Cross-lingual Transfer Learning and Multitask Learning for Capturing Multiword Expressions

Shiva Taslimipoor, Omid Rohanian, Le An Ha
Research Group in Computational Linguistics
University of Wolverhampton, UK
{shiva.taslimi,omid.rohanian,l.a.ha}@wlv.ac.uk

Abstract

Recent developments in deep learning have prompted a surge of interest in the application of multitask and transfer learning to NLP problems. In this study, we explore for the first time, the application of transfer learning (TRL) and multitask learning (MTL) to the identification of Multiword Expressions (MWEs). For MTL, we exploit the shared syntactic information between MWE and dependency parsing models to jointly train a single model on both tasks. We specifically predict two types of labels: MWE and dependency parse. Our neural MTL architecture utilises the supervision of dependency parsing in lower layers and predicts MWE tags in upper layers. In the TRL scenario, we overcome the scarcity of data by learning a model on a larger MWE dataset and transferring the knowledge to a resource-poor setting in another language. In both scenarios, the resulting models achieved higher performance compared to standard neural approaches.

1 Introduction

Multiword Expressions (MWEs) are combinations of two or more lexical components that form non/semi-compositional meaning units. Due to their idiosyncratic behaviour, MWEs have been studied using various statistical and machine learning approaches including supervised classification (Diab and Bhutada, 2009), tagging (Schneider et al., 2014), and unsupervised prediction (Fazly et al., 2009). Studies have focused on both their syntactic (Constant and Nivre, 2016) and semantic (Van de Cruys and Moirón, 2007) features.

Recently, the PARSEME project provided an extensive multilingual dataset of verbal MWEs (Ramisch et al., 2018). Datasets of certain languages in this resource are rich with a huge number of tagged sequences while others are considerably smaller. Several notable systems have been proposed to train sequence labelling models on this dataset including neural (Taslimipoor and Rohanian, 2018) and non-neural systems (Moreau et al., 2018). MWE prediction for some of these languages has proved to be more challenging due to several reasons including scarcity of data, higher percentage of unseen MWE instances in the test set, and prevalence of discontinuous or variable MWEs.

In this paper, we focus on one of those languages for which the results were collectively low (interestingly it was English) and explore two neural approaches in order to address the shortcomings of the current neural models and enhance learning. The two approaches are: multitask learning and transfer learning, with two different motivations.

Syntactic and semantic idiosyncrasies in MWEs call for special treatment, with models that take them into account from different perspectives. Syntactic and semantic information are commonly fed to the models as input features. However, we consider an alternative way to exploit this information. Specifically, in a supervised setting, we add dependency syntax information as auxiliary supervision. Therefore we perform multitask learning between MWE and dependency parse tags.

Syntactic dependency information has been previously proven to be successful in identifying MWEs (Constant and Nivre, 2016). However, neural processing methodologies are yet to be deeply explored for MWE modelling (Constant et al., 2017). In multitask learning we have several different prediction tasks over the same input. The idea is that the process of learning features for one task can be helpful for another.

In order to deal with data scarcity in the English dataset, in another setting we train our model on a language with a larger data and transfer the learned knowledge for predicting MWE tags in English.

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 155–161
Florence, Italy, August 2, 2019.

In this study we build upon recent neural network systems that have proved to be successful in representing syntactic and semantic features of text and design novel multitask and transfer learning architectures for MWE identification. The contributions of this work are: 1) we propose a neural model that improves MWE identification by jointly learning MWE and dependency parse labels; 2) We show that MWE identification models, when multitasked with dependency parsing, outperform the models which naively add dependency parse information as additional features; 3) we propose, to the best of our knowledge for the first time, a cross-lingual transfer learning method for processing MWEs, thus making a contribution towards the study of low-resource languages.

2 Related Work

Constant and Nivre (2016) proposed joint syntactic and lexical analysis in which the syntactic dimension of their structure is represented by a dependency tree, and the lexical dimension is represented by a forest of trees. The two dimensions share token-level representations. They use a transition-based system that jointly learns both lexical and syntactic analysis resulting in an improvement for the task of MWE identification.

The idea of multitask learning (MTL) in neural networks was popularised by the work of Collobert et al. (2011). They improved the performance of chunking by jointly learning it with POS tagging. Søgaard and Goldberg (2016) discuss the idea further by pinpointing that supervising different tasks on different layers is beneficial. Specifically, in their work, for an input sequence, $w_{1:n}$ they have several RNN layers l for each task, t, and their task-specific classifier is defined as: $task_t(w_{1:n}, i) = f_t(v_i^{l(t)})$ where $1 \leq i \leq n$, v_i is the output representation of RNN for word i and f_t is the tagger/classification function. This way, different tasks might be applied to different RNN layers (i.e. there are layers shared by several tasks, and layers that are specific to some tasks). We use this idea here, by having some specific layers for final MWE prediction which are not shared with the auxiliary parsing task.

Using an LSTM-based model, Bingel and Søgaard (2017) performed a study to find beneficial tasks for the purpose of MTL in a sequence labelling scenario. In their work, the MWE model benefited from most auxiliary tasks such as chunking, CCG parsing, and Super-sense tagging. A similar finding is reported in Changpinyo et al. (2018) where performance of an MWE tagger was consistently improved when jointly trained with any of the 10 different auxiliary tasks in various MTL settings.

Transfer learning (TRL) has seen a flurry of interest with the advent of pre-trained language models, transformers, and contextualised embeddings (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2018). Transfer learning is particularly helpful where data scarcity can be an issue, and a related task with more data can be used to alleviate the issue. Liu et al. (2018) is an example of the use of task-aware language models to enhance sequence labelling using an LSTM-CRF architecture powered by a language model.

A related scenario in TRL is when tasks remain the same but models are designed to transfer knowledge across languages. In NLP, cross-lingual transfer learning has been extensively explored in the context of representation learning where monolingual spaces are mapped into a common embedding space through methods like retrofitting (Faruqui et al., 2015), matrix factorization (Vyas and Carpuat, 2016) or similar. Outside representation learning, there have been many attempts to use TRL in NLP tasks. For sequence labelling, Kim et al. (2017) trained POS tagging models cross-lingually without access to parallel resources. The model consisted of two LSTM components where one is shared between the languages and the other is private (language-specific).

Yang et al. (2017) is a notable example of cross-lingual transfer learning under low-resource settings where sequence labelling models were trained to transfer knowledge between English, Spanish, and Dutch for POS tagging, chunking, and Named Entity Recognition (NER) through the use of shared and private parameters. In that work, three different architectures were explored for cross-domain, cross-application, and cross-lingual transfer. The core of their proposed models is similar to Lample et al. (2016), with minor differences including the incorporation of GRU instead of LSTM and a training objective based on the max-margin principle.

3 Methodology

The core of our model is a neural architecture that incorporates CNN and LSTM layers which are

commonly employed in sequence tagging models.[1] We adapt the architecture to the two scenarios of multitask and transfer learning. The details of the layers and input representations for these models are further explained in Section 4 and depicted in Figure 1.

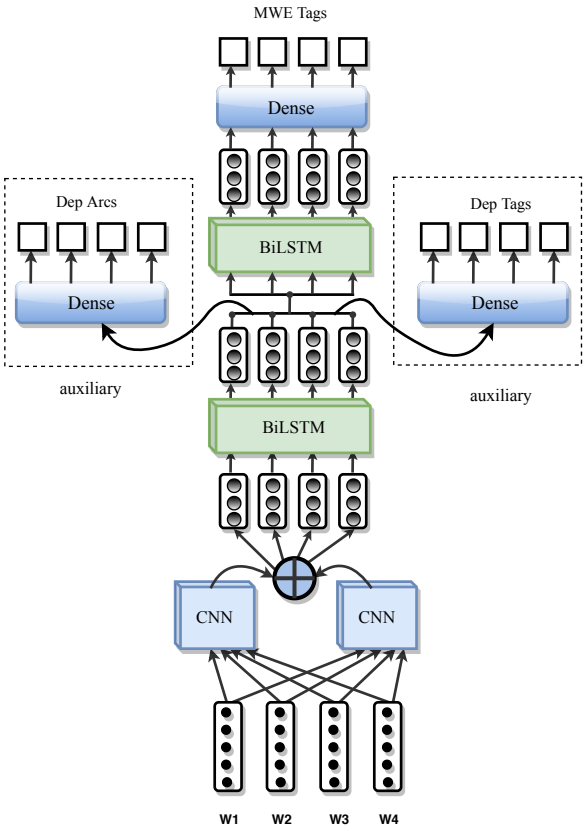

Figure 1: Overall architecture of the model (consisting of two auxiliary tasks in case of MTL)

3.1 Multitask Learning

In the multitask learning scenario, the models are required to simultaneously predict MWE tags, dependency parse arcs and dependency parse labels. A sample of all three-fold labels that the model should predict for a sentence is depicted in figure 2. In order to learn the main output, MWE tag, the model computes loss values for two auxiliary outputs, Dep arc and Dep tag, and add them to the main output loss.

Similar to the idea of Søgaard and Goldberg (2016), we introduce the supervision of dependency parsing in lower layers and aim to boost the performance of the final MWE tagging layer. To this end, the parallel CNNs and the first BiLSTM layer is shared between the two tasks. On top of this, two layers with independent auxiliary losses are applied to predict dependency tags. Parallel to this, we add a single BiLSTM before the main output layer for predicting MWE tags (Figure 1). In this study, we simply add the main loss to the two auxiliary losses (which are all computed using categorical cross-entropy).

[1]Two CNN layers without pooling act like feature extractors. Their results are then concatenated and given to the next BiLSTM layer.

INPUT	INPUT	AUX-OUT	AUX-OUT	OUTPUT
Word	POS	Dep arc	Dep tag	MWE tag
Worse	ADJ	10	advmod	*
yet	ADV	1	advmod	*
,	PUNCT	1	punct	*
what	PRON	6	nsubj	*
is	AUX	6	aux	*
going	VERB	10	csubj	2:VPC
on	ADV	6	compound :prt	2
will	AUX	10	aux	*
not	PART	10	advmod	*
let	VERB	0	root	1:VID
us	PRON	10	obj	*
alone	ADJ	10	xcomp	1
.	PUNCT	10	punct	*

Figure 2: Annotation of one sample sentence containing one VPC and a verbal idiom in the English data for the Parseme shared task edition 1.1.

3.2 Transfer Learning

In transfer learning, also known as domain adaptation, information from a source task is retained to enhance learning for another related task. In this study, we use TRL in a multilingual scenario. Since our target language is low-resource, the aim is to benefit from richer data of another language. To this end, a model which is trained on the domain of one language is transferred to the domain of another target language.

The two languages have the same sets of POS and dependency parse tags. Therefore, one-hot encoded POS and dependency inputs are shared between the trained and the transferred models. When loading pretrained contextualised embeddings as inputs, the sentences of individual languages have their own sets of weights. On the other hand, we also have a setting in which our model starts with a trainable embedding layer. In this case, the vocabularies of both languages are combined and indexed together. This way, common vocabularies or proper nouns of the two languages receive the same indices.

In this study, we first train the model on the German data, and then transfer the weights to an identical model which is re-trained on English for a fewer number of iterations.

4 Experiments

We experiment with the multilingual dataset from the PARSEME project (Savary et al., 2018) which was made available for the shared task on identification of verbal MWEs (Ramisch et al., 2018). Verbal MWEs in the dataset include idioms, verb particle constructions, and light verb constructions, among others. MWE tags in the dataset are similar to IOB labels, since there is a distinction between the beginning and other components of an MWE. We target the data for English which is surprisingly small in this dataset (with $3,471$ training and $3,965$ test sequences) and try to use MTL and TRL to improve MWE identification.

The inputs to our system are combinations of ELMo embeddings which are trained on our data using the implementation provided by Che et al. (2018) and one-hot encoded POS tags. In cases where we add dependency parse information as inputs, the representation for dependency arcs and labels are as follows. In order to represent arcs, we use adjacency matrix representation for each sentence. In the adjacency matrix, each token is assigned a row in which all cells are zero except for the one corresponding to the head of the token in dependency tree. Dependency labels, though are one-hot encoded.

We set hyperparameters based on the ones used in a similar architecture proposed by Taslimipoor and Rohanian (2018) which was implemented for a single task and mono-lingual setting. The CNN layers have 200 neurons, one with filter size 2 and the other with size 3, both with $relu$ activation. BiLSTM layers have both 300 neurons, dropout 0.5, and recurrent dropout of 0.2. We use the $Adam$ optimizer for all settings. Figure 1 shows the whole architecture for MTL. The model architecture for standard setting and TRL is the same excluding the auxiliary components.

4.1 Evaluation

In the MTL setting, we make comparison between the case when the model is trained only on MWE tags (single-task, STL) to when jointly trained to predict MWE and dependency parsing tags in a multitask scenario (MTL). We also compare the results of joint prediction with the case when dependency information is directly fed as additional input. In the TRL setting, we first train our model on the German data which has $6,734$ training sequences.[2] We finally compare the results from TRL with all other results.

We evaluate the models using F1-score in two settings: 1) strict matching (MWE-based) in which all components of an MWE are considered as a unit that should be correctly classified; and 2) fuzzy matching (token-based) in which any correctly predicted token of the data is counted (Savary et al., 2017).

4.2 Results

The results are reported in Table 1. We report the average F1-score over five separate runs along with standard deviation. The first two rows show the baseline results when we use the neural model in the standard setting. For the second row, we use dependency parsing tags as well as ELMo and POS tags for the input to the system.

In the third and the fourth rows (MTL), we observe that the results improve when dependency parse information is predicted as auxiliary output. In particular, we observe these improvements when adding the dependency loss outputs at one layer before the outermost BiLSTM. We also see that the addition of POS to the input is not necessarily effective in the MTL setting (i.e. according to the third row, the MTL setting without POS results in a better performance). Our best MTL system outperforms the systems that participated in open track of the Parseme shared task (Ramisch et al., 2018) for English data. However, it performs slightly worse than the neural system proposed by Rohanian et al. (2019), which deals with discontinuous MWEs using graph convolutional network and attention mechanism.

The models are trained on `google colab` with GPU: 1xTesla K80, having 2496 CUDA cores, compute 3.7, and 12GB GDDR5 VRAM. While the MTL model might seem to be complicated, it does not add much to the time complexity of the model. Specifically it takes, on average, 45 minutes to train the MTL model compared to 43 minutes to train STL both for 100 epochs.

The performance of TRL is only slightly better than STL and lower than MTL. This is not to our surprise, because ELMo vectors, that are one of the inputs to all the models, are pre-trained on huge amount of data and bring enough knowledge to the low resource.

[2]The idea is to train on a Germanic language which is a category that English also belongs to.

setting	inputs	Token-based F1	MWE-based F1
STL	ELMo	34.86 ± 1.66	32.27 ± 1.36
	ELMo+POS+DEP	36.08 ± 2.41	33.68 ± 2.99
MTL	ELMo	**40.18 ± 1.52**	35.96 ± 1.09
	ELMo+POS	38.86 ± 1.63	**36.61 ± 1.27**
TRL	ELMo+POS	37.55 ± 1.42	35.69 ± 1.99
	ELMo+POS+DEP	38.44 ± 1.92	35.84 ± 2.39

Table 1: Comparing the performance of the CNN-biLSTM model (in terms of average F1 over 5 runs with standard deviation) in single (STL), multitask (MTL) and transfer learning (TRL) scenarios.

setting	Token-based F1	MWE-based F1
closed STL	30.34 ± 1.36	28.12 ± 1.37
TRL	33.31 ± 0.75	30.40 ± 0.66

Table 2: Comparing the performance of transfer learning (TRL) with the standard setting (STL).

Furthermore, in the case of TRL, we hypothesize a scenario in which we do not have access to a huge amount of data and avoid using ELMo as the input. We perform a preliminary experiment with a randomly initialized embedding layer as the first component of the network to be trained with other layers. We report the results of this experiment in Table 2. This way the model is not using any extensive external data (hence the name closed STL). Here we can better see the benefits of transferring the model cross-lingually. More investigations need to be done to discover the limits of this approach (e.g. through the application of different language models and experimentation with other architectures of the same kinds).

4.3 The Effect of Learning Rate in TRL

When transferring from the source to the target domain, the model is prone to overfitting on the new data, losing the potentially beneficial information from the high-resource model. This problem is sometimes referred to as catastrophic forgetting. One way to mitigate this issue is to control for the hyperparameters of the source and target language, specially setting the learning rate in a way that domain adaptation occurs incrementally. Ongoing research explore various regularization and ensemble methods to preserve and transfer knowledge between tasks (Chronopoulou et al., 2019; Lee et al., 2017; Rusu et al., 2016). These methods, however, introduce varying degrees of computational complexities.

Even though the sensitivity of TRL to the learning rate is largely acknowledged in the literature, previous work is indecisive as to what learning rate scheduling achieves the best result. Bowman et al. (2015) lower the starting learning rates after transfer, in order to preserve pre-transfer information in early training. Kocmi and Bojar (2018) however, found that, in TRL between language pairs in the task of neural machine translation, changing hyperparameters from the parent to the child model harmed performance. Mou et al. (2016) set the best hyperparameters from the source task during the validation phase and transferred them to the target domain. They acknowledged that the hyperparameters can potentially become biased towards the source domain. The conclusion was that the best hyperparameters are ready to be transferred during the epoch range when the performance peaks in the source domain.

In this work we refrained from altering the learning rate, since, consistent with some of the previous work, we noticed a sharp decline in performance when changing this value.

5 Conclusions and Future Work

In this work we explored two neural architectures to improve identification of MWEs through learning of related linguistic tasks. [3] We experimented with cross-lingual transfer learning between two Germanic languages, and in a separate scenario, we designed and tested a multitask learning approach to tag MWEs while concurrently training on dependency arcs and labels as auxiliary tasks. Our results show that the models prove promising and outperform the standard baseline. In future we plan to study these techniques in more detail, and make extensive comparisons between them in order to understand to what extent and under what circumstances they help MWE identification.

[3] The code for the experiments is available at https://github.com/shivaat/VMWE-Identification

References

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2965–2977.

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.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. *arXiv preprint arXiv:1902.10547*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Mathieu Constant, Glen Eryiit, 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.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 161–171, Berlin, Germany. Association for Computational Linguistics.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 17–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in neural information processing systems*, pages 4652–4662.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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 *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) at the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 241–247.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 479–489.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. 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 2692–2698, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebes kind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. 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*. Language Science Press, Berlin, Germany.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47.

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.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

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*.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1187–1197.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.