Comparing Fine-Tuning Objectives for Sentence Transformer Embeddings in Idiomatic Contexts

Jack Xiao
Stanford University
jackxiao@

Avi Gupta
Stanford University
agupta07@

Jack Michaels
Stanford University
jackfm@

Abstract

In this paper, we leverage the non-compositional nature of idioms to probe the efficacy of modern sentence encoders. Our experiments fine-tune a pre-trained embedding model with isolated and in-context idiomatic data across three varying loss objectives with limited contextual information being used to improve the models ability to recognize idiom significance. Our results show that our baseline transformer produces idiom embeddings that show a slight sensitivity to idiomatic material. Yet, by fine-tuning this model we find promising areas to explore that may lead to improved embedding similarities beyond the baseline performance, though we note this effect varies and is sensitive to data and training strategy. Across the board, we observe that a triplet loss scheme has the most promising impact on model performance, particularly when fine-tuned on the idiomatic-sentence/paraphrased sentence dataset, perhaps due to the consideration of both positive and negative examples as well as the larger initial distance between idiom and definition embeddings. Our few-shot contextual learning task yields inconclusive results and leaves room for further study.

1 Introduction

Idiomatic expressions (IEs) are a category of multi-word expressions (MWEs). MWEs, such as *good morning*, can be a compound, a sentence fragment, or an entire sentence. IEs typically surface as collocations and exhibit semantic idiomaticity, also understood as semantic non-compositionality, where the meaning of the IE is not logically inductible. NLP applications historoically have struggled with IEs for reasons including their non-compositionality and semantic ambiguity.

Embeddings are lower-dimensional spaces in which higher-dimensional vectors can be translated into. Embeddings are also meant to encode semantic similarity; similar inputs should be closer in the embedding space, and dissimilar inputs should be further apart. While word embeddings capture the meaning of individual words, sentence embeddings aim to capture the meaning of the sentence without encoding individual words. Textual embeddings for IEs are commonly generated with pre-trained encoding models in the NLP literature. Less frequently, however, are they fine-tuned for quality improvements. In this paper, we set out to understand how various fine-tuning training objectives impact the quality of idiom-related embeddings generated by a pre-trained sentence transformer.

To perform this study, we merge and process three existing datasets to produce a dataset of unique pairs of idioms in isolation and their definitions (referred to as I/D). In addition, we generate a dataset of sentences with idioms and corresponding sentences with paraphrased IEs (referred to as S/P). Before comparing fine-tuning approaches, we establish a baseline using the pre-trained Sentence Transformers model on both datasets. We then experiment with contrastive loss, triplet loss, and similarity loss as fine-tuning objectives for the baseline transformer model. In addition, we re-purpose our datasets to test if the model can better decipher the meaning of an IE by exposure to contextual information.

2 Prior Literature

The contextual embeddings provided by encoding models are critical for downstream NLP tasks. For instance, BART (Lewis et al., 2019), a denoising autoencoder for pretraining sequence-to-sequence models, is applicable to comprehension, translation, and natural language text generation tasks. With regard to idiom detection, pre-trained encoding models are commonly used to generate embeddings. (Skvorc et al., 2020) use the contextual embeddings of BERT (Bidirectional Encoder Representations from Transformers) and ELMo (Embeddings from Language Models) on deep neural networks to edge

out existing approaches for Slovenian idiom detection. (Chu et al., 2022) also make use of large-scale pre-trained language models, multilingual BERT and XLM-RoBERTa, to train a classification model that detects whether a MWE in the sentence is an idiomatic usage. (Tan and Jiang, 2020) propose a BERT-based dual embedding model to encode Chinese idioms. However, (Shwartz and Dagan, 2019) show with six distinct lexical composition tasks that although textual pre-trained language models, capable of polysemy, were more performant than their non-contextual predecessors at detecting meaning shift (i.e. semantic non-compositionality) in idiomatic MWEs, their efficacy continued to be insufficient relative to human performance.

As a remedy, (Tayyar Madabushi et al., 2021) make an argument for fine-tuning a pre-trained model in their study of idiomaticity representation. They produce a metric ρ to measure the consistency of a model in capturing similarities between sentences containing idioms and corresponding sentences that are fully compositional. Given an example E, the example with the MWE repalced with a correct paraphase E_c, and the example with the MWE repalced with an incorrect paraphase E_i, the task requires the model to generate similarity scores under the criteria that:

$$\forall_{i \in I}(sim(E, E_c) = 1; sim(E, E_i) = sim(E_c, E_i))$$

(Tayyar Madabushi et al., 2021) divide this task into Subtask A, only pre-training, and Subtask B, fine-tuning. In both subtasks, Sentence BERT (Reimers and Gurevych, 2019) produces embeddings that are be compared using cosine similarity. For their "select replace" tokenization model, in which a given instance of an MWE is replaced only when a one-shot model predicts that the MWE in a given sentence has an idiomatic meaning, the English language case's development set and test set similarity scores in Subtask A were $\rho = 0.848$ and $\rho = 0.805$ respectively, whereas in Subtask B they were $\rho = 0.851$ and $\rho = 0.825$. Their results demonstrate that fine-tuning could provide an effective method of learning representations of sentences containing MWEs. Other fine-tuning approaches for idiomatic MWEs in the literature include (Gamage et al., 2022), who propose an idiom detection model that also uses BERT (specifically DistilBERT) and additionally uses a token classification approach to fine-tune the pre-trained model.

3 Data

One of the key aspects of our experiment is the combination of several existing idiom datasets in order to form a more comprehensive dataset containing idioms, corresponding definitions, examples of the idioms in-context (i.e. sentences containing the idioms), and corresponding paraphrased non-idiomatic sentences with the same overall meaning. For our quasi-trasnfer learning task of fine-tuning idiom embeddings on either sentence/paraphrase data or idiom/definition data and exploring the effects on the other dataset, we found that existing idiom datasets did not contain the necessary information organized in the right manner, and so extensive data-processing was needed. In general, our dataset is a re-organized, extended version of specific components from the EPIE dataset (Saxena and Paul, 2020).

3.1 Idiom-Definition Dataset

One of the two key datasets we required for our task was a dataset mapping idioms to definitions. There are many existing datasets containing this information, and so we processed and combine three idiom oriented datasets to create an extensive corpus of idiom-definition pairs.

The EPIE dataset (Saxena and Paul, 2020) contains 717 unique idiomatic expressions labeled with definitions. Said idiomatic expressions come from two distributions: static and formal idioms. Static idiomatic expressions represent idioms that stay the same across instances while formal idiomatic expressions represent idioms which may undergo lexical changes depending on the context. For instance, the static idiom "never mind" never changes form between usages though the formal idiom "keep an eye open" may change to "keep your eye open" without losing meaning. EPIE contains 359 static idioms and 358 formal idioms. and both of these idiomatic types are utilized during training and testing are important to consider. In order to utilize the formal idioms, we converted the formal idioms to static idioms through a trial-and-error pronoun substitution. Formal idioms in EPIE contained $[pron]$ tokens to stand-in for situations with multiple possible pronouns. For each $[pron]$ token in the formal idiom set, we substituted it from a set of possible pronouns and validated the correctness of the idiom by cross-checking the idiom with the two non-EPIE datasets (described below). If a match is found, we can be confident that the formal idiom

with the given substituted pronoun is valid and can contribute to our overall idiom-definition dataset. The EPIE dataset also contains corresponding definitions for most of the static and formal idioms, obscured in a text file full of HTML formatting. We carefully extracted definitions from the definition text files using several sets of regular expressions, and manually verifying the outputs to ensure that the definitions matched-up. Our careful preprocessing steps guaranteed that all idiom-definition pairs added from EPIE were unique and valid.

The second dataset is the Idioment dataset provided by (Williams et al., 2015), containing 599 static idiom-definition pairs. The third dataset was found on Kaggle (Saied, 2021), an online resource that provides reliable datasets for machine learning applications. It contains 615 static idiom-definition pairs. These two separate datasets both provided valuable cross-reference material for the idioms in the EPIE dataset, while also allowing us to futher increase the size of our idiom-definition set with unique idiom-definition pairs not present in EPIE.

After the extensive preprocessing and combining these three datasets (taking care to remove any overlapping idiomatic expressions and/or idiomatic expressions with no associated definition), we were left with 1,349 unique idiom-definition pairs for use in our experiment.

To better leverage this dataset in triplet-loss training, we also gather additional "incorrect" definitions[1]. To do so, we gather definitions from (Data, 2021) which provides a complete dictionary of 176,010 definitions. To purge degenerate definitions (too small or large), we remove any definition less than 30 characters or more than 200 characters, resulting 112,062 random word definitions. During training time (specifically for the triplet-loss scheme), for a given idiom-description example we randomly sample from these 112,062 definitions to get an incorrect definition for a given idiom. This broad dictionary dataset is also used in part of the evaluation, to establish an (ideally low) baseline of similarity between idioms and incorrect definitions with which to compare the relative similarity strength of idioms and correct definitions under a certain embedding model.

[1]We can use these incorrect descriptions to push idiom embeddings away from embeddings of incorrect definitions/phrases. For more information on our triplet loss scheme, see Section 5.

3.2 Sentence-Paraphrase Dataset

In addition to an idiom-definition dataset, a sentence-containing-idiom/paraphrased-sentence-without-idiom dataset is also crucial. Much like the idiom-definition dataset generated above, we generate a dataset of sentence-paraphrase pairs starting from the the EPIE dataset (Saxena and Paul, 2020). , the 'sentence' in the sentence-paraphrase pair refers to an example sentence containing some idiomatic expression, and the 'paraphrase' in the pair is a sentence with similar meaning but with the idiomatic expression paraphrased to its actual non-idiomatic meaning. For instance, consider the usage of the idiom 'keeping tabs on' in the example sentence 'Keeping tabs on the motor trade'. An alternative paraphrasing may be 'Routinely monitoring the motor trade', thus completing the sentence-paraphrase pair.

For the formal idioms in the EPIE dataset, 3,136 sentence-paraphrase pairs already exist. For static idioms, 21,892 idiomatic example sentences exist, but corresponding paraphrased sentences don't. Thus, to include static idioms in this dataset, we can generate paraphrased sentences by processing examples for each static idiom and complete the pairs. Within EPIE we preprocess each example sentence by finding the underlying idiom and its location. Using the idiom-description dataset generated above, we can then replace each idiom with its definition, generating an alternative wording. Though some direct insertions of definitions may not be elegantly worded, we found most insertions to make general sense. By combining both formal and static sentence-paraphrase pairs, we are left with a dataset of 25,028 sentence-paraphrase pairs. Our training pipeline consisted of an 80/20 train/test split amongst both the sentence-paraphrase and idiom-description dataset.

4 Model

4.1 Sentence Transformer

Our experimentation revolves around a pre-trained sentence transformer model (Reimers and Gurevych, 2019), and specifically the "all-MiniLM-L6-v2" pretrained model found on Hugging Face (Face, 2023). This model encodes sentences/phrases into a 384-dimensional embedding space, a vector that captures the semantic information of the encoded text, and which can be used for similarity tasks such as ours. We apply the pre-trained sentence transformer (both with and

without pre-training) to both our datasets (idiom-definition and sentence-paraphrase), and we can evaluate the resulting embeddings using with similarity metrics [2] to determine how effective the model is at encoding the meaning of idiomatic material.

4.2 Loss Schemes

To fine-tune the pre-trained sentence transformer model in the hopes of improving performance on similarity metrics, we consider three different loss functions that can be used to train the model on our datasets. The first loss we employ is contrastive loss (with cosine similarity as the distance metric). With two inputs that are either the same label (e.g. idiom and correct definition) or a different label (e.g. sentence and incorrect paraphrase), contrastive loss seeks to either reduce the distance between the resulting embeddings (if they are the same label) or increase the distance (if they are different labels). When employed on our datasets, it should ideally bring the embeddings of idioms closer to their correct definitions and the embedding of idiomatic sentences closer to their correct paraphrases, while distancing them from incorrect definitions/paraphrases.

Similar to contrastive loss, triplet loss also seeks to reduce embedding distance between similar inputs and to increase distance between dissimilar inputs. The difference is that triplet loss receives three inputs: an anchor, a positive example, and a negative example. With the resulting embeddings, triplet loss minimizes the distance between the anchor and the positve example, while maximizing the distance between the anchor and the negative example. In a sense, this is like performing both a positive and negative contrastive loss step simultaneously.

The final loss function we utilize is cosine similarity loss, which (like contrastive loss) includes two inputs, along with along with a "ground truth" similarity score against which the two resulting embeddings are compared. Instead of maximizing or minimizing the distances of embeddings relative to each other, this loss scheme seeks to come as close as possible to the label similarity provided for each example pair during training. In our case, for example, when we pass in an idiom-definition where the definition is correct, we assign this a gold label similarity of 1 since we would like the model to encode the idiom in a similar manner to the definition. We understand that a label of 1 (perfect similarity) may not be the most accurate measure of the true similarity in this situation, but it serves the purpose of encouraging the model to generate embeddings for idioms that are more similar to their definitions, and to generate embeddings for idiomatic sentences that are more similar to their corresponding paraphrases.

[2]For most of our project we used cosine similarity, though it is worth mentioning the variety of other potential similarity metrics which exist such as Euclidean distance or TS-SS.

5 Methods

5.1 Baseline

Our baseline model, as described in Section 4, is a pre-trained "all-MiniLM-L6-v2" sentence transformer from Hugging Face with no fine-tuning applied. Applying this model (which is designed for encoding semantics and to be used in semantic similarity tasks) without fine-tuning will allow us to get an idea for how accurately idioms are represented sans additional training. We apply this model on idioms, definitions, idiomatic sentneces, and paraphrased sentences, and compare the idiom-defintion embeddings and the sentence-paraphrase embeddings using cosine similarity to establish a baseline prior to fine-tuning.

5.2 Fine-Tuning Experiments

Once our baseline is established, we explore various fine-tuning schemes using the loss functions described above on both the idiom/definition dataset and the idiomatic-sentence/paraphrase dataset, with the goal of improving embedding performance across the board even when only training on one of the datasets at time.

Given that the sentence transformer model is already pretrained on large amounts of text and posses existing information on semantics, we want to preserve as much of that as possible while still incorporating meaningful fine-tuning using the idiom data. For contrastive loss, we attempted two variations: one which included only positive examples (in which the only inputs were either correct idiom-definition pairs or correct sentence-paraphrase pairs), and another that included both positive and negative examples (with probability 0.5, instead of providing a correct pair, an random incorrect definition/paraphrase is inputted instead). We found that using only positive examples for contrastive loss generally resulted all embeddings becoming very similar to each other regardless of

correctness (since all inputs during fine-tuning are labeled as similar). Thus, we proceeded using both positive and negative examples, and results for this scheme are shown in Section 6.1.

For fine-tuning using triplet loss, our strategy slightly differs between our two datasets. For the idiom-definition dataset, we included a dictionary of random non-relevant definitions that can be used as negative examples, so for every triplet loss input, we can randomly sample a "negative" incorrect definition from the dictionary. This should prevent our embeddings from all converging to a similarity of close to 1, and preserve aspects of semantic embedding structure while still bringing idiom embeddings closer to the embeddings of their definitions. For sentence-paraphrase dataset, we were not able to find an adequate "negative" dataset from which we could sample negative examples, and so we instead randomly sampled a different (incorrect) paraphrase from the same sentence-paraphrase dataset. This would still push embeddings for idiomatic sentences closer to their paraphrased counterparts while distancing them from incorrect paraphrases.

For the cosine similarity loss (which ended up being quite similar to contrastive loss with a cosine similarity metric and so we did not apply cosine similarity loss to the larger sentence-paraphrase dataset due to time/resource constraints), we experimented with a similar strategy to contrastive loss, using all-positive examples with a gold similarity label of 1 or including incorrect examples with a gold similarity label of 0. As with contrastive loss, we found the including incorrect examples yielded better results, and so those are the ones we display in Section 6.1.

One other task we experiment with blurs the boundaries between the two curated datasets: fine-tuning on idiom/context sentences and evaluating on idiom/definitions. This is similar to the framework of a few-shot learning problem, and mimics a framework of contextual learning that might model how humans learn language. We see an unfamiliar or challenging phrase (the idiom), but do not have access to its precise definition. Instead, we observe the phrase in context, and use our knowledge of the contextual information combined with our broad general knowledge of language to decipher the meaning of a phrase. So, by taking a small random sample of 200 idiom/context sentence pairs, fine-tuning on the small training sample, and and evaluating the similarity of embeddings on the same idioms from the training sample and their unseen definitions, we can observe whether or not the fine-tuning allows the model to learn meanings/representations from a small amount of context with no explicitly provided meaning (either from direct definition or paraphrase).

I/D	S/P	I/Incor.D	S/Incor.P
0.272	0.869	0.108	0.075

Table 1: Average embedding cosine similarities of baseline pre-trained sentence transformer.

To keep results consistent across all our fine-tuning experiments, we trained for 10 epochs with a batch size of 16. We did not experiment much with hyperparameter tuning because the focus of our experiment is not so much to develop and optimal embedding model, but to establish patterns and understand the effects of certain fine-tuning schemes on embedding performance across multiple datasets, which are insights that may be important for future research on similar topics.

5.3 Metrics and Evaluation

Our primary evaluation for all models, both pre-trained baseline and fine-tuned, is cosine similarity, which is a standard metric for evaluating semantic similarity. This quantitative measure of similarity between embeddings will provide a useful, clear, and easily interpretable foundation of understanding.

It is also important that we evaluate our results in a qualitative manner. We can do so by examining individual examples of idioms/definitions and sentences/paraphrases to see how the model treats certain inputs, as well as by plotting embeddings using techniques such as t-SNE to visualize the broader structure of the data.

6 Results

6.1 Baseline

The average cosine similarity score of the baseline embeddings on I/D (Idiom/Definition) in Table 1 is 0.272, notably higher than the 0.108 average cosine similarity score of the baseline embeddings on I/Incor.D (Idiom/Incorrect Definition), so there is some reaction toward correct definitions. S/P (Sentence/Paraphrase) also shows significantly higher similarity compared to S/Incor.P (Sentence/Incorrect Paraphrase).

Loss	I/D	S/P	I/Incor.D	S/Incor.P
Contr.	0.476	0.895	0.287	0.199
Triplet	0.336	0.877	-0.102	0.077
CosSim	0.477	0.895	0.291	0.192

Table 2: Average embedding cosine similarities of sentence transformer fine-tuned on idiom-definition data.

6.2 Idiom/Definition Fine-tuning

Table 2 displays the average cosine similarity results across all the data of a sentence transformer fine-tuned in idiom-definition data (which is a much smaller dataset than the sentence/paraphrase set). As expected, the similarities for I/D are much higher than the baseline, as each of the loss functions would have directly pushed the idioms closer to their definitions. Interestingly, we do also observe a consistent slight increase in average consine similarity in S/P as well, potentially indicating that the increased similarity in I/D as a direct result of fine-tuning also has a slight indirect result in the increased similiarity between idiomatic sentences and correct paraphrases. However, we also do note that the similarity between idioms and incorrect definitions (I/Incor.D) also increased significantly for contrastive loss and cosine similarity loss, as well as did the similarity between sentences and incorrect paraphrases, indicating that generally all the embeddings may have been brought closer together.

Interestingly, for the triplet loss fine-tuning in Table 2, we see a promising effect on I/Incor.D and S/Incor.P, I/Incor.D decreasing to a slightly negative value and S/Incor.P remaining virtually the same is the baseline model. This might indicate a stronger preservation of existing semantic relationships, while still slightly improving the similarity between idioms/definitions and sentences/paraphrases.

6.3 Sentence/Paraphrase Fine-tuning

The effects of fine-tuning on the much larger idiom-sentence/paraphrase dataset was even more apparent (Table 3). The contrastive loss resulting on all the similarities nearing 1, indicating that all the embeddings, correct and incorrect, have moved very close to each other. While the similarity in I/D and S/P are extremely high at above 0.99, so are the similarities of I/Incor.D and S/Incor.P, essentially eliminating the nuanced semantic information present in the sentence embedding.

Loss	I/D	S/P	I/Incor.D	S/Incor.P
Contr.	0.992	0.999	0.983	0.994
Triplet	0.797	0.992	0.727	0.002

Table 3: Average embedding cosine similarities of sentence transformer fine-tuned on sentence-paraphrase data.

Loss	I/D	I/Incor.D
Baseline	0.297	0.107
Contr.	0.303	0.111
Triplet	0.308	0.104

Table 4: Embedding cosine similarities of sentence transformer fine-tuned on small sample of idiom-sentence data

Again, we see that triplet loss yields somewhat more promising results. We do observe (as expected) a very high similarity for S/P, and a near 0 similarity for S/Incor.P, which still preserves some semantic information and is better than the results from contrastive loss. While I/D is not as high as it was with contrastive loss, it is still much higher at 0.797 than the baseline at 0.272, indicating that the fine-tuning on sentences/paraphrases did have a notable effect on the embeddings of idioms/definitions. However, again we see that the similarity in I/Incor.D also rose significantly. It is still lower than I/D, but right around the same region at 0.727.

6.4 Idiom/Sentence Fine-tuning

Lastly, looking at Table 4, we see that our few-shot contextual learning task yielded underwhelming results. None of the results improved in any discernable way on the baseline, and on the case of contrastive loss, despite the slightly higher similarity scores in I/D, it is counteracted by a virtually equal undesired increase in similarity in I/Incor.D. We do see that the triplet loss scheme (again) appears to perform the best, with a slight increase in I/D and a slight decrease in I/Incor.D.

7 Analysis

7.1 Baseline

Our baseline results from Table 1 demonstrate that the pre-trained sentence transformer model, even without any additional fine-tuning, already posses

some degree of attunement to idiomatic meanings. The already higher similarity between idioms and their corresponding definitions as opposed to idioms and incorrect definitions indicates some sort of sensitivity toward idiomatic meaning, even if small. The high similarity between idiomatic sentences and paraphrases given the longer nature of context sentences and the large amount of other similar non-idiomatic words that the idiomatic sentence and paraphrase might share. The higher similarity in I/D might also be partially caused by more commonly shared words between idioms and their definitions, even if they are generally different (hence the relatively low similarity, but still higher than an incorrect definition).

Looking at the t-SNE plot in Figure 1, can generally see some vague denser clusters around the idioms and the definitions, with the sentences and paraphrase embeddings lining up closely but dispersed more sparsely throughout the scatter plot. This certainly is reflective of the high similarity between the sentences and paraphrases, and clearly reflects some disconnect between idioms and definitions, separated by a sparser gap between the two blue (idiom) and orange (definition) clusters.

7.2 Idiom/Definition Fine-tuning

The t-SNE plot in Figure 2 for the triplet-loss fine-tuned model on idioms/definitions reflects the slightly improved results of this model as seen in Table 2. We see the idiom sentences and paraphrases again lining up almost exactly, but spread out across the plot, with two denser clusters for idioms and definitions. However, we can clearly see in Figure 2 that the idiom and definition clusters are denser and closer together than they are in Figure 1. This reflects the (expected) closer similarity we observe between idiom and definition after fine-tuning. The slight improvements here are promising, especially considering they preserve more semantic meaning than the other two loss schemes (which essentially raised the similarity of everything).

However, looking at specific examples from a smaller set of idiom/definition/sentence/paraphrases, we see room for improvement. In the t-SNE plot in Figure 3 (corresponding idiom/definition/sentence/paraphrases labeled with the idiom), while some idioms are very close to their definitions (and sentences very close to paraphrases), there are regions of definitions very distant from idioms and regions of idioms very distant from definitions. Perhaps running the triplet model for additional epochs may help address this issue. With these plots, it may also be possible to observe specific relationships between idioms/definitions and idiom sentences/paraphrases, although the current plots don't appear to show any clear trends besides clustering. Future analysis of relative directions between embedding positions might yield insightful results.

7.3 Sentence/Paraphrase Fine-tuning

Given the already similar nature (as we observed in the baseline) if sentences/paraphrases, it is sensible that this fine-tuning task did not yield results as promising. While it did certainly show that fine-tuning on sentences/paraphrases strongly affects performance on idioms/definitions, it does so for the wrong reasons, eliminating semanting meaning and making all embeddings more similar rather than preserving existing relationships.

7.4 Idiom/Sentence Fine-tuning

More experimentation may be warranted here if the slight improvement yielded by the triplet-loss model consistently holds true over multiple samples becomes more apparent with more training epochs. However, it is likely that these slight differences may well be due to sampling differences (since the data for this task was a random sample of idioms/contexts rather than a full dataset, and so the results here are not conclusive. As we can see in the t-NSE plots in figures 1, 2, and 3, the nature of idiom embeddings and idiom-sentence embeddings might be fundamentally different, especially due to the many other words present in idiom-sentences besides the idiom on its own. While there may be possible exploration into correlations between idiom embeddings and context embeddings, we might conclude that attempting to bring the idiom embeddings closer to embeddings of sentences containing idioms might negatively affect the semantic representations.

8 Conclusion

Modern pre-trained sentence transformer models, even smaller models such as the "all-MiniLM-L6-v2" model used in this paper, have demonstrated promise at recognizing and representing surface level idiomatic expressions, though at a limited

level. Different fine-tuning schemes, as we observed with our different choices for loss functions, have a significant effect on the quality of embeddings. While not every fine-tuning sceheme will yield positive results, our experiments show that there is certainly room for exploring how specific methods (such as triplet loss) might improve the capability of pre-trained models to represent idioms, an extremely challenging and nuanced aspect of language. The results in this paper are limited, and there is little evidence that strongly confirms that fine-tuning in a manner similar to our experiment will substantially improve idiom representation. There is certainly more to be done, whether it is in terms of making our experiments more comprehensive or a more detailed qualitative analysis between idioms/definitions and sentences/paraphrases. But, ultimately, we believe that these experiments are necessary in order to move beyond the current state-of-the-art in natural language understanding. Being able to learn unfamiliar phrases from context, and to represent phrases with meanings that are greater than a sum of their parts without massive amounts of data will be valuable in the development of smaller, more powerful, and more personable language models.

Known Project Limitations

First and foremost, training a good embedding model requires a large amount of data that interacts with each other, and despite our efforts in collating multiple idiom datasets, our data is still rather limited and not representative of the full spectrum of language. Fine-tuning schemes on this limited data (as we observed in our results) results in either severe overfitting and/or minimal benefit to the desired task. In general, idiom understanding and idiom representation is an extremely challenging task, and our approach, while promising in certain aspects, is far too simplistic and straightforward to encompass the nuances of idiomatic language.

Moreover, due to the ever changing nature of language with a specific emphasis on the creation of new slang, a limitation our embedding model faces is a lack of exposure. Without an ability to dynamically train on future idioms our model will eventually become obsolete. Since our model leverages context to inform embeddings, we hope this mitigates the extent of this limitation.

One limit we did not explore in our project are the limits of the transfer learning capabilities provided by our embedding model. Having fine-tuned our sentence transformer to effectively embed idioms, we may have unknowingly decreased the efficacy of the embeddings for other, non-idiomatic phrases for use in alternative training contexts. Future work is warranted to mitigate and otherwise quantitatively look into the adverse affects of idiomatic fine-tuning, though we hypothesize that the extent of damage is rather negligible due to the comparatively small fine-tuning conducted.

A final limitation which dominates the ultimate performance of our model is the underlying similarity metric. Since we use similarity metrics as a guide to improve our embeddings, it has final say into the form of embeddings and the possible interactions between them. Though cosine similarity is widely used in the field for contrastive loss and for general purpose embedding discovery, perhaps more sophisticated similarity metrics may have resulted in more nuanced embedding derivations. Due to the complexity and context dependent nature of idioms, carefully walking this line is vital to creating an effective embedding architecture leading us to believe our model may have been limited by the choice of our similarity metric.

Authorship Statement

On the technical side: Jack Xiao took ownership of executing the fine-tuning experiments discussed in this paper and visualizing the results, Jack Michaels collected and pre-processed the datasets, and Avi Gupta implemented baseline models. On the authorship side: all authors collaborated across sections, though each author took the lead on a particular section. Jack Xiao led the Methods, Results, and Analysis; Jack Michaels led the Data, Model, Conclusion, and Known Project Limitations; and Avi Gupta led the Abstract, Introduction, Prior Literature, Conclusion, and Authorship sections.

References

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. HIT at SemEval-2022 task 2: Pre-trained language model for idioms detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 221–227, Seattle, United States. Association for Computational Linguistics.

DFY Data. 2021. The online plain text english dictionary (opted).

Hugging Face. 2023. Hugging face sentence transformer model.

Gihan Gamage, Daswin De Silva, Achini Adikari, and Damminda Alahakoon. 2022. A bert-based idiom detection model. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–5.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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*. Association for Computational Linguistics.

Soha Saied. 2021. English idioms.

Prateek Saxena and Soma Paul. 2020. EPIE dataset: A corpus for possible idiomatic expressions. *CoRR*, abs/2006.09479.

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

Tadej Skvorc, Polona Gantar, and Marko Robnik-Sikonja. 2020. MICE: mining idioms with contextual embeddings. *CoRR*, abs/2008.05759.

Minghuan Tan and Jing Jiang. 2020. A bert-based dual embedding model for chinese idiom prediction. *CoRR*, abs/2011.02378.

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.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

A Appendix: Embedding Visualizations

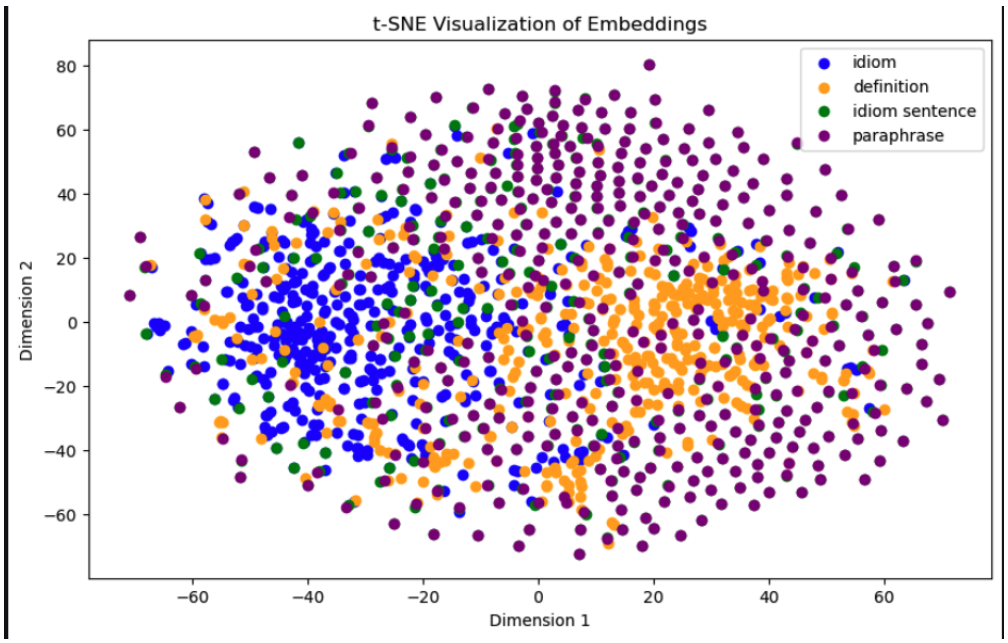

Figure 1: t-SNE plot of all embeddings, baseline model.

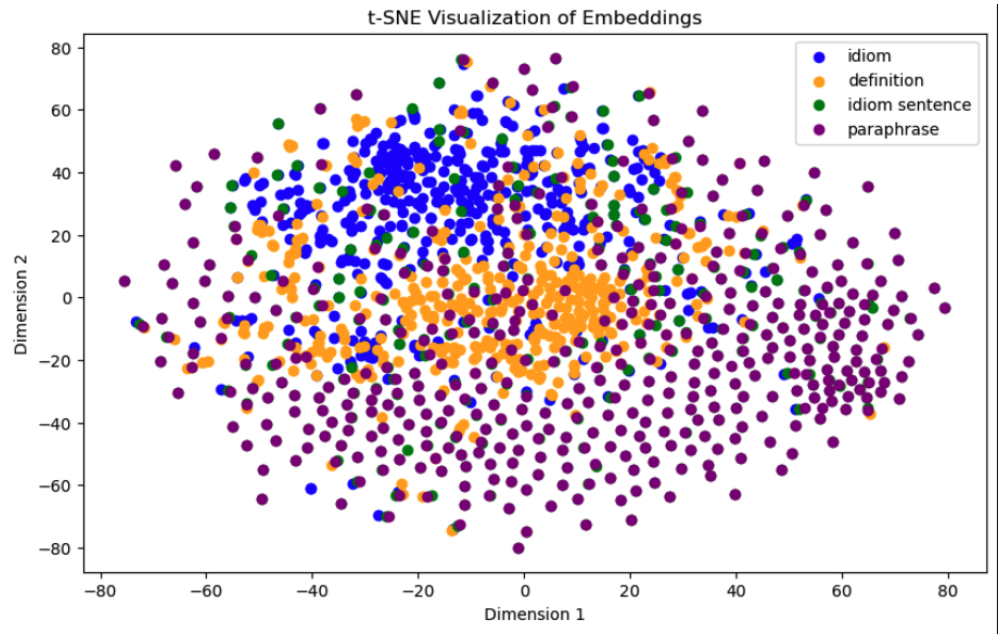

Figure 2: t-SNE plot of all embeddings, triplet-loss model on idiom/definitions.

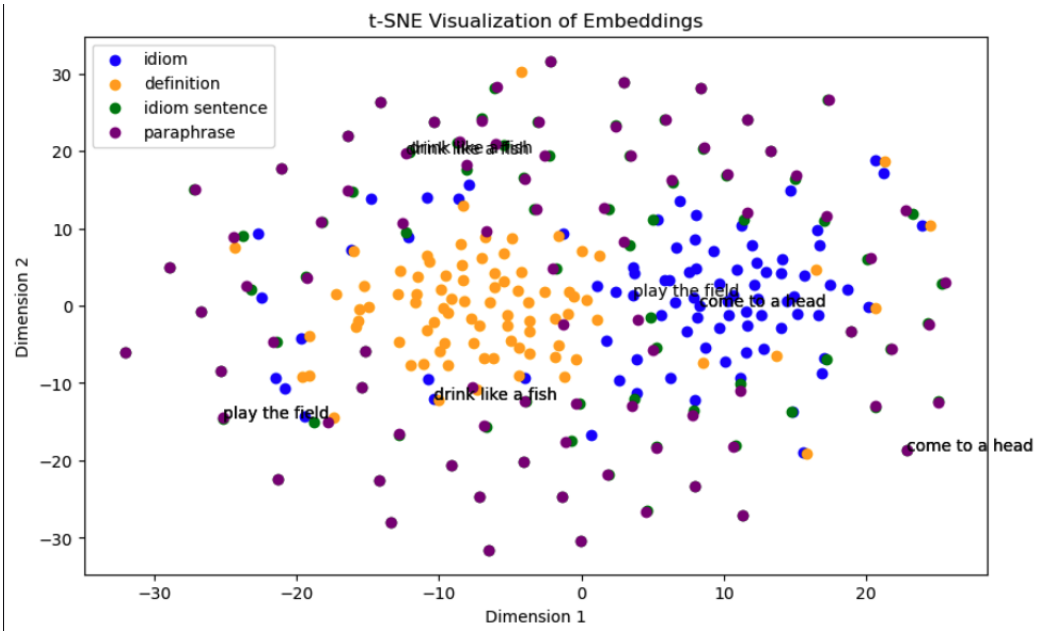

Figure 3: t-SNE plot of formal idioms, triplet-loss model on idiom/definitions.