# CardiffNLP-Metaphor at SemEval-2022 Task 2: Targeted Fine-tuning of Transformer-based Language Models for Idiomaticity Detection

**Joanne Boisson** and **Luis Espinosa Anke** and **Jose Camacho Collados**
School of Computer Science and Informatics
Cardiff University
`BoissonJC@cardiff.ac.uk`

## Abstract

This paper describes the experiments ran for SemEval-2022 Task 2, subtask A, zero-shot and one-shot settings for idiomaticity detection. Our main approach is based on fine-tuning transformer-based language models as a baseline to perform binary classification. Our system, CardiffNLP-Metaphor, ranked 8th and 7th (respectively on zero- and one-shot settings on this task. Our main contribution lies in the extensive evaluation of transformer-based language models and various configurations, showing, among others, the potential of large multilingual models over base monolingual models. Moreover, we analyse the impact of various input parameters, which offer interesting insights on how language models work in practice.

## 1 Introduction

Idiomatic language identification is an important task for language understanding. Recent language models are surprisingly accurate at distinguishing literal and figurative use of language, but very little work has been done on measuring their ability to generalize across languages. Even for mainstream languages such as English, there is still little understanding on the way language models process idiomatic expressions (IE's). This SemEval task (Tayyar Madabushi et al., 2022) focuses, in particular, on multi-word expressions (MWE), adding the challenge of representing such expressions in models.

The Subtask A of SemEval Task 2 invited participants to extend the range of existing experiments for multilingual idiomatic language detection. Data were provided in English, Portuguese, and Galician. The task is framed as a binary classification of MWE between an idiomatic and a literal usage. In the zero-shot setting, no Galician example is provided in the training and development set. The MWE of the development, evaluation and test sets are unseen in the training set. In the one-shot setting, exactly one example of the MWE encountered respectively in the practicing and test phase is added to the training data. Therefore, these settings provide a challenging framework in which language models have to learn from few or no examples.

As part of the CardiffNLP-Metaphor team, we used a simple strategy similar to the method employed in the original paper releasing the dataset Tayyar Madabushi et al. (2021). In particular, we assessed the performance of monolingual and multilingual language models on the task. To this end, we compared the performance of these models using different input formats and training parameters. The best results are obtained with a XLM-RoBERTa large (Lample and Conneau, 2019) with 7 epochs, 8 instances per batch, a maximum sequence length of 350, the longest three-sentence context, and including target information (i.e., the embedding and the position of the target in the sentence). Our submitted model was based on the best performance in the development set across both tasks, using a wide range of different inputs and parameters.

Our system ranked 8th with a best f1-macro score of 0.7378 for the zero-shot competition and 7th with a score of 0.8934 for the one-shot competition.[1] The main contributions of this paper are the following:

- We show that the multilingual large RoBERTa model (Liu et al., 2019) performs better than monolingual and base models on the one-shot track, which differs from what was found in the original paper (Tayyar Madabushi et al., 2021).

- We found that XLM-RoBERTa base and large can be unstable, also in comparison with simi-

[1] The script written for our experiments is available in a GitHub repository: `https://github.com/Mionies/CardiffNLP-SemEval-2022-Task2`

*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 169 - 177
July 14-15, 2022 ©2022 Association for Computational Linguistics

| Data | Total | EN | PT | GL | %Id. |
|---|---|---|---|---|---|
| tr. 0-shot | 4491 | 3327 | 1164 | 0 | 56 |
| tr. 1-shot[2] | 140 | 87 | 53 | 0 | 43 |
| dev. | 739 | 466 | 273 | 0 | 45 |
| eval. | 762 | 483 | 279 | 0 | ? |
| test | 2342 | 916 | 713 | 713 | ? |

Table 1: Dataset description. Number of example for each language and percentage of idiomatic MWE expressions. The labels of the evaluation and test sets are unknown.

lar models. This could also explain the difference in our conclusion and that of Tayyar Madabushi et al. (2021) after exploratory runs with large models.

- We confirm the importance of providing the embeddings of the MWE separately to the model, and running a relatively large amount of epochs (up to 9 leads to improvements). Our best model is obtained with seven epochs.
- We test various input formats, including maximum sequence length and context length, and the impact of shuffling the training data on the results, allowing us to discuss the results obtained in previous experiments with this dataset.

## 2 Related Work

In this task, idiomatic expressions are either frozen (well-known) metaphors, or frozen noun compounds involved in longer metaphors. This dataset relates to other datasets labelled for metaphorical usage of words such as the VU Amsterdam corpus (VUAC) (Steen, 2010) used in a SemEval 2020 task (Leong et al., 2020). However, such datasets are not restricted to idioms or compounds. All the words occurring in texts are labeled. This could ultimately lead to a design of NLP tasks focusing on idioms, but has in the main been used for the predictions of metaphors at the word level. Other metaphor datasets built for NLP such as the LCC corpus (Mohler et al., 2016) may contain some MWE but are not focusing on the specific issues posed by idioms, and also include creative metaphors in their scope.

To the best of our knowledge, there are other five datasets particularly designed for the study of the compositionality of MWE in context in English. The idioms in context (IDIX) corpus (Sporleder et al., 2010) includes idiomatic constructions with non consecutive words (e.g. *raise* one's *eyebrows*) and the phrasal verb corpus (Tu and Roth, 2012) is restricted to *V+PRP* constructions. The SemEval 2013 Task 5b on phrasal semantics is very similar to the task addressed this year, with a division between *known phrases* and *unknown phrases* settings within the binary classification task, but restricted to English. More recently, the MAGPIE corpus (Haagsma et al., 2020), a large repository of 56,622 sentences containing potential idiomatic expressions has been shared with the NLP community. The selection of its initial list of idioms differ from our dataset: after a semi automatic selection of idiomatic expressions, a crowdsourced annotation approach is adopted to determine whether the expression is used metaphorically or literally. In a similar design than the dataset used for Subtask B, Zhou et al. (2021) constructed a curated dataset of sentences pairs: one element containing an idiomatic expression and the second element being the same sentences with the IEs replaced by its literal paraphrase.

As for its connection with language models, Garcia et al. (2021) compared various language models for probing idiomaticity in vector space models. In this work, we go beyond the capabilities of vector space models and test the capabilities of fine-tuning multilingual language models on the task. The most related work to our analysis is perhaps that done by Zeng and Bhat (2021). They proposed a neural architecture that uses attention flow, designed for the task of detecting whether a sentence has an idiomatic expression and localizing it when it occurs in a figurative sense.

## 3 Data

Our team participated in Subtask A (zero and one-shot tracks) of the SemEval-2022, Task 2 on Idiomaticity Detection (Tayyar Madabushi et al., 2022). The tasks tackles binary classifiation of MWE in three languages, with variable amount and type of data seen in the training set by the model. Table 1 summarizes the distribution of the instances per language and label. The MWE are all noun compounds, sourced from the Noun Coumpound Senses dataset (Cordeiro et al., 2019).The examples consists of excerpts of text of the Web.

As shown in Table 2, literal instances in our task

[2] One-shot addition designed for the development and the evaluation sets

| Example | label | Orig. label |
|---|---|---|
| To avoid a **blood bath**, prison officials ordered the gate to be opened. | 0 | idio. |
| Remind me *to shed a* **crocodile tear** or two over't. | 0 | meta usage |
| **Marketing consultant** Katy Williams saw the potential of social media. | 1 | non-idio. |
| Deborah Loomis is [...] known for [...] Foreplay (1975) and **Blood Bath** (1976). | 1 | prop. n. |

Table 2: Examples of labelled instances with their original four labels in Tayyar Madabushi et al. (2021) and grouping to two labels idiomatic/non-idiomatic for the SemEval binary classifiation Subtask A.

include *non-idiomatic* use of MWE and *proper nouns*. Idiomatic instances gathers *idiomatic* use and *literal use within a longer metaphor*.

The experiments are organized along six splits of the data (c.f. Table 1) : training zero-shot, training one shot for evaluation phase, training one shot for test phase, development, evaluation and test sets. Labels were provided to the participants for the training and development sets. The practice and test phase were run on Codalab.

## 4 System overview and experiments

### 4.1 System configurations

We test two different configurations to address this binary classification task: one multilingual classifier trained on all the training data at once, and one monolingual classifier per language. For the zero-shot setting in the monolingual classification configuration, we do not have any training examples of Galician. Therefore, we replace the Galician model by a multilingual model trained on the English and Portuguese examples.

We use well-known transformer-based language models: English, Portuguese BERT and Multilingual BERT (Devlin et al., 2019), XLM-RoBERTa base and large (Conneau et al., 2020). For the monolingual models of Galician, we use Bertinho Vilares et al. (2021) model[3], trained on Wikipedia. The cased version of the language models is used in all our experiments, following Tayyar Madabushi et al. (2021) and because the target MWE contain proper nouns.

### 4.2 Preprocessing

The data are preprocessed to find all the occurrences of the expressions and record their positions in the three sentences provided for each instance of the datasets.

We search for lower case and upper case occurrences, with words separated by a space or a hyphen. Only in the cases where an exact match cannot be found, we also rely on their lemmata to identify MWEs in plural form. For this, we relied on Stanza[4], which covers the three languages of the experiments including Galician.

We find 80% instances with only one occurrence, and 20% with multiple occurrences in the training and development sets. We considered contexts of one or three sentences (the previous and following sentence in the latter case). The positions of the target are recorded for both contexts length. We then generate two versions of tagged sentences, one where only the first occurrence of the target in the core sentence is marked and one with all the occurrences are marked, using special tokens.

### 4.3 Experiments

All the experiments are done using the Simple Transformers library[5] with a Quadro RTX 8000 GPU. In order to analyse the effect of several variables in the performance, we performed the following experiments on the development set.

**Experiment 1: Shuffling the training set.** We study the variation of the performances for three seeds (1,2,3), after three shuffles of the training set (A, B, C), for different batch sizes (8, 16, 32, 64). Our goal is to distinguish the variations in the performances due to various parameters modifications from the variation induced by the order in which instances are fed into the model during training.

**Experiment 2: Context and input format.** A context limited to the core sentence provided for each example (noted *core-sent* in Table 4) is compared to the concatenation of this core sentence with its previous and following sentence (noted *3-sent*). Different maximum sequence lengths (128, 300, 350, 400, 512) are also tested.

We further test the various ways to encode information about the target and its position in the

[3] Huggingface ID: dvilares/bertinho-gl-base-cased

[4] https://stanfordnlp.github.io/stanza/

[5] Version 0.62.0, https://simpletransformers.ai/

sentence: tagging only the first occurrence of the target in the core sentence (*first*) is compared to tagging all the occurrences of the target within the input text (*multiple*); one option allows the embedding of the target expression to be passed to the model independently from the sentence (*pair*).[6]

When the *tagged* and the *pair* parameters are both set to *False*, the sentence is provided to the model without any indication concerning the target. This baseline configuration is very interesting in order to evaluate the impact of the topic of the text on idiom detection. For example, in the training data, all the occurrences of *blood bath* are idiomatic except for one occurrence of a proper noun (c.f. Table 2). *Blood bath* is more likely to be used idiomatically than literately in many corpora, as long as they are not rare domain-specific archives on vampires relaxing habits. On the contrary, all 21 occurrences of *marketing consultant* are literal.

**Experiment 3: Monolingual and multilingual models.** Monolingual and multilingual language models are compared with the two configurations introduced in Section 4.1. In this experiment, we measure the ability of the multilingual models to transfer knowledge across English and Portuguese with a comparison between two additional training methods, bringing the experiment to a comparison between three configurations:

1. Fine-tuning monolingual BERT models for English and Portuguese, and Galician for the one-shot setting. Data are split by language, three classifiers are trained.

2. Fine-tuning three multilingual models using the same settings as in 1.

3. Fine-tuning one single monolingual model, with all the data in the two languages for the zero-shot track and three languages for the one-shot track

**Experiment 4: Language models size.** Previous initial experiments from Tayyar Madabushi et al. (2021) concluded that large models were not performing better than base models, after a few attempts. We explore further the performance of large models in comparison with base ones under various classifier parameters and for different shuffles of the training set.

| Zero-shot | | | | | |
|---|---|---|---|---|---|
| train set | seed | Batch size | | | |
| | | 8 | 16 | 32 | 64 |
| A | 1 | 0,70 | **0,75** | 0,74 | 0,69 |
| | 2 | 0,70 | 0,73 | 0,73 | 0,38 |
| | 3 | 0,31 | 0,73 | 0,71 | 0,73 |
| B | 1 | 0,31 | 0,73 | 0,74 | 0,71 |
| | 2 | 0,72 | 0,74 | 0,72 | 0,72 |
| | 3 | 0,74 | 0,72 | 0,73 | **0,75** |
| C | 1 | 0,73 | 0,73 | 0,72 | 0,71 |
| | 2 | 0,72 | **0,75** | 0,71 | 0,68 |
| | 3 | 0,74 | 0,74 | 0,68 | 0,71 |
| One-shot | | | | | |
| train set | seed | Batch size | | | |
| | | 8 | 16 | 32 | 64 |
| A | 1 | 0,73 | **0,80** | 0,73 | 0,70 |
| | 2 | 0,75 | 0,74 | 0,73 | 0,70 |
| | 3 | 0,31 | 0,75 | 0,74 | 0,73 |
| B | 1 | 0,31 | 0,75 | 0,76 | 0,73 |
| | 2 | 0,71 | 0,75 | 0,76 | 0,72 |
| | 3 | 0,73 | **0,78** | 0,71 | 0,71 |
| C | 1 | 0,61 | 0,71 | 0,72 | 0,72 |
| | 2 | 0,71 | 0,73 | 0,71 | 0,69 |
| | 3 | 0,70 | **0,76** | 0,75 | 0,71 |

Table 3: Experiment 1. Results of XLM-RoBERTa base with 1 epoch, max-seq-length=128, for 3 data shuffles and 3 random seeds. A context of 1 sentence is used, with multiple occurrences of the target tagged, and the MWE embedding provided separately to the classifier (pair). Displayed scores are F1-macro for the development set, aggregated for both English and Portuguese.

## 5 Results

During the exploratory phase, we tested 111 different parameter configurations, shuffling the data before each run. The twenty best models (sorted according to their performances in the one-shot setting) are shown in Table 8 in the Appendix[7]. These results are used in complement to the following experiments for drawing our conclusions.

**Experiment 1: Shuffling the training set.** With XLM-RoBERTa-base, Table 3 shows that the classifier is very sensitive to the order in which the input data are passed to the model. When the model does not attribute the same label to all instances of the development set, it may vary by 2 points for a given random seed. The model fails to converge for

[6] Table 7 in the Appendix includes more details about the input formats.

[7] The complete results are available in the GitHub repository of this paper `https://github.com/Mionies/CardiffNLP-SemEval-2022-Task2/blob/main/param_optimization_shuffe/data.csv`.

| Input parameters | | | | Zero-shot | | | One-shot | | |
|---|---|---|---|---|---|---|---|---|---|
| Context | Occ. | Tagged | Pair | EN | PT | EN,PT | EN | PT | EN,PT |
| 3-sent | first | True | True | 0.758 | 0.646 | 0.737 | 0.851 | 0.818 | 0.846 |
| 3-sent | multiple | True | True | 0.743 | 0.627 | 0.722 | 0.661 | 0.361 | 0.644 |
| 3-sent | - | False | True | 0.749 | 0.624 | 0.723 | 0.806 | 0.789 | 0.810 |
| core-sent | first | True | True | 0.735 | 0.650 | 0.718 | 0.855 | 0.832 | 0.850 |
| core-sent | multiple | True | True | 0.744 | 0.603 | 0.708 | 0.866 | 0.841 | **0.863** |
| core-sent | - | False | True | **0.769** | 0.564 | 0.724 | 0.826 | **0.853** | 0.841 |
| 3-sent | first | True | False | 0.740 | **0.688** | **0.741** | **0.872** | 0.773 | 0.845 |
| 3-sent | multiple | True | False | 0.281 | 0.361 | 0.313 | 0.788 | 0.686 | 0.768 |
| core-sent | first | True | False | 0.764 | 0.513 | 0.706 | 0.716 | 0.541 | 0.69 |
| core-sent | multiple | True | False | 0.774 | 0.58 | 0.724 | 0.777 | 0.799 | 0.794 |
| Below, the target not indicated to the model : | | | | Zero-shot | | | One-shot | | |
| 3-sent | - | False | False | 0.695 | **0.652** | 0.699 | 0.649 | 0.361 | 0.611 |
| core-sent | - | False | False | **0.753** | 0.588 | **0.711** | **0.688** | **0.579** | **0.667** |

Table 4: Experiment 2. Contextual and input format parameters. This experiment is run with XLM-RoBERTa-base, 3 epochs, a batch size=8, max-seq-length=512, a lr=4e-05, on 3 seeds with training set shuffle A (c.f. Experiment 1). The results obtained with the best seed is displayed. An average over the three seeds was impossible because the model often does not converge. The metric used is F1 macro, computed on the development set.

| Languages | | Zero-shot | | |
|---|---|---|---|---|
| Pre-train | Fine-tune | EN | PT | EN,PT |
| mono | mono | 0.786 | 0.645 | 0.747 |
| multi | mono | **0.793** | 0.664 | **0.764** |
| multi | multi | 0.76 | **0.686** | 0.748 |
| Languages | | One-shot | | |
| Pre-train | Fine-tune | EN | EN | EN,PT |
| mono | mono | **0.897** | **0.873** | **0.892** |
| multi | mono | 0.835 | 0.783 | 0.829 |
| multi | multi | 0.851 | 0.809 | 0.843 |

Table 5: Experiment 3. Mono and multilingual training data configurations for pretrained models and fine-tuning. XLM-RoBERTa base is used. The experiment ran with 4 epochs, a batch size=8, a lr=2e-05 using one seed [3] and training set shuffle A. The metric used is F1 macro, computed on the development set.

some seeds and shuffle combinations. The problem arises more often with a small batch size of 8, but it also fails to converge once with batch sizes as large as 64 in our experiment. The issue does not disappear for a larger number of epochs. In the exploratory phase, we tried a broad range of training hyper-parameters, and encountered this issue for models trained with 6, 7 and 8 epochs, both with XLM-RoBERTa base and XLM-RoBERTa large.

The multilingual BERT language model shows more stability. With the same datasets and parameters as those used in Table 3, it always obtains a f-score >0.70 in the zero-shot track, and >0.72 in the one shot track. BERT and XLM-RoBERTa perform comparably in the zero-shot experiment but XLM-RoBERTa obtains the best performance in the one-shot setting.

**Experiment 2: Context and input format.** The results are presented in Table 4. With the experimental settings chosen, it is difficult to draw any conclusion on which context window (*core-sentence* or *3-sentences*) or tagging scheme (*first* or *multiple*) is better for the task. Both Table 8 in the Appendix and the results obtained by Tayyar Madabushi et al. (2021) suggest that providing the embedding of the target MWE separately to the model (*pair*) improves the performance.

Among the two configurations which input the sentences (*core-sentence* or *three-sentences* contexts) to the model without giving any information about the target, one performs consistently better than random for English examples in the development set zero-shot, with F1-scores of 75.3. It suggests that performances of the model may not mainly be due to the discrimination between compositional and non compositional interaction between the target and the context. The topic of the sentence may also have an important influence, which we did not fully analyze in this work.

**Experiment 3: Monolingual and multilingual models.** The base monolingual and multilingual settings show similar performance, in preliminary experiments (Table 8) and Experiment 3 (Table

| Training parameters | | | | Zero-shot | | | One-shot | | |
|---|---|---|---|---|---|---|---|---|---|
| shuffle | length | batch size | model size | EN | PT | EN/PT | EN | PT | EN/PT |
| A | 350 | 8 | base | 0.8 | 0,677 | 0,77 | 0,877 | 0,867 | 0,879 |
| A | 350 | 8 | large | 0.785 | 0.673 | 0.761 | **0.902** | 0.902 | **0.905** |
| B | 350 | 8 | base | 0.795 | 0.677 | 0.768 | 0.888 | 0.882 | 0.89 |
| B | 350 | 8 | large | 0.776 | 0.698 | 0.762 | 0.892 | **0.903** | 0.9 |
| C | 350 | 8 | base | 0.774 | 0.667 | 0.749 | 0.868 | 0.825 | 0.859 |
| C | 350 | 8 | large | 0.782 | 0.677 | 0.756 | 0.863 | 0.825 | 0.857 |
| A | 350 | 16 | base | 0.794 | 0.673 | 0.764 | 0.868 | 0.885 | 0.879 |
| A | 350 | 16 | large | **0.807** | 0.689 | **0.778** | 0.891 | 0.893 | 0.896 |
| A | 350 | 32 | base | 0.797 | 0.683 | 0.771 | 0.871 | 0.857 | 0.871 |
| A | 350 | 32 | large | 0.775 | **0.723** | 0.768 | 0.89 | 0.882 | 0.891 |
| A | 256 | 8 | base | 0.768 | 0.641 | 0.737 | 0.898 | 0.768 | 0.895 |
| A | 256 | 8 | large | 0.777 | 0.719 | 0.768 | 0.884 | 0.886 | 0.888 |
| A | 128 | 8 | base | 0.787 | 0.675 | 0.761 | 0.866 | 0.897 | 0.882 |
| A | 128 | 8 | large | 0.784 | 0.685 | 0.764 | 0.867 | 0.839 | 0.862 |

Table 6: Experiment 4. Base and Large XLM-RoBERTa models comparison. The results are averaged over three seeds. All the models are trained with 7 epochs. The input parameters are set to pair=True, multiple=True, context=paragraph, lr.=2e-05. The metric used is F1 macro, computed on the development set.

5). Experiment 3 is a comparison of the models, for one fixed set of parameters and one fixed shuffle of the training set. In this case, monolingual pre-training or fine-tuning with BERT outperforms the exclusive usage of the multilingual XLM-RoBERTa configuration. Overall, XLM-RoBERTa large obtains higher scores than monolingual BERT models, base and large[8], for the two settings and languages. In conclusion, XLM-RoBERTa base is outperformed by monolingual BERT models for some parameters and shuffles, but XLM-RoBERTa large attains 7 of the 10 best overall scores in the one-shot settings, and of 5 of the 10 best results in the zero-shot settings.

**Experiment 4: Language Model Size.** Table 6 and Table 8 in Appendix A both show that the best performances reached are obtained by XLM-RoBERTa large. The gap between the models is clear with the one-shot track, and unclear for the zero-shot. The pairwise comparison of the base and large models for the zero-shot track shows that the base model often outperforms the large one.

In the one-shot setting, a closest look at the results per seed reveals base and large models show similar results only when a large standard deviation between seeds affect the overall performance of the large model [9].

**Tracks and optimal number of epochs.** The evolution of the scores between Table 3 and Table 6 shows that the one-shot setting needs more epochs to reach its highest performances than than the zero-shot setting. The F1-macro score increases by 2 points in the Zero-shot between 1 and 7 epochs when it gains 10 points in the one-shot training configuration.

## 6 Conclusion

In this system description paper, we explained our method to fine-tune transformer-based language models for the task of idiomaticity detection. Beyond the implementation, we also attempted to answer a few practical questions on how these models learn the task, and particularly their optimal parameters and input settings.

As future work, we would like to explore unsupervised approaches (e.g. sentence embeddings especially tuned on in-domain data such as news corpora of English, Portuguese and Galician). We are also planning to explore various methods to input the three contextual sentences, beyond simple concatenation as explored in this paper. Another interesting topic for further research would be to explore the complex compositionality relations occurring also withing the idiomatic expression, as exemplified sometimes in the examples labelled *meta-usage* in this dataset.

[8] Large and base models are both tested for the English classifier during the preliminary experiments (c.f. Table 8).

[9] The F1-macro for EN and PT and seed [1,2,3] in shuffle C are 0.907, 0.764 and 0.9, the standard deviation is 0.081.

## References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

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.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. 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. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gerard Steen. 2010. *A method for linguistic metaphor identification: from MIP to MIPVU*, volume v. 14 of *Converging evidence in language and communication research*. John Benjamins Pub. Co., Amsterdam.

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)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 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. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 65–69, Montréal, Canada. Association for Computational Linguistics.

David Vilares, Marcos García, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT representations. *CoRR*, abs/2103.13799.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

## A Appendix

| Context | Pair input | Tag type | Instance input format |
|---|---|---|---|
| core-sent | F | first | **text**: If you sell your **<idiom> insurance company </idiom>**'s stock privately or publicly to raise capital, you're considered a stock insurance company. |
| core-sent | F | mult. | **text**: If you sell your **<idiom> insurance company </idiom>**'s stock privately or publicly to raise capital, you're considered a stock **<idiom> insurance company </idiom>**. |
| core-sent | T | - | **text a**: If you sell your insurance company's stock privately or publicly to raise capital, you're considered a stock insurance company.<br>**text b** : insurance company |
| core-sent | T | mult. | **text a**: If you sell your **<idiom> insurance company </idiom>**'s stock privately or publicly to raise capital, you're considered a stock i**<idiom> insurance company </idiom>**.<br>**text b**: insurance company |
| core-sent | F | - | **text**: If you sell your insurance company's stock privately or publicly to raise capital, you're considered a stock insurance company. |
| 3-sent | T | first | **text a**: For example, River Stone Insurance Limited, a domestic insurance company in the United Kingdom, is an alien insurance company in the U.S. Alien insurance companies must file financial statements, auditor's reports and meet the requirements of the National Association of Insurance Commissioners (NAIC) International Insurers Department before being permitted to sell policies in the U.S. If you sell your **<idiom> insurance company </idiom>**'s stock privately or publicly to raise capital, you're considered a stock insurance company. Stock insurance companies are owned by their stockholders.<br>**text b**: insurance company |
| 3-sent | T | mult. | **text a**: For example, River Stone Insurance Limited, a domestic **<idiom> insurance company </idiom>** in the United Kingdom, is an alien **<idiom> insurance company </idiom>** in the U.S. Alien insurance companies must file financial statements, auditor's reports and meet the requirements of the National Association of Insurance Commissioners (NAIC) International Insurers Department before being permitted to sell policies in the U.S. If you sell your **<idiom> insurance company </idiom>**'s stock privately or publicly to raise capital, you're considered a stock **<idiom> insurance company </idiom>**. Stock insurance companies are owned by their stockholders.<br>**text b**: insurance company |

Table 7: Examples of various input formats obtained with the variation over three parameters : Context, pair-input, and and tag type. When tag type is set to first, the first occurrence of the target MWE in the core sentence is marked, even in a three sentences context. Occurrences of the target MWE in its plural form are not marked in this instance, because the singular form of the MWE is found in the sentence.

| 0 shot EN | 0 shot PT | 0 shot EN,PT | 1 shot EN | 1 shot PT | 1 shot EN,PT | ep. | batch size | l.r. | tag | pair | ctxt. | occ. | max sq length | split lang. | mod. EN | mod. PT/GL | mod. ML |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,807 | 0,677 | 0,772 | 0,922 | 0,911 | 0,921 | 7 | 8 | 2e-05 | T | T | par. | mult. | 350 | F | - | - | xlm rob. large |
| 0,756 | 0,678 | 0,740 | 0,909 | 0,922 | 0,916 | 8 | 8 | 2e-05 | T | T | par. | mult. | 350 | F | - | - | xlm rob. large |
| 0,728 | 0,673 | 0,722 | 0,934 | 0,876 | 0,915 | 9 | 32 | 4e-05 | T | T | par. | mult. | 512 | T | bert-base | bert-base | bert-base |
| 0,793 | 0,726 | 0,783 | 0,919 | 0,892 | 0,911 | 7 | 16 | 2e-05 | T | T | par. | mult. | 512 | F | - | - | xlm rob. large |
| 0,793 | 0,698 | 0,767 | 0,896 | 0,911 | 0,904 | 9 | 8 | 2e-05 | T | T | par. | mult. | 400 | F | - | - | xlm rob. base |
| 0,791 | 0,706 | 0,773 | 0,895 | 0,911 | 0,904 | 9 | 8 | 2e-05 | T | T | par. | mult. | 512 | F | - | - | xlm rob. large |
| 0,799 | 0,705 | 0,774 | 0,905 | 0,884 | 0,900 | 8 | 8 | 2e-05 | T | T | par. | mult. | 512 | F | - | - | xlm rob. large |
| 0,796 | 0,687 | 0,771 | 0,879 | 0,922 | 0,899 | 4 | 16 | 4e-05 | T | T | par. | mult. | 512 | F | - | - | xlm rob. large |
| 0,757 | 0,667 | 0,740 | 0,900 | 0,876 | 0,895 | 9 | 32 | 4e-05 | F | T | par. | first | 512 | T | bert-base | bert-base | bert-base |
| 0,769 | 0,710 | 0,763 | 0,903 | 0,867 | 0,894 | 5 | 16 | 4e-05 | T | T | par. | mult. | 512 | F | - | - | xlm rob. large |
| 0,281 | 0,361 | 0,313 | 0,886 | 0,891 | 0,891 | 7 | 8 | 2e-05 | T | T | par. | mult. | 300 | F | - | - | xlm rob. large |
| 0,768 | 0,659 | 0,743 | 0,882 | 0,895 | 0,891 | 6 | 8 | 2e-05 | T | T | par. | mult. | 350 | F | - | - | xlm rob. base |
| 0,740 | 0,683 | 0,736 | 0,889 | 0,875 | 0,889 | 9 | 16 | 4e-05 | F | T | par. | first | 512 | T | bert-large | bert-base | bert-base |
| 0,765 | 0,691 | 0,752 | 0,905 | 0,848 | 0,888 | 4 | 8 | 4e-05 | T | T | par. | mult. | 512 | T | bert-base | bert-base | bert-base |
| 0,736 | 0,718 | 0,746 | 0,903 | 0,847 | 0,888 | 9 | 16 | 4e-05 | F | T | par. | mult. | 512 | T | bert-large | bert-base | bert-base |
| 0,762 | 0,690 | 0,747 | 0,886 | 0,881 | 0,888 | 5 | 32 | 4e-05 | T | T | par. | mult. | 512 | T | bert-base | bert-base | bert-base |
| 0,757 | 0,602 | 0,722 | 0,860 | 0,922 | 0,886 | 4 | 16 | 4e-05 | F | T | par. | first | 512 | F | - | - | xlm rob. large |
| 0,767 | 0,699 | 0,757 | 0,895 | 0,861 | 0,886 | 9 | 32 | 4e-05 | F | T | par. | mult. | 512 | T | bert-base | bert-base | bert-base |
| 0,750 | 0,669 | 0,736 | 0,895 | 0,862 | 0,886 | 9 | 16 | 4e-05 | T | T | par. | first | 512 | T | bert-large | bert-base | bert-base |
| 0,763 | 0,701 | 0,754 | 0,891 | 0,860 | 0,885 | 9 | 32 | 2e-05 | F | T | par. | first | 512 | T | bert-base | bert-base | bert-base |

Table 8: Top twenty scores obtained during initial parameter optimization on the development set, sorted according to the F1 macro for one shot on English and Portuguese. The training data is shuffled between each run. All the scores provided for English and Portuguese in the zero-shot and one-shot settings are F1 macro scores.