATILF-LLF system for Parseme shared task: a transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 127–132, Valencia, Spain. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, pages 267–292. CRC Press.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages

55–64, Brussels, Belgium. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2017. Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 54–64, Vancouver, Canada. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Joël Legrand and Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016)*, Berlin, Germany.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515.

Karine Megerdoomian. 2004. A semantic template for light verb constructions. In *Proceedings of the First Workshop on Persian Language and Computers. Tehran University, Iran*, pages 25–26.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, and Carl Vogel. 2018. CRF-Seq and CRF-DepTree at PARSEME Shared Task 2018: Detecting verbal MWEs using sequential and dependency-based approaches. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 241–247. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Jean-Yves Antoine, and Carlos Ramisch. 2018. Towards a variability measure for multiword expressions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 426–432.

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

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. In *the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK. Springer-Verlag.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL*, 2:193–206.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *AAAI Conference on Artificial Intelligence*.

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*.

Shiva Taslimipoor, Omid Rohanian, Ruslan Mitkov, and Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth.*

Extended papers from the MWE 2017 workshop, pages 299–317. Language Science Press, Berlin, Germany.

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.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.