PIE: Parallel Idiomatic Expression Corpus for Idiomatic Sentence Generation and Paraphrasing

Anonymous ACL-IJCNLP submission

Abstract

Idiomatic expressions (IE) play an important role in natural language, and have long been a "pain in the neck" for NLP systems. Despite this, text generation tasks related to IEs remain largely under-explored. In this paper, we propose two new tasks of idiomatic sentence generation and paraphrasing to fill this research gap. We introduce a curated dataset of 823 IEs, and a parallel corpus with sentences containing them and the same sentences where the IEs were replaced by their literal paraphrases as the primary resource for our tasks. We benchmark existing deep learning models, which have state-of-the-art performance on related tasks using automated and manual evaluation with our dataset to inspire further research on our proposed tasks. By establishing baseline models, we pave the way for more comprehensive and accurate modeling of IEs, both for generation and paraphrasing.

1 Introduction

Idiomatic expressions (IEs) make language natural. These expressions, more broadly called a multiword expressions (MWEs) are (non-compositional) phrases whose meaning differs from the literal meaning of their constituent words taken together (Nunberg et al., 1994). Their use imparts naturalness and fluency (Wray and Perkins, 2000; Sprenger, 2003; Pawley and Syder, 2014; Schmitt and Schmitt, 2020), is prompted by pragmatic and topical functions in discourse (Simpson and Mendis, 2003) and often conveys a nuance in expression (stylistic enhancement) using imagery that is beyond what is available in the context (Nunberg et al., 1994). Idiomatic expressions, including phrasal verbs (e.g., carry out), idioms (e.g., pull one's leg) are also an essential part of a native speakers vocabulary and lexicon (Jackendoff, 1995).

English	Vote them out!
Spanish	¡Vote para sacarlos!
Arabic	!التصويت لهم
Chinese	投票给他们！
Hindi	उन्हें वोट दें!
French	Votez-les!
German	Stimmen Sie sie ab!
Korean	투표하세요!
Russian	Проголосуйте за них!

Figure 1: State-of-the-art machine translations of "Vote them out!" into different languages mean the opposite.

IEs constitute a ubiquitous part of daily language and social communication, primarily used in conversation, fiction and news (Biber et al., 1999), frequently used by teachers when presenting their lessons to students (Kerbel and Grunwell, 1997) and occur cross-lingually (Baldwin et al., 2010; Nunberg et al., 1994). Their non-compositionality is the reason for their classical standing as "a pain in the neck" (Sag et al., 2002) and "hard going" (Rayson et al., 2010) for NLP.

The Oxford English dictionary defines the phrasal verb (an IE) *vote out* as 'To turn (a person) out of office.' Using Google translate[1] to translate the topical slogan "vote them out!" into eight of the world's most spoken and relatively resource-rich languages yielded the results shown in Figure 1. As native speakers will attest, other than in Spanish, all the translations mean just the opposite, "vote for them!" This, and other studies on computational processing of idioms and metaphors in (Salton et al., 2014; Shao et al.; Shutova et al., 2013) reinforce the need for nuanced language processing—a grand challenge for NLP systems.

Gaining a deeper understanding of IEs and their literal counterparts is an important step toward

[1] https://translate.google.com/. Accessed November 19, 2020

this goal. In this paper, we introduce two novel tasks related to paraphrasing between literal and idiomatic expressions in unrestricted text: (1) Idiomatic sentence simplification (ISS) to automatically paraphrase idiomatic expressions in text, and 2) Idiomatic sentence generation (ISG) to replace a literal phrase in a sentence with a synonymous but more vivid phrase (e.g., an idiom). ISS directly addresses the need for performing text simplification in several application settings, including summarizers (Klebanov et al., 2004) and parsing (Constant et al., 2017). Moreover, ISS may actually be helpful when an idiomatic expression does not have an exact counterpart in a target language. This is akin to the 'translation by paraphrase' strategy recommended for human translation when the source language idiom is obscure and non-existent in the target language (Baker, 2018). On the other hand, ISG advances the area of text style transfer (Jhamtani et al., 2017; Gong et al., 2019) bringing the as yet unexplored dimension of nuanced language to style transfer.

A second important component of this paper is the introduction of a new curated dataset of parallel idiomatic and literal sentences, created for the purpose of advancing progress in nuanced language processing and serving as a testbed for the proposed tasks. Recent literature has explored several aspects of figurative and nonliteral language processing, including detecting and interpreting metaphors (Shutova, 2010b; Shutova et al., 2013), disambiguating IEs for their figurative or literal in a given context (Constant et al., 2017; Savary et al., 2017; Liu and Hwa, 2019) and analyzing sarcasm (Muresan et al., 2016; Joshi et al., 2017; Ghosh et al., 2018), by using curated datasets of sentences with linguistic processes in the wild. These datasets are ill-suited for the proposed tasks because they consist of specific figurative constructions (metaphors) (Shutova, 2010a), do not cover multiple IEs (Cook et al., 2008; Korkontzelos et al., 2013), or are not parallel (Haagsma et al., 2020; Savary et al., 2017) underscoring the need for a new dataset.

The newly constructed dataset permits us to benchmark the performance of several state-of-the-art neural network architectures (seq2seq and pretrained+fine-tuned models, with and without copy-enrichment) that have demonstrated competitive performance in the related tasks of simplification, and style transfer. Using automatic and manual evaluations of the outputs for the two tasks, we find that the existing models are inadequate for the proposed tasks. The sequence-to-sequence models clearly suffer from data sparsity, the added copy mechanism helps preserve the context that is not replaced, and despite their prior knowledge of the pretrained models, they are still limited in their ability to paraphrase and generate. This leads us to discussing novel insights, applications and future directions for related research.

The main contributions of this work are summarized as follows.

1. We propose two new tasks related to idiomatic expressions—idiomatic sentence simplification and idiomatic sentence generation;

2. We introduce a curated dataset of 823 idiomatic expressions, replete with sentences containing these IEs in the wild and the same sentences where the IEs were replaced by their literal paraphrases.

3. We use the combination of the new dataset and the proposed tasks as a lens through which we gain novel insights about the capabilities of deep learning models for processing nuanced language generation and paraphrasing.

2 Task Definition

We propose two new tasks: **idiomatic sentence generation** transforms a literal sentence into a sentence involving idioms. Used frequently in everyday language, idioms are known to add color to expressions and improve the fluency of communication. The idiomatic rewriting improves the quality of text generation in that it could enhance the textual diversity and convey abstract and complicated ideas in a succinct manner. For example, the idiomatic sentence *BP cut corners and violated safety requirements.* conveys the same idea as its literal counterpart *BP saved time, money and energy and violated safety requirements*, but in a more vivid and succinct manner.

The second task is **idiomatic sentence paraphrasing**, simplifying sentences with idioms into literal expressions. As an example, the sentence– *It is certainly not a sensible move to cut corners with national security*– has the idiom *cut corners* replaced the literal counterpart *save money*. By paraphrasing the idioms from which machine translation often suffers, our task of idiomatic sentence paraphrasing can also benefit machine translation.

In this work, we distinguish our task of idiomatic sentence generation from idiom generation. While the latter task creates new idioms with novel word combinations, our study is to use existing idioms in a sentence and preserve the semantic meaning.

The task of idiomatic sentence paraphrasing is closely related to text simplification that has mostly been studied as related tasks of lexical paraphrasing and syntactic paraphrasing (Xu et al., 2015). A significant departure of this task from that of these related tasks that centrally address style is that (i) we aim for local synonymous paraphrasing by transforming not the entire sentence but a phrase in the sentence, (ii) the transformation is not related to syntactic structures, but related to the complexity in meaning[2]. We propose doing joint monolingual translation with simplification and is similar in spirit to (Agrawal and Carpuat, 2020).

There are many technical challenges to performing these tasks. The task of idiomatic sentence paraphrasing involves first identifying that an expression is an idiom and not a literal expression (e.g. *black sheep*) (Fazly et al., 2009; Korkontzelos et al., 2013; Liu and Hwa, 2019). Once identified, the IE may have multiple senses (e.g. *tick off*) and its appropriate sense will need to be identified before paraphrasing it. Third, an appropriate literal phrase will have to be generated to replace the IE. Finally, the literal phrase will have to be fit in the surrounding sentential context for a fluent construction. For idiomatic sentence generation, the context of the literal phrase could permit more than one candidate idiom (e.g. keep quiet). In this study, we assume that we have an idiomatic sentence and leave it to future work to explore the task in conjunction with this step.

3 Related Work

The theme of this paper is naturally connected to three streams of text generation tasks—paraphrasing, style transfer and metaphoric expression generation. We will discuss these tasks and also the datasets used in these tasks to study their similarities and differences to our dataset and tasks.

3.1 Paraphrase

The aim of paraphrasing is to rewrite a given sentence while preserving its original meaning. Being widely studied in the recent research, many datasets have been constructed to facilitate the task. PPDB (Ganitkevitch et al., 2013), MRPC [3], Twitter URL Corpus (Lan et al., 2017), Quora [4] and ParaNMT-50M (Wieting and Gimpel, 2017) have been the most commonly used datasets. The most commonly used Seq2Seq models have been successfully applied to paraphrasing Prakash et al. (2016); Gupta et al. (2018); Iyyer et al. (2018); Yang et al. (2019). Besides the end-to-end models, a template-based pipeline model was proposed to divide paraphrase generation into template extraction, template transforming and template filling (Gu et al., 2019).

However, unlike paraphrasing a sentence or a literal-to-literal paraphrasing task, our proposed tasks are more constrained given the existence of idiomatic expressions. This renders the datasets used for the task of paraphrasing and the associated paraphrasing models inadequate for our task. Our dataset is created to fill this need to advance a fundamental understanding of idiomatic text generation and paraphrasing. Therefore, research into our tasks and dataset can also be used for paraphrasing when only part of the sentence need to be paraphrased or idiom need to be paraphrased.

3.2 Style Transfer

The task of style transfer can be defined as rewriting sentences into those with a target style. Recent research has primarily focused sentiment manipulation and changes in writing styles (Jhamtani et al., 2017; Gong et al., 2019). Our proposed tasks are different from the nature of style transfer studies in recent works because (i) our tasks retain a large portion of the input sentences while style transfer may need to completely change the input sentences, and (ii) our tasks explore the nuance component of style, an aspect heretofore unexplored. To test different models' performance on style transfer, several non-parallel corpora have been used (Yelp (Shen et al., 2017), Grammarly's Yahoo Answers Formality Corpus (Rao and Tetreault, 2018), Amazon Food Review dataset (McAuley and Leskovec, 2013) and Product Review dataset (He and McAuley, 2016)). Despite their size, they lack the focus on IEs and are all non-parallel. This has led to the the study of unsupervised methods for style transfer, including cross-aligned auto-encoder (Hu et al., 2017),

[2]The consideration of whether idioms are semantic- or pragmatic- or discourse-level phenomena is important, but beyond the scope of this paper.

[3]https://www.microsoft.com/en-us/download/details.aspx?id=52398

[4]https://www.kaggle.com/aymenmouelhi/quora-duplicate-questions

VAE (Hu et al., 2017), Generative Adversarial Network (Zeng et al., 2020), reinforcement learning for constraints in style transfer (Xu et al., 2018; Gong et al., 2019) and pipeline models (Li et al., 2018; Sudhakar et al., 2019). Owing to the essential departure of our tasks from those of previously studied style transfer tasks, and the limitation of non-parallel corpus, we create our own parallel dataset which focuses on IEs.

3.3 Metaphoric Expression Generation

Prior work on automated metaphor processing has primarily focused on their identification, interpretation and also generation. (Shutova, 2010b; Shutova et al., 2013; Abe et al., 2006). Also, data for this task is extremely sparse: there are not any large scale parallel corpora containing literal and metaphoric paraphrases which aims for metaphor generation. The most useful one is that of (Mohammad et al., 2016). However, their dataset has a small number (171) of metaphoric sentences extracted from WordNet. Early works on metaphor generation mainly focus on phrase level metaphor and template-based generation (Terai and Nakagawa, 2010; Ovchinnikova et al., 2014). Recent works also explore the power of neural networks (Mao et al., 2018; Yu and Wan, 2019; Stowe et al., 2020). However, most of the research on metaphor generation suffer from the lack of parallel corpora.

Our proposed tasks share some similarities with metaphor generation but also have differences. Instead of focusing on paraphrase of single word like most metaphor generation work, our tasks often require a mapping between two multi-word expressions, which makes our tasks more challenging.

3.4 Text Simplification

Text simplification aims to rewrite input sentences into lexically and/or syntactically simplified forms. The Simple Wikipedia Corpus (Zhu et al., 2010) and more recently, the Newsela dataset (Xu et al., 2015) and the WikiLarge dataset (Zhang and Lapata, 2017) dominate the research area. The use of different machine learning models have also been explored for this task, including statistical machine translation model (Wubben et al., 2012), the Seq2Seq architecture (Nisioi et al., 2017) and the Transformer architecture (Zhao et al., 2018).

Departing from previous attempts at lexical or syntactic simplification, our proposed task of idiomatic sentence paraphrasing aims to simplify the nuance of non-compositional and figurative expressions thereby permitting a more literal understanding of the sentence.

We summarize the datasets of the related tasks in Table 1.

4 Building the Dataset

We describe the details of the data collection, data annotation, corpus analyses and comparisons with other existing corpora.

4.1 Data Collection

The Parallel Idiomatic Expression Corpus (PIE), consists of idiomatic expressions (IEs), their definitions, sentences containing the IEs and corresponding sentences where the IEs are replaced with their literal paraphrases. One instance of the dataset is shown in Figure 2.

We collected a list of 1042 popular IEs and their meanings from an educational website [5] that has a broad coverage of frequently used IEs including phrasal verbs, idioms and proverbs. For a broad coverage of IEs we did not limit them to a specific syntactic category. The list was then split between the members of the research team consisting of a native English speaker, and three near-native English speakers. Some IEs such as "tick off" (Figure 2) have multiple senses. The annotators labeled the sense of IEs in given sentences according to the sense information from reliable sources including the Oxford English Dictionary[6], the Webster Dictionary [7] and the Longman Dictionary of Contemporary English[8]. IEs that were not available in any of the popular dictionaries were excluded from dataset as were proverbs that are independent clauses (e.g., *the pen is mightier than the sword*). To guarantee each sense is well represented, the annotators collected at least 5 sentences for each sense of an IE from online sources (e.g., the Contemporary corpus of American English, and examples listed in dictionaries).

The data collection step yielded the corpus with a total of 823 IEs and 5170 sentence-pairs using these IEs (an average of 6.3 sentence-pairs per idiom). We also note that every instance (idiomatic-literal pair) is only one sentence long. The corpus statistics are summarized in Table 2.

Idiom	Tick Off	
Sense	to complete an item on a list	to make someone angry or offended
Idiomatic Sentence	I would like to tick off some more items on my list before going home	My decision is going to tick off my entire family.
Idiomatic Labels	O O O O **B I** O O O O O O O O O	O O O O O **B I** O O O.
Literal Sentence	I would like to cross out some more items on my list before going home	My decision is going to anger my entire family.
Literal Labels	O O O O **B I** O O O O O O O O O	O O O O O **B** O O O.

Figure 2: An example from our dataset. Idioms are highlighted in blue, and their literal paraphrases are in red.

Dataset	Parallel	Task	Size	# idioms	Sent Len (original)	Sent Len (target)
PIE (ours)	✔	Idiom Generation/Paraphrasing	3,524/823/823	823	18.5	19.0
Para-NMT	✔	Paraphrase	5,370,128	-	11.43	10.56
WikiLarge	✔	Text Simplification	296,402/992/359	-	24.1	15.51
Metaphor	✔	Metaphor Generation	171	-	7.30	7.37

Table 1: Comparison of our dataset with other related datasets. Training, validation and testing size splits are provided when applicable. Data in all these datasets is a combination of collection from the wild and manual generation. In our corpus, original sentences are idiomatic sentences and target sentences are literal sentences.

Statistics	# of instances	Avg. # of words
Idioms	823	3.2
Sense	862	7.9
Idiomatic sent	5170	19.0
Literal sent	5170	18.5

Table 2: Statistics of our parallel corpus.

% n-grams	PIE	Para-NMT	Wiki-Large	Metaphor
uni-grams	13.86	46.34	36.2	16.88
bi-grams	23.60	71.24	52.56	36.59
tri-grams	30.19	82.26	58.75	59.61
4-grams	36.51	86.46	62.79	74.41

Table 3: The percentage of n-grams in source sentences which do not appear in the target sentences. In our case, it is the percentage of n-grams in literal sentences which do not appear in the idiomatic sentences.

4.2 Data Annotation

In order to create the parallel dataset of idiomatic and literal sentences for the proposed tasks, a native English speaker was asked to rewrite each idiomatic sentence into its literal form, where the IE was replaced by a literal phrase. As part of this manual paraphrasing, the annotator was asked to paraphrase only the IE so as not to alter its meaning in the context of the sentence, preserving the phrases syntactic function and to conform to the sense definition. The rest of the sentence was to be left unchanged. The annotator is free to use original sense definition when rewriting or use paraphrases of sense definition. After the first annotation pass, the researchers checked the literal sentences generated by the first annotator and corrected any errors.

To specify the span of the IE in each idiomatic sentence and that of the literal paraphrase in the corresponding literal sentence, **BIO** labels were used; **B** marks the beginning of the idiom expressions (resp. the literal paraphrases), **I** the other words in the IE (resp. words in the literal paraphrases) and **O** all the other words in the sentences. This labeling was done automatically considering that the only difference between a given idiomatic sentence and its literal sentence is the replacement of idiom with literal phrase. An example of the **BIO** labeled sentence pair is shown in Figure 2.

4.3 Corpus Analyses

We summarize the statistics of our PIE dataset in Table 2 and compare it with existing datasets in Table 1. We notice that the parallel sentences in our dataset are comparable in terms of sentence length, while simple sentences are much shorter in the text simplification dataset. This suggests that the tasks we propose may not result in significantly shorter sentences compared to their inputs, and this constitutes a core departure from the task of text simplification. Moreover, the sentences in our dataset are longer on an average compared to the

[5]www.theidioms.com
[6]https://www.oxfordlearnersdictionaries.com
[7]https://www.merriam-webster.com
[8]https://www.ldoceonline.com

sentences in existing datasets (with the exception of text simplification data). This can pose challenges to the text generation model performing the tasks proposed in the paper.

We also report the percentage of n-grams in the literal sentences which do not appear in the idiomatic sentences as a measure of the difference between the idiomatic and literal sentences. As shown in Table 3, there is smaller variation between the source sentences and the target sentences in our dataset. This is again due to the nature of our task, which calls for a local paraphrasing (rewriting only a part of the sentence).

We note that IEs may be naturally ambiguous due to the existence of both figurative and literal senses, as also pointed out in previous works.A small portion of IEs in our dataset have multiple senses, and one example is "tick off " in Figure 2. Table 4 presents the distribution of the senses in the IEs in our dataset, and the average number of senses is 1.05, suggesting that the majority IEs in our dataset are monosemous.

4.4 Dataset quality

Noting that the idiomatic to literal sentences were manually created, the quality of our dataset may be called into question. We point out that in an effort to quickly use sentences of good quality and in line with existing datasets for related tasks with idiomatic expressions (Haagsma et al., 2020; Korkontzelos et al., 2013) we collected idiomatic expressions in the wild. However, as acknowledged by previous dataset creation efforts, not all IEs occur equally frequently, which can result in a representation bias. In addition, finding true paraphrases of IEs in the wild is hard. In light of these practical data-related concerns, we resorted to a manual paraphrasing of the IEs as a trade-off between naturalness and representation. This idea of using non-natural instances is also influenced by successful recent approaches to training data collection and data augmentation using synthetic methods reported in severely resource-constrained domains such as machine translation (Sennrich et al., 2016) and clinical language processing (Ive et al., 2020).

5 Experiments

5.1 Experimental setup

Considering that our tasks of idiomatic sentence generation and paraphrasing have never been studied before and the fact that they are both text generation tasks, we first choose some basic end-to-end models which have shown state-of-the-art performance on other related text generation tasks. Accordingly, we used the LSTM-based Seq2Seq model (Sutskever et al., 2014) and the transformer architecture (Vaswani et al., 2017). These will be alluded to as the **models that translate.** Based on the observation that the idiomatic sentences and literal sentences share much of the context, which remains unchanged during generation, we also use the copy-enriched seq2seq model (Jhamtani et al., 2017) and the transformer model with a copy mechanism (Gehrmann et al., 2018) [9] (hereafter collectively called the **models that copy**). Moreover, considering the similarity between our tasks and paraphrasing, we also choose the pretrained BART (Lewis et al., 2019), used for text simplification and paraphrasing, which was fine-tuned on our training instances. Finally, we used a retrieve-delete-generate pipeline model (Sudhakar et al., 2019) that showed a competitive performance for style transfer. More details are provided in Appendix.

# senses	# of idioms	# pairs	Avg. # of words
1	788	4788	3.2
2	31	322	2.6
3	4	60	2.0

Table 4: Statistics of sense distribution. An idiom has an average of 1.05 senses.

5.2 Evaluation

For automatic evaluation, Rouge (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and SARI (Xu et al., 2016) are used to compare the similarity between the generated sentences and the references. These metrics has been widely used in various text generation tasks such as paraphrasing, style transfer and text simplification. To measure linguistic quality, we use a pre-trained language model BERT to calculate perplexity scores and a recently proposed measure, GRUEN (Zhu and Bhat, 2020).

Considering that automatic evaluation cannot fully analyze the results, we use human evaluation as a complement to the automatic evaluation metrics. For each task, We randomly sampled 100 input sentences and the corresponding outputs of all baselines. Human annotations were collected with respect to context, style and fluency of generated sentences based on the following criteria.
(1) **Context preservation** measures how well the context surrounding the idiomatic/literal phrase is

[9] https://github.com/lipiji/TranSummar

Model	BLEU		SARI		GRUEN	
	s2i	i2s	s2i	i2s	s2i	i2s
Seq2Seq	25.16	42.96	24.13	33.89	32.25	33.45
Seq2Seq with copy	38.02	47.58	43.02	49.69	27.79	32.84
Transformer	45.58	46.65	36.67	38.62	44.05	44.06
Transformer with copy	59.56	57.91	39.93	45.10	59.27	52.25
Pretrained BART	**79.32**	**78.53**	62.30	61.82	**77.49**	**78.03**
Pipeline	65.56	70.03	**67.64**	**62.45**	67.27	74.16

Table 5: Automatic evaluation results for the task of idiomatic sentence generation (**s2i**) and idiomatic sentence paraphrasing (**i2s**).

Model	Context		Target		Fluency		Overall	
	s2i	i2s	s2i	i2s	s2i	i2s	s2i	i2s
Seq2Seq	1.3	1.2	1.1	1.1	1.1	1.0	1.7	1.7
Seq2Seq with copy	3.8	3.8	1.6	1.7	2.1	3.4	3.5	3.6
Transformer	4.2	4.3	1.3	1.2	3.3	3.4	3.4	3.3
Transformer with copy	5.4	5.3	1.2	1.6	4.6	4.6	3.9	4.2
Pretrained BART	**5.9**	**5.9**	1.5	2.1	**5.9**	**5.9**	4.4	5.0
Pipeline	5.6	5.8	**1.7**	**2.2**	5.1	5.3	**4.5**	**5.1**

Table 6: Human evaluation results for the two tasks.

preserved in the output.
(2) **Target inclusion** checks whether the correct IE or literal phrase is used in the output.
(3) **Fluency** evaluates the fluency and readability of the output sentence including how appropriately the verb tense, noun and pronoun forms are used.
(4) **Overall meaning** evaluates the overall quality of the output sentence.

For each output sentence, two annotators with native-speaker-level English proficiency were asked to rate it on a scale from 1 to 6 in terms of the context preservation, fluency and overall meaning. Higher scores indicate better quality. As for the target inclusion, they were asked to rate it on a scale from 1 to 3. Score 1 denotes that the target phrase is not included in the input at all, 2 denotes partial inclusion, and 3 is for the complete inclusion. We report the average score over all samples for each baseline in each aspect.

5.3 Results and Discussion

Results. We report the automatic and human evaluation results in Table 5 and 6. More detailed results with all the metrics considered are in the appendix. On both tasks, going by the automatic metrics, copy-enriched transformer, pretrained BART model and the pipeline model perform better than other baselines. Pretrained BART achieved the best performance in BLEU and GRUEN, and the pipeline model does best in SARI. As for human evaluation, BART and the pipeline again achieve the best performance among the baselines. While BART is the best in preserving contexts and achieving fluency, the pipeline is the best in idiom paraphrasing and generation.

Model competence. BART and the pipeline model outperform other baselines in that they leverage auxiliary information (large pretaining corpora and selective idiomatic expression information, respectively) which is not available to the other models. The benefit of the copy mechanism by explicitly retaining the contexts as required by our tasks, is shown in the corresponding gains in automatic and manual evaluation scores for both Seq2Seq and transformer models.

When it comes to the comparison between BART and the pipeline, BART does better in retaining the contexts surrounding idiomatic expressions given its high context score in human evaluation while the pipeline is better at handling the idiomatic part, i.e., target inclusion. Despite the reported superior performance of BART in related text generation tasks (Lewis et al., 2019), our experiments show that BART has limited capability in idiom paraphrasing and generation. The pipeline method, by virtue of error propagation from its retrieval and deletion modules suffers in terms of both the context preservation and fluency. For task of idiomatic sentence generation, the accuracy for retrieval module is 0.27 and F1 score for deletion module is 0.68. For task of idiomatic sentence paraphrasing, the accuracy for retrieval module is 0.96 and F1 score for deletion module is 0.85.

Comparison between two tasks. According to human evaluation results in Table 6, both BART and the pipeline received higher scores for idiomatic sentence paraphrasing than idiomatic sentence generation, suggesting that paraphrasing is

Corr	Context		Target		Fluency		Overall	
	s2i	i2s	s2i	i2s	s2i	i2s	s2i	i2s
BLEU	0.27	0.17	0.56	0.28	0.09	0.02	0.64	0.29
SARI	0.21	0.17	0.61	0.40	-0.02	-0.01	0.61	0.39
GRUEN	-0.18	-0.07	-0.11	0.12	0.23	0.15	-0.18	0.11

Table 7: Instance-level Spearman's correlations between human and automatic evaluation for pretrained BART.

relatively easier among the two tasks. This resonates with our intuitions as language users in that given a lexical resource, paraphrasing an IE is easier than finding the right IE to replace a phrase.

Limitation of automatic metrics. Table 7 presents the correlation between automatic metrics and human judgements. All the correlation scores between automatic metrics and human evaluate scores are not high enough. For BLEU and SARI which mainly measure overlapping tokens, some synonymous idioms or literal phrases are ignored while they are still appropriate. For GRUEN metric aiming to measure text quality, its correlation scores with fluency and overall meaning are quite low. Therefore, more reliable automatic evaluation methods are needed.

Error analysis. For task of idiomatic sentence generation, the primary challenge is in identifying the appropriate IE, which is the hardest when the IE is highly non-compositional (e.g., *bird of passage* in Table 10). The examples are presented in Table 10 in the Appendix. For the task of idiomatic sentence paraphrasing, one challenge is the difficulty of choosing the correct sense of the idiom. As is shown in Table 11 in Appendix, all the baseline models were unable to generate the correct literal phrases for "alpha and omega", which have two senses: the beginning and the end; the principal element. Also, we noticed that strong baseline models of pretrained BART and the pipeline model tend to use a short but inaccurate literal phrase when the correct one is long. Paraphrasing of "the bird of passage" in Table 11 is an example.

Applications: Research in the proposed tasks has many potential practical applications. 1) An idiomatic sentence paraphrasing tool would be of importance in several language processing settings encountered by by humans and machines. The non-literal and stylized meaning of multi-word expressions (MWE) in general and idioms in particular, pose two broad kinds of challenges. First, they affect readability in target populations. For instance, despite their intact structural language competence, individuals with Asperger syndrome and more broadly those with autism spectrum disorder are known to experience significant challenges understanding figurative language (idioms) in their native language (Kalandadze et al., 2018). It is also widely acknowledged that idiomatic expressions are some of the hardest aspects of language acquisition and processing for second language learners (Liontas, 2002; Ellis et al., 2008; Canut et al., 2020). Moreover, natural language processing systems are known to be negatively impacted by idioms in text ((Salton et al., 2014; Shao et al.; Shutova et al., 2013) shown the negative impact of idioms and metaphors on machine translation leading to awkward or incorrect translations from English to other languages). Fruitful results of this task can lead to a system capable of recognizing and interpreting IEs in unrestricted text in a central component of any real-world NLP application (e.g., information retrieval, machine translation, question answering, information extraction, and opinion mining).2) A realistic application of the idiomatic sentence generation task would be for computer-aided style checking, where a post-processing tool could suggest a list of idioms to replace a literal phrase in a sentence. 3) True integration with an external NLP application would require combining the first step of IE identification followed by paraphrasing as done in (Shutova et al., 2013), which will require a combination of the paraphrasing with identification, and can be a future direction for research.

6 Conclusions

To conclude, in this paper, we proposed two new tasks: idiomatic sentence generation and paraphrasing. We also presented PIE, the first parallel idiom corpus. We benchmark existing end-to-end trained neural network models and a pipeline method on PIE and analyze their performance for our tasks. Our experiments and analyses reveal the competence and shortcomings of available methods, underscoring the need for continued research on processing idiomatic expressions.

There are many possibilities for improving performance through more extensive exploration of richer model architectures and using more reliable evaluation methods, which we leave as future work.

References

Keiga Abe, Kayo Sakamoto, and Masanori Nakagawa. 2006. A computational model of the metaphor generation process. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Sweta Agrawal and Marine Carpuat. 2020. Multitask models for controlling the complexity of neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 136–139.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Robert Baldwin, Martin Cave, and Martin Lodge. 2010. *The Oxford handbook of regulation*. Oxford university press.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. Longman grammar of written and spoken english. *Harlow: Longman*.

Emmanuelle Canut, Juliette Delahaie, and Magali Husianycia. 2020. Vous avez dit falc? pour une adaptation linguistique des textes destinés aux migrants nouvellement arrivés. *Langage et societe*, (3):171–201.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

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.

Nick C Ellis, RITA Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol. *Tesol Quarterly*, 42(3):375–396.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

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.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 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)*, pages 3168–3180.

Yunfan Gu, Zhongyu Wei, 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)*, pages 109–114.

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

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.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digital Medicine*, 3(1):1–9.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 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)*, pages 1875–1885.

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

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Tamar Kalandadze, Courtenay Norbury, Terje Nærland, and Kari-Anne B Næss. 2018. Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

Debra Kerbel and Pam Grunwell. 1997. Idioms in the classroom: An investigation of language unit and mainstream teachers' use of idioms. *Child Language Teaching and Therapy*, 13(2):113–123.

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text simplification for information-seeking applications. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 735–747. Springer.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *arXiv preprint arXiv:1708.00391*.

Alon Lavie and Abhaya Agarwal. 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*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Juncen Li, Robin Jia, He He, and Percy Liang. 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)*, pages 1865–1874.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

John Liontas. 2002. Context and idiom understanding in second languages. *EUROSLA yearbook*, 2(1):155–185.

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.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. 2016. Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology*, 67(11):2725–2737.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Ekaterina Ovchinnikova, Vladimir Zaytsev, Suzanne Wertheim, and Ross Israel. 2014. Generating conceptual metaphors from proposition stores. *arXiv preprint arXiv:1409.7619*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Andrew Pawley and Frances Hodgetts Syder. 2014. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication*, pages 203–239. Routledge.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begona Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1-2):1–5.

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.

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.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Norbert Schmitt and Diane Schmitt. 2020. *Vocabulary in language teaching*. Cambridge university press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. Evaluating machine translation performance on chinese idioms with a blacklist method.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Ekaterina Shutova. 2010a. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037. Association for Computational Linguistics.

Ekaterina Shutova. 2010b. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Rita Simpson and Dushyanthi Mendis. 2003. A corpus-based study of idioms in academic speech. *Tesol Quarterly*, 37(3):419–441.

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

Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych. 2020. Metaphoric paraphrase generation. *arXiv preprint arXiv:2002.12854*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 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)*, pages 3260–3270.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Asuka Terai and Masanori Nakagawa. 2010. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*, pages 142–147. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Alison Wray and Michael R Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication*, 20(1):1–28.

Sander Wubben, EJ Krahmer, and APJ van den Bosch. 2012. Sentence simplification by monolingual machine translation.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Qian Yang, Dinghan Shen, Yong Cheng, Wenlin Wang, Guoyin Wang, Lawrence Carin, et al. 2019. An end-to-end generative architecture for paraphrase generation. 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 3123–3133.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 861–871.

Kuo-Hao Zeng, Mohammad Shoeybi, and Ming-Yu Liu. 2020. Style example-guided text generation using generative adversarial transformers. *arXiv preprint arXiv:2003.00674*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint arXiv:1810.11193*.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 94–108.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.