Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss

Wei He
University of Sheffield
wei.he1@sheffield.ac.uk

Marco Idiart
Federal University of Rio Grande do Sul
marco.idiart@gmail.com

Carolina Scarton
University of Sheffield
c.scarton@sheffield.ac.uk

Aline Villavicencio
University of Exeter and University of Sheffield
A.Villavicencio@exeter.ac.uk

Abstract

Accurately modeling idiomatic or non-compositional language has been a longstanding challenge in Natural Language Processing (NLP). This is partly because these expressions do not derive their meanings solely from their constituent words, but also due to the scarcity of relevant data resources, and their impact on the performance of downstream tasks such as machine translation and simplification. In this paper we propose an approach to model idiomaticity effectively using a triplet loss that incorporates the asymmetric contribution of components words to an idiomatic meaning for training language models by using adaptive contrastive learning and resampling miners to build an idiomatic-aware learning objective. Our proposed method is evaluated on a SemEval challenge and outperforms previous alternatives significantly in many metrics. Our code is available at our project[1].

1 Introduction

Among multiword expressions (MWEs), idiomatic expressions (IEs) are difficult to model as their meaning is often not straightforwardly related to the meaning of the component words (Sag et al., 2002). These expressions, which are also commonly referred to as non-compositional expressions, often take on figurative meanings. For example, *eager beaver* has a figurative meaning of *an enthusiastic person who works very hard* different from the literal meanings of its component words like *impatient rodent* (Sag et al., 2002; Villavicencio and Idiart, 2019). They are a common occurrence across various genres (Haagsma et al., 2020).

Accurately understanding idiomatic expressions has posed a significant challenge, as word and phrase representations may favor inherently compositional usages at the levels of both words and subwords to minimize their vocabulary (Gow-Smith et al., 2022). Indeed recent models are mainly driven by compositionality, which is at the core of tokenization (Sennrich et al., 2016) and self-attention mechanism (Vaswani et al., 2017). Pre-trained language models including static and contextualised embeddings do not seem to be well-equipped to capture the meanings of IEs, as IEs with similar meanings are not close in the embedding space (Garcia et al., 2021b). This reveals a need for models that can accurately capture idiomatic language. Ensuring precise representation of IEs is crucial for their precise handling in various downstream applications, such as sentiment analysis (Liu et al., 2017; Biddle et al., 2020), dialog models (Jhamtani et al., 2021) and text simplification (He et al., 2023).

To address this issue, previous methods often rely on new datasets with human annotations or on data augmentation (Liu et al., 2023a; Dankers and Lucas, 2023). However, the use of alternative training processes has also been effective, including regression objective functions with a siamese network (Tayyar Madabushi et al., 2021) or substitute objectives (Liu et al., 2022) to break the compositionality of idiomatic phrases, as finding an objective to stand for idiomatic representation is difficult.

Our work focuses on the development of idiomatic-aware language models, which are designed to better represent MWEs of various degrees of idiomaticity in natural language text. To achieve this, we adopt the definition of idiomatic-aware models from SemEval 2022 task 2 (Tayyar Madabushi et al., 2022) that when using the model, the semantic similarity between an IE and its incorrect paraphrase equals the semantic similarity between a correct and an incorrect paraphrase. Our approach involves fine-tuning a pre-trained model using a bespoke triplet loss function that is specifically designed for capturing the asymmetry between the surface forms of the component

[1] https://github.com/risehnhew/Enhancing-Idiomatic-Representation-in-Multiple-Languages

Findings of the Association for Computational Linguistics: ACL 2024, pages 12473–12485
August 11-16, 2024

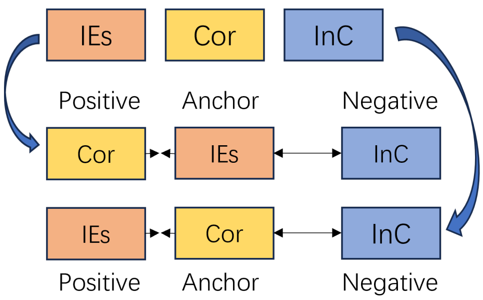

Figure 1: Triplet Resampling by using a specifically designed miner. For a triplet, it can generate 2 samples by treating the sentence containing IEs (IEs) and a correct (Cor) paraphrase sentence as positive and anchor (and vice-versa) interchangeably and the Incorrect (InC) sentence as a negative sample.

words and their semantic contribution to the meaning of the expression. To build this idiomatic-aware language model, we use in-batch positive-anchor-negative triplets (Balntas et al., 2016). Our model is trained on extracted triplets, where sentences with the idiomatic expressions and their synonyms correspond to positive and anchor respectively, and vice-versa, Figure 1. The aim of this training is to enable the model to learn the difference between the literal meanings of the component words of an MWE when used in isolation and their idiomatic meanings as part of the MWE. We use the "learn-to-compare" paradigm of contrastive learning (CL), which has been successfully adopted for obtaining better text embeddings (Ni et al., 2022b,a; Wang et al., 2022) including for polysemous words (Liu et al., 2019a). This framework fits well with our objective of distinguishing between the figurative and literal meanings of MWEs.

To evaluate the approach, a set of models with varying sizes and pre-training strategies is trained using this novel training method that we proposed. The best models achieved new state-of-the-art results in the dataset containing expressions of varying levels of idiomaticity, and our best model demonstrated a substantial improvements in both idiom-only performance and overall performance compared to the previous best results. Our contributions are:

An efficient approach for creating language models that can represent MWEs of varying levels of idiomaticity. This is achieved through a specialized training process using a triplet loss function and in-batch positive-anchor-negative triplets.

An idiomatic-aware loss function tailored to directly optimize the representation of idiomatic language and the potentially asymmetric and non-compositional contributions of the component words. This function plays a crucial role in training to discern the nuanced differences between idiomatic and literal meanings of MWEs.

New state-of-the-art performance models that enable the understanding of idiomatic language. This advancement represents a major leap forward opening up new possibilities for more nuanced and accurate language understanding.

The paper starts with an overview of previous work on idiomaticity representation in Section 2. It also introduces contrastive learning in NLP and IE evaluation methods. Section 3 presents our method using a triplet loss and data mining to do efficient training. Section 4 describes our experiments, and Section 5 analyzes the results.

2 Related Work

Idiomaticity representation can be challenging even for large language models (King and Cook, 2018; Nandakumar et al., 2019; Cordeiro et al., 2019a; Hashempour and Villavicencio, 2020; Garcia et al., 2021b; Klubička et al., 2023). For instance, GPT-3 (Brown et al., 2020) reaches only 50.7% accuracy in idiom comprehension (Zeng and Bhat, 2022a). This may be possibly due to idiomatic expressions being non-compositional and having figurative meanings that go beyond its individual words (Baldwin and Kim, 2010). Methods that have been used for representing idiomaticity include combining compositional components with adaptive weights (Hashimoto and Tsuruoka, 2016; Li et al., 2018a), representing MWEs with single tokens (Yin and Schütze, 2015; Li et al., 2018b; Cordeiro et al., 2019b; Phelps, 2022) and creating phrase embeddings that effectively capture both compositional and idiomatic expressions (Hashimoto and Tsuruoka, 2016). The latter involves an adaptive learning process that adjusts to the nature of the phrases to generate accurate representation. An adapter-based approach is proposed that augmenting the BART model with an "idiomatic adapter" trained on dedicated idiom datasets (Zeng and Bhat, 2022b). This adapter acts as a lightweight expert, enhancing BART's ability to capture figurative meanings alongside literal interpretations.

PIER (Zeng and Bhat, 2023), a language model based on BART, specifically addresses the challenge of representing non-compositional expressions, such as idioms, in natural language. Traditional compositionality-based models often struggle with these expressions, as their meaning cannot be simply derived from the sum of their parts. PIER overcomes this by incorporating an "idiomatic adapter" which learns to represent figurative meanings alongside literal ones. Additionally, Liu et al. (2023b) proposed a novel approach to idiomatic machine translation through retrieval augmentation and loss weighting, which significantly improves the translation quality of idiomatic expressions by leveraging context retrieval mechanisms and adjusting loss functions to better handle idiomaticity.

Contrastive Learning Contrastive learning is a method in machine learning that trains a model to distinguish between similar and dissimilar pairs of data. In recent times, significant progress has been made in sentence embeddings through contrastive learning (Gao et al., 2021; Giorgi et al., 2021a; Kim et al., 2021; Wu et al., 2022; Zhang et al., 2022; Xu et al., 2023). It also has been widely applied in other NLP research fields, such as text classification (Fang et al., 2020), machine translation (Pan et al., 2021), information extraction (Qin et al., 2021), question answering (Karpukhin et al., 2020) and text retrieval (Xiong et al., 2020). Despite their shared goal of acquiring high-quality text representations (Reimers and Gurevych, 2019; Gao et al., 2021; Neelakantan et al., 2022; Giorgi et al., 2021b), the exploration of idiomatic representation and related research through contrastive learning is still yet to be fully explored. Contrastive learning with triplet loss involves training the model on triplets: an anchor sample, a positive sample (similar to the anchor), and a negative sample (dissimilar to the anchor). The goal is to minimize the distance between anchors and positive samples while maximizing the distance between anchors and negative samples. This approach has recently been applied to tasks such as idiom usage recognition and metaphor detection (Zhou et al., 2023).

Idiomaticity Representation Evaluation Assessing idiomatic representation in language models has included both extrinsic and intrinsic evaluations. Extrinsic methods evaluate how well the model's idiomaticity representation impacts downstream tasks, such as machine translation (Dankers et al., 2022), sentence generation (Zhou et al., 2021) or conversational systems (Adewumi et al., 2022). Intrinsic methods evaluate the model's understanding of idiomaticity itself, using approaches like probing to investigate and understand the linguistic information encoded in the representation (Garcia et al., 2021a). Datasets like AStitchInLanguageModels (Tayyar Madabushi et al., 2021) and Noun Compound Type and Token Idiomaticity (NCTTI) dataset (Garcia et al., 2021a) offer labelled examples for intrinsically testing how much the similarities perceived by a model are compatible with human judgements about similarity. More broadly, SemEval-2022 task 2B (Tayyar Madabushi et al., 2022), evaluates idiomaticity representation in multilingual text while also requiring models to predict the semantic text similarity (STS) scores between sentence pairs, regardless of whether or not either sentence contains an idiomatic expression. The main objective of this task is to address the shortcomings of existing state-of-the-art models, which often struggle to handle idiomaticity. We use this dataset to evaluate our methods.

3 Idiomaticity-aware Objective

Our strategy for improving IE representation in language models utilizes a contrastive triplet loss adapted to prioritize idiomaticity and employs a miner to generate positive-anchor-negative triplets for training the model.

3.1 Triplet Loss

Triplet loss is a powerful tool for training language models to learn representations of data that are useful for a variety of NLP tasks (Neculoiu et al., 2016). It has also been widely used in training models for tasks such as image retrieval, and face recognition (Schroff et al., 2015; Khosla et al., 2020).

Triplet loss is a distance-based loss function defined as

$$L_{a,p,n} = max(d(a_i, p_i) - d(a_i, n_i) + m, 0), \tag{1}$$

where the triplets $(a_i, p_i, n_i), i = 1 \cdots N$, correspond to $anchor$, $positive$ and $negative$ examples, where a_i and p_i are semantically identical and n_i is semantically dissimilar from them. $d(x, y)$ is a distance measure and in our method we use cosine similarity (denoted here by *sim*)

$$d(x, y) = sim(x, y). \tag{2}$$

Finally, the margin m controls the minimum distance between anchor-positive pairs and anchor-negative pairs.

Selecting the right margin is crucial for our method. If it is too small, the task becomes too easy, lacking meaningful distinctions. Conversely, if it is too large, it can slow down convergence or yield suboptimal solutions (Schroff et al., 2015). The margin is a hyperparameter and its tuning requires experimentation based on the specific dataset and application.

In this paper we use a miner to build triplets for learning idiomaticity more efficiently.

3.2 Modelling IEs with Adaptive Contrastive Tripet Loss

This section explains how to improve the language model's ability to understand IEs in text without STS scores by adapting triplet loss to the IE-aware training strategy. We will describe the process step-by-step and discuss its benefits.

3.2.1 Task Definition

One widely used approach for measuring idiomaticity is by calculating the distance between a dedicated representation for the MWE as a single token and a compositional representation of its components using operations like sum or multiplication (Mitchell and Lapata, 2008; Cordeiro et al., 2019b). A good idiomatic expression representation, as framed by Madabushi et al. (2022), should have the following property:

$$\begin{aligned} sim(S_{MWE}, S_{\to c}) &= 1 \\ sim(S_{MWE}, S_{\to i}) &= sim(S_{\to c}, S_{\to i}) \end{aligned} \tag{3}$$

where S_{MWE} denotes a sentence containing the idiomatic expression and $S_{\to c}$ and $S_{\to i}$ represent sentences with the idiomatic expression replaced by its correct and incorrect paraphrases, respectively. Ensuring these properties hold for all MWEs during training using standard loss functions can be challenging.

Previous studies need annotated similarity scores of pairs as labels for building the training set (Tayyar Madabushi et al., 2021; Phelps, 2022). Their objective functions are as follows:

$$\begin{aligned} sim(S_{MWE}, S_{\to c}) &= 1 \\ sim(S_{MWE}, S_{\to i}) &= score_1 \\ sim(S_{\to c}, S_{\to i}) &= score_2 \end{aligned} \tag{4}$$

where $score_1$ and $score_2$ are STS scores used to measure the similarity between two pieces of text, with scores typically ranging from 0 (no similarity) to 1 (identical meaning). In previous methods, language models were trained to predict STS scores between text containing IEs and those without IEs, in order to improve their ability to understand IEs.

In our method, we will utilize a triplet loss in combination with a miner to extract triplets without STS scores, approximating the definition in equations (3). It is worth noting that without using STS scores, training data can be acquired more easily.

3.3 Mining to Extract Triplets

The original dataset only has IEs, the sentences with IE and their correct and incorrect paraphrases. To extract triplets for our idiomatic-aware training we use a semantic meaning miner. We use batch negatives approach that leverages the other samples present in the same mini-batch for serving as negative instances. However, not all negatives in a batch are useful for our training. Thus, we introduce a special preprocessing step in our method.

Relabel Training Data For a triplet to be valid, it must meet certain requirements. We first categorize sentences into different groups. Each group contains a sentence with the IE (s), its correct (c) and incorrect (i) paraphrases, such as examples in Table 2. New labels will be assigned in each group based on IEs and their paraphrases. Our approach assigns identical labels to sentences with the same meaning (original sentence and correct paraphrases). Firstly, s and c must have the same label, which means they represent the same meaning. Secondly, each i must have different labels and differ from the label of s and c, which means they represent different meanings.

It also needs labels in different groups to be distinct to others. For example in Table 2, as sentences with index 4 and 5 are a pair of s and c, they are assigned with the same label **en3**. Other sentences in Group 2 are assigned different labels because they are incorrect paraphrases. The labels in Group 2 are distinct from labels in Group 1.

In this way, a triplet can be acquired easily since $anchor$, $positive$ are sentences with the same labels, and a $negative$ is a sentence with different labels.

Selected Multi-negatives Negative instances refer to sentences whose labels differ from the anchor and positive in a batch. In the case of Multi Negative Ranking Loss (Sun et al., 2020) with triplet formation, there are multiple negatives $[n_1, n_2, ...n_k]$

Group	Index	Label	Instance
1	1	en1	So Aaron faced the same brutal racism other Black players of the era experienced, especially as the slugger approached Ruth's **IDhomerunID** record.
	2	en1	So Aaron faced the same brutal racism other Black players of the era experienced, especially as the slugger approached Ruth's **baseball run** record.
	3	*en2*	So Aaron faced the same brutal racism other Black players of the era experienced, especially as the slugger approached Ruth's **house run** record.
2	4	en3	Robinhood is supposed to be the revolutionary trading app that made it possible for the **IDsmallfryID** to get together and crush the big boys.
	5	en3	Robinhood is supposed to be the revolutionary trading app that made it possible for the **insignificant** to get together and crush the big boys.
	6	*en4*	Robinhood is supposed to be the revolutionary trading app that made it possible for the **little fry** to get together and crush the big boys.
	7	*en5*	Robinhood is supposed to be the revolutionary trading app that made it possible for the **little kid** to get together and crush the big boys.

Table 2: Examples of training data. Sentences that have the same meaning are given the same labels. We treat IE expressions as a single token and preprocess it as shown. For example, the IE *home run* is replaced as *IDhomerunID*.

for each anchor-positive pair, and the objective is to ensure that the anchor is closer to the positive than to any of the negatives by a margin.

$$\begin{aligned} &\mathcal{L}_{\text{multi-negative}}(a, p, [n_1, \ldots, n_k]) \\ &= \sum_{i=1}^{k} \max(d(a,p) - d(a,n_i) + m, 0) \end{aligned} \tag{5}$$

We take the SemEval 2022 task 2B training set as our source to build our training data. The dataset comprises approximately $8,600$ annotated examples in multiple languages, including English and Portuguese. It was divided into 4,840 training sentences, 739 development sentences, 483 evaluation sentences and 2,342 test sentences. The original training data already includes information on correct and incorrect paraphrases. The context sentences help disambiguate the IE's meaning. This annotated data is crucial for training machine learning models to detect idiomatic expressions in varied linguistic contexts, facilitating multilingual natural language understanding and processing.

After relabeling, the training dataset will be a list of sentences with their corresponding label. We do not shuffle the training data to maintain its order, as sentences that belong to a triplet are adjacent in the training set. The batch size is set to 64, which is a balance between easy training and ensuring a sufficient number of sentences to build triplets.

However, not all negatives contribute equally to our learning. Some triplets may already satisfy the constraint (easy triplets), such as triplets with negatives that are sentences from other groups. They provide little to no information of IEs understanding for the model to learn from.

Mine Triplets In our methods, a semantic similarity miner calculates Euclidean distance between all possible pairs of embeddings in a batch and selects according to its similarity margin. The miner similarity margin is the difference between the anchor-positive distance and the anchor-negative distance. It is also a hyperparameter in our method. The miner select the triplets that violate the miner similarity margin to make the model learn nuanced differences between figurative and literal meanings of IEs. For instance, a triplet of sentences could include an idiomatic expression as the anchor, its paraphrases as the positive, and a sentence with a literal meaning as the negative.

Table 2 illustrates the newly build training data. In this case, S_{MWE} and $S_{\to c}$ can act as anchor and positive to each other, and $S_{\to i}$ can only be treated as the negative in a triplet. It is worth noting that S_{MWE} and $S_{\to c}$ are interchangeable to form pairs (a_i, p_i), which can build more triplets for our training. For example, in Table 2, with the miner, it will only take sentences in the same group because the semantic meanings of different groups are not similar. In this way, sentences in Group 1 can build 2 triplets with index 1 and 2 being the anchor and positive interchangeably. Sentences in Group 2 can build 4 triplets.

3.4 Objective Transformation

In our approach, both S_{MWE} and $S_{\to c}$ can serve as anchors. However, since we assign different labels to various incorrect paraphrases, no positive sentence in a group can be associated with any $S_{\to i}$ as the anchor. As a result, there are only two possible scenarios in our approach.

If S_{MWE} is the anchor,

$$sim(S_{MWE}, S_{\to c}) - sim(S_{MWE}, S_{\to i}) \leq m_a \tag{6}$$

if $S_{\to c}$ is the anchor,

$$sim(S_{\to c}, S_{MWE}) - sim(S_{\to c}, S_{\to i}) \leq m_b \tag{7}$$

The margin m is a predefined fixing value. If we set $m_a = m_b$, then combining Eq. (6) and Eq. (7), the objective function can be transformed to:

$$\begin{aligned} sim(S_{MWE}, S_{\to c}) - sim(S_{MWE}, S_{\to i}) \approx \\ sim(S_{\to c}, S_{MWE}) - sim(S_{\to c}, S_{\to i}) \end{aligned} \tag{8}$$

The similarity measure is symmetric, therefore $sim(S_{MWE}, S_{\to c}) = sim(S_{\to c}, S_{MWE})$. In this way, our objective function equivalent to:

$$sim(S_{MWE}, S_{\to i}) \approx sim(S_{\to c}, S_{\to i}) \tag{9}$$

Equation (9) approximates the definition of the good idiomatic aware model in Equation (3). In this way, by using our specific triplet loss, we can train a model to be idiomatically aware more directly without STS scores.

4 Experiment

This section presents the comprehensive methodology employed to our model training. We begin by detailing the experiment implementation, including the hyperparameter setting, models used, evaluation method, and the overall setup.

4.1 Implementation Details

The method is implemented by using the Transformers (Wolf et al., 2020) and PyTorch Metric Learning (Musgrave et al., 2020) libraries. Some of the pre-trained models are fetched from Sentence-transformer library[2] and HuggingFace Model repositories[3].

We calculate sentence similarity using the cosine similarity of the mean pooling of the last two hidden layers. Empirically, we set the similarity margin for the miner to 0.4, and the training loss margin to 0.3. Given the limited availability of idiomatic text data, relying solely on the training signal from our contrastive objective is insufficient for learning general semantic representations. Therefore, we initialize our model with other pre-trained semantic-aware models (Reimers and Gurevych, 2019). Our best model uses a pre-trained multilingual model, '*paraphrase-multilingual-mpnet-base-v2*'[4], and fine-tunes it with our method to fit the task. It is pre-trained with millions of paraphrases, so it can represent sentence semantic meanings well (Reimers and Gurevych, 2019).

4.2 Evaluation

We perform intrinsic evaluation (Reimers et al., 2016) using the SemEval-2022 task 2 Subtask B[5] (Sem2B). We use Spearman's rank correlation (ρ) between model-generated scores and human judgment scores to see how well models understand idioms in sentences. Instead of comparing exact scores, this method focuses on how the sentence pairs are ranked based on predicted similarity compared to human judgments. A higher correlation means the model is better at understanding relationships, including those involving idioms, even if the exact predicted scores themselves aren't always perfect matches.

4.3 Comparative Analysis

We compare our method with well-performed Semantic Textual Similarity models and recent large language models (LLMs). Some training-based methods are from SemEval-2022 task 2 Fine Tune solutions (Madabushi et al., 2022). Here are brief descriptions:

YNU-HPCC (Liu et al., 2022) is the previous best method, which uses contrastive learning approaches in sentence representation. However, it treats negatives in a batch equally.

drsphelps (Phelps, 2022) introduces a method for improving idiom representation in language models by incorporating idiom-specific embeddings using BERTRAM into a BERT sentence transformer.

baseline is the SemEval task's baseline results. It is fine-tuned using multilingual BERT (Devlin et al., 2019) and adding single tokens for each MWE in the data.

GTE large [6] is a powerful text embedding model trained with multi-stage contrastive learning, delivers impressive performance across NLP

[2]https://www.sbert.net/docs/pretrained_models.html

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

[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[5]https://codalab.lisn.upsaclay.fr/competitions/8121

[6]https://huggingface.co/thenlper/gte-large

Method	Model Size	Subset		All
		Idiom	STS	
YNU-HPCC	183M	0.428	0.664	0.665
drsphelps	420M	0.412	**0.819**	0.650
baseline	110M	0.399	0.596	0.595
GTE large	334M	0.236	0.806	0.465
E5 large	334M	0.252	0.807	0.514
LLama2	13,000M	0.171	0.486	0.399
Our best	558M	**0.548**	0.716	**0.690**

Table 3: Test results of Task 2 on Spearman's rank correlation coefficient between the two sets of STS scores.

and code tasks despite its modest size (Li et al., 2023).

E5 large [7] uses weakly-supervised contrastive pretraining for text embeddings that achieves excellent for general-purpose text representation (Wang et al., 2022).

LLama2 (Touvron et al., 2023) achieved excellent performance in a series of NLP tasks. We select the LLama2-13B for comparison.

5 Results and Analysis

In this section, we report results and analyze them in different settings.

5.1 Overall Results

Table 3 demonstrates that our method outperforms all other models both overall and in the Idiom Only subset. The "STS only" score refers to the performance of systems on Semantic Text Similarity data that does not necessarily contain idioms. In contrast, the "Idiom only" score pertains to the performance on idiom STS data. "All" represents the overall performance of a model across the entire dataset. In the Idiom Only subset, our method achieves a score of 0.548, which is higher than the score of the next best model, YNU-HPCC (0.428). In the overall performance, it achieves a score of 0.690, exceeding the score of the next best model, YNU-HPCC (0.665). In the STS subset, drsphelps achieves the highest score of 0.819. These results suggest that our method is a powerful and effective idiom-aware text embedding model that can be used for a variety of idiomatic expressions related NLP tasks.

GTE large and E5 large both show a similar pattern of lower performance in the Idiom task (0.236 and 0.252, respectively) but strong results in the STS task (0.806 and 0.807, respectively). Their overall scores (0.465 and 0.514, respectively) suggest that while they are proficient in semantic textual similarity, their capacity to handle idiomatic expressions is not as developed. LLama2 has the lowest scores across all three categories, with a particularly low score for Idiom (0.171). It reveals a surprising lack of ability to represent idiomatic expressions for such recent general large language model.

[7] https://huggingface.co/intfloat/e5-large

Language	Subset		All
	Idiom Only	STS Only	
EN	0.560	**0.759**	**0.757**
PT	**0.570**	0.657	0.707
GL	0.515	-	0.515
3L	0.548	0.716	0.690

Table 4: Our test results of Task 2 on Spearman's rank correlation coefficient in English (EN), Portuguese (PT), and Galician (GL) separately. 3L is the combination of 3 languages.

5.2 Performance on Different Languages

Method	Lang(s)	Idiom	STS	ALL
drsphelps	EN	**0.486**	0.834	**0.764**
	PT	0.464	**0.791**	0.731
	GL	**0.286**	-	0.286
	3L	0.412	0.819	0.650
E5 large	EN	0.242	**0.919**	0.607
	PT	0.276	0.646	0.551
	GL	0.247	-	0.247
	3L	0.252	0.807	0.514
LLama2	EN	0.156	0.512	0.391
	PT	0.206	0.496	0.510
	GL	0.185	-	0.185
	3L	0.171	0.486	0.399

Table 5: Spearman's rank correlation coefficients for drsphelps, E5 large, and LLama2 methods across idiomatic only, STS only, and overall results in English (EN), Portuguese (PT), Galician (GL), and their combination (3L).

The results in Table 4 show that our best model performed well on Sem2B and in all three languages. The best results were achieved, with overall ρ values of 0.757 for English, 0.707 for Portuguese, and 0.515 for Galician. The best overall results on Sem2B were achieved for English, and the best Idiom Only score was achieved for Portuguese. There is no STS-only score for Galician in the test set. The models performed best on English,

Model	Subset		All
	Idiom Only	STS Only	
Original			
roberta-base	0.184	0.626	0.492
x-r-large	0.138	0.284	0.444
p-v2	0.225	0.838	0.532
After Training			
roberta-base	0.454	0.622	0.613
x-r-large	0.484	0.465	0.639
p-v2	0.548	0.716	0.690

Table 6: Test results across three models *roberta-base*, *xlm-roberta-large* (x-r-large) and *paraphrase-multilingual-mpnet-base-v2* (p-v2) before and after training.

followed by Portuguese and Galician. This is due to the fact that there is more training data available for English than for Portuguese or Galician. The results also show that the models were able to generalize well, even when the amount of training data was limited. For example, the models achieved ρ values of 0.707 and 0.515 for Portuguese and Galician, even though the training data for these two languages was smaller than the training data for English. Compared to other methods in Table 5, our model excels particularly in handling idiomatic expressions, outperforming other models in the Idiom Only subset. Additionally, while drsphelps and E5 large show strong results in the STS subset, our model maintains a balanced performance across all datasets, demonstrating its robustness.

5.3 Impact of Our Training

Table 6 presents comparative performance results of three language models, roberta-base (Liu et al., 2019b), xlm-roberta-large (Conneau et al., 2020) (x-r-large) and *paraphrase-multilingual-mpnet-base-v2* (p-v2), across three different subsets of data: Idiom Only, STS Only, and All. The first two models are widely used language models with general and multilingual properties, respectively. The third model is the base model we used in our best model. The results are split into two categories: 'Original', which indicates the performance before additional training, and 'After Training', showing the performance post-training.

For the Idiom Only subset, the original scores were 0.184 for roberta-base, 0.138 for x-r-large, and 0.225 for p-v2. After training, these scores improved significantly to 0.454 for roberta-base, 0.484 for x-r-large and 0.548 for p-v2. When looking at the overall performance, the x-r-large model's performance originally was 0.444 and increased to 0.639 after training. Similarly, the p-v2 model's performance was initially 0.532 and rose to 0.690 after training. In the STS Only subset, there have been declines at p-v2 from 0.838 to 0.716. It is because our training only focuses on improving idiom representation, and it may slightly sacrifice the performance of specific fully-trained models.

The size of our model's parameters is slightly larger than most, but it significantly outperforms others, demonstrating the effectiveness of our proposed method beyond just using a larger model. As shown in Table 4, our method achieves superior results in idiomatic representation even when compared with implementations using the same model sizes.

The results in Table 7 showcase that as the number of epochs increases, the overall performance as well as the performance on the "Idiom Only" subset generally improves. This suggests that the model is learning and improving its IE understanding ability during our training. The performance on the "Idiom Only" subset starts very low at epoch 0, with an accuracy of 0.225, which is expected since the model has not learned much IE representation yet. There is a significant improvement between epoch 0 and epoch 8, with the score nearly doubling to 0.499. The improvement in performance starts to plateau after epoch 10, with only minor increases observed at epochs 15 and 25. The "STS Only" subset starts with a high performance even at epoch 0, with an accuracy of 0.838. This is because the model has already been pre-trained with STS tasks. Unlike the "Idiom Only" subset, the performance on the "STS Only" subset decreases as the number of epochs increases, dropping to 0.716 by epoch 25. This could indicate that the model is becoming more specialized in the idiom task at the expense of the STS task. In summary, while the model is improving in its ability to understand idioms with more training, this comes at the cost of its performance on STS tasks. This trade-off can be addressed by adjusting the training process.

In summary, our models were able to generalize well to different settings, even when the amount of training data was limited. This suggests that the models are learning to capture the underlying properties of idiomatic expressions, rather than simply memorizing a list of idiomatic expressions.

Epoch	Subset		All
	Idiom Only	STS Only	
0	0.225	0.838	0.532
8	0.499	0.785	0.670
10	0.531	0.740	0.682
15	0.539	0.740	0.688
25	0.548	0.716	0.690

Table 7: Test Results with different training epochs by using same *p-v2* model.

6 Discussion

The proposed model for training requires the identification of idiomatic expressions (IEs) in each sentence beforehand. This step is crucial for reducing the difficulty of the training process. Without identifying the IEs beforehand, the model may not perform optimally, and its accuracy may be compromised. Therefore, it is essential to ensure that the text has IEs identified to achieve the best results.

7 Conclusion

Idiom representations have always been a challenge due to the non-compositional nature of idiomatic expressions. The performance of downstream tasks, such as translation and simplification, is dependent on the quality of the representations. This paper proposes a new method to train language models using adaptive contrastive learning with triplets and resampling miners. In this way, our method can build a better optimization objective, which makes the training very efficient.

The proposed method, evaluated on the idiomatic semantic text similarity tasks, significantly outperforms previous methods. With limited idiomatic text data, the sole training signal of the contrastive objective is not sufficient to learn general semantic representations. Therefore, the model is initialized with other pre-trained semantic-aware models. A series of base models in different sizes and pre-training strategies are trained in the proposed training loss. The best models achieve new state-of-the-art results with a significant improvement in overall over the previous best in the evaluation task.

8 Future Work

In the future, we plan to use the idiomatic-aware model in other NLP tasks that require sensitivity to idiomatic expressions, such as machine translation. Additionally, we aim to improve the model's training by adding more supervision, which will help it focus on contextual information. This will allow the model to better understand multiword expressions based on different contexts.

9 Limitations

In order to train our model, we require triplets that consist of three distinct parts: a sentence that contains IEs, a correct paraphrase of those IEs, and an incorrect paraphrase of those IEs. The quality of triplets is crucial to the development of our model and requires intensive human expert involvement to ensure accuracy.

10 Acknowledgments

The work was partly supported by EPSRC grant EP/T02450X/1, by NAF/R2/202209 grant and by Brazilian CNPq Grant 311497/2021-7.

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. Vector representations of idioms in conversational systems. *Sci*, 4(4):37.

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

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In *Proceedings of the web conference 2020*, pages 1217–1227.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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. 2019a. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

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

Verna Dankers and Christopher Lucas. 2023. Non-compositionality in sentiment: New data and analyses. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5150–5162, Singapore. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

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.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. 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.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021a. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021b. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenisation by alternative treatment of spaces. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11430–11443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany. Association for Computational Linguistics.

Wei He, Katayoun Farrahi, Bin Chen, Bohua Peng, and Aline Villavicencio. 2023. Representation transfer and data cleaning in multi-views for text simplification. *Pattern Recognition Letters*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. Association for Computational Linguistics.

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.

Filip Klubička, Vasudevan Nedumpozhimana, and John Kelleher. 2023. Idioms, probing and dangerous things: Towards structural probing for idiomaticity in vector space. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 45–57, Dubrovnik, Croatia. Association for Computational Linguistics.

Bing Li, Xiaochun Yang, Bin Wang, Wei Wang, Wei Cui, and Xianchao Zhang. 2018a. An adaptive hierarchical compositional model for phrase embedding. In *IJCAI*, pages 4144–4151.

Minglei Li, Qin Lu, Dan Xiong, and Yunfei Long. 2018b. Phrase embedding learning based on external and internal context with compositionality constraint. *Knowledge-Based Systems*, 152:107–116.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023a. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023b. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Kuanghong Liu, Jin Wang, and Xuejie Zhang. 2022. YNU-HPCC at SemEval-2022 task 2: Representing multilingual idiomaticity based on contrastive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 211–216, Seattle, United States. Association for Computational Linguistics.

Ninghao Liu, Qiaoyu Tan, Yuening Li, Hongxia Yang, Jingren Zhou, and Xia Hu. 2019a. Is a single vector enough? exploring node polysemy for network embedding. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 932–940.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. 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. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos García, 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.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning. *ArXiv*, abs/2008.09164.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 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. Association for Computational Linguistics.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with Siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022b. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Dylan Phelps. 2022. drsphelps at semeval-2022 task 2: Learning idiom representations using bertram. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 158–164.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. Association for Computational Linguistics.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407.

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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. SimCSE++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12028–12040, Singapore. Association for Computational Linguistics.

Wenpeng Yin and Hinrich Schütze. 2015. Discriminative phrase embedding for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1368–1373, Denver, Colorado. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2022a. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng and Suma Bhat. 2022b. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng and Suma Bhat. 2023. Unified representation for non-compositional and compositional expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11696–11710, Singapore. Association for Computational Linguistics.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11730–11738.

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.

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023. CLCL: Non-compositional expression detection with contrastive learning and curriculum learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–743, Toronto, Canada. Association for Computational Linguistics.