TRAVIS at PARSEME Shared Task 2020: How good is (m)BERT at seeing the unseen?

Murathan Kurfalı
Linguistics Department, Stockholm University
Stockholm, Sweden
murathan.kurfali@ling.su.se

Abstract

This paper describes the TRAVIS system built for the PARSEME Shared Task 2020 on semi-supervised identification of verbal multiword expressions. TRAVIS is a fully feature-independent model, relying only on the contextual embeddings. We have participated with two variants of TRAVIS, TRAVIS$_{multi}$ and TRAVIS$_{mono}$, where the former employs multilingual contextual embeddings and the latter uses monolingual ones. Our systems are ranked second and third among seven submissions in the open track, respectively. Thorough comparison of both systems on eight languages reveals that despite the strong performance of multilingual contextual embeddings across all languages, language-specific contextual embeddings exhibit much better generalization capabilities.

1 Introduction

Multiword expressions (MWEs) are, most commonly, defined as a group of words which act as a single lexical unit and display idiomaticity at lexical, syntactic, semantic or pragmatic levels (Baldwin and Kim, 2010). As the name suggests, verbal MWEs (VMWEs) are MWEs with a verb as the head in their canonical form. Identification of VMWEs tend to be more challenging than that of other MWEs, as they exhibit more syntactic/morphological variation (due to inflection of the verb), their components can be interrupted by other words (he **made** a serious **mistake**) and furthermore their order may vary (the **decision** was hard to **take**) (Savary et al., 2017). Yet, their identification is equally important as it is a prerequisite to fully address a number of downstream tasks, such as machine translation, information retrieval or syntactic parsing.

This year's shared task is built upon the observation that the existing models fail when it comes to identify the VMWEs which are not seen during the training. Hence, the aim of this year's shared task is updated to identify the *unseen* VMWEs in running text and the organizers provide annotated corpora with varying sizes in 14 different languages.

In this paper, we present two variants of TRAVIS, TRAVIS$_{multi}$ and TRAVIS$_{mono}$, which were submitted to the open track of the shared task, where additional resources were allowed. TRAVIS follows the tradition of approaching VMWE identification as a token classification task. To this end, it employs the, now standard, contextual embeddings model, BERT (Devlin et al., 2019), which has seen very limited application to this task. Due to the multilingual nature of the shared task, we also pay special attention to the performance on different languages. The variants of TRAVIS are named after this concern, highlighting the type of the contextual embeddings in terms of their pre-training languages: TRAVIS$_{multi}$ only uses the multilingual-BERT, which is trained on 104 languages, whereas TRAVIS$_{mono}$ uses the available language-specific BERT model for each language. Hence, the aim of the current submission is twofold: (i) we investigate the generalizability capabilities of pre-trained language models on identification of VMWEs (ii) we provide a thorough comparison of the multilingual-BERT against language-specific BERTs to understand the limitations of the former, if there is any, hoping to guide the future multilingual research on VMWEs.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 136–141
Barcelona, Spain (Online), December 13, 2020.

2 Background

Treating MWE identification as a sequence tagging problem has been one of the most popular approaches (Zampieri et al., 2019). To this end, Schneider et al. (2014) propose new tagging schemes for VMWE by extending the BIO format to allow the annotation of discontinuous and nested MWEs. Gharbieh et al. (2017) constitutes the first study which adopts this approach and applies deep learning models including feedforward, recurrent and convolutional networks. Later, a number of studies adopting a recurrent neural network with an optional CRF classifier have been proposed (Klyueva et al., 2017; Taslimipoor and Rohanian, 2018; Zampieri et al., 2018; Berk et al., 2018). Zampieri et al. (2019) further study the effects of different word representations on this architecture by using the Veyn model of (Zampieri et al., 2018). Rohanian et al. (2019) specifically target the challenge caused by discontinuity of verbal MWEs and propose a neural model which combines convolutional network and self-attention mechanism to deal with long-range relations.

3 System Description

Below, we briefly introduce the BERT language model, which constitutes the backbone of our model, followed by the introduction of the proposed models.

3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based deep bidirectional language model which has quickly become a new standard in NLP. It is pre-trained on a large unannotated corpus with two training objectives: (i) prediction of the missing words in a context (ii) given a sentence pair, determine if the second sentence follows the first one. These general pre-training objectives allow BERT to learn general enough representations which can be adjusted to any particular task through *fine-tuning*. In fine-tuning, a task-specific classification layer is added on the top of BERT and the whole model is, further, trained on the target task, updating the parameters of the BERT as well.

Originally, two different BERT models were released with different number of layers in their architecture: BERT-base (12-layer, 768-hidden, 12-heads, 110M parameters) and BERT-large (24-layer, 1024-hidden, 16-heads, 340M parameters). Additionally, a multilingual BERT (henceforth mBERT) was released, sharing the same architecture with the BERT-base model but trained on the concatenation of Wikipedias of 104 languages. Yet, since mBERT does not have any cross-lingual objectives nor trained on aligned data, its cross-lingual abilities and its limitations have, since, become a research topic (Karthikeyan et al., 2019).

3.2 Proposed Model(s)

We approach identification of VMWEs as a token classification problem. Our architecture follows the standard fine-tuning strategy employed for similar sequence tagging problems as described in the original BERT paper (Devlin et al., 2019). Briefly, we use BERT as our encoder with a linear layer connected to its hidden states on top to perform token level classification. In cases where the input token is split into several sub-tokens by the BERT's internal tokenizer, we pass the representation of the first sub-token to the linear layer classifier as the representation of the input token.

As stated earlier, there are two variants of TRAVIS where the only difference between them is the BERT model employed, otherwise completely identical. The first variant, *TRAVIS-multi*, uses mBERT as the encoder whereas the second variant, *TRAVIS-mono*, employs language specific BERT models and covers the following 8 languages:DE, FR, IT, PL, RO, SV, TR, ZH [1].

The motivation behind these two variants is the general finding that the monolingual models usually outperform mBERT (Nozza et al., 2020); yet, most languages still lack their own monolingual model

[1]Our original sshared task submission for TRAVIS$_{mono}$ also included predictions for EL and HI. However, we later discovered that the predictions for these languages were completely erroneous due to an error in the tokenization process. Therefore, we confine ourselves to the remaining eight languages in the current system description paper and we ask reader to dismiss the published results on the web-site for these two languages.

System	Langs	Unseen MWE-based			Global MWE-based			Global Token-based		
		P	R	F1	P	R	F1	P	R	F1
MTLB-STRUCT	14/14	36.24	41.12	38.53	71.26	69.05	70.14	77.69	70.9	74.14
TRAVIS-multi	13/14*	28.11	33.29	30.48	60.65	57.62	59.1	70.39	60.08	64.83
TRAVIS-mono	10/14	24.33	28.01	26.04	49.5	43.48	46.3	55.92	45.01	49.88
Seen2Unseen	14/14	16.14	11.95	13.73	63.36	62.69	63.02	66.33	61.63	63.89
FipsCo	3/14	4.31	5.21	4.72	11.69	8.75	10.01	13.26	8.51	10.37
HMSid	1/14	1.98	3.81	2.61	4.56	4.85	4.7	4.74	4.84	4.79
MultiVitamBooster	7/14	0.05	0.07	0.06	0.19	0.09	0.12	3.49	1.26	1.85
TRAVIS-multi	13/13	30.27	35.85	32.83	65.31	62.05	63.64	75.81	64.70	69.82
TRAVIS-mono	8/8[1]	43.86	49.91	46.69	74.87	74.77	74.82	80.76	76.94	78.80

Table 1: The official results of the all participating teams in the open track, ranked according to the F-score on unseen MWE identification. The bottom part presents our results when averaged over the languages covered. *The missing language is Portuguese, for which we failed to submit a result by the time of the shared task deadline due to a bug in the script.

as training such a language-specific BERT is computationally expensive. Hence, we believe that it is important to compare these models to gain insight regarding their performance for the future multilingual research, especially on low resource languages.

As for labelling, we follow a procedure similar to (Taslimipoor and Rohanian, 2018) and convert the PARSEME annotations into IOB-like labels. The PARSEME labels consist of VMWE's consecutive number in the sentence and its category, e.g. *2:LVC.full* denotes that the token with this tag is the first token of the 2^{nd} VMWE in that sentence, which is a light verb construction, whereas the other components of that VMWE are labeled with merely *2*. We modify these labels so that the initial token receives *B-* and the respective category and other tokens receives *I-* plus that category. All other tokens, which are not a part of any VMWE, receive *O* tag.

4 Experimental Design

Our implementation is based on the Transformers library of Huggingface (Wolf et al., 2019). All monolingual BERT models as well as mBERT are obtained through Huggingface's model hub[2]. In languages with several available BERT models, we opted for the most downloaded cased one. The complete list of the models used in the first submission are provided in Appendix A.

We train all models for four epochs using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5e-5. The sequence length is set to 400 during training, but at the time of the inference we use BERT's maximum sequence limit which is 512. As the fine-tuning procedure is prone to high variance, we run all our models four times and used the run with the best development performance to obtain the final predictions for the test sets.

5 Results and Discussion

TRAVIS variants ranked 2^{nd} and 3^{rd} in the general ranking, according to the target metric of the unseen MWE-based F-score. Table 1 summarizes the official results of all participating teams in the open track. Additionally, global MWE- and token-based scores are presented in order to give an overall idea about the participating teams[3]. In what follows, we discuss our results with a focus on performance in the discovery of the unseen VMWEs, following the main aim of the shared task.

Although TRAVIS$_{multi}$ ranks higher than TRAVIS$_{mono}$ in Table 1, it is because the official results are obtained by averaging the performance of the systems over all languages, independent of the number of the languages covered in the submission. When the results are averaged only over the languages

[2]https://huggingface.co/models

[3]The details of each measure can be found in (Savary et al., 2017).

System	Langs	Unseen MWE-based			Seen MWE-based			Global MWE-based		
		P	R	F1	P	R	F1	P	R	F1
TRAVIS-multi	8	35.22	40.19	37.54	90.44	81.42	85.69	73.30	71.03	72.15
TRAVIS-mono	8	40.28	48.27	43.91	90.98	83.68	87.17	74.87	74.77	74.82

Table 2: Average performance comparison of our two submissions on the following set of languages: DE, FR, IT, PL, RO, SV, TR, ZH. Following the updated definition of the shared task, any VMWE in the training *or* development set are regarded as seen.

covered in each submission (last two rows of Table 1), it becomes clear that TRAVIS$_{mono}$ performs better on average, achieving an increase of 14 F-score. However, to draw a more healthy comparison, we also compared the averaged performances of these variants on the same set of languages, which is provided in Table 2. The results show that TRAVIS$_{mono}$ still significantly outperforms TRAVIS$_{multi}$ by 6 F-score even when evaluated on the same set of languages. It must also be noted that this difference in performance is not due to a significant increase in one or several languages but consistent across all the common eight languages where TRAVIS$_{multi}$ only achieves better performance for Swedish by 1.4 F-score, otherwise outperformed by 8 points in F-score on average.

These results are in line with the previous findings that language specific BERT models perform better on the respective language. However, it must be highlighted that the biggest gain of the language-specific models is in the discovery of the unseen VMWEs. As far as those eight languages are concerned, both models show similar performance for the seen VMWEs with TRAVIS$_{mono}$ 87.17 and TRAVIS$_{multi}$ 85.69 F-score, respectively (Table 2). Hence, it is evident that language-specific BERTs are particularly better at generalizing to unseen VMWEs.

However, the performance of mBERT cannot be simply dismissed, as the TRAVIS$_{multi}$ also achieves consistent results across languages. Although the results are not directly comparable as the set of the languages is different and the datasets have, possibly, been modified over time, TRAVIS$_{multi}$ achieves an average of 10% increase in F-score over SHOMA (Taslimipoor and Rohanian, 2018), the best performing system of the previous PARSEME Shared Task (2018), in the identification of the unseen MWEs.

Language-wise, TRAVIS$_{mono}$ achieved the best performance in the open track of the shared task for the six of the eight languages it covers which are: FR, IT, PL, RO, TR, ZH. These languages represent various language families suggesting that the performance of TRAVIS is stable typologically. As for TRAVIS$_{multi}$, Irish (GA) turns out to be the most challenging language with only 2.6% F-score. However, that is probably due to the size of the dataset that contains only 100 VMWEs in the training portion which is too limited to fine-tune mBERT in a meaningful way. The language-wise comparison of our submissions with the best performing system is provided in Figure 1.

Finally, a general advantage of employing contextual embeddings is being completely feature-independent. Hence, the proposed model only requires a training data annotated for the positions of the target VMWEs, rendering it easily adaptable to other low resource languages where obtaining other linguistics features, such as POS-tags or dependency trees, can be challenging.

6 Conclusion

In this paper, we try to answer two questions: (i) how generalizable is the performance of contextual embeddings in VMWE identification, and (ii) if the pre-training language plays an important role or, in other words, the multilingual contextual embeddings are good enough. To this end, we offer a computational model, TRAVIS, which treats VMWE identification as a sequence classification task and employs various BERT models.

The results indicate that language-specific models perform particularly well on the identification of the *unseen* VMWEs by outperforming the multilingual embeddings by 6% in F-score when compared on the same set of languages. Yet, the multilingual-BERT also exhibits strong multilingual abilities, suggested by the average of 32% F-score in identification of the unseen VMWEs which is significantly higher than results obtained in the previous editions of the PARSEME shared tasks.

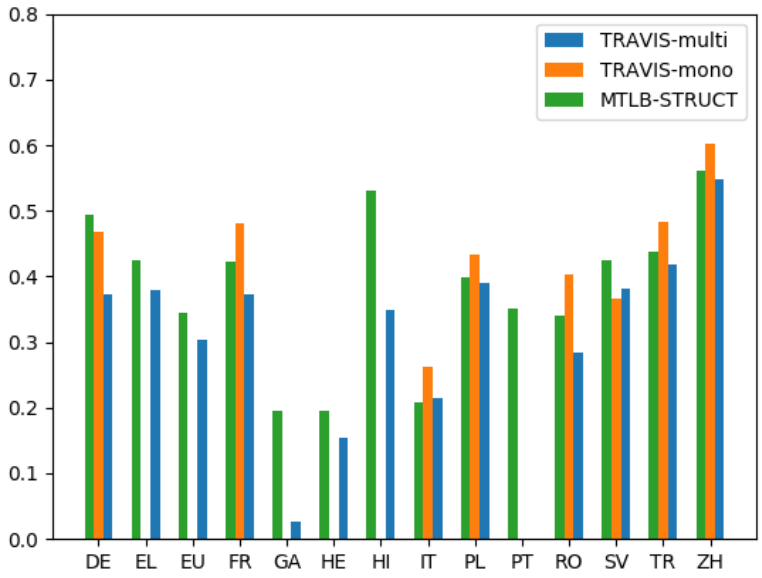

Figure 1: Language-wise comparison of our submissions and the first-ranked system (MTLB-STRUCT) on unseen MWEs.

Acknowledgments

I would like to thank Johan Sjons and anonymous reviewers for their valuable comments and NVIDIA for the GPU grant.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018. Deep-bgt at parseme shared task 2018: Bidirectional lstm-crf model for verbal multiword expression identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 248–253.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Waseem Gharbieh, Virendrakumar 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.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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

Natalia Klyueva, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden–making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Ruslan Mitkov, et al. 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.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Nicolas Zampieri, Manon Scholivet, Carlos Ramisch, and Benoit Favre. 2018. Veyn at parseme shared task 2018: Recurrent neural networks for vmwe identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 290–296.

Nicolas Zampieri, Carlos Ramisch, and Géraldine Damnati. 2019. The impact of word representations on sequential neural mwe identification.

Appendix A

Language	**Model name**
German	bert-base-german-cased
French (Martin et al., 2019)	camembert-base
Italian	bert-base-italian-cased
Polish	dkleczek/bert-base-polish-cased-v1
Romanian	bert-base-romanian-cased-v1
Swedish (Malmsten et al., 2020)	bert-base-swedish-cased
Turkish	bert-base-turkish-128k-cased
Chinese	bert-base-chinese

Table 3: The list of the monolingual BERT models used in the experiments. The model name denotes the model identifier of the corresponding model on Huggingface's model hub (huggingface.com/models)