White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/190380/

Version: Published Version

Proceedings Paper:
Phelps, D., Fan, X.-R., Gow-Smith, E. et al. (2022) Sample efficient approaches for idiomaticity detection. In: Bhatia, A., Cook, P., Taslimipoor, S., Garcia, M. and Ramisch, C., (eds.) Proceedings of The 18th Workshop on Multiword Expressions @LREC2022. The 18th Workshop on Multiword Expressions @LREC2022, 20-25 Jun 2022, Marseille, France. European Language Resources Association (ELRA), pp. 105-111. ISBN: 9791095546900.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Sample Efficient Approaches for Idiomaticity Detection

Dylan Phelps, Xuan-Rui Fan, Edward Gow-Smith,
Harish Tayyar Madabushi, Carolina Scarton, Aline Villavicencio

Department of Computer Science,
University of Sheffield
United Kingdom
{drsphelps1, lhsu1, egow-smith1, h.tayyarmadabushi, c.scarton, a.villavicencio}
@sheffield.ac.uk

Abstract

Deep neural models, in particular Transformer-based pre-trained language models, require a significant amount of data to train. This need for data tends to lead to problems when dealing with idiomatic multiword expressions (MWEs), which are inherently less frequent in natural text. As such, this work explores *sample efficient* methods of idiomaticity detection. In particular we study the impact of Pattern Exploit Training (PET), a few-shot method of classification, and BERTRAM, an efficient method of creating contextual embeddings, on the task of idiomaticity detection. In addition, to further explore generalisability, we focus on the identification of MWEs not present in the training data. Our experiments show that while these methods improve performance on English, they are much less effective on Portuguese and Galician, leading to an overall performance about on par with vanilla mBERT. Regardless, we believe sample efficient methods for both identifying and representing potentially idiomatic MWEs are very encouraging and hold significant potential for future exploration.

Keywords: Idiomaticity Detection, Sample Efficient MWE Detection, Pre-Trained Language Models

1. Introduction and Motivation

The handling of idiomaticity is an important part of natural language processing, due to the ubiquity of idiomatic multiword expressions (MWEs) in natural language (Sag et al., 2002). As such, it is an area where the performance of state-of-the-art Transformer-based models has been investigated (Yu and Ettinger, 2020; Garcia et al., 2021b; Nandakumar et al., 2019), with the general finding being that, through pre-training alone, these models have limited abilities at handling idiomaticity. However, these models are extremely effective at transfer learning through fine-tuning, and thus are able to perform much better on supervised idiomatic tasks (Fakharian and Cook, 2021; Kurfalı and Östling, 2020), where significant amounts of labelled data is provided.

Unfortunately, individual MWEs tend to occur infrequently in natural text, making it harder to train models to capture the idiomatic meaning due to the lack of available training data. As such it is important to be able to find methods of identifying potentially idiomatic MWEs using relatively less data. To address this question we focus on *sample efficient* methods for the task, taking two perspectives. The first is an evaluation of a few-shot method on the task of zero-shot idiomaticity detection. In particular we evaluate Pattern Exploit Training (PET) (Schick and Schütze, 2021a), which has been shown to be an effective few-shot method on other tasks (Schick and Schütze, 2021b). The second is an evaluation of the effectiveness of better representations of MWEs, created using a sample efficient strategy, namely BERTRAM (Schick and Schütze, 2020). Both of these are explored in the zero-shot context, where training data does not include MWEs present in the test data. So as to ensure reproducibility and to enable others to build upon this work, we make the programme code and models publicly available[1].

1.1. Research Questions and Contributions

Given the need for sample efficient methods when dealing with idiomaticity, this work is aimed at exploring the following questions:

- How effective are few-show methods on the task of zero-shot idiomaticity detection? In particular we evaluate Pattern Exploit Training (PET) (Schick and Schütze, 2021a), which has been shown to be an effective few-shot method on other tasks (Schick and Schütze, 2021b).
- Given that prior work has shown pre-trained language models do not adequately capture multiword expressions, in particular those which are idiomatic, how effective is improving their representations on the task of detecting idiomaticity? In particular, we use BERTRAM (Schick and Schütze, 2020) as a sample efficient strategy for creating representations of MWEs.

From our experiments, we find that both BERTRAM and PET are able to outperform mBERT (Devlin et al., 2019) significantly on the English portion of the test

[1] https://github.com/drsphelps/idiom-bertram-pet

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 105–111
Marseille, 25 June 2022. © European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

data, which is a promising result. However, both of these models perform worse overall due to their significantly lower performance on Portuguese. We explore potential reasons for this poor performance on non-English languages: for PET our patterns are all in English and a multilingual model is used instead of a language specific one. However, an error analysis (Section 5.1) suggests that these are not the reasons for the lower performance on non-English languages. In BERTRAM, however, a monolingual model is used for each language which might have contributed to the drop in performance. We believe that these results point to the need for further exploration in languages other than English.

Additionally, our exploration using BERTRAM is, to the best of our knowledge, the first work to explore the relation between the representation and detection of idiomaticity.

The rest of this paper is structured as follows. We begin in Section 2 by presenting a quick overview of work related to MWE identification, before presenting more details of the methods we make use of in this work. We then provide an overview of the data and task we use for our evaluation in Section 3, before presenting the methods in Section 4. We then present our results and a discussion of what these results imply in Section 5, before concluding in Section 6.

2. Related Work

Despite idiomaticity detection being a problem that has been widely explored (Constant et al., 2017), the impact of better MWE representations, especially within contextualised models, has not been well studied. To this end we use BERT for Attentive Mimicking (BERTRAM) (Schick and Schütze, 2020), which has been shown to perform well on idiom representation tasks (Phelps, 2022), to evaluate the effect idiom representations have on detection. Additionally, we apply a few-shot learning technique Pattern Exploit Training (PET) (Schick and Schütze, 2021a), to assess whether the relatively new paradigm of few-shot learning can be applied to this task successfully.

2.1. PET

PET (Schick and Schütze, 2021a; Schick and Schütze, 2021b) is a semi-supervised training method that improves performance in few-shot settings by integrating task descriptions into examples.

A Pattern is used to map each example into a cloze-style question with masked out tokens, for example *'X. It was [MASK]'*, where X is the input example, could be used for a sentiment classification task. A Verbaliser maps the task classes into outputs from the masked language model (MLM), for example positive/negative labels map to the words 'good'/'bad' in the MLM's vocabulary (label tokens), and is combined with the pattern to form a Pattern Verbaliser Pair (PVP). The probability of each class is then calculated using softmax over the logits for each label token.

For each PVP, an MLM can be fine-tuned on the small amount of labelled data. Knowledge is distilled from multiple PVPs by combining the predictions on the unlabelled data and using it as a larger labelled dataset to train another classifier. This allows for multiple patterns and verbalisers to be used without having to choose the best performer for each task, which may also change depending on the data split.

2.1.1. iPET

iPET (Schick and Schütze, 2021a) is a variation where each PVP's model is trained iteratively using a gradually increasing training set made up of labelled examples from another model's predictions in the previous iteration. Despite using the same PVPs and MLMs, iPET has been shown to improve the performance on a number of tasks (Schick and Schütze, 2021b).

2.2. BERTRAM

BERTRAM (Schick and Schütze, 2020) is a model for creating embeddings for new tokens within an existing embedding space, from a small number of contexts. To create an embedding for a token with a number contexts, a form embedding is first created using embeddings trained for each of the n-grams in the token. This form embedding is then passed as an input, alongside the embeddings for words in the context, into a BERT model. An attention layer is then applied over the contextualised embedding output from BERT for each context to create the final embedding for the token.

The model is trained using embeddings for common words as 'gold standard' embeddings, with the distance from the embedding created by the model and the 'gold standard' embedding being used as the loss function.

3. Dataset and Task Description

In evaluating the models presented in this work we use the Task 2 of SemEval 2022: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022). This task aims at stimulating the development and evaluation of improved methods for handling potentially idiomatic MWEs in natural language. While there exist datasets for evaluating models' ability to identify idiomaticity (Haagsma et al., 2020; Korkontzelos et al., 2013; Cook et al., 2008; Cordeiro et al., 2019; Garcia et al., 2021b; Shwartz and Dagan, 2019), these are often not particularly suited to investigating a) the transfer learning capabilities across different data set-ups b) the performance of pre-trained contextualised models.

The task consists of two subtasks: Subtask A, which is focused on the detection of idiomaticity, and Subtask B, which is focused on the representation of idiomaticity. In this work we are interested in the task of idiomaticity classification, since we wish to investigate how our models can identify idiomaticity in text without having to generate semantic similarity scores. As such, we restrict our attention to Subtask A. We also want to see how our models perform when MWEs

Pattern Number	Pattern	Literal Token	Idiom Token
P1	X: ____	literal	phrase
P2	(____) X	literal	phrase
P3	X. [IDIOM] is ____ literal.	actually	not
P4	X. ____, [IDIOM] is literal.	yes	no
P5	X. [IDIOM] is ____ $[IDIOM]_2$	actually	not

Table 1: Pattern Verbaliser Pairs used in the task. X represents the example sentence, [IDIOM] is the idiom found in the example, and $[IDIOM]_2$ represents the n^{th} component word of the idiom

in the test data are disjoint from those in the training data, as we argue this means the models cannot so easily leverage statistical information garnered from the training data, but must instead have some 'knowledge' of idiomaticity in general. As such, we also restrict our attention to the zero-shot setting of the SemEval task. The dataset consists of three languages: English, Portuguese and Galician. In the training data there are 3,327 entries in English, and 1,164 entries in Portuguese. There is no Galician training (or development) data in the zero-shot setting, to test the ability of models at cross-lingual transfer. In the test set, there are 916 English, 713 Portuguese, and 713 Galician examples, and macro F1 score is used as an evaluation metric.

It should be noted that the dataset provided by Tayyar Madabushi et al. (2022) consists of four data splits: The training set, two development sets and the test set. Of the two development sets, the first - called the 'dev' split - includes gold labels and the second - called the 'eval' split - does not include gold labels but requires submission to the competition website. We report our results on the 'eval' set to maintain consistency with the SemEval task.

4. Methods

In this section we detail our use of PET, iPET and BERTRAM for the task of idiomaticity detection.

4.1. PET and iPET

During our experiments with PET and it's variants, we define and test 5 Pattern Verbaliser Pairs, shown in table 2. P1 and P2 are generic prompts which do not give the model much more information about the example, whereas P3, P4, and P5 include the whole idiom within the prompt. We hypothesise that this will allow the model to understand which part of the example it should be focusing on. Each of the patterns we define is in English, even when the example sentence and idiom are in Portuguese or Galician — we will investigate the effect that this has on the final performance across the languages, as we hypothesise this may not have an impact given our use of a multilingual model.

For each PVP, we train a classification model using mBERT as the MLM. Furthermore, we train a standard PET model using all of the patterns. An iPET model is also trained, however to evaluate how using only generic prompts affects the results, we only train our iPET model using PVPs P1 and P2, for 2 iterations. Each of the model setups is trained 3 times using different random seeds, and the final distilled model is then used to produce the presented results.

Additionally, we investigate how the number of labelled examples affects the achieved performance for each of the model setups discussed. We train the models using 10, 100, and 1000 labelled examples separately, with the examples chosen randomly across English and Portuguese, but with the split of idiomatic and literal uses being kept at 50/50. The PET and iPET models then have access to 3,000 unlabelled examples to use within their training tasks.

We evaluate each model setup and labelled example set size combination on the *eval* set, before choosing the best-performing combination for each PET variant to evaluate on the test set. The results from the *eval* set can be seen in Table 2. Here we see that PET-all trained on 1000 labelled examples performs best overall, beating the individual pattern models, a result also seen in the original paper (Schick and Schütze, 2021a). The lack of example specific prompts causes iPET to perform poorly when compared to the individual task specific patterns, and when compared to the best PET-all model. The highest scoring PET model (PET-all) and our iPET model are evaluated on the test dataset in Section 5.

4.2. BERTRAM

To evaluate the effect that improved idiom representations have on this idiom detection task, we use the same BERTRAM setup as presented in Phelps (2022), that was shown to give greatly improved performance over the baseline system for Subtask B, the task of representing idiomaticity. We use the same BERTRAM models: the English model presented in the original BERTRAM paper (Schick and Schütze, 2020), and the Portuguese and Galician models that were trained for Subtask B from data in the CC100 corpus. Unlike the English BERTRAM model, Phelps (2022) does not use one token approximation when training the Portuguese and Galician models. Embeddings for each of the idioms in the task datasets were generated with the appropriate BERTRAM model using 150 examples scraped from the CC100 dataset. 150 examples were chosen as this was shown to have the highest performance on Subtask B. It should be noted that the BERTRAM models were used to create representations of MWEs in the test set. While this does not require labelled data asso-

Model	EN	PT	Overall
mBERT (Tayyar Madabushi et al., 2021)	0.7420	0.5519	0.6871
PET-all (10 labelled)	0.4365	0.2901	0.4267
PET-all (100 labelled)	0.5908	**0.5718**	0.5888
PET-all (1000 labelled)	**0.7820**	0.5619	**0.7164**
PET-P1 (1000 labelled)	0.6386	0.5507	0.6278
PET-P2 (1000 labelled)	0.6905	0.5495	0.6607
PET-P3 (1000 labelled)	0.7493	0.5474	0.6981
PET-P4 (1000 labelled)	0.7441	0.5315	0.6860
PET-P5 (1000 labelled)	0.7551	0.5680	0.7032
iPET (1000 labelled) [P1 & P2]	0.6701	0.5648	0.6522

Table 2: The F1 Score (Macro) on the *eval* set, broken down into each language, for each of the models. Highest score for each language (or overall) shown in bold.

Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446
iPET (1000 labelled) [P1 & P2]	0.6604	0.5676	0.4735	0.5879

Table 3: The F1 Score (Macro) on the *test* set, broken down into each language, for each of the models. Highest score for each language (or overall) shown in bold.

ciated with MWEs (thus remaining a zero-shot task), it does require knowledge of which phrases need to have explicit representations created.

As we have separate BERTRAM models for each language that are trained to mimic embeddings from single language BERT models, we split the system and data into English, Portuguese and Galician. The English model uses BERT base (Devlin et al., 2019), and is trained on the 3,327 English training examples found in the training set. The Portuguese model uses BERTimbau (Souza et al., 2020), and Galician uses BERTinho (Vilares et al., 2021), and as there is no Galician training data available, both are trained on the 1,164 Portuguese examples. Each model has the MWEs from the relevant language added to its embedding matrix.

5. Results and Discussion

Table 3 presents the results of our best PET-based models alongside our BERTRAM-based model on the test set, as well as the mBERT system presented in (Tayyar Madabushi et al., 2022), for comparison. For each model we present the F1 macro score on the test set for each language, as well as the overall F1 macro score.

An increase in performance over mBERT by our BERTRAM model is seen for the English split, with the score on the Galician split not seeing a significant change. The overall score for BERTRAM is brought down by a much lower score on the Portuguese data, however, meaning no overall increase in performance is seen. A similar picture is seen for the PET-all (1000 examples) model, with a higher F1 score in both English and Galician, and a lower score in Portuguese, leading to an overall lower F1 score across the entire test dataset. As found on the example data, the iPET model which was only trained on the non-example specific prompts (P1 and P2) performs very poorly.

The significant boost from using BERTRAM on English seems to indicate that the improved representations also lead to better classification, despite the lacklustre performance on Galician and Portuguese. We believe that this drop in performance is either because one-token approximation was not used in creating the non-English BERTRAM models, or because mBERT, trained on all three languages simultaneously, is trained on more data than each of our monolingual models. This lack of training data does not affect our English model as there is a more training data in English than in Portuguese and none at all in Galician. We perform a language specific error analysis to explore the causes of this drop in performance (Section 5.1).

It is interesting to note that pre-trained language models can identify idiomaticity in a zero-shot and sample efficient context *even when prior work has shown that they do not encode idiomaticity very well* (Garcia et al., 2021a). We believe that this implies that, while these models do not encode idiomaticity, they encode enough related information to be able to *infer* idiomaticity from relatively little data.

Unsurprisingly, 'highlighting' the phrase that is potentially idiomatic by adding the phrase to the pattern, as in patterns P3, P4 and P5 (see Table 2), significantly improves a model's ability to identify idiomatic-

Language	Pattern	Literal Token	Idiom Token
EN	X. ____, [IDIOM] is literal.	yes	no
PT	X. ____, [IDIOM] é literal.	sim	não
GL	X. ____, [IDIOM] é literal.	si	non

Table 4: The translations of P4 into Portuguese and Galician

Model	Prompt Language	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	N/A	0.7070	0.6803	0.5065	0.6540
PET-P4 (1000 labelled)	EN	0.7161	0.6373	0.5365	0.6581
PET-P4 (1000 labelled)	PT	0.6994	0.6260	0.4964	0.6283
PET-P4 (1000 labelled)	GL	0.7040	0.5997	0.5154	0.6279

Table 5: The F1 Score (Macro) on the *test* set, broken down into each language, for PET using prompts in each of the task languages.

ity, which is consistent with results presented by Tayyar Madabushi et al. (2021).

Research Questions The results presented herein suggest that few-shot learning methods are indeed effective on the task of idiomaticity detection despite the lower accuracy on Portuguese and Galician. Similarly, our results support the conclusion that improved MWE representations does have an impact on improved detection.

5.1. Error Analysis

The effectiveness of PET on the English split of the task suggests that pre-trained language models can effectively identify idiomatic MWEs in a sample efficient manner. However, the overall drop in performance on the task can be attributed to lower performance on non-English languages when compared to the results achieved by Tayyar Madabushi et al. (2021).

One possibility for the decrease in performance is the use of English prompts across all the languages. This leads to the inputs for English examples being monolingual and the inputs for non-English examples to be multilingual, which may cause confusion in the output logits for the verbalizer tokens from which PET draws it's predictions.

To investigate this further we translate one of our patterns, P4, into both Portuguese and Galician and evaluate the performance on the entire *test* split. P4 was chosen as it was one of the better performing patterns for English in our initial experiments (Table 2), and was easily translated into the two languages. The translations can be seen in table 4.

As shown in table 5, the use of Portuguese and Galician prompts does not increase the performance in the respective language. For Portuguese the model with Portuguese prompts achieves 0.6260 F1 score compare to 0.6373 for that with English prompts. Galician shows similar results, with 0.5154 F1 score for the model with prompts in Galician and 0.5365 for that in English.

Additionally, we use multilingual BERT which was trained on a lot more English training data than Portuguese or Galician language. To investigate the impact of this on our results, we extract only the Portuguese section of the training and test data and compare the performance of multilingual BERT with Portuguese BERT (Souza et al., 2020). Surprisingly, we find that the there isn't a significant difference between the performance of multilingual BERT and Portuguese BERT, with overall F1 (macro) scores of 0.4541 and 0.4621, respectively.

6. Conclusions and Future work

This work presented our exploration of *sample efficient* methods for idiomaticity detection, crucial given the infrequent occurrence of specific MWEs in natural language text. Our experiments show that these methods are extremely promising and have great potential.

In future work, we intend to raucously evaluate and find solutions to the problem of lower performance on non-English test splits. We also intend to explore other variations of BERTRAM (e.g. one-token approximation) in bridging the performance gap between English and the other languages.

As noted earlier, we show that pre-trained language models can identify idiomaticity in a zero-shot and sample efficient context *even when prior work has shown that they do not encode idiomaticity very well*. As such, an important avenue of future exploration is the generalisation of these methods to develop models capable of identifying *the notion of idiomaticity*, much like humans are able to grasp that certain phrases are clearly non-compositional.

Acknowledgements

This work is partially supported the Healthy Lifespan Institute (HELSI) at The University of Sheffield and is funded by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/T517835/1]. This work was also partially supported by the UK EPSRC grant EP/T02450X/1 and the CDT in Speech and Language Technologies and their Applications funded by UKRI [grant number EP/S023062/1].

7. Bibliographical References

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892, December.

Cook, P., Fazly, A., and Stevenson, S. (2008). The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Fakharian, S. and Cook, P. (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.

Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., and Villavicencio, A. (2021a). Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online, April. Association for Computational Linguistics.

Garcia, M., Vieira, T. K., Scarton, C., Idiart, M., and Villavicencio, A. (2021b). Probing for idiomaticity in vector space models. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Haagsma, H., Bos, J., and Nissim, M. (2020). Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 279–287.

Korkontzelos, I., Zesch, T., Zanzotto, F. M., and Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.

Kurfalı, M. and Östling, R. (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, December. Association for Computational Linguistics.

Nandakumar, N., Baldwin, T., and Salehi, B. (2019). How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA, June. Association for Computational Linguistics.

Phelps, D. (2022). drsphelps at semeval-2022 task 2: Learning idiom representations using bertram.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Schick, T. and Schütze, H. (2020). BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online, July. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021a). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021b). It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online, June. Association for Computational Linguistics.

Shwartz, V. and Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Tayyar Madabushi, H., Gow-Smith, E., Scarton, C., and Villavicencio, A. (2021). AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Tayyar Madabushi, H., Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M., and Villavicencio, A. (2022). SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Vilares, D., García, M., and Gómez-Rodríguez, C. (2021). Bertinho: Galician BERT representations. *CoRR*, abs/2103.13799.

Yu, L. and Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. *arXiv preprint arXiv:2010.03763*.