Token-level Identification of Multiword Expressions using Pre-trained Multilingual Language Models

Raghuraman Swaminathan and **Paul Cook**
Faculty of Computer Science
University of New Brunswick
{rswamina,paul.cook}@unb.ca

Abstract

In this paper, we consider novel cross-lingual settings for multiword expression (MWE) identification (Ramisch et al., 2020) and idiomaticity prediction (Tayyar Madabushi et al., 2022) in which systems are tested on languages that are unseen during training. Our findings indicate that pre-trained multilingual language models are able to learn knowledge about MWEs and idiomaticity that is not language-specific. Moreover, we find that training data from other languages can be leveraged to give improvements over monolingual models.

1 Introduction

Multiword expressions (MWEs) are combinations of lexical items that exhibit some degree of idiomaticity (Baldwin and Kim, 2010). For example, *ivory tower* exhibits semantic idiomaticity because its meaning of a place where people are isolated from real-world problems is not transparent from the literal meanings of its component words.

Multiword expressions can be ambiguous in context with similar-on-the-surface literal combinations. For example, *red flag* is ambiguous between an MWE meaning a warning sign and a literal combination. Knowledge of MWEs can enhance the performance of natural language processing systems for downstream tasks such as machine translation (Carpuat and Diab, 2010) and opinion mining (Berend, 2011). Much work has therefore focused on recognizing MWEs in context, by identifying which tokens in a text correspond to MWEs (e.g., Schneider and Smith, 2015; Gharbieh et al., 2017; Ramisch et al., 2018, 2020) and by distinguishing idiomatic and literal usages of potentially-idiomatic expressions (e.g., Fazly et al., 2009; Salton et al., 2016; Haagsma et al., 2018; Liu and Hwa, 2018; King and Cook, 2018; Kurfalı and Östling, 2020).

One interesting line of investigation in such work is the ability of models to generalize to expressions that were not observed during training. For example, this was a focus in the evaluation of Ramisch et al. (2020). Fakharian and Cook (2021) further explore the ability of language models to encode information about idiomaticity that is not specific to a particular language by considering cross-lingual idiomaticity prediction, in which the idiomaticity of expressions in a language that was not observed during training is predicted. In this paper we further consider cross-lingual idiomaticity prediction.

SemEval 2022 task 2 subtask A (Tayyar Madabushi et al., 2022) is a binary sentence-level classification task of whether a sentence containing a potentially-idiomatic expression includes an idiomatic or literal usage of that expression. In this subtask, the training data consists of English and Portuguese, while the model is evaluated on English, Portuguese, and Galician. As such, the shared task considered evaluation on Galician, which was not observed during training. In this paper, we examine cross-lingual settings further, conducting experiments which limit the training data to one of English or Portuguese, to further assess the cross-lingual capabilities of models for idiomaticity prediction.

PARSEME 1.2 is a sequence labelling task in which tokens which occur in verbal MWEs, and the corresponding categories of those MWEs (e.g., light-verb construction, verb-particle construction), are identified (Ramisch et al., 2020). This shared task considered a monolingual experimental setup for fourteen languages; separate models were trained and tested on each language. In this work, we consider two different experimental setups: a multilingual setting in which a model is trained on the concatenation of all languages, and a cross-lingual setting in which, for each language, a model is trained on training data from all other languages, and is then tested on that language that was held out during training.

For each task considered, we use models based

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 1–6

on multilingual language models (e.g., mBERT). Our findings in cross-lingual experimental setups indicate that language models are able to capture information about MWEs that is not restricted to a specific language. Moreover, we find that knowledge from other languages can be leveraged to improve over monolingual models for MWE identification and idiomaticity prediction.

2 Models

For SemEval 2022 task 2 subtask A we apply BERT (Devlin et al., 2019) models for sequence classification. In the initial shared task, a multilingual BERT (mBERT) model is used for the baseline. We consider this, and also more-powerful models, including XLM-RoBERTa (Conneau et al., 2019) and mDeBERTa (He et al., 2021).

For PARSEME 1.2, we use the MTLB-STRUCT system (Taslimipoor et al., 2020), which performed best overall in the shared task. MTLB-STRUCT simultaneously learns MWEs and dependency trees by creating a dependency tree CRF network (Rush, 2020) using the same BERT weights for both tasks.

3 Materials and methods

In this section, we describe our datasets and experimental setup (Section 3.1), implementation and parameter settings (Section 3.2), and evaluation metrics (Section 3.3).

3.1 Datasets and experimental setup

The SemEval 2022 task 2 subtask A dataset is divided into train, dev, eval, and test sets. We train models on the train set and evaluate on the test set, which was used for the final evaluation in the shared task. The dataset includes instances in three languages: English (en), Portuguese (pt) and Galician (gl). We only consider the "zero-shot" setting from the shared task in which models are evaluated on MWE types that are not seen in the training data. For this setting, the training data consists of English and Portuguese, while the test data includes these languages and also Galician. In this work, we consider further cross-lingual experiments in which a model is evaluated on expressions in a language which was not observed during training. Specifically, we explore models that are trained on one of English or Portuguese. We evaluate on the test dataset, and focus on results for languages that were not observed during training (e.g., when training on English, we focus on results for Portuguese and Galician). The train data consists of 3327 English instances and 1164 Portuguese instances. The test data consists of 916, 713, and 713 English, Portuguese and Galician instances, respectively.

For PARSEME 1.2, the shared task dataset contains sentences with token-level annotations for verbal MWEs (VMWEs) in fourteen languages. (The set of languages is shown in Table 2.) The data for each language is divided into train, dev, and test sets. The average number of sentences in the train and test sets, over all languages, is roughly $12.5k$ and $6k$, respectively. In the initial shared task, experiments were conducted in a monolingual setting, i.e., models were trained on the train set for a particular language, and then tested on the test set for that same language. In this work, we consider further multilingual and cross-lingual settings. In the first setting, referred to as "all", we train a multilingual model on the concatenation of the training data for all languages, and then test on each language. In the second setting, referred to as "heldout", for each language, a model is trained on training data from all other languages, and is then tested on that language that was held out during training.

3.2 Implementation and parameter settings

We use Huggingface (Wolf et al., 2020) implementations of mBERT, XLM-RoBERTa and mDeBERTa. Specifically, we use the bert-base-multilingual-cased, xlm-roberta-base and mdeberta-v3-base implementations. mBERT is pre-trained on the 104 languages with the largest Wikipedias. XLM-RoBERTa and mDeBERTa are pre-trained on 2.5TB of CommonCrawl data covering 100 languages. We use mBERT, XLM-RoBERTa, and mDeBERTa for the SemEval task and mBERT for the PARSEME task.

For the SemEval task, for testing, since the gold standard for the test data was not publicly available when we conducted our experiments, we uploaded our models' predictions to the competition website to obtain results over the test data.

For the MTLB-STRUCT system for the PARSEME task, we use the "multi-task" setting, where the loss of the model is back-propagated based on learning of MWE and dependency parse tags (Taslimipoor et al., 2019). For both the multilingual and cross-lingual settings (described in Section 3.1), we use the default parameter settings of MTLB-STRUCT, where the number of epochs

Model	Train	Test			
		en	pt	gl	ALL
mBERT	en	0.717	0.583	0.420	0.587
	pt	0.355	0.578	0.478	0.482
	en+pt	0.700	0.662	0.550	0.665
RoBERTa	en	0.697	0.590	0.390	0.571
	pt	0.555	0.553	0.440	0.531
	en+pt	0.706	0.668	0.526	0.651
mDeBERTa	en	0.700	0.523	0.304	0.526
	pt	0.582	0.567	0.499	0.556
	en+pt	0.720	0.644	0.495	0.635
Baseline		0.345	0.391	0.434	0.389

Table 1: Macro F1 score for each model, training and testing on the indicated language(s). Results for a most-frequent class baseline are also shown.

is 10 and the batch size is 3×10^{-5}.

3.3 Evaluation metrics

For the SemEval task, the classes are imbalanced. We follow the shared task and evaluate using macro F1 score.

For the PARSEME task, we also use the shared task evaluation metrics: global token-based F1 score, global MWE-based F1 score, and unseen MWE-based F1 score. The global token-based evaluation measures the precision and recall of the predicted VMWE boundaries. The global MWE-based evaluation measures the precision and recall of complete VMWEs, including their type (e.g., LVC, VPC). The unseen MWE-based evaluation considers only VMWEs that are not observed in the training (or development) data. Note that in the case of cross-lingual experiments in the heldout setting, in which systems are evaluated on expressions in a language that was not observed during training, all test expressions are unseen during training.

For both tasks we compare against a most-frequent class baseline. For the PARSEME task, for each language, we label each token as the most-frequent class of VMWE observed in the training data for that language. Although this most-frequent class baseline performs relatively poorly for the PARSEME task, it provides a point of comparison to determine whether cross-lingual models capture information about idiomaticity.

4 Results

Here we present results on the SemEval (Section 4.1) and then PARSEME (Section 4.2) tasks.

4.1 SemEval

Results are shown in Table 1. We focus on cross-lingual settings, i.e., when the model is tested on a different language than it is trained on.

When testing on English, and training on Portuguese, each model improves over the most-frequent class baseline, although the difference is quite small for mBERT. When testing on Portuguese, and training on English, the findings are similar in that all models again improve over the baseline. It is also interesting to note that for mBERT and RoBERTa, results for training on English and testing on Portuguese are in fact higher than for training and testing on Portuguese. This somewhat counter-intuitive finding could be due to the larger number of training instances for English compared to Portuguese (Section 3.1). When testing on Galician, results for models trained on English do not improve over the baseline. Models trained on Portuguese perform better than those trained on English, and show small improvements over the baseline. Despite differences in training data size for English and Portuguese, models trained on Portuguese could perform better on Galician than those trained on English because Portuguese and Galician are both Romance languages. Training on the concatenation of the English and Portuguese training data gives the best results on Galician, and improves over the results for models trained on only Portuguese for mBERT and RoBERTa. This finding suggests that models for predicting idiomaticity can be improved with additional training data from other languages.

Overall, these findings indicate that the models are able to learn information about idiomaticity that is not language-specific. These findings are in line with those of Fakharian and Cook (2021).

4.2 PARSEME

Results on the PARSEME task are shown in Table 2. The monolingual approach ("Mono" in Table 2) is our reproduction of the MTLB-STRUCT system on the shared task. In this setting, a monolingual model is trained and tested on each language. In the "all" setting, a model is trained on the concatenation of the training data for all languages. For "heldout", for a given target language, a model is trained on all other languages, and then evaluated on the target language, which was held out during training. When calculating the unseen MWE-based F1 score ("Unseen" in Table 2), for each setting,

Language	Setting	MWE	Token	Unseen
DE	Mono	0.699	0.734	0.398
	All	0.729	0.738	0.434
	Heldout	0.269	0.423	0.207
EL	Mono	0.732	0.776	0.420
	All	0.743	0.776	0.423
	Heldout	0.407	0.415	0.147
EU	Mono	0.804	0.832	0.346
	All	0.815	0.839	0.380
	Heldout	0.194	0.258	0.112
FR	Mono	0.802	0.830	0.431
	All	0.797	0.825	0.437
	Heldout	0.501	0.560	0.196
GA	Mono	0.311	0.465	0.210
	All	0.422	0.483	0.301
	Heldout	0.111	0.133	0.069
HE	Mono	0.482	0.527	0.215
	All	0.491	0.536	0.219
	Heldout	0.141	0.146	0.064
HI	Mono	0.729	0.785	0.504
	All	0.759	0.796	0.549
	Heldout	0.376	0.452	0.278
IT	Mono	0.632	0.673	0.227
	All	0.618	0.656	0.200
	Heldout	0.376	0.437	0.160
PL	Mono	0.815	0.826	0.400
	All	0.808	0.815	0.380
	Heldout	0.361	0.382	0.144
PT	Mono	0.736	0.758	0.358
	All	0.807	0.821	0.397
	Heldout	0.486	0.500	0.183
RO	Mono	0.903	0.908	0.299
	All	0.898	0.900	0.275
	Heldout	0.481	0.502	0.092
SV	Mono	0.721	0.731	0.425
	All	0.769	0.751	0.467
	Heldout	0.303	0.413	0.215
TR	Mono	0.701‘	0.716	0.430
	All	0.708	0.718	0.457
	Heldout	0.394	0.416	0.189
ZH	Mono	0.696	0.725	0.605
	All	0.705	0.732	0.618
	Heldout	0.121	0.188	0.148
Average	Mono	0.699	0.738	0.380
	All	0.722	0.746	0.400
	Heldout	0.331	0.381	0.169
	Baseline	0.002	0.067	0.001

Table 2: MWE-based, token-based, and unseen F1 score for the monolingual (mono), “all”, and “heldout”, experimental settings, for each language.

we report results over the instances that are unseen based on the monolingual training and development data. This enables comparisons between settings for this evaluation metric. However, in the heldout setting, all test instances are in fact unseen during training.

For each of the three evaluation metrics, we see that the average F1 score for the all setting is higher than that for the monolingual setting. This indicates that information from other languages can be leveraged to give improvements over a monolingual approach. This is inline with the findings on the SemEval task from Section 4.1. We also see that, for all languages, and all evaluation metrics, the F1 score for the heldout setting is less than that for the monolingual setting. This is perhaps unsurprising; a model that has access to language-specific training data is able to outperform one that does not. However, the results in the heldout setting are higher than the baseline on average (Table 2) and for each language (results not shown). This indicates that models are able to learn information about MWEs that is not language specific. This is again inline with the findings on the SemEval task from Section 4.1 and the findings of Fakharian and Cook (2021).

Category	Mono	All	Heldout
IAV	0.4929	0.5408	0.0000
IRV	0.6945	0.7188	0.3135
LS.ICV	0.0000	0.0000	0.0000
LVC.cause	0.3965	0.4429	0.0994
LVC.full	0.6392	0.6661	0.3495
MVC	0.4707	0.4853	0.0000
VID	0.5147	0.5335	0.2320
VPC.full	0.5799	0.5825	0.0565
VPC.semi	0.4363	0.4712	0.0052

Table 3: Per-category MWE-based F1 score across languages which have instances of these categories.

In an effort to better understand the performance in the heldout setting and the knowledge about idiomaticity that is learned, we report results for each category of VMWE in Table 3. The best results for the heldout setting are for (full) light-verb constructions (LVC.full), inherently-reflexive verbs (IRV), and verbal idioms (VID). Although not all languages have instances of all of these categories, they are by far the most frequent categories of VMWEs in the PARSEME 1.2 data (Ramisch et al., 2020), which could be why the model performs relatively well on these categories in the heldout setting.

5 Conclusions

In this paper, we considered new cross-lingual settings for the SemEval 2022 task 2 subtask A and PARSEME 1.2 shared tasks, in which models are evaluated on languages that are not seen during training. Our findings indicate that language models are able to learn information about MWEs and idiomaticity that is not language-specific. Our findings further show that additional training data from other languages can be leveraged to give improvements over monolingual models for identifying MWEs and predicting idiomaticity.

In future work, we intend to further explore the influence of language families and categories of multiword expressions on the ability of idiomaticity prediction and MWE identification models to generalize to unseen languages. We further plan to explore the ability of these models to generalize to languages that were unseen during language model pre-training (Muller et al., 2021).

Acknowledgements

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada and the University of New Brunswick.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition. CRC Press, Boca Raton, USA.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

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

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, Vancouver, Canada. Association for Computational Linguistics.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2018. The other side of the coin: Unsupervised disambiguation of potentially idiomatic expressions by contrasting senses. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 178–184, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu 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 *Proceedings of* *the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Alexander Rush. 2020. Torch-struct: Deep structured prediction library. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 335–342, Online. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Shiva Taslimipoor, Omid Rohanian, and Le An Ha. 2019. Cross-lingual transfer learning and multitask learning for capturing multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 155–161, Florence, Italy. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.