Getting BART to Ride the Idiomatic Train: Learning to Represent Idiomatic Expressions

Ziheng Zeng and **Suma Bhat**

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Champaign, IL USA
{zzeng13, spbhat2}@illinois.edu

Abstract

Idiomatic expressions (IEs), characterized by their non-compositionality, are an important part of natural language. They have been a classical challenge to NLP, including pre-trained language models that drive today's state-of-the-art. Prior work has identified deficiencies in their contextualized representation stemming from the underlying compositional paradigm of representation. In this work, we take a first-principles approach to build idiomaticity into BART using an adapter as a lightweight non-compositional language expert trained on idiomatic sentences. The improved capability over baselines (e.g., BART) is seen via intrinsic and extrinsic methods, where idiom embeddings score 0.19 points higher in homogeneity score for embedding clustering, and up to 25% higher sequence accuracy on the idiom processing tasks of IE sense disambiguation and span detection.

1 Introduction

Natural language has a common yet special class of multi-word expressions (MWEs) called idiomatic expressions (IEs) that exhibit *semantic non-compositionality*, where the meaning of the expression cannot be inferred from that of its constituent words (e.g., the idiom *break a leg*) (Baldwin and Kim, 2010). They are commonly used for specific communicative intents (Moon, 1998; Baldwin and Kim, 2010) and are individually rare but collectively frequent, appearing frequently across genres (Moon, 1998; Haagsma et al., 2020). They have been classically regarded as a ''pain in the neck'' to NLP systems (Sag et al., 2002) not only because of their non-compositionality, but also because of their contextual semantic ambiguity (used in idiomatic or literal meaning depending on the context). Challenges posed by the presence of IEs have been identified across multiple NLU tasks even with state-of-the-art (SOTA) solutions, including sentiment analysis (Liu et al., 2017; Biddle et al., 2020), paraphrase generation (Zhou et al., 2021), natural language inference (Chakrabarty et al., 2021), and dialog models (Jhamtani et al., 2021).

Even the the flagship NLP model GPT-3 (Brown et al., 2020) finds idioms challenging. We tested for its idiom comprehension over 75 idioms, covering a spectrum of the most to the least frequent idioms (based on their frequency of occurrence in the BNC (Haagsma et al., 2020)). We do this in question-answering mode where we ask GPT-3[1] simple questions about idioms that can only be answered correctly if their meanings are understood in context (a sample of questions and answers are shown in Table 1) and noted 37 incorrect answers (50.7% accuracy), showing its inability to comprehend many idioms.

Despite the relative success of recent studies using pre-trained language models (PTLMs) for IE processing, including idiom token and type classification (Liu, 2019), IE span detection (Zeng and Bhat, 2021; Škvorc et al., 2022), and idiomatic to literal paraphrasing (Zhou et al., 2021), the emerging knowledge is that PTLMs are unable to produce meaningful representation of IEs in task-specific contexts (e.g., for IE sense disambiguation) (Tayyar Madabushi et al., 2021). Noting that solutions to improve IE representation in a task-agnostic manner are largely missing, this study aims to develop targeted solutions to make language models (LMs) idiom-aware with the immediate objective of improving IE representation in large PTLMs yet without relying on

[1] We used OpenAI's playground at https://gpt3demo.com/.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1120–1137, 2022. https://doi.org/10.1162/tacl_a_00510
Action Editor: Doug Downey. Submission batch: 5/2022; Revision batch: 7/2022; Published 10/2022.

Idiom	**Question**	**Answer by GPT-3**
Never say die	If I have a never say die attitude, would I run the marathon injured or forfeit?	I would never run a marathon injured.
All at sea	If I am all at sea with my math assignment, am I making progress or am I lost?	You are making progress.
Every trick in the book	If I use every trick in the book to guarantee my grade, am I a magician or a determined student?	A magician.
Kill two birds in one stone	If I wanted to kill two birds in one stone, what kind of a workplace should I work in?	A slaughterhouse.
Ahead of the game	If I want to be ahead of the game, would I study early or procrastinate?	Procrastinate.

Table 1: Questions used to probe the understanding of idioms and the answers provided by GPT-3.

astronomically large training corpora and parameters. In this work, we focus on improving IE representation and seek answers to the following questions about IE embeddings.

Q1: *Are the current PTLMs capable of generating semantically meaningful contextualized representations for IEs?* Answering this question, we examine the IEs embeddings produced by a representative SOTA LM, for example, BART (Lewis et al., 2020). Specifically, we perform an intrinsic evaluation of IE embeddings by grouping them into semantic classes and observing how they cluster.

Observing the low quality of IE representation, we ask,

Q2: *How can we expand the capability of these LMs to produce high quality IE embeddings?* As a solution, we propose the **G**eneration of **I**diom **E**mbedding with **A**dapter (GIEA) approach that extends the capabilities of the current SOTA LMs by producing quality IE embeddings.

Concretely, unlike prior work that treats each idiom as a new token (Hashempour and Villavicencio, 2020), GIEA refrains from new tokenization to represent IEs and uses an adapter (Houlsby et al., 2019; Pfeiffer et al., 2020a) as a parameter-constrained learner of IE embeddings. Finally, we devise a denoising auto-encoder-style learning objective and train the network to reconstruct selective masked sentence parts. Our use of symbolic knowledge (Yu et al., 2021) of IEs to aid the learning of their embeddings results in the model needing a significantly small amount of data ($\sim$60MB) compared to that required for LM pre-training ($\sim$160GB of text for BART).

Our main contributions are as follows.

(1) We demonstrate the limited ability of SOTA PTLMs for generating semantically meaningful embeddings for IEs via a simple probing task.

(2) We propose a lightweight solution, GIEA, an adapter built over BART, to produce quality IE embeddings without altering input sentences.

(3) We evaluate the resulting IE embeddings using intrinsic and extrinsic methods to show that they are meaningful in the embedding space and are task-agnostic and generalizable across different idiom processing tasks (IE sense disambiguation and IE span detection). Compared to BART, GIEA gains 0.19 in homogeneity score (intrinsic evaluation), performs competitively on IE sense disambiguation, and gains 25% in sequence accuracy for IE span detection.

(4) We conduct detailed analyses on the performance and limitations of GIEA system to provide meaningful insights and future directions.[2]

2 The Inability to Represent Idiomatic Expressions

Compositionality is a dominant paradigm driving the SOTA in NLP both at the tokenization and architectural levels. The tokenization of most LMs, for example, Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and WordPiece (Wu et al., 2016), assumes compositionality not only

[2] The code for GIEA framework can be found at `https://github.com/zzeng13/GIEA`.

Idiom	BART	ITI+SF+SI
in the final analysis	in the long run in the works in light of	at the end of the day in light of all things being equal
see red	see the light see stars go down like a lead balloon	go spare fly off the handle do someone's head in
quick as a flash	flash in the pan keen as mustard thin as a rake	in the blink of an eye like a bat out of hell thick and fast

Table 2: The top-3 closest idioms ranked by cosine similarity by IE embeddings generated by BART and ITI+SF+SI (our GIEA method). While the IE embeddings from GIEA are grouped by semantic meaning, BART's IE embeddings are grouped together mostly by surface-level token and/or syntactic similarity.

at the phrase-level but also at the word level. This suggests that the meaning of a word is deduced from that of the subword components. At the architectural level, transformer-based LMs implicitly consider all phrases (or even words) as compositional. The self-attention mechanism in transformers considers the embedding of a word to be an attention weighted sum of the word embeddings in its context. This design leads to phrase or even sentence embeddings to be overall compositional. In addition, each IE is individually rare, compounding the difficulty for obtaining good IE representation. This leads us to hypothesize that the inherent notion of compositionality and the rarity of IEs are a hindrance to the representation of the IEs that are inherently non-compositional. We test the validity of this hypothesis by analyzing PTLMs' representation of IEs.

IE Embedding Generation: We first obtain the embeddings for the IEs in the MAGPIE dataset (Haagsma et al., 2020), a collection of potentially idiomatic expressions (PIEs), that is, idioms used in a literal and idiomatic sense, and the sentences in which they occur. Focusing on the IEs used idiomatically (thus ensuring their non-compositionality), we first retrieve all the sentences in which they occur. Then, for each sentence, we extract the BART base embeddings corresponding to the IE tokens in the sentence. We then apply mean pooling across the tokens and across all the sentences in which the IE appears. In this manner we generate the embeddings for 1,480 idioms from an average of 22 sentences per idiom.

Group	Idioms
Success	home and dry; bear fruit; hit the mark
Quick	in two shakes; full tilt; quick as a flash
Death	kick the bucket; drop like flies
Happy	on cloud nine; over the moon; ride high

Table 3: Example meaning groups and sampled idioms from the groups.

We then list IEs most similar to a set of IEs in the embedding space produced by the base BART model, computed using the cosine similarity. Table 2 shows examples of this listing including three most similar IEs (second column) to a sample of IEs (first column). As noted from the examples, IEs with superficial token-level (*see red* vs. *see stars*) and/or syntactic-level (*quick as a flash* vs. *keen as mustard*) matches tend to be most similar according to BART's embeddings without accounting for their semantic congruence. This suggests that BART considers IEs mostly compositionally, an inadequate approach for representing the non-compositionality of the IEs.

Synonymous IE Groups Creation: To quantify the above qualitative finding, we manually assigned 129 idioms into 20 distinct meaning groups—''in summary'', ''anger/upset'', ''easy/relax'', ''quick'', ''exactly'', ''death'', ''punish/criticize'', ''impress'', ''happy'', ''to understand'', ''fail'', ''success'', ''close to'', ''decline/worsen'', ''grief/sad'', ''confront/deal with'', ''persevere'', ''great effort'', ''unimportant'', ''careful''—averaging 6.4 idioms per group (see Table 3 for example groups and their idioms).

The idiom groups must satisfy the following two requirements: (1) Any two idioms from the same group must have a similar meaning though the idioms may not necessarily be interchangeable; and (2) any two idioms from different groups must not overlap in their meanings, that is, the boundaries between any groups should be clear. Moreover, we selected idioms that are idiomatically monosemous (excluding their literal interpretations) according to our dictionaries.[3] To group the idioms, we first created a few candidate groups based on commonly occurring idiom meanings, such as ''anger/upset'' and ''happy''. Then, for each idiom we either assigned it to an existing group or to a newly created meaning group. We only retained groups with more than three idioms and stopped the process once we had 20 groups. Using the aforementioned requirements, the validity of the groups and the idiom assignments were verified by two annotators, one with native and the other with near-native English abilities (one of whom was not associated with this study), using an idiom dictionary as needed. Only idiom assignments that were judged as correct by both the annotators were considered.

Clustering Embeddings: First, we generate the embeddings for these idioms based on their dictionary definitions using a pre-trained MPNet[4] (Song et al., 2020) for sentence embeddings, referred to as *definition embeddings*. As a contrast, we generate their BART IE embeddings, referred to as *BART embeddings*, following the procedure discussed above. Then, we run agglomerative clustering[5] to produce 20 clusters with complete linkage using the pairwise cosine similarity between the embeddings (definition and BART embeddings separately) as the distance metric. Finally, we measure the clustering quality using the homogeneity score as an index of the embedding quality, which is 1.0 if all the clusters contain only data points that are members of a single class. The homogeneity score for definition embeddings is 0.68, whereas the score for BART embedding is only 0.45. This suggests that BART embeddings are more scattered in the embedding space with less than half of the IEs from each cluster having the same meaning.

3 Learning Representation for Idiomatic Expressions

Toward producing higher quality IE embeddings by PTLMs, we propose GIEA; given a set of idiomatic sentences (i.e., sentences that each contains an IE), GIEA freezes the base PTLM and trains an adapter that specializes in IE representations. This is done by reconstructing idiomatic sentences that are corrupted with an idiom-aware noising function and meeting a dictionary definition-aided objective. GIEA's overall framework is illustrated in Figure 1. In this work, we select BART as our base PTLM.

Noising Function. Following the pre-training for BART, our training has a *text corruption* stage with novel noising functions and a *text reconstruction* stage. In the text corruption stage, we introduce three noising functions such that one permits predicting masked IEs using the context words—the *idiom-aware text infilling* transformation—and the other two permit the model to use IEs to predict context words, namely, the *copy* and the *span infilling* transformation. In the idiom-aware text infilling transformation, given a sentence containing an IE, the entire IE is replaced with a single `[MASK]` token. During training, the model is asked to reconstruct the masked IE using the context words. Yet the masking of IEs alone is not sufficient for learning meaningful IE embeddings because the model sees IEs only in the decoder's input but never in the input sentences, leaving the encoder's adapter parameters unreachable by the reconstruction loss.

The two additional noising functions, the *copy* and the *span infilling* transformation, alleviate this shortcoming by allowing the model to learn to use IEs to infer the context words. In the *copy* transformation, for each sentence with its IE masked, we also supply its original, uncorrupted sentence as input and thus the model only has to copy the input sentence to the output. In the *span infilling* transformation, we mask a span of consecutive tokens *excluding* the IE tokens with a single `[MASK]`, effectively asking the model to reconstruct the masked span using the IE and the remaining context. As in BART pre-training, span lengths are drawn from a Poisson distribution ($\lambda = 3$). However, our 0-length spans correspond

[3] The definitions were obtained from the Google dictionary and Wiktionary. The idiom groups can be retrieved from `https://bit.ly/3R2pomM`.

[4] The checkpoint used is ''all-mpnet-base-v2'' from Sentence-Transformers (Reimers and Gurevych, 2019).

[5] Implemented by Scikit-Learn (Pedregosa et al., 2011).

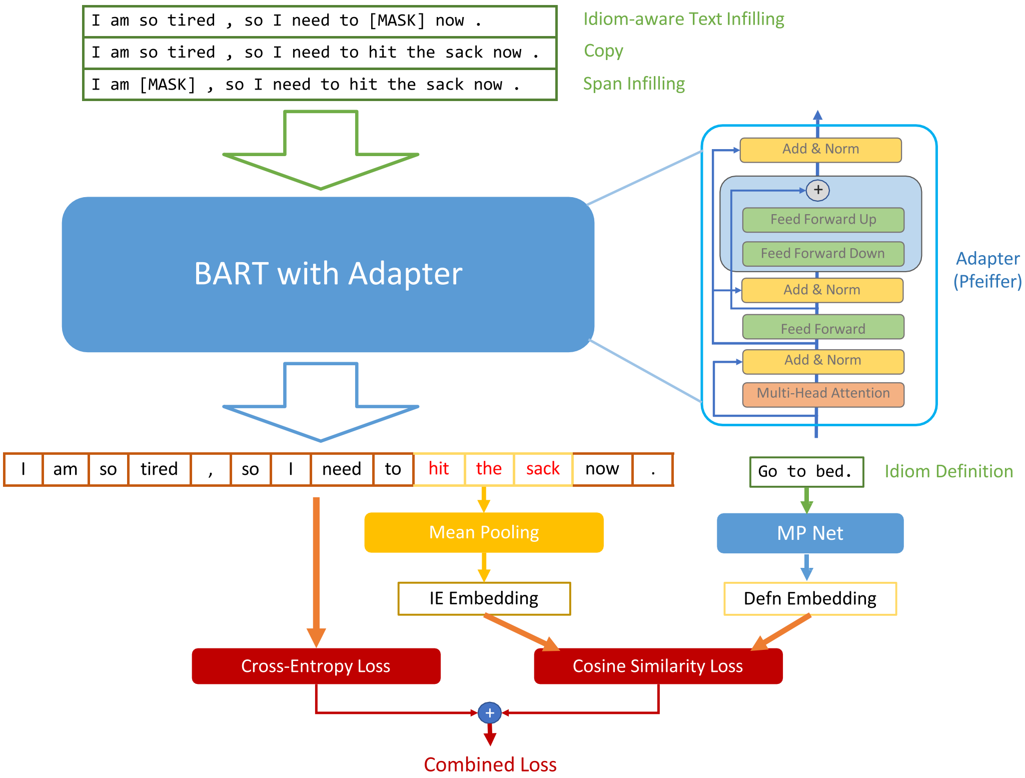

Figure 1: Overview of the GIEA training framework.

to the original (input) sentence, identical to that of the *copy* transformation. Hence, the span infilling technically subsumes the copy transformation.

Ideally, we would like the model to use an IE to predict masked context words that are directly related to the meaning of the IE. For example, as shown in Figure 1, masking the sequence ‘‘so tired’’ helps the learning of the IE, ‘‘hit the sack’’. However, since the masked spans are randomly chosen, to guarantee that reconstructing the masked spans contributes to the IE meaning acquisition and inspired by prior success in prompting methods (Liu et al., 2021), we inject manually created templates for span infilling (e.g., *When people say hit the sack, they mean that* `[MASK]`*.*) by connecting each IE to its dictionary definition as a sentence. We create four such templates per idiom with variations.[6]

During training, that is, the reconstruction stage, we randomly apply the *idiom-aware text infilling* transformation to 50% of sentences, while applying the *copy* or *span infilling* transformation to the remaining sentences in each epoch, and the model is asked to reconstruct the uncorrupted sentences. We experiment with and analyze the use of both the *copy* and *span infilling* in Section 5.

Similarity Forcing. We leverage the dictionary definitions of IEs to aid the learning of semantically rich IE embeddings and supplement the small number of idiomatic sentences. To give an idea of the relative paucity of available idiomatic sentences, the number of idiomatic sentences in MAGPIE, the largest dataset for idiomatic sentences to date, is less than 30K, which is several orders of magnitude smaller than the BART pre-training corpus. Although collecting more sentences with IEs from other corpora is a way to directly enlarge the existing collection, isolating the truly idiomatic instances of potentially idiomatic expressions requires manual annotation, an exercise that we leave for future work.

Specifically, during training, we use MPNet to generate definition embeddings for each IE as before. MPNet is used because it empirically outperforms BART, as we will show in Section 5. We also generate IE embeddings by mean pooling

[6] The templates for a given [IE] are:

(1) ‘‘The idiom [IE] means `[MASK]`.’’,

(2) ‘‘When people say [IE] , they mean `[MASK]`.’’,

(3) ‘‘[IE] is used to mean `[MASK]`.’’,

(4) ‘‘If someone says [IE] , they mean that `[MASK]`.’’

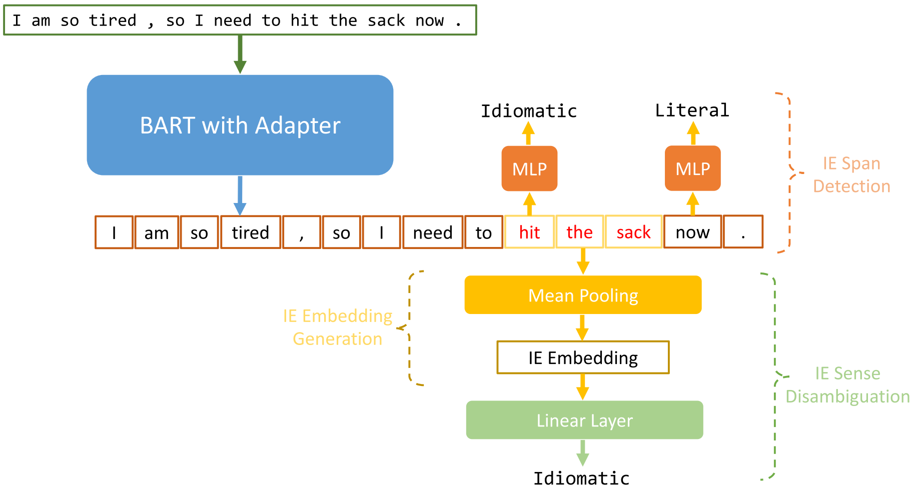

Figure 2: Illustration of the intrinsic and extrinsic evaluation tasks, including the generation of IE embeddings, IE sense disambiguation, and IE span detection.

the BART's final layer output embeddings corresponding to the IE tokens. Note that these IE embeddings are generated from BART and the adapter being trained and thus correspond to a non-compositional representation. We then include the learning objective of increasing the cosine similarity between the IE embeddings and their corresponding definition embeddings. We refer to this learning objective as *similarity forcing*, which is intended to facilitate the learning of the IE embeddings by making the embedding space be more semantically meaningful, that is, locating IEs with similar meanings closer to each other.

The final loss during training is the weighted sum of the cross-entropy loss from reconstruction and the cosine similarity loss from similarity forcing. In our experiments, we set the two losses to be equally weighted and leave other weighting schemes for future explorations.

Non-Compositional Language Adapter. Instead of fine-tuning the full model on our new learning objective, we added an adapter with the Pfeiffer architecture (Pfeiffer et al., 2020a) to the base BART model for conditional generation. This is so that during training, only the parameters of the adapter are trainable while those of the underlying language model are fixed, thus making our solution lightweight. Intuitively, because the added adapter is trained with the added objective of producing meaningful embeddings for non-compositional phrases (IEs), the adapter can be considered to be an expert in processing non-compositional language.

4 Experiments

Datasets. We use MAGPIE (Haagsma et al., 2020), a recent and the largest-to-date dataset of potentially idiomatic expressions in English, to train GIEA and evaluate the baseline models. We sample a subset of the dataset by selecting idioms with a single idiomatic meaning according to our IE dictionary (referencing Google dictionary and Wiktionary) and their corresponding sentences that are unambiguously labeled as being idiomatic (indicated by a perfect confidence score). The resulting collection has sentences drawn from a diverse set of genres from the British National Corpus (BNC) with 1,480 idioms with 32,693 sentences (77.4% idiomatic) in the train set and 1,001 idioms with 4,102 (77.57% idiomatic) sentences in the test set.

Evaluation Tasks. The overview of the intrinsic and extrinsic evaluation tasks are illustrated in Figure 2. The first task is an intrinsic evaluation of IE embeddings.

Embedding Clustering. We follow the same procedure as described in Section 2 to perform clustering on the 20 distinct idiom groups with IE embeddings from the testing models. Note that we only use the sentences from the test set here to generate the IE embeddings. We use agglomerative clustering with complete linkage and pairwise embedding cosine similarity as the affinity metric.

The following two idiom-related tasks serve as extrinsic evaluations of the IE embeddings.

IE Sense Disambiguation. This is a common probing task used to probe if IE embeddings can differentiate the literal (compositional) from the idiomatic (non-compositional) uses of the IEs (Tayyar Madabushi et al., 2021; Adewumi et al., 2021). Many IEs can be used both figuratively or literally depending on the context. For example, the phrase ‘‘behind closed doors’’ can be interpreted literally as in *The valuable items are locked behind closed doors* and can be understood figuratively as in *They avoided any publicity and made all deals behind closed doors*. To account for this contextual ambiguity, these phrases are often refer to as potentially idiomatic expressions (PIEs) (Haagsma et al., 2020). The IE sense disambiguation task aims to classify each IE usage into idiomatic and literal class. To create a disambiguation classifier, we appended a single linear layer after the trained baseline embedding model. Given a sentence with a PIE and the location of the tokens belonging to the PIE, the baseline embedding model generates the embeddings for every token in the sentence. Then, the token embeddings corresponding to the PIE are mean pooled and fed to the linear layer to generate a binary classification. Only the linear layer is trainable when training the classifier. Given that nearly 78% percent of IEs are used figuratively in MAGPIE test data, the majority-class baseline predicts *idiomatic* label for all instances.

IE Span Detection. This is a more demanding task compared to IE sense disambiguation and studies focusing on this task are only emerging (Zeng and Bhat, 2021). Given a sentence with a PIE, a model is expected to classify every token as *idiomatic* or *literal*; when the PIE is used idiomatically, the tokens from the PIE will be tagged as idiomatic; when the PIE is used literally, all its tokens will be tagged as literal. To succeed in this task, a model must identify the presence of an IE and then precisely predict its boundary. To create such a classifier, we append a two-layer MLP that reduces the number of hidden neurons by a factor of 2 after each layer and uses ReLU activation between the layers. Only the MLP is trainable. Because the tokens are overwhelmingly literal, the majority-class baseline predicts each token to be *literal*.

Note that for both the above tasks, more powerful classifiers exist, as shown in prior works (Liu and Hwa, 2017, 2019; Zeng and Bhat, 2021; Škvorc et al., 2022). However, we deliberately constrain the complexity of the classifiers to linear layers (or MLPs) to ensure the performance differences reflect primarily the effect of different IE embeddings rather than that of additional modeling.

Evaluation Metrics. For intrinsic evaluation, that is, the embedding clustering task, we evaluate performance using *homogeneity score* to evaluate the clustering quality. Given that two idioms from different groups should have distinct meanings, we also measure the *mean cosine distance* between the embeddings for IEs from different groups; the larger the distance the better. For the IE sense disambiguation task, because it is a binary classification problem, we use *accuracy* and *F1* score to evaluate the performance. For IE span detection, given that this is a sequence tagging task, we use three evaluation metrics, namely, *sequence accuracy*, *token-level recall* score, and *token-level accuracy*. In sequence accuracy, an instance is considered as correct if and only if all the tokens in the sequence are tagged correctly, making this the strictest metric. However, by only considering sequence accuracy, one may underestimate the performance of models that can tag most of the tokens from the positive (idiomatic) class correctly. Hence, we also consider the token-level recall and the accuracy score to complement the strict sequence accuracy metric. For token-level recall and accuracy, we compute the recall and accuracy for each predicted sequence and the final scores are averaged across all sequences.

Baseline Models. Due to the lack of directly related prior work, we include only the majority-class baseline, BART, and variations of GIEA to demonstrate the effect of different components of our method detailed below.

Majority-class is a naïve baseline that chooses the majority class for any classification problem.

BART is the original pre-trained BART-base model.

BART-FT is the fine-tuned full pre-trained BART-base model using dictionary definition template sentences mentioned in Section 3 in addition to the MAGPIE train data with the idiom-aware text infilling and span infilling objective.

Idiom-aware Text Infilling (ITI) Model is a baseline that trains the adapter with only the idiom-aware text infilling transformation.

Idiom-aware Text Infilling + Span Infilling Model (ITI+SI) is a baseline that trains the adapter with both the idiom-aware text infilling and span infilling transformations.

Idiom-aware Text Infilling + Similarity Forcing (ITI+SF) is GIEA that trains the adapter with the idiom-aware text infilling transformation and similarity forcing learning objective.

Our Models. We include two competing versions of GIEA using different noising functions:

Idiom-aware Text Infilling + Similarity Forcing + Copy Model (ITI+SF+Copy) is GIEA that trains the adapter with the similarity forcing objective and both the idiom-aware text infilling and copy transformations.

Idiom-aware Text Infilling + Similarity Forcing + Span Infilling Model (ITI+SF+SI) is GIEA that trains the adapter with the similarity forcing objective and both the idiom-aware text infilling and span infilling transformations.

Experimental Setup. For the adapters in all baseline models, our adapter implementation is based on Pfeiffer et al. (2020a). The BART-base model is implemented and maintained by Huggingface (Wolf et al., 2020). The definition embeddings are generated by an MPNet hosted and maintained by the Sentence-Transformers package (Reimers and Gurevych, 2019). For the adapters, we trained all baseline GIEA models for 220 epochs with a batch size of 16. We trained a set of IE sense disambiguation and IE span detection classifiers for each baseline model except for the majority-class baseline. For IE sense disambiguation, we trained the classifier for 55 epochs with a batch size of 32 and for IE span detection, we trained it for 100 epochs with a batch size of 16. The linear layer and the MLP in the respective classifiers were trained with a dropout rate of 0.2. For all training, we used the Adam optimizer with a learning rate of 1e-5. For all models, checkpoints with the best validation performances were used in the experiments. All the other hyperparameters were in their default values. We only use MAGPIE's idiomatic sentences to train GIEA and the baseline models, but we use both the idiomatic and the literal sentences to train the probing models for evaluation.

Method	**Score** (Norm.)	**Dist.** (Norm.)
BART	0.4546 (0.0)	0.0379 (0.0)
BART-FT	0.4659 (4.97)	0.0681 (14.99)
ITI	0.4597 (2.26)	0.0397 (0.876)
ITI+SI	0.4483 (−2.76)	0.0514 (6.71)
ITI+SF	0.4357 (−8.31)	0.0411 (1.64)
ITI+SF+Copy	0.5906 (59.92)	0.1980 (79.47)
ITI+SF+SI	0.6450 (83.86)	0.2284 (94.54)
Definition	0.6816 (100.0)	0.2394 (100.0)

Table 4: Results of intrinsic evaluation via clustering. *Score* is the homogeneity scores. *Dist.* is the averaged cosine distance between idioms from different groups. Values are normalized (*Norm.*) using BART and Definition embeddings are used as lower and upper bound. Higher values are better.

5 Results and Analyses

Intrinsic Evaluation. One of the defining characteristics of a good representation is that the embedding space should be semantically meaningful, that is, the embeddings of similar meaning IEs should be closer to each other in the embedding space via some distance metric (e.g., cosine similarity). As shown in Table 2, it is clear that after training with our ITI+SF+SI objective, the IE embeddings no longer cluster based on mere superficial similarities, instead, their meaning is the driving factor in determining their proximity in the embedding space. As shown in Table 4, the ITI+SF+SI method achieves the best homogeneity score and is significantly higher than the original BART embeddings by 0.19. Also, the mean cosine distance between the embeddings for the IEs from different meaning groups is merely 0.0379 for BART, indicating the BART embeddings are inadequate in discriminating between meanings; yet, the averaged distance is 0.2284 for ITI+SF+SI, which is very close to the distance of 0.2394 by the definition embeddings. To provide a more direct comparison, we also normalized the baseline performances using the BART embedding score as the lower bound and the definition embedding score as the upper bound. Comparing ITI+SF+SI and ITI+SF+Copy reveals that the more sophisticated SI noising function enabled the model to learn an embedding space that is

Model	Disambiguation		Span Detection		
	F1	**Acc**	**Seq Acc**	**Tkn Recall**	**Tkn Acc**
Majority Class	87.37	77.57	22.43	0.0	91.18
BART	95.89	93.71	50.76	75.45	96.51
BART-FT	96.46	94.49	61.53	84.98	97.24
ITI	96.04	93.88	55.07	79.16	96.82
ITI+SI	**96.53**	**94.61**	60.29	84.39	97.15
ITI+SF	95.81	93.52	54.97	76.75	96.69
ITI+SF+Copy	95.73	93.30	**76.35**	89.48	98.12
ITI+SF+SI	95.73	93.25	76.01	**90.75**	**98.17**

Table 5: Results of IE embedding extrinsic evaluation via IE disambiguation—evaluated using F1 score (F1) and Accuracy (Acc%), and IE span detection—evaluated using sequence accuracy (Seq Acc%), and token-level recall (Tkn Recall) and accuracy (Tkn Acc%). Best performances are **boldfaced**.

semantically richer, as the normalized homogeneity score and cosine distance of ITI+SF+SI is higher than that of ITI+SF+Copy by 23.9 and 15.1.

Performance on IE Sense Disambiguation. Though commonly used by prior work, IE sense disambiguation is a relatively simple probing task in idiom processing. As shown in Table 5, though ITI+SI achieves the best performance numerically, all methods compared achieve competitive performances with respect to F1 and accuracy. This shows that BART embeddings already capture the idiosyncratic properties of IEs, in line with the findings from recent papers (Tayyar Madabushi et al., 2021; Adewumi et al., 2021). However, we believe that one cannot judge the quality of IE embeddings via this task alone, because IE senses can be distinguished correctly without the semantic knowledge of IEs. As evidence, under the same setting, we trained another disambiguation classifier with BART but replaced all the IEs from the sentences with single mask tokens for the classifier to make predictions based on just the embeddings of the mask tokens, thus removing all possible IE-related semantic information. We found that such a classifier still performs with an 86% accuracy, operating only on non-IE contextual information. So, IE comprehension ability and IE embedding quality cannot be fully assessed by probing the IE sense disambiguation ability, suggesting that the intrinsic embedding quality and performances on more difficult IE processing tasks must also be considered.

Performance on IE Span Detection. IE span detection is more difficult than IE sense disambiguation as it requires detecting the presence of IEs and precisely identifying their locations. The performance in this task showcases the superiority of our IE embedding methods. ITI+SF+Copy achieves the best performance that is 25.6 points higher than BART in sequence accuracy, our strictest metric. For token-level recall and token-level accuracy, ITI+SF+SI achieves the best performance with a 15-point gain in recall and 1.66 higher in accuracy than BART. The gain in token accuracy is small because the tokens are overwhelmingly literal; the majority-class baseline already achieves a 91% accuracy. The fact that ITI+SF+SI has better token-level performance than ITI+SF+Copy signifies that though ITI+SF+SI detects the span less precisely, it recovers the tokens from within IEs better than ITI+SF+Copy does.

Effect of Copy and Span Infilling. We next examine the usefulness of the copy transformation and span infilling transformation in the noising function. Without copy and span infilling, the ITI+SF suffers in both intrinsic and extrinsic evaluation. For embedding clustering, the homogeneity score of ITI+SF is lower than ITI+SF+SI by 0.15 and lower than ITI+SF+Copy by 0.21, performing even slightly worse than the original BART's embeddings. For IE span detection, ITI+SF's sequence accuracy is lower than that of ITI+SF+SI and ITI+SF+Copy by 21.0% and 21.4%, respectively. Notably, without copy and span infilling transformation, ITI+SF performs barely better than BART, gaining only 4.2% in sequence accuracy. To a lesser degree, ITI+SI also demonstrates the usefulness of the span infilling

Base Model	Sent Emb	Clustering	Disambiguation		Span Detection		
		Homogeneity	F1	Acc	Seq Acc	Tkn Recall	Tkn Acc
BART	MPNet	0.6450	95.73	93.25	76.01	90.75	98.17
BART	BART	0.4671	95.75	93.29	74.55	88.66	98.02
BERT	MPNet	0.4879	91.42	86.36	56.05	78.19	97.34

Table 6: Alternative models' evaluation performances with different LM base models and sentence embedding models (Sent Emb). All models are trained with the same ITI+SF+SI objective.

transformation when compared with ITI, gaining 5.22% in sequence accuracy. Thus, copy or span infilling transformation is necessary and beneficial during the training of the embedding model. Moreover, even though ITI+SF+Copy and ITI+SF+SI performs competitively on the extrinsic evaluation tasks, ITI+SF+SI outperforms ITI+SF+Copy in the intrinsic evaluation task by a meaningful margin demonstrating ITI+SF+SI's superiority over ITI+SF+Copy.

Effect of Similarity Forcing. By comparing ITI+SF and ITI or ITI+SF+SI and ITI+SI, we examine the effect of similarity forcing. While ITI+SF performs similarly or even slightly worse than ITI on evaluation tasks, the performance gain of ITI+SF+SI over ITI+SI is noteworthy, for example, it gains 15.8% in sequence accuracy for IE span detection and 0.20 points in homogeneity score for embedding clustering. Considering the effect of copy and span infilling noising function, we see that ITI+SF+SI shows better performance than either ITI+SI or ITI+SF. This leads us to infer that similarity forcing is only useful when combined with the copy and span infilling transformation. In addition, we also compare the performance between ITI+SF+SI and BART-FT to demonstrate the usefulness of similarity forcing. BART-FT is a BART model fine-tuned on the same training data as ITI+SI. Though BART-FT has significantly more trainable parameters and the same access to external knowledge from the IE definition template sentences during training, BART-FT under-performs ITI+SF+SI by 14.48 points in sequence accuracy for span detection and 0.18 points in homogeneity score for embedding clustering. Therefore, we conclude that using similarity forcing in combination with *copy*- or *span infilling* transformation can boost the performance by a significant margin.

MPNet vs. BART for Definition Embedding. Though the MPNet's definition embeddings and BART's IE embeddings are in different spaces, we believe minimizing the cosine similarity between them to improve IE embeddings' semantic meanings is a valid exercise because (1) the idiomatic meanings of IEs and the meaning of their component words are not related; hence relating their idiomatic meanings to the definition meanings from MPNet's space will not affect the embeddings of the original words; and (2) prior research suggests that minimizing cosine similarity can even help relate the meanings between image embeddings and natural language embeddings (clearly not in the same embedding space) (Radford et al., 2021), hence the space difference between MPNet and BART should not present a problem. Moreover, using MPNet for the definition embedding results in an overall better empirical performance because MPNet produces higher-quality sentence embeddings than BART. We experimented training the ITI+SF+SI model but replaced the MPNet's definition embeddings with that from BART. Comparing the results of the resulting model with those of ITI+SF+SI with MPNet embeddings, shown in the second row of Table 6, we see the resulting model achieves competitive performance for disambiguation but inferior performances in both span detection and embedding clustering with a sequence accuracy that is lower by 1.46% and a homogeneity score that is lower by 0.18. In fact, even the *definition embedding*, when generated by BART, only obtains a homogeneity score of 0.55 (not shown in tables) which is even lower than the ITI+SF+SI by around 0.10. This justifies our use of MPNet for definition embeddings.

Effect of Base Language Models. In our case, encoder-decoder LMs (e.g., BART) are more suitable than an encoder-only LMs (e.g., BERT) because the decoder allows the use of the idiom-aware text infilling objective that asks the model to reconstruct the entire idiom from a *single* mask token. To empirically demonstrate the benefit, we trained an ITI+SF+SI model with BERT as the base LM and modified the idiom-aware text infilling objective by using one mask token per idiom token. As shown in the third row of Table 6, the BERT-based model under-performs its BART-based counterpart in all evaluation tasks by large margins.

Error Analysis on IE Embeddings. Here, we further examine the quality of the definition embeddings and ITI+SF+SI's IE embeddings (named *GIEA embeddings*). We compute *precision at k* (P@k) score for each idiom from the 129 idioms in the 20 meaning groups as follows. Given the embedding for an IE, E, we first find the $k = 3$ closest IEs using pairwise cosine similarity and n, the number of k closest IEs that are from the same group as E; then, P@3 is computed as n/k. The mean score for definition embeddings is 0.64. Meanwhile, the mean score for GIEA embeddings is 0.52, that is, each IE has about half of the 3-closest IEs from the same group. We found a large disparity among the groups with respect to the mean score for each meaning group. While most groups have a mean score around 0.5, groups such as ''anger/upset'', ''quick'', and ''success'' have scores higher than 0.6, and those of others, such as ''punish/criticize'', ''decline/worsen'', ''persevere'' are lower than 0.2.

Also, we found that the per group P@3 scores of the definition embedding are positively correlated with those of GIEA embedding with a Pearson correlation coefficient of 0.76. Based on these observations, we infer that the difficulty of learning IE meanings depends on the specific meaning group and the quality of the definition embedding directly affects the learned GIEA embedding. Improving the definition embeddings through better sentence embedding methods (e.g., by training specifically on dictionary definitions) may further improve the performance of our method. We also leave the important aspect assessing the quality of original compositional embeddings after learning IE embeddings to a follow-up study.

Error Analysis on Extrinsic Evaluation Tasks. Here, we analyze the error of the best performing ITI+SF+SI model on the tasks of span detection and disambiguation. For span detection, we sampled 300 incorrect instances with imperfect sequence accuracies (30.5% of all incorrect samples) and categorized them into the six error types defined in Zeng and Bhat (2021). Among the sampled errors, we found that 3.7% were attributable to identifying one of the IEs when multiple IEs are present, 57% to detecting only a portion of the idiom span, 1% to identifying figurative expressions other than the ground truth idiom, 25% to identifying a PIE as idiomatic when actually used in the literal sense, 8.3% for failing to recognize the presence of an idiom, and another 5% for returning random tokens that are not meaningful nor part of any PIEs, that is, over 60% of the errors were in the detection of figurative tokens. In fact, over 40.8% of test idioms had their spans precisely tagged in all of their test instances. For disambiguation, over 82.8% of the test PIEs were classified with 100% accuracy and only less than 6% of the test PIEs had an accuracy less than 50%. For both disambiguation and span detection, the per-idiom accuracies were weakly correlated with the number of training instances per idiom (Pearson correlation coefficient of -3.84e-4 for disambiguation and 0.26 for span detection), suggesting that the performance discrepancy among idiom types is caused by factors other than their frequency in the train set. Future studies should consider the characteristics of the hard-to-learn idioms to improve the embeddings of the under-performing idioms.

Limitations. An obvious limitation of GIEA is that it cannot generalize its representation ability to idioms unseen during training. From the results in Section 2 and Section 5, it is evident that the meanings of IEs cannot be learned from general corpora alone (even when there is a collection of sentences with IEs), rather, external knowledge (e.g., IE definitions) is a fundamental to providing the strong supervising signal (i.e., similarity forcing loss) needed for training. Taking this into consideration, we believe that it is impractical to generalize the representation ability to the unseen idioms because (1) intuitively, each IE has a unique origin, metaphorical linkage, and interpretation, so, the *meaning* of IEs have to be learned on a case-by-case basis; and (2) from our error analysis, even with the same training data and objective, the learning difficulty is highly idiom dependent, a point that is also corroborated by Nedumpozhimana et al. (2022). Therefore, we do not currently see a practical way to generalize GIEA to idioms that are unseen. However, we argue that this does not hinder the utility of GIEA, since our training data, MAGPIE, already contains idiomatic sentences for idioms (and metaphors) that occur in sources such as the Oxford Dictionary of English Idioms (Ayto and Press, 2009) and Wiktionary. Thus, we expect GIEA to cover most frequently used idioms. Besides, even though expanding an IE lexicon to include new idioms may be easy, gathering idiomatic sentences for those new idioms requires human input. So, an important future study is to consider methods that generalize GIEA to idioms with known identities but with limited or no idiomatic sentences.

6 Related Work

IE Processing Tasks. Classically, two main idiom-related processing tasks, namely, *idiom type classification* and *idiom token classification*, have been studied (Cook et al., 2008; Liu and Hwa, 2019; Liu, 2019). Idiom type classification aims to decide if a set of MWEs can be used as IEs without considering additional context (Westerståhl, 2002; Fazly and Stevenson, 2006; Tabossi et al., 2008, 2009; Shutova et al., 2010; Reddy et al., 2011; Cordeiro et al., 2016). Idiom token classification determines if a given PIE is used in a literal or figurative sense in a sentence and solutions include those that mostly assume the knowledge of the location and/or identify of the PIEs (Fazly et al., 2009; Feldman and Peng, 2013; Peng and Feldman, 2016; Salton et al., 2016; Taslimipoor et al., 2018; Peng et al., 2014; Liu and Hwa, 2019), build per-idiom classifiers (Liu and Hwa, 2017), extract embeddings based on PIE positions (Liu and Hwa, 2019), or focus on only PIEs with specific syntactic structures (Taslimipoor et al., 2018). Due to the impracticality of acquiring this prior knowledge in real-world applications, most recent works (Zeng and Bhat, 2021; Škvorc et al., 2022) study the *idiomatic expression identification problem*, jointly the detecting and localizing a PIE without requiring PIE identity or position. This problem is related to the MWE identification task in STREUSLE (Schneider and Smith, 2015) but with a focus on expressions with semantic idiomaticity. In-line with prior state-of-the-art, we use the IE token classification and IE identification, dubbed as *IE sense disambiguation* and *IE span detection*, as the extrinsic evaluation tasks to our IE embeddings.

Impact of IE Presence. Since Sag et al.'s (2002) study on the impact of MWE, not only have studies identified the influence of IEs across various NLP applications (Salton et al., 2014; Fadee et al., 2018; Ganitkevitch et al., 2013; Liu et al., 2017; Biddle et al., 2020), recent efforts have also sought ways to mitigate them (Jhamtani et al., 2021; Chakrabarty et al., 2021). However, the techniques used either simply enlarge the training data by including idiomatic sentences or paraphrase idiomatic sentences into equivalent literal sentences, completely ignoring the fundamental issue of IE representation. Other works (Tayyar Madabushi et al., 2021) have probed how idiomaticity is handled in PTLMs but offer no solution to improve their representation. Efforts to improve IE span detection or IE sense disambiguation include transforming the original representations from pre-trained LMs by incorporating static word embeddings alone (Liu and Hwa, 2017), with additional syntactic information (Zeng and Bhat, 2021), utilizing contrastive loss to make literal and figurative speech embeddings more distinctive (Lin et al., 2021), treating IEs as new tokens during training (Hashempour and Villavicencio, 2020), or combining representations from multiple pre-trained LMs (Škvorc et al., 2022). Taking a different approach in this work, instead of creating task-specific representations or altering tokenization at the input, we first train an LM that produces better IE embeddings in general and then show their benefit in the idiom processing tasks. In principle, our trained GIEA can be plugged into the prior works for idiom processing tasks, replacing their embedding models and improving their performances, an aspect we leave to future explorations.

Adapter. Originally developed for computer vision applications (Rebuffi et al., 2017, 2018), adapters are new modules of simple projection layers added between the trained transformer layers, used in NLP as a parameter-efficient and fast fine-tuning method to adapt pre-trained LMs to new tasks or domains (Houlsby et al., 2019; Bapna and Firat, 2019). Recently, adapters have shown effectiveness in multi-task and multi-lingual transfer learning as well (Pfeiffer et al., 2020b; Ansell et al., 2021). In this work, we utilize an adapter as a lightweight non-compositional language expert that is trained on idiomatic sentences and thus can expand upon the base LM to generate semantically meaningful IE embeddings. The compact Pfeiffer adapter architecture (Pfeiffer et al., 2020a) is used in GIEA.

(Non-)Compositional Phrase Embedding. The core idea for works on non-compositional phrase embeddings is to avoid treating phrases as purely compositional (by aggregating word embeddings) or non-compositional (treating phrases as single units), but consider both aspects. The approaches have adaptive weights and consider different compositions within a phrase (Li et al., 2018a; Hashimoto and Tsuruoka, 2016; Li et al., 2018b) or utilize hypernymy information and represent phrases in special embedding spaces (Jana et al., 2019). Although related, these embedding methods cannot produce the contextualized phrase embeddings as transformer-based models do, nor can they be combined with PTLMs to aid downstream tasks.

Embedding Evaluation. The evaluation of word and phrase embeddings (Hashimoto and Tsuruoka, 2016; Jana et al., 2019) is typically via *intrinsic* methods (e.g., similarity and analogy) and *extrinsic* methods, e.g., downstream NLP tasks (Schnabel et al., 2015; Ghannay et al., 2016; Hupkes and Zuidema, 2018; Wang et al., 2019). A popular alternative evaluation method is *probing*, where a simple diagnostic classifier is trained to extract information from frozen embeddings and determine the extent to which desired linguistic properties are encoded in the representations (Adi et al., 2016; Warstadt et al., 2019; Alt et al., 2020; Ravichander et al., 2021). Our intrinsic and extrinsic evaluation of embeddings follow these prior works.

7 Conclusion and Future Work

In this work, we first demonstrate current BART's inability produce semantically meaningful representations for idioms, then, we propose GIEA, that uses a lightweight adapter, a set of denoising auto-encoder-style learning objectives, and a similarity forcing objective to produce quality IE embeddings without altering the input tokenization. Through both intrinsic evaluation of embedding quality and extrinsic evaluation on their usefulness on idiom-processing tasks, we find that GIEA greatly improves upon embedding quality and usefulness compared to the original pre-trained BART's embeddings.

Future work should explore means to improve embedding quality for hard-to-learn idioms based on observed performance, IEs other than idioms (e.g., phrasal verbs), and the use of GIEA with other SOTA idiom processing models. Lastly, applying idiom-aware PTLMs to downstream applications that require the IE comprehension, such as dialog modeling and machine translation, would be fruitful pursuits.

References

Tosin P. Adewumi, Saleha Javed, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Simistira Liwicki, and Marcus Liwicki. 2021. Potential idiomatic expression (PIE)-english: Corpus for classes of idioms. *ArXiv*, abs/2105.03280.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, Online. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.findings-emnlp.410`

John Ayto and Oxford University Press. 2009. *Oxford Dictionary of English Idioms / [edited] by John Ayto*, 3rd edition. Oxford University Press [Oxford].

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1165

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. https://doi.org/10.1145/3366423.3380198

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.297

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

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997. https://doi.org/10.18653/v1/P16-1187

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103. https://doi.org/10.1162/coli.08-010-R1-07-048

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 435–446. Springer. https://doi.org/10.1007/978-3-642-37247-6_35

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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 472–80, 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. https://doi.org/10.18653/v1/P16-1020

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5617–5621. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/796

Abhik Jana, Dima Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee. 2019. On the compositionality prediction of noun phrases using poincaré embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3263–3274, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1316

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. https://doi.org/10.18653/v1/2021.emnlp-main.592

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Bing Li, Xiaochun Yang, Bin Wang, Wei Wang, Wei Cui, and Xianchao Zhang. 2018a. An adaptive hierarchical compositional model for phrase embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4144–4151. International Joint Conferences on Artificial Intelligence Organization.

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. https://doi.org/10.1016/j.knosys.2018.04.009

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changsheng Liu. 2019. *Toward Robust and Efficient Interpretations of Idiomatic Expressions in Context*. Ph.D. thesis, University of Pittsburgh.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31. https://doi.org/10.1609/aaai.v31i1.10998

Changsheng Liu and Rebecca Hwa. 2019. A generalized idiom usage recognition model based on semantic compatibility. In *Proceedings of* *the AAAI Conference on Artificial Intelligence*, volume 33, pages 6738–6745. https://doi.org/10.1609/aaai.v33i01.33016738

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuan-Jing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*, Oxford University Press.

Vasudevan Nedumpozhimana, Filip Klubička, and John D. Kelleher. 2022. Shapley idioms: Analysing BERT sentence embeddings for general idiom token identification. *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.813967, PubMed: 35360661

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jing Peng and Anna Feldman. 2016. Automatic idiom recognition with word embeddings. In *Information Management and Big Data - Second Annual International Symposium, SIMBig 2015, Cusco, Peru, September 2-4, 2015, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, September 1-3, 2016, Revised Selected Papers*, volume 656 of *Communications in Computer and Information Science*, pages 17–29. Springer. https://doi.org/10.1007/978-3-319-55209-5_2

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1216

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.7

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.617

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.295

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. 2018. Efficient parametrization of multi-domain deep neural networks. In *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8119–8127.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218. Asian Federation of Natural Language Processing.

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. https://doi.org/10.18653/v1/D19-1410

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer. https://doi.org/10.1007/3-540-45715-1_1

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1007

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204. https://doi.org/10.18653/v1/P16-1019

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1036

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1177

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. https://doi.org/10.18653/v1/P16-1162

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):313. https://doi.org/10.1037/0278-7393.34.2.313, PubMed: 18315408

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2009. Why are idioms recognized fast? *Memory & Cognition*, 37(4):529–540. https://doi.org/10.3758/MC.37.4.529, PubMed: 19460959

Shiva Taslimipoor, Omid Rohanian, Ruslan Mitkov, and Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*, volume 2, page 299. Language Science Press.

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. https://doi.org/10.18653/v1/2021.findings-emnlp.294

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19. https://doi.org/10.1017/ATSIP.2019.12

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1286

Dag Westerståhl. 2002. On the compositionality of idioms. In *Proceedings of LLC8. CSLI Publications*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

W. Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2021. Dict-BERT: Enhancing language model pre-training with dictionary. *ArXiv*, abs/2110.06490. https://doi.org/10.18653/v1/2022.findings-acl.150

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562. https://doi.org/10.1162/tacl_a_00442

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2021. Idiomatic expression paraphrasing without strong supervision. https://doi.org/10.1609/aaai.v36i10.21433

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606. https://doi.org/10.1016/j.knosys.2021.107606