It Takes One to Know One?
Idiomaticity Detection using Zero and One Shot Learning

Ashwin Pathak*
apathak60@gatech.edu

Raj Shah*
rajsanjayshah@gatech.edu

Vaibhav Kumar*
vaibhavk@gatech.edu

Yash Jakhotiya*
yashjakhotiya@gatech.edu

Abstract

Large Language Models have been successful in a wide variety of Natural Language Processing tasks by capturing the compositionality of the text representations. In spite of their great success, these vector representations fail to capture meaning of idiomatic multi-word expressions (MWEs). In this paper, we focus on the detection of idiomatic expressions by using binary classification. We use a dataset consisting of the literal and idiomatic usage of MWEs in English and Portuguese. Thereafter, we perform the classification in two different settings: zero shot and one shot, to determine if a given sentence contains an idiom or not. N shot classification for this task is defined by N number of common idioms between the training and testing sets. In this paper, we train multiple Large Language Models in both the settings and achieve an F1 score (macro) of 0.73 for the zero shot setting and an F1 score (macro) of 0.85 for the one shot setting.

1 Introduction

Transformer based Large Language Models (LLMs)(Kant et al., 2018) like BERT, DistilBERT, RoBERTa and their variants show state of art performance for a large number of NLP tasks, yet, they fail to capture multi-word expressions such as idioms. This is because contextualized pre-trained models learn compositional representations of text at sub-word and word level to reduce the vocabulary size. Therefore, we evaluate how well do LLMs identify idiomaticty by formulating the problem as a classification task.

For our task, we use data provided by SemEval 2022 task 2[†]. We treat the development data as held-out development data, as the test data of the SemEval task is not publicly available.

To evaluate how well LLMs identify idiomaticity, we use two different settings to determine the generalizability of the LLMs: zero shot and one shot setting. The zero shot setting is defined such that the MWEs in the train set are mutually exclusive of the MWEs found in the test set. For the one shot setting, there is only one Idiomatic and Literal training example for one MWE in the development set. This is different from traditional definitions of zero shot and one shot classification.

The paper briefly describes the related works in section II and the dataset used in Section III. Section IV gives the methodology used in zero shot and one shot learning. Section V describes the performed experiments and Section VI discusses the results. Section VII concludes the paper with a discussion on future research prospects and directions.

2 Related Work

MWEs pose a wide range of problems for detection (Constant et al., 2017). Idiomaticity identification was initially addressed by using statistical properties of text (Lin, 1999), symbolic methods (Baldwin and Villavicencio, 2002), and latent semantic analysis (Baldwin et al., 2003).

Further, constituent word embeddings were used for semantic similarity between the distributional vectors associated with an MWE and its parts (Katz and Giesbrecht, 2006). These were improved by an explicit disambiguation step (Kartsaklis et al., 2014) and by the joint learning of compositional and idiomatic embeddings (Hashimoto and Tsuruoka, 2016). However, these methods have their shortcomings due to low frequency of MWEs and limited non-contextual type level representations of MWEs multiple meanings.

Transformer-based pre-trained models do not benefit MWE identification without a new token representation (Hashempour and Villavicencio,

*Equal contribution
[†]sites.google.com/view/semeval2022task2-idiomaticity

Language	Split	MWEs	Samples
English	development	30	466
English	one-shot	60	87
English	zero-shot	163	3327
Portuguese	development	20	273
Portuguese	one-shot	40	53
Portuguese	zero-shot	73	1164

Table 1: Language distribution in the dataset

2020). (Madabushi et al., 2021) introduces new tokens for MWEs into a contextual pre-trained language model. However, they do not explore the relationship of potential MWEs in a sentence. We present a contextual and compositional network incorporating latent semantic significance of MWEs in a sentence.

3 Data

The dataset used in this report is the one provided by (Madabushi et al., 2021) as part of SemEval 2022 task 2 Subtask A. Each of the train and development splits of this dataset consists of samples containing a 'target' sentence, it's language information, a multiword expression (MWE), two contextual sentences that occur before and after the target sentence, and a label associated with the target. The label represents whether the multiword expression was used in an idiomatic sense or not.

The train split is further divided into 'zero-shot' and 'one-shot' data, containing 4491 and 140 samples each, consisting of 236 and 100 distinct MWEs respectively. Similarly, the development data contains 739 samples made from 50 different MWEs. One-shot MWEs have no overlap with zero-shot ones however, development data MWEs are a proper subset, as can be expected in a one-shot classification scenario. The language distribution of the dataset is summarized in Table 1.

Even though all the MWEs present in the development set are present in one-shot data, not all of them have both idiomatic and non-idiomatic samples. 17 out of the 50 development set MWEs have only a non idiomatic sample present in one shot data, and 12 development set MWEs have only an idiomatic sample present. Additionally, no MWE-Label pair is ever repeated in one shot data, making this dataset and the task one-shot with respect to distinct MWEs.

4 Method

We aim to solve the task to evaluate the extent to which models can identify idiomaticity in text through a coarse-grained classification into an "Idiomatic" or "Non-idiomatic" class. To better evaluate a model's ability to generalise and learn in a sample efficient fashion, we report scores in the zero-shot and one-shot setups.

Zero-shot learning

For the zero shot learning task, we use the train data to build a classifier using large language models like BERT-multilingual-uncased, DistilBERT-multilingual-uncased, XLM-RoBERTa-large and XLM-net. This task is "zero-shot" in nature as the idioms used in the train set and the development set are distinct. Therefore, we capture the discrepancy in the contextual meaning for idiomaticity, that is, we aim that our classifier distinguishes on the basis of lack of semantic correctness of literal meaning in the presence of an idiom in a sentence.

To make sure that idioms are not used explicitly while pre-training in large language models, we run a natural language inference task on BART-Large-MNLI and RoBERTa-Large-MNLI with the hypothesis as "idiom". The macro F1 score for both approaches is 0.51 and 0.50 respectively, which proves that there is no semantically learnt concept of "idiomaticity" by the model. It is important to note that no training data has been used for this step.

We therefore use multilingual LLMs to build classifiers for this setting. We need multilingual classifiers as the data consists of idioms in two languages: English and Portuguese. We further analyse the majority voting approach on the predictions of trained classifiers (inference based ensembling).

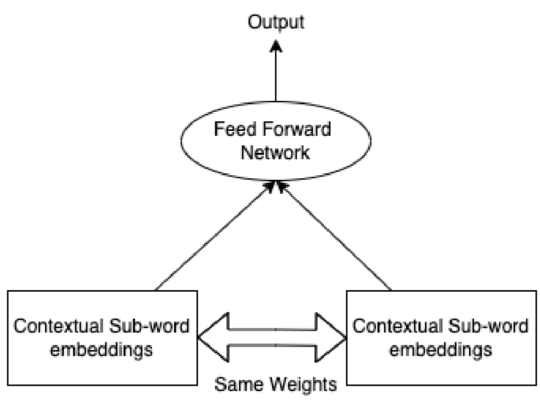

Figure 1: One-shot learning framework

One-shot learning
In the 'one-shot' setting, we use the only positive and/or the only negative training example, as available for each MWE in the development set. Note that the actual examples in the training data are different from those in the development set in both settings.

As shown in Fig 1, our model relies on finding similarity or relation scores between two input sentences. We first train this model on the pretext task of predicting whether two sentences with the same MWE belong to the same class. To achieve this goal, we employ contextual word embeddings to encode two sentences into feature vectors via an embedding function f_θ. The feature vectors are then combined with an operator $\mathcal{O}(.,.)$ to output $\mathcal{O}(f_\theta(x_i), f_\theta(x_j))$ on two inputs x_i and x_j. This is finally passed to a similarity/relation function g_ϕ to give score $s_{i,j}$ as,

$$s_{i,j} = g_\phi(\mathcal{O}(f_\theta(x_i), f_\theta(x_j)))$$

We test this framework with two underlying models - a Siamese Neural Network (Koch et al., 2015) and a Relation Network (Sung et al., 2018). With the Siamese Network, the operator $\mathcal{O}(.,.)$ is the element-wise difference between the two input feature vectors. The function g_ϕ is a fully connected layer followed by sigmoid activation. The loss in this case can be defined as, $L(s_{i,j}) = \Sigma_{i,j} 1_{y_i=y_j} \log(s_{i,j}) + (1 - 1_{y_i=y_j}) \log(1 - s_{i,j})$. Similarly, for the Relation Network, $\mathcal{O}(.,.)$ becomes the concatenation operator, g_ϕ becomes three fully connected layers with non linear activations followed by a sigmoid activation function. The loss in this case is the MSE loss, $L(s_{i,j}) = \Sigma_{i,j}(s_{i,j} - 1(y_i == y_j))^2$. In both of the models, x_i, x_j pairs are samples with matching MWEs.

We propose a novel inference methodology for our binary classification problem, where we also consider a dissimilarity score $1 - s_{i,j}$, with x_i, x_j belonging to support and query sets respectively. Support set is defined to be all samples with the same MWE as the query. We find the maximum of similarity and dissimilarity scores for all examples in the support set, and assign the same label or the opposite depending on whether the maximum was the similarity or the dissimilarity score. This helps us with (Madabushi et al., 2021) dataset where one-shot training data doesn't have samples for both the classes (idiomatic and non-idiomatic) for all MWEs.

5 Experiments

Zero-shot learning
We run our experiments on pre-trained models for zero-shot classification. We use Multilingual BERT, Multilingual DistilBERT, BERT-Portuguese, XL-Net and XLM-RoBERTa for exhaustive comparison and evaluation. We ensemble XL-NET, XLM-RoBERTa, and Multilingual DistilBERT in a majority vote based setting. As per the SemEval task, our baseline is Multilingual BERT for classification.

One-shot learning
For the contextual embeddings, we run our experiments on pre-trained compositional multilingual base models BERT, DistilBERT and XLM-RoBERTa for exhaustive comparison and evaluation. We run Siamese networks with cross entropy loss and Relation Networks with an MSE loss.

Our hyperparameter search pointed towards a dropout rate of 0.5, a learning rate of 2e-5 and we found AdamW to be the best performing optimizer.

6 Results

LN	Emb Model	Siamese F1	Relation F1
EN	BERT	0.79	**0.85**
EN	DistilBERT	0.79	0.83
EN	XLM-RoBERTa	0.83	**0.85**
PT	BERT	0.81	0.84
PT	DistilBERT	0.80	**0.85**
PT	XLM-RoBERTa	0.85	**0.85**
EN-PT	BERT	0.80	**0.85**
EN-PT	DistilBERT	0.79	0.84
EN-PT	XLM-RoBERTa	0.84	**0.85**

Table 2: One-shot evaluation results

Zero-shot learning
Table 3 shows F1-scores for different configurations, both ensemble and individual language models, with the baseline model being Multilingual BERT. We observe that the ensemble model performs better than the baseline in case of EN (0.71 F1 score) and EN-PT (0.68 F1 score) as compared to PT (0.53 F1 score) data. We further observe that XL-NET outperforms other models in case of English and Portuguese inputs while BERT performs the best in mixed setting.

One-shot learning
Table 2 reports F-1 scores for one-shot learning.

LN	Model	Dev F1
EN	BERT	0.65
EN	DistilBERT	0.70
EN	XLM-RoBERTa	**0.73**
EN	XL-NET	**0.73**
EN	Ensemble	0.71
PT	BERT	**0.64**
PT	DistilBERT	0.58
PT	XLM-RoBERTa	0.63
PT	XL-NET	0.62
PT	Ensemble	0.53
EN-PT	BERT	0.67
EN-PT	DistilBERT	0.70
EN-PT	XLM RoBERTa	0.71
EN-PT	XL-NET	**0.73**
EN-PT	Ensemble	0.68

Table 3: Zero-shot evaluation results

We found the best results of our Siamese and Relation network with XLM-RoBERTa (0.85 F1-score). We also observed a better score for Portuguese dataset as compared to English dataset on all of our models.

7 Analysis and Conclusion

In this paper we analyzed the effectiveness of large Language Models towards identifying idiomaticity in a given phrase using zero shot and one-shot classification tasks. Our experiments highlight the efficacy of introducing a new token representation for MWE as input to the pre-trained language model (Madabushi et al., 2021).

In zero shot classification, we use inference-level ensembling of different language models and observe that it outperforms BERT baseline in cases where the input language consists of English. This highlights a high degree of disagreement amongst the language models w.r.t Portuguese input, highlighting their brittleness.

For one shot classification, through Siamese and Relation Networks, we are able to represent the latent semantic relationship among MWEs leading to a much better F1 score than zero shot classification. Future work for one-shot classification could aim at breaking the barrier of 0.85 F1 score we seem to have hit with all embedding base models.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. *In Proceedings of the ACL 2003 Workshop on Multiword Expressions*, pages 89–96.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable. *In COLING-02: The 6th Conference on Natural Language Learning 2002*.

Mathieu Constant, Gülsen Eryigit, Johanna Monti and Lonneke van der Plas, Carlos Ramisch, and Michael Rosner. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. *In Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1:205–215.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *CoRR*, abs/1812.01207.

Dimitri Kartsaklis, Nal Kalchbrenner, and Mehrnoosh Sadrzadeh. 2014. Resolving lexical ambiguity in tensor regression models of meaning. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2:212–217.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. *In Proceedings of the Workshop on Multiword Expressions*, pages 12–19.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. *In Proceedings of the 32nd International Conference on Machine Learning*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317—-324.

Madabushi, Harish Tayyar, Gow-Smith, Edward, Scarton, Carolina, Villavicencio, and Aline. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *arXiv preprint arXiv:2109.04413*.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning.