Idiomatic Expression Paraphrasing without Strong Supervision

Jianing Zhou[1], Ziheng Zeng[1], Hongyu Gong[1,2], Suma Bhat[1]

[1] University of Illinois at Urbana-Champaign
[2] Facebook AI
[1] {zjn1746, zzeng13,spbhat2}@illinois.edu
[2] hygong@fb.com

Abstract

Idiomatic expressions (IEs) play an essential role in natural language. In this paper, we study the task of idiomatic sentence paraphrasing (ISP), which aims to paraphrase a sentence with an IE by replacing the IE with its literal paraphrase. The lack of large-scale corpora with idiomatic-literal parallel sentences is a primary challenge for this task, for which we consider two separate solutions. First, we propose an unsupervised approach to ISP, which leverages an IE's contextual information and definition and does not require a parallel sentence training set. Second, we propose a weakly supervised approach using back-translation to jointly perform paraphrasing and generation of sentences with IEs to enlarge the small-scale parallel sentence training dataset. Other significant derivatives of the study include a model that replaces a literal phrase in a sentence with an IE to generate an idiomatic expression and a large scale parallel dataset with idiomatic/literal sentence pairs. The effectiveness of the proposed solutions compared to competitive baselines is seen in the relative gains of over 5.16 points in BLEU, over 8.75 points in METEOR, and over 19.57 points in SARI when the generated sentences are empirically validated on a parallel dataset using automatic and manual evaluations. We demonstrate the practical utility of ISP as a preprocessing step in En-De machine translation.

Introduction

Idiomatic expressions (IEs) are multi-word expressions whose meaning cannot be inferred from that of their constituent words, a property known as non-compositionality (Nunberg, Sag, and Wasow 1994). These expressions have varied forms, ranging from fixed expressions such as *by the way* to figurative constructions such as *born with a silver spoon in one's mouth*. Not only are IEs an essential component of a native speakers' lexicon (Jackendoff 1995), they also render language more natural (Sprenger 2003). Their non-compositionality has been the classical 'pain in the neck' for NLP applications (Salton, Ross, and Kelleher 2014) and studies to make these applications idiom-aware, either by identifying them before or during the task (Nivre and Nilsson 2004; Nasr et al. 2015) suggest that IE paraphrasing as a preprocessing step holds promise for NLP. Despite this, research on IE paraphrasing remains largely under-explored (Zhou, Gong, and Bhat 2021a). While most IE processing studies have focused on their identification and detection (Gong, Bhat, and Viswanath 2017; Liu and Hwa 2018; Biddle et al. 2020), in this paper, we study the task of idiomatic sentence paraphrasing (ISP), i.e., automatically paraphrasing IEs into literal expressions. We refer to a sentence with an IE as an *idiomatic sentence* and to its corresponding sentence where the IE is replaced with a literal phrase as the *literal sentence*. Table 1 shows examples of idiomatic and literal sentences between which we expect to paraphrase. Ideally, an ISP system would have an IE span detection stage to detect the presence and span of IEs (Zeng and Bhat 2021) and feeds only idiomatic sentences to ISP. Here we study the ISP task on its own and assume the input sentence is idiomatic and the IE span is available.

Idiomatic sentences	Literal sentences
Nature conservation **<u>runs against the grain</u>** of current political doctrine.	Nature conservation **<u>is contrary to</u>** current political doctrine.
Putting him **<u>behind bars</u>** won't serve any purpose, will it?	Putting him **<u>in prison</u>** won't serve any purpose, will it?

Table 1: Examples of idiomatic sentences and corresponding literal sentences. Idioms and their corresponding literal paraphrases are in bold underlined.

Semantic simplification using ISP can be used to many ends, including for making reading more inclusive for populations that struggle to comprehend figurative expressions in everyday text (e.g., children with the autistic spectrum disorder (Norbury 2004)). Based on prior studies (Nivre and Nilsson 2004; Nasr et al. 2015), it could also serve as a preprocessing step for downstream applications—an aspect we explore in this study.

Successful ISP involves overcoming at least two challenges: (1) The linguistic challenge of handling semantic ambiguity, i.e., ensuring that the meaning of the IE and that of the literal phrase match when an IE is polysemous, e.g. the idiom *give her a hand* can mean both "applaud her" and "help her," and (2) the related resource-challenge of the lack

*The work was done while Hongyu Gong was at UIUC of large-scale parallel literal and idiomatic expressions for training, because a small training set leads to the input being unchanged at the output (Zhou, Gong, and Bhat 2021a).

Addressing the second challenge is the main focus of this study, whose contributions are summarized below.

1. Given the paucity of large-scale parallel datasets of idiomatic-literal sentence pairs, we study ISP in two machine learning settings. The first is *unsupervised*, where we consider a zero-resource scenario with neither access to a parallel dataset nor to a lexicon of IEs *during training*, and the second is *weakly-supervised,* where we consider a low-resource scenario with access to a limited but high quality parallel dataset and a large corpus of idiomatic sentences. Our training strategy relies on a back-translation-based augmentation that yields a large parallel dataset.
2. Compared to competitive supervised baselines the proposed weakly-supervised method shows performance gains of over 5.16 points in BLEU and over 19.57 points in SARI (automatic evaluation) and superior generation quality (manual evaluation). Despite the lack of supervision, the unsupervised method's performance compares favorably to that of the supervised baselines.
3. Our weakly-supervised method yields a large parallel dataset of idiomatic sentences and their literal counterparts with 1,169 IEs and their 15,627 sentence pairs, which we share for future research.[1]
4. We demonstrate the gains to machine translation only using ISP as a pre-processing step via an English-German challenge set (Fadaee, Bisazza, and Monz 2018); translating idiomatic sentences after paraphrasing them to their literal counterparts yielded a gain of 0.6 points in BLEU.

Related Work

ISP was explored as idiomatic expression substitution in Liu and Hwa (2016) using a set of pre-defined heuristic rules to extract portions of the idiom's definitions to replace the IE and then applying various post-processing steps to render the sentence. Going beyond this study, ISP relates to three distinct streams of text generation tasks: *paraphrasing*, *style transfer* and *IE processing*.

Paraphrasing is to rewrite a given sentence while preserving its original meaning; prior studies include several sequence-to-Sequence (Seq2Seq) models (Gupta et al. 2018) and other controlled generation methods via template (Gu, Wei et al. 2019), syntactic structures (Huang and Chang 2021), or versatile control codes (Keskar et al. 2019). Unlike paraphrasing, which is unconstrained, ISP is more stylistically constrained given the paraphrasing of an IE to its literal meaning.

Style Transfer rewrites sentences into those that conform to a target style. This has been studied as distinctive lexical patterns and syntactic constructions by Krishna, Wieting, and Iyyer (2020), and as sentiment, formality or authorship manipulation (Jhamtani et al. 2017; Gong et al. 2019). Our study is different from these prior methods, including the supervised (Li et al. 2018; Sudhakar, Upadhyay, and Maheswaran 2019) and unsupervised ones (Gong et al. 2019; Zeng, Shoeybi, and Liu 2020), in that our task retains a large portion of the input sentence in the transferred sentence. Besides, we consider a heretofore unexplored nuanced stylist element that is marked by figurative and non-literal phrases.

IE processing tasks consider idiom type classification and idiom token classification (Liu 2019): idiom type classification (Cordeiro et al. 2016) determines if a phrase could be used as an IE; and idiom token classification (Liu and Hwa 2017, 2019) disambiguates if a given potentially idiomatic expression is used literally or idiomatically in a given context (sentence). Most prior works require the knowledge of the IE (Liu and Hwa 2017, 2019) but recent efforts on idiom span detection (Zeng and Bhat 2021) have removed the need for IEs' identity. Our study is in line with the traditional set-up where the IE positions are assumed to be known.

The Unsupervised Approach

For the zero-resource ISP scenario where no parallel datasets are available during training, we train a masked conditional sentence generation model such that given a sentence with a masked word, the model fills the mask using the masked word's definition and part-of-speech (POS) tag. The word's definition and POS tags as inputs account for the semantic and the syntactic properties of the filled word. During inference, we mask the IE in the sentence to perform ISP while providing the definition of the IE[2]. The definitions of the masked word (or the IE during inference) and its POS tag are available from linguistic resources such as dictionaries and POS taggers. Our model, denoted as BART-UCD, is unsupervised because its training does not rely on knowing the IEs nor the direct supervision from a parallel dataset.

Although conceptually similar to Liu and Hwa (2016)'s setup, BART-UCD (1) does not modify or operate on the definitions using pre-determined dictionary-specific rules; (2) inserts phrases based on the context instead of inserting a fixed chunk from the definition; (3) is naturally applicable to words and IEs with multiple definitions; and (4) generates fluent and grammatically correct sentences without burdensome post-processing steps. We exclude the unsupervised method of Liu and Hwa (2016) as a baseline in our experiments owing to its unavailability and poor replicability.

Model Architecture

The overall architecture of BART-UCD is illustrated in Figure 1 and it consists of three stages: (1) the *embedding stage*, (2) the *fusion stage*, and (3) the *generation stage*. In this section, we describe each stage in detail.

The Embedding stage. This stage generates the contextualized word- and sentence embeddings for the definitions. Specifically, given $[I, \langle \mathtt{sep} \rangle, t]$, where I is the masked sentence and t is the POS tag, the model uses a pre-trained BART (Liu et al. 2020) encoder to produce contextualized word embeddings $E^I \in \mathbb{R}^{(L+2)\times D^B}$, where $|I| = L$. Then,

[1]The code and dataset are available at https://github.com/zhjjn/ISP.git.

[2]A dictionary for accessing the IE definitions is available to the model during inference; the users only provide the sentences.

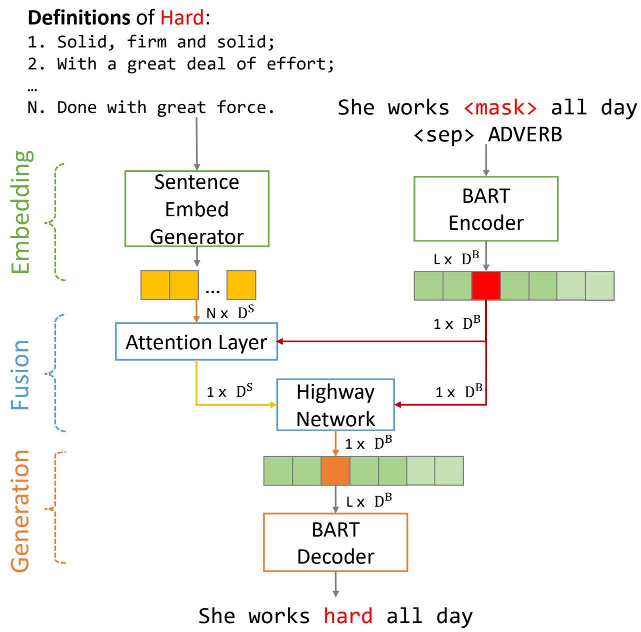

Figure 1: An overview of the unsupervised method. In this example of a *training* instance, the input sentence has a masked word "hard". The model takes as input the sentence, the definitions of the word "hard" and its POS tag "ADVERB" and generates a sentence with the mask filled.

given a list of N definitions for masked word, the model employs a pre-trained RoBERTa-based (Liu et al. 2019) sentence embedding generator to generate definition sentence embeddings $E^D \in \mathbb{R}^{N \times D^S}$. During training, both the BART encoder and the sentence embedding generator are pre-trained and frozen.

The Fusion stage. This stage combines the definition embeddings E^D and the word embedding E_w^B for the masked token i_w and replaces E_w^B with the combined embedding. Specifically, the model first transforms E^D into a single vector $\hat{E}^D \in \mathbb{R}^{1 \times D^S}$ using an attention mechanism (Luong, Pham, and Manning 2015) with E_w^B as the query to generate the attention weights. Then, the model fuses $\hat{E}^D$ and E_w^B using a highway network (Srivastava, Greff, and Schmidhuber 2015) followed by a linear layer to produce the definition-aware contextualized embedding for w, $\tilde{E}_w^D \in \mathbb{R}^{1 \times D^B}$. Based on an empirical observation of improved performance, we replace the original linear + $tanh$ part of the attention mechanism with the highway network. Finally, the model replaces E_w^B from E^B with $\tilde{E}_w^D$ to produce $\tilde{E}^D$.

The Generation stage. Here, the model decodes the output sentence S from $\tilde{E}^D$ using a pre-trained BART decoder that is fine-tuned during training with the rest of the model.

Model Training and Inference

Training data preparation. Acquiring training data for our masked conditional sentence generation model described above is relatively easy as any well-formed sentence can be converted into a training instance. We do this by first identifying a masked word, which can be any verb, adjective, and adverb from the sentence because IEs mostly assume these roles in a sentence. Then, we retrieve the definitions of the masked word from dictionaries. To increase the diversity in definitions and prevent the model from becoming dictionary-specific, we access the masked word's definitions randomly from WordNet (Miller 1995), Wiktionary[3], or Google Dictionary[4]. Finally, we use a BERT-based (Devlin et al. 2019) POS tagger to predict the POS tag for the masked word. Inspired by Hegde and Patil (2020)'s way of improving the fluency of generated sentences we drop stop words from the input sentences and ask the model to reconstruct them. Hence, in each batch of our training, 80% of the sentences have their stop words removed and 40% of the sentences have their words lemmatized (these two operations can happen simultaneously). For our case, these sentence corruptions have the additional benefit of allowing the model to generate more than one word in place of the masked token, which is critical for generating substitutions for several IEs.

Inference. During inference, given an IE, I, it is replaced by the masked token i_w. Then, the POS tag of i_w is predicted with a pretrained POS tagger and fed to the model with the masked IE's definition. The model then generates the output S with the masked IE replaced by a literal phrase. It is important to note that the ISP task is performed in a zero-shot manner in that the model is trained to fill in a masked word, but during inference its knowledge and function are transferred to predict the literal meaning of IEs.

The Weakly Supervised Method

For the low-resource scenario, we use a small parallel dataset $\mathbb{P} = \{(I_1, S_1), (I_2, S_2), \cdots, (I_N, S_N)\} = \{\mathbb{I}; \mathbb{S}\}$ of N pairs of sentences, where (I_k, S_k) is a pair of idiomatic sentence and its literal counterpart. to create a weakly supervised end-to-end model for ISP. Like BART-UCD above, it takes an idiomatic sentence as input (without the IE's definition/identity during training) and generates the entire paraphrased literal sentence as output. Drawing a parallel between ISP and that of machine translation, our weakly supervised approach relies on an iterative back-translation mechanism to (generate and) augment the limited training data and improve the performance of a vanilla BART model, which we refer to as BART-IBT.

The limited size of $\mathbb{P}$ prompts us to generate a much larger $\mathbb{I}_M$ by iteratively training two models simultaneously: (1) an ISP model that translates an idiomatic sentence $\mathbb{I}$ to a literal sentence $\hat{\mathbb{S}}$, and (2) an Idiomatic Sentence Generation (ISG) model that translates a literal sentence $\mathbb{S}$ into an idiomatic sentence $\hat{\mathbb{I}}$. Note that besides our main objective of training an ISP model, acquiring a competent ISG model and a larger parallel dataset are both welcome byproducts.

Each training iteration consists of three stages—*Model training*, *Data generation*, and *Data selection*. The iterative process is described in Figure 2 and Algorithm 1.

[3]https://en.wiktionary.org/

[4]https://dictionaryapi.dev/

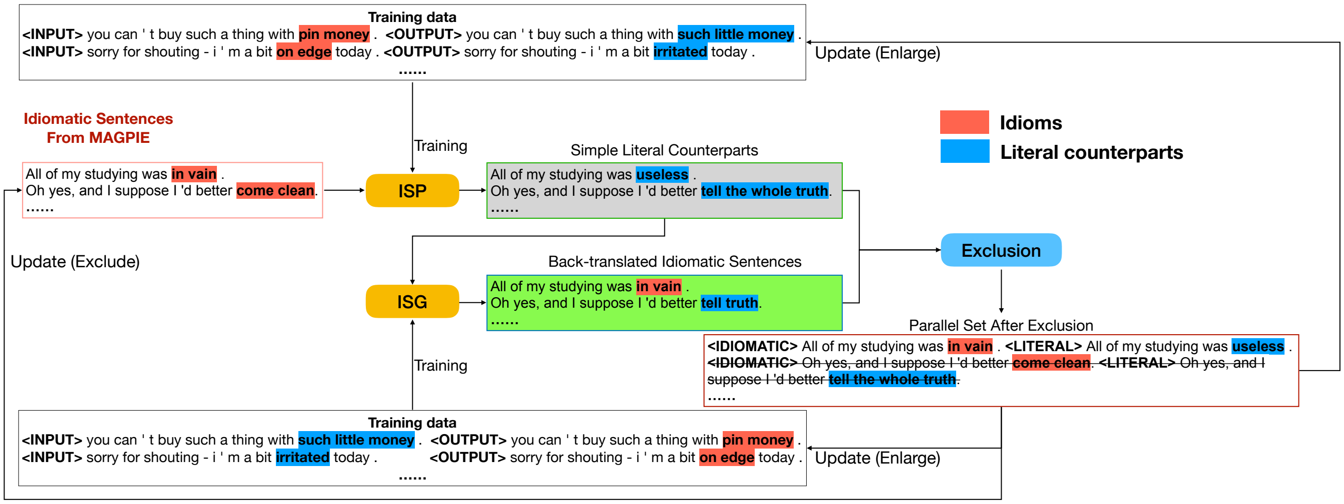

Figure 2: The overview of the weakly supervised method. In each iteration, the method (1) uses the parallel dataset to train an ISP and an ISP model; (2) constructs augmented parallel pairs; (3) enlarges the parallel dataset with the augmented pairs.

Model Training

We use the parallel dataset ($\mathbb{P}$ to begin with and the augmented set—described below— during subsequent iterations) to fine-tune two separate pretrained BART models yielding the ISP and the ISG model.

Data Generation

In this stage, the trained ISG model and the ISP model from the previous stage generate more idiomatic-literal sentence pairs that augment the initial training set. First, the ISP model generates literal counterparts $\hat{\mathbb{S}}_M$ for all the idiomatic sentences in $\mathbb{I}_M$. Then the ISG model is used to transform the literal sentences back into the idiomatic form, whose collection is $\hat{\mathbb{I}}_M$.At the end of this stage, we gather $\hat{\mathbb{S}}_M$ and $\hat{\mathbb{I}}_M$ to produce the set of candidate pairs $\mathbb{D}_M$ for the next stage.

Data Selection

Note that there may be low quality pairs in $\mathbb{D}_M$ resulting from, e.g., IEs not replaced in the generated literal sentences or IEs omitted from the back-translated idiomatic sentences. Toward excluding these pairs from the collection $\mathbb{D}_M$ we propose two rules: (1) For any example $(I^j_M, \hat{S}^j_M, \hat{I}^j_M) \in \mathbb{D}_M$, if the literal sentence $\hat{S}^j_M$ still contains the IE in I^j_M, the example will be excluded; and (2) for any example $(I^j_M, \hat{S}^j_M, \hat{I}^j_M) \in \mathbb{D}_M$, if the back-transformed idiomatic sentence $\hat{I}^j_M$ is different from the original idiomatic sentence I^j_M, the example will be excluded. After filtering, we get $\mathbb{D}^*_M \in \mathbb{D}_M$ such that $= \{\mathbb{I}^*_M; \hat{\mathbb{S}}^*_M\}$, where $\mathbb{I}^*_M \in \mathbb{I}_M$ and $\hat{\mathbb{S}}^*_M \in \hat{\mathbb{S}}_M$. Finally, the parallel dataset $\mathbb{P}$ is enlarged to $\mathbb{P} \cup \mathbb{D}^*_M$. Also, $\mathbb{I}_M$ is shrunk to $\mathbb{I}_M \setminus \mathbb{I}^*_M$. The enlarged parallel dataset and the updated set of idiomatic sentences are used in the next iteration.

After all the iterations, we obtain an enlarged parallel dataset with idiomatic/literal sentence pairs and the well-trained models for ISG and ISP.

Algorithm 1: WeaklySupervisedModel

Input: Original parallel dataset $\mathbb{P}$, Idiomatic sentences $\mathbb{I}_M$ and number of iterations N
Output: **ISP** and **ISG** Model, Enlarged parallel dataset $\mathbb{P}$

$\mathbb{P}_1 = \mathbb{P}, \quad \mathbb{I}^1_M = \mathbb{I}_M$;
for $n = 1; n \leq N$ **do**
  $\mathbb{D}_M = \varnothing$;
  $\mathbf{ISP}_n \Leftarrow \mathit{TRAIN}(\mathbb{P}_n), \quad \mathbf{ISG}_n \Leftarrow \mathit{TRAIN}(\mathbb{P}_n)$;
  for $I_M \in \mathbb{I}^n_M$ **do**
    $\hat{S}_M = \mathbf{ISP}_n(I_M), \quad \hat{I}_M = \mathbf{ISG}_n(\hat{S}_M)$;
    $\mathbb{D}_M \Leftarrow \mathbb{D}_M \bigcup \{(I_M, \hat{S}_M, \hat{I}_M)\}$;
  end
  $\mathbb{D}^*_M = \varnothing$;
  for $(I_M, \hat{S}_M, \hat{I}_M) \in \mathbb{D}_M$ **do**
    if $I_M \neq \hat{S}_M \wedge \hat{I}_M = I_M$ **then**
      $\mathbb{D}^*_M \Leftarrow \mathbb{D}^*_M \bigcup \{(I_M, \hat{S}_M)\}$;
    end
  end
  $\mathbb{I}^{n+1}_M = \mathbb{I}^n_M$;
  for $(I_M, \hat{S}_M) \in \mathbb{D}^*_M$ **do**
    $\mathbb{I}^{n+1}_M \Leftarrow \mathbb{I}^{n+1}_M \setminus I_M$;
  end
  $\mathbb{P}_{n+1} \Leftarrow \mathbb{P}_n \bigcup \mathbb{D}^*_M$;
end
return $\mathbf{ISP}_N, \mathbf{ISG}_N, \mathbb{P}_{N+1}$;

Experiments

In this section, we evaluate the performances of the proposed BART-UCD and BART-IBT against competitive baselines, while later in the paper, we show an application of ISP in a downstream NLP task.

Baselines

We study the following competitive text generation baselines for ISP—the **Seq2Seq** model (Sutskever, Vinyals, and

Le 2014), the **Transformer** model (Vaswani et al. 2017), the copy-enriched Seq2Seq (**Seq2Seq-copy**) model (Jhamtani et al. 2017), the copy-enriched Transformer (**Transformer-copy**) model (Gehrmann, Deng, and Rush 2018), and the **T5** model (Raffel et al. 2020). To validate the effectiveness of BART-IBT, we also use a fine-tuned BART (**BART**) model without back-translation as a baseline.

Our baselines do not include standard paraphrasing and style-transfer models due to the lack of a large-scale parallel corpus and the ISP requirement of changing only a single phrase in the sentence. Moreover, we also exclude pretrained language models mainly to highlight the overall difficulty of ISP.

Datasets

In this section we first introduce the training sets for the proposed methods followed by the test sets used by the proposed methods and the baselines.

Training Set Recall that any corpus of well-formed sentences can be used to train BART-UCD. Accordingly, we choose two large news datasets—AG News (Zhang, Zhao, and LeCun 2015) and CNN-Dailymail (See, Liu, and Manning 2017)— and the GLUE datasets MRPC and COLA (Wang et al. 2018). This choice is guided by the rationale that they are well-formed and less likely to contain IEs owing to their being sentences from the news and the scientific domain (to minimize the likelihood that the model may generate IEs). For AG News and CNN-Dailymail, we randomly sampled 1 million sentences from each sentence-tokenized dataset. Considering each sentence with a masked word as a data instance, our final training corpus has 1.97 million instances, 11,071 unique masked words, and 17 unique POS tags. Even though including more training instances, as with all models, can improve the model's performance, we found our current training corpus to yield satisfactory results.

Toward training BART-IBT (i.e., fine-tuning the backbone pretrained BART models for our task), we used the parallel dataset constructed by Zhou, Gong, and Bhat (2021a) (henceforth termed **PIL**) with a training set of 3,789 manually created idiomatic and literal sentence pairs from a list of 876 IEs and their definitions, with at least 5 idiomatic sentences per IE. The idiomatic sentences (without literal counterparts) used for BART-IBT training are from the MAGPIE corpus (Haagsma, Bos, and Nissim 2020) collected from the BNC. Choosing sentences with figurative IEs yielded 27,582 idiomatic sentences from 1,644 IEs to form the idiomatic sentence set $\mathbb{I}_M$. Among the 1,644 IEs, 208 overlap with those in PIL.

All **baselines** were trained using only the PIL training set.

Test Set. For a fair comparison across the methods, we used two types of test sets to evaluate all the methods. The first was the test split of PIL for both automatic and manual evaluation. This includes 876 idiomatic-literal sentence pairs with each idiomatic sentence containing a unique IE that occurred in the training set. We leave it to future work to examine generalization to IEs unseen during training.

To afford a different perspective of the models' capabilities with naturally occurring idiomatic instances, we used a second test set constructed from the MAGPIE dataset (**MIL**; only for manual evaluation) consisting of 100 idiomatic sentences unseen in the training set of BART-IBT. The literal counterparts were provided by one annotator and then verified by a second annotator, both native English speakers and proficient users of IEs and not part of the research team. To ensure compatibility between the set of IEs in MIL and PIL, we verified that the same IEs were used in the idiomatic sentences of the two test sets.

Experimental Setup

Here we introduce the basic settings for the models.

Unsupervised Method. We use the pretrained BART-large model, the BERT-based POS tagger and their respective checkpoints as implemented and hosted by Huggingface's Transformers library. The RoBERTa-based sentence embedding generator and its checkpoint are implemented and hosted by (Reimers and Gurevych 2020).

Weakly Supervised Method. We used two independent pretrained BART-large models as the ISP model and the ISG model in BART-IBT. These pretrained models were also implemented as hosted by Huggingface's Transformers library. The maximum length for a sentence, the learning rate and the number of iterations were 128, $5e-5$, and 5 respectively. The other hyper-parameters were their default values.

Baselines. For the Seq2Seq, the Transformer, the Seq2Seq-copy, and the Transformer-copy, we followed the experimental settings described in (Zhou, Gong, and Bhat 2021a,b); the baseline pretrained BART model is identical to that used in BART-IBT, and the T5 model is that hosted by Huggingface and trained under the same settings as the BART model. The model was trained for 5 epochs. During inference, we used a beam search with 5 beams with top-*k* set to 100 and top-*p* set to 0.5. The other hyper-parameters were set to their default values.

Evaluation Metrics

Automatic Evaluation. We used metrics widely used in text generation tasks such as paraphrasing and style transfer—ROUGE (Lin 2004), BLEU (Papineni et al. 2002) and METEOR (Lavie and Agarwal 2007)—to compare the generated sentences with the references. Due to the similarity between ISP and text simplification, we also used SARI (Xu et al. 2016), the metric for text simplification. To measure linguistic quality, we use a pretrained GPT-2 (Radford et al. 2019) to calculate perplexity scores and a recently proposed measure of linguistic quality, GRUEN (Zhu and Bhat 2020). These scores were collected on the PIL test set.

Human Evaluation. For a qualitative measure of ISP we use human evaluation to complement the automatic evaluation. We used 100 instances from the PIL test set and the entire MIL test set, and collected the outputs from the 3 best methods ranked by automatic evaluation. For each output sentence, two native English speakers, who were blind to the systems being compared, were asked to rate the output sentences with respect to meaning, style and fluency using the following scoring criteria:

(1) **Meaning preservation** measures on a binary scale how well the meaning of the input is preserved in the output.

Model - ISP	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	SARI	GRUEN	PPL
Seq2Seq	42.96	62.43	40.46	62.54	59.36	33.89	33.45	11.54
Transformer	46.65	60.90	43.34	61.39	69.82	38.62	44.06	10.59
Seq2Seq-copy	47.58	71.67	50.20	76.77	77.23	49.69	32.84	9.85
Transformer-copy	57.91	68.44	54.97	69.59	79.17	45.10	52.25	4.61
T5	55.36	77.79	67.66	77.63	74.19	54.63	61.74	6.22
BART	*78.53	84.64	77.21	84.95	85.36	61.82	*78.03	5.35
BART-UCD (ours)	76.58	*84.92	*77.99	*85.31	*87.80	*74.50	77.13	*5.11
BART-IBT (ours)	**83.69**	**87.82**	**82.47**	**88.19**	**87.92**	**81.39**	**83.06**	**3.12**

Table 2: Performance comparison for ISP on the PIL test set. The best performance for each metric is in bold and the second best has an asterisk (*).

Idiomatic sentence		But dear Caroline's got an almighty hangover, ***sick as a dog***, so I brought him over on the back of the bike.
Literal sentence		But dear Caroline's got an almighty hangover, **very ill**, so I brought him over on the back of the bike.
ISP	**Seq2Seq**	but caroline got, as as, so I brought him over .
	Transformer	but dear caroline's got an almighty hangover, ***sick as a dog***, so I brought him over.
	Seq2Seq-copy	but dear caroline's got an an, sick as as, so I brought him over on on the back.
	Transformer-copy	but dear caroline's got an almighty hangover, ***sick as a dog***, so I brought him over on the back of the bike.
	T5	But dear Caroline's got an almighty hangover, ***sick as a dog***, so I brought him over on the back of the bike.
	BART	But dear Caroline's got an almighty hangover, ***sick as a dog***, so I brought him over on the back of the bike.
	BART-IBT (Ours)	But dear Caroline's got an almighty hangover, **feeling sick**, so I brought him over on the back of the bike.
	BART-UCD (Ours)	But dear Caroline's got an almighty hangover, **<u>sick</u>**, so I brought him over on the back of the bike.

Table 3: A sample of generated literal sentences. Text in bold and italics represents the IEs, text in bold represents the correct literal counterparts in the outputs, and text in bold underlined represents the near-correct literal phrases.

(2) **Target inclusion** shows on a scale of 1-4 if the correct literal phrase was used in the output (1: the target phrase was not included in the output at all, 2: partial inclusion, 3: complete inclusion of a different phrase but with similar meaning with the target, and 4: complete inclusion).
(3) **Fluency** evaluates the naturalness and the readability of the output, including the appropriate use of the verb tense, noun and pronoun forms, on a scale of 1 to 4, ranging from "highly nonfluent" to "very fluent."
(4) **Overall** evaluates the overall quality of output on a scale of 0 to 2 like that used to evaluate paraphrases (Iyyer et al. 2018), jointly capturing meaning preservation and fluency: a score of 0 for a sentence that was clearly wrong, grammatically incorrect or does not preserve meaning; a score of 1 for a sentence with minor grammatical errors or meaning largely preserved from the original but not completely; score 2 denotes that the sentence is grammatically correct and the meaning is preserved.

Results and Discussion

BART-UCD. As shown in Table 2, without training on PIL, BART-UCD outperforms the supervised baselines in 6 out of 8 metrics for the task of ISP and achieves a competitive performance with the strongly supervised BART outperforming it by 2.44 (METEOR) and 12.68 (SARI) points.

BART-IBT. As shown in Table 2, BART-IBT achieves the best performance across all metrics, even though its actual performance may be underrepresented by the automatic metrics that fail to capture meaning equivalences despite differences in surface form.

Model Comparison. Overall, the pretrained BART model, our BART-IBT and BART-UCD perform competitively on ISP going by the metrics METEOR and ROUGE-1. However, a qualitative analysis shows that BART tends to copy the input sentence in the output 15% of time and on an average only modifies 9% of the tokens from the input sentences, suggesting an *overrepresentation* of its performance by the automatic metrics. On the contrary, while being good at copying context words (a desirable feature), BART-IBT outperforms the other models showing the best SARI score (a measure of the novelty in the generated output compared to the input). This underscores the importance of the iterative back-translation mechanism without which the performance gains would have been impossible. Moreover, we note that BART performs better on PIL while BART-UCD performs better on MIL. A plausible explanation for this divergence is that PIL is synthetically created idiomatic sentences whereas MIL is in-the-wild ones. Thus, MIL is an out of distribution, yet more general test data for BART that was trained on PIL. However, BART-UCD, being agnostic to PIL, is indifferent to the distribution shift in MIL.

Human Evaluation. The results of human evaluation are presented in Table 4. We note that the output of BART-IBT was rated the best across all the dimensions. It appears that the fine-tuned BART performs on par with BART-IBT in meaning preservation and fluency. However, BART's tendency to copy its input artificially inflates its meaning preservation and fluency scores. It is worth noting that when tested on MIL, both BART-UCD and BART-IBT outperform the pretrained BART in corresponding tasks, which speaks of the generalizability of BART-UCD and BART-IBT to the naturally occurring idiomatic sentences in MIL. Averaged over the four dimensions, the inter-annotator agree-

Model	PIL Test Set								MIL Test Set							
	Meaning		Target		Fluency		Overall		Meaning		Target		Fluency		Overall	
	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.	Scr.	Agr.
BART	0.73	0.88	2.56	0.57	**3.85**	0.80	1.30	0.56	0.53	0.92	1.70	0.54	2.37	0.80	0.92	0.58
BART-UCD	0.48	0.74	2.25	0.42	3.43	0.59	1.13	0.56	0.64	0.74	2.21	0.42	3.16	0.57	0.98	0.54
BART-IBT	**0.81**	0.83	**3.11**	0.47	**3.85**	0.80	**1.63**	0.47	**0.80**	0.89	**2.48**	0.47	**3.36**	0.63	**1.28**	0.47

Table 4: Human evaluation results for ISP based on the PIL and MIL test sets. The best performance is in bold. Scr. represents the humam evaluation scores and Agr. represents the human evaluation inter-annotator agreement.

English Idiomatic Sentence	I do not know if she is present , but I would like to **pass on** my deepest condolences to her .
German Translation (no ISP)	Ich weiß nicht, ob sie anwesend ist, aber ich möchte mein tiefstes Beileid
English Literal Sentence	I do not know if she is present , but I would like to **express** my deepest condolences to her .
German Translation (with ISP)	Ich weiß nicht, ob sie anwesend ist, aber ich möchte ihr mein tiefstes Beileid aussprechen

Table 5: Example that shows how ISP helps En-De machine translation.

ment score was 0.58 for BART-UCD and 0.62 for BART-IBT.

Error Analysis. The main challenge for all the models seems to be generating long informative literal phrases based on the correct sense of the IE. For example, BART-IBT replaces the IE *blow hot and cold* with *fluctuate*, which is inaccurate and the reason for annotators to diverge on Target inclusion and Overall scores.

Byproducts from BART-IBT. The back-translation mechanism used in BART-IBT leads to an ISG model (in addition to the ISP model) after training. To evaluate its competence, we perform the same automatic and human evaluations against the same set of baseline models. From the results, we found that BART-IBT outperforms all the baselines across all automatic metrics by wide margins, ranging from 11.76 higher in BLEU, 12.92 higher in ROUGE-2 and 16.32 higher in SARI over the next best model, while achieving the best performance across all human metrics as well. Besides, we also obtain a large scale parallel dataset, which includes 1,169 IEs with 15,627 idiomatic/literal sentence pairs. Table 3 shows the sentences generated.

Application

The challenges posed by IEs to machine translation owing to inadequate handling of non-compositional phrases has been documented by Fadaee, Bisazza, and Monz (2018) who also provide a challenge set of idiomatic sentences. Here we explore the extent to which using ISP as a preprocessing step to remove all the IEs from the input sentences can reduce the negative influence of IEs in machine translation. Performing ISP as a preprocessing step is inexpensive and flexible since it does not require the expansive development or retraining of new models to handle IE specifically and can be widely used in any downstream application.

Specifically, we use BART-IBT to first transfer the idiomatic sentences into literal sentences in the source language. Then, we use a state-of-the-art NMT system to translate the resulting literal sentences into the target language. We run experiments using the challenge test set for English-to-German translation constructed by (Fadaee, Bisazza, and Monz 2018) that consists of idiomatic sentences in English and their corresponding translations in German. There were 1,500 En-De pairs in the test set, using a total of 132 IEs. We used a pre-trained mBART (Liu et al. 2020) as the NMT system with all the parameters set to their default values.

As a result of the pre-processing using BART-IBT, the BLEU score on the challenge set improved from 10.1 to 10.7, which shows the effectiveness of the ISP in a downstream NLP application. Though this improvement may not seem substantial, we stress that this gain comes with just a preprocessing step and no other change in training. Table 5 shows an example of how ISP helps the translation of idiomatic sentences. In the original translation, the main verb *aussprechen* is missing. However, when the IE 'pass on' is replaced with 'expressed', the translation is complete.

Conclusion

In this paper, we studied the task of idiomatic sentence paraphrasing (ISP) in a zero- and low-resource setting. We proposed an unsupervised method that utilizes contextualized word embeddings and word definition sentence embeddings for ISP. In addition, we explored the use of a weakly supervised method based on an iterative back-translation mechanism. Our experiments and analyses demonstrate that unsupervised and weakly supervised methods show competitive paraphrasing performance in low-resource settings, with the weakly supervised method outperforming available baseline methods in all evaluation dimensions. Furthermore, the weakly supervised approach yields an ISG model and a large-scale parallel dataset.

The limitations of this study include conducting the study without a large parallel dataset of high quality, assuming one sense for IEs (Hümmer and Stathi 2006), limiting each sentence to have only one IE and using a list of IEs that did not account for the diversity of World Englishes (PITZL 2016). Future work should address these limitations.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant number R305A180211 to the Board of Trustees of the

University of Illinois. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Biddle, R.; Joshi, A.; Liu, S.; Paris, C.; and Xu, G. 2020. Leveraging Sentiment Distributions to Distinguish Figurative From Literal Health Reports on Twitter. In *Proceedings of The Web Conference 2020*, 1217–1227.

Cordeiro, S.; Ramisch, C.; Idiart, M.; and Villavicencio, A. 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)*, 1986–1997.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fadaee, M.; Bisazza, A.; and Monz, C. 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)*.

Gehrmann, S.; Deng, Y.; and Rush, A. M. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Gong, H.; Bhat, S.; and Viswanath, P. 2017. Geometry of compositionality. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gong, H.; Bhat, S.; Wu, L.; Xiong, J.; and Hwu, W.-m. 2019. Reinforcement Learning Based Text Style Transfer without Parallel Training Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3168–3180.

Gu, Y.; Wei, Z.; et al. 2019. Extract, Transform and Filling: A Pipeline Model for Question Paraphrasing based on Template. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 109–114.

Gupta, A.; Agarwal, A.; Singh, P.; and Rai, P. 2018. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Haagsma, H.; Bos, J.; and Nissim, M. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 279–287.

Hegde, C.; and Patil, S. 2020. Unsupervised paraphrase generation using pre-trained language models. *arXiv preprint arXiv:2006.05477*.

Huang, K.-H.; and Chang, K.-W. 2021. Generating Syntactically Controlled Paraphrases without Using Annotated Parallel Pairs. In *EACL*.

Hümmer, C.; and Stathi, K. 2006. Polysemy and vagueness in idioms: A corpus-based analysis of meaning. *International Journal of Lexicography*, 19(4): 361–377.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1875–1885. New Orleans, Louisiana: Association for Computational Linguistics.

Jackendoff, R. 1995. The boundaries of the lexicon. *Idioms: Structural and psychological perspectives*, 133–165.

Jhamtani, H.; Gangal, V.; Hovy, E.; and Nyberg, E. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Keskar, N. S.; McCann, B.; Varshney, L.; Xiong, C.; and Socher, R. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Krishna, K.; Wieting, J.; and Iyyer, M. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 737–762. Online: Association for Computational Linguistics.

Lavie, A.; and Agarwal, A. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, 228–231.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1865–1874.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

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

Liu, C.; and Hwa, R. 2016. Phrasal Substitution of Idiomatic Expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 363–373. San Diego, California: Association for Computational Linguistics.

Liu, C.; and Hwa, R. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Liu, C.; and Hwa, R. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1723–1731.

Liu, C.; and Hwa, R. 2019. A Generalized Idiom Usage Recognition Model Based on Semantic Compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6738–6745.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11): 39–41.

Nasr, A.; Ramisch, C.; Deulofeu, J.; and Valli, A. 2015. Joint dependency parsing and multiword expression tokenisation. In *Annual Meeting of the Association for Computational Linguistics*, 1116–1126.

Nivre, J.; and Nilsson, J. 2004. Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Norbury, C. F. 2004. Factors Supporting Idiom Comprehension in Children With Communication Disorders. *Journal of Speech, Language, and Hearing Research*, 47(5): 1179.

Nunberg, G.; Sag, I. A.; and Wasow, T. 1994. Idioms. *Language*, 70(3): 491–538.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

PITZL, M.-L. 2016. World Englishes and creative idioms in English as a lingua franca. *World Englishes*, 35(2): 293–309.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Salton, G.; Ross, R.; and Kelleher, J. 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)*, 36–41.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *CoRR*, abs/1704.04368.

Sprenger, S. A. 2003. *Fixed expressions and the production of idioms*. Ph.D. thesis, Radboud University Nijmegen Nijmegen.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Training Very Deep Networks. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Sudhakar, A.; Upadhyay, B.; and Maheswaran, A. 2019. "Transforming" Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. 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)*, 3260–3270.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4: 401–415.

Zeng, K.-H.; Shoeybi, M.; and Liu, M.-Y. 2020. Style Example-Guided Text Generation using Generative Adversarial Transformers. *arXiv preprint arXiv:2003.00674*.

Zeng, Z.; and Bhat, S. 2021. Idiomatic Expression Identification using Semantic Compatibility. *arXiv preprint arXiv:2110.10064*.

Zhang, X.; Zhao, J. J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.

Zhou, J.; Gong, H.; and Bhat, S. 2021a. From Solving a Problem Boldly to Cutting the Gordian Knot: Idiomatic Text Generation. *arXiv preprint arXiv:2104.06541*.

Zhou, J.; Gong, H.; and Bhat, S. 2021b. PIE: A Parallel Idiomatic Expression Corpus for Idiomatic Sentence Generation and Paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 33–48. Online: Association for Computational Linguistics.

Zhu, W.; and Bhat, S. 2020. GRUEN for Evaluating Linguistic Quality of Generated Text. *arXiv preprint arXiv:2010.02498*.