

# MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding

Tuhin Chakrabarty<sup>1</sup> Xurui Zhang<sup>3\*</sup>, Smaranda Muresan<sup>1,4</sup> and Nanyun Peng<sup>2</sup>

<sup>1</sup>Department of Computer Science, Columbia University

<sup>2</sup>Computer Science Department, University of California, Los Angeles,

<sup>3</sup>Tsinghua University, <sup>4</sup>Data Science Institute, Columbia University

{tuhin.chakr, smara}@cs.columbia.edu  
thuzxr@gmail.com, violetpeng@cs.ucla.edu

## Abstract

Generating metaphors is a challenging task as it requires a proper understanding of abstract concepts, making connections between unrelated concepts, and deviating from the literal meaning. In this paper, we aim to generate a metaphoric sentence given a literal expression by replacing relevant verbs. Based on a theoretically-grounded connection between metaphors and symbols, we propose a method to automatically construct a parallel corpus by transforming a large number of metaphorical sentences from the Gutenberg Poetry corpus (Jacobs, 2018) to their literal counterpart using recent advances in masked language modeling coupled with commonsense inference. For the generation task, we incorporate a metaphor discriminator to guide the decoding of a sequence to sequence model fine-tuned on our parallel data to generate high quality metaphors. Human evaluation on an independent test set of literal statements shows that our best model generates metaphors better than three well-crafted baselines 66% of the time on average. Moreover, a task-based evaluation shows that human-written poems enhanced with metaphors proposed by our model are preferred 68% of the time compared to poems without metaphors.

## 1 Introduction

Czech novelist Milan Kundera in his book “The unbearable lightness of being” said

*“Metaphors are not to be trifled with. A single metaphor can give birth to love.”*

Metaphors allow us to communicate not just information, but also feelings and complex attitudes (Veale et al., 2016). While most computational work has focused on metaphor detection (Gao et al., 2018; Stowe et al., 2019; Shutova et al., 2010; Tsvetkov et al., 2014; Veale et al., 2016; Stowe and Palmer, 2018), research on metaphor generation is

Literal Input1	The wildfire <b>spread</b> through the forest at an amazing speed.
GenMetaphor1	The wildfire <b>danced</b> through the forest at an amazing speed.
Literal Input2	The window panes were <b>rattling</b> as the wind blew through them
GenMetaphor2	The window panes were <b>trembling</b> as the wind blew through them

Table 1: Examples of two generated metaphors GenMetaphor1 and GenMetaphor2 by our best model MERMAID from their literal inputs.

under-explored (Yu and Wan, 2019; Stowe et al., 2020). Generating metaphors could impact many downstream applications such as creative writing assistance, literary or poetic content creation.

Relevant statistics demonstrate that the most frequent type of metaphor is expressed by verbs (Steen, 2010; Martin, 2006). We therefore focus on the task of generating a metaphor starting from a literal utterance (Stowe et al., 2020), where we transform a literal verb to a metaphorical verb. Table 1 shows examples of literal sentences and the generated metaphors.

To tackle the metaphor generation problem we need to address three challenges: 1) the lack of training data that consists of pairs of literal utterances and their equivalent metaphorical version in order to train a supervised model; 2) ensuring that amongst the seemingly endless variety of metaphoric expressions the generated metaphor can fairly consistently capture the same general meaning as the literal one, with a wide variety of lexical variation; and 3) computationally overcome the innate tendency of generative language models to produce literal text over metaphorical one.

In an attempt to address all these challenges, we introduce our approach for metaphor generation called MERMAID (MEtaphor geneRation with syMbolism And dIscriminative Decoding), making the following contributions:

- A method to automatically construct a corpus

\*Work down when the author is interning at UCLA.

that contains 93,498 parallel [*literal sentence, metaphorical sentence*] pairs by leveraging the theoretically-grounded relation between *metaphor* and *symbols*. Barsalou et al. (1999) showed how *perceptual symbols* arising from perception are used in conceptual tasks such as representing propositions and abstract concepts. Philosopher Susanne Langer in her essay “Expressiveness and Symbolism” stated “A metaphor is not language, it is an idea expressed by language, an idea that in its turn functions as a *symbol* to express something”. Our approach has two steps: 1) identify a set of sentences that contains *metaphorical verbs* from an online poetry corpus; 2) convert these metaphorical sentences to their literal versions using Masked Language Models and structured common sense knowledge achieved from COMET (Bosselut et al., 2019), a language model fine-tuned on ConceptNet (Speer et al., 2017). For the later, we exploit the SymbolOf relation to make sure the generated sentence that contains the literal sense of the verb has the same symbol as the metaphorical sentence. For example, for the metaphorical sentence “The turbulent feelings that *surged* through his soul” our method will generate “The turbulent feelings that *continued* through his soul” maintaining the common symbolic meaning of (*love, loss, despair, sorrow, loneliness*) between the two (Section 2).

- A metaphor discriminator that guides the decoding of a sequence-to-sequence model fine-tuned on our parallel data to generate high quality metaphors. Our system MERMAID, *fine-tunes* BART (Lewis et al., 2019) – a state of the art pre-trained denoising autoencoder built with a sequence to sequence model, on our *automatically collected parallel corpus* of [*literal sentence, metaphorical sentence*] pairs (Sec. 3.1) to generate metaphors. A discriminative model trained in identifying metaphors is further used to complement our generator and guide the decoding process to improve the generated output (Sec. 3.2). Human evaluations show that this approach generates metaphors that are better than two literary experts 21% of the time on average, better 81% of the time than two well-crafted baselines, and better 36% of the time than fine-tuned BART (Lewis et al., 2019) (Section 5).

- A task-based evaluation to improve the quality of human written poems using metaphorical rewriting. Evaluation via Amazon Mechanical Turk shows that poems enhanced with metaphors generated by MERMAID are preferred by Turkers 68% of the times compared to poems without metaphors, which are preferred 32% of the times (Section 6).<sup>1</sup>

## 2 Dataset Creation with Symbolism

Datasets for metaphors are scarce. To our knowledge, there is no large scale parallel corpora containing literal and metaphoric paraphrases. The closest and most useful work is that of Mohammad et al. (2016). However the size of this data-set is small: 171 instances, which is not sufficient to train deep learning models. Recently, Stowe et al. (2020) rely on available metaphor detection datasets to generate metaphors by a metaphor-masking framework, where they replace metaphoric words in the input texts with metaphor masks (a unique “metaphor” token), hiding the lexical item. This creates artificial parallel training data: the input is the masked text, with the hidden metaphorical word, and the output is the original text (e.g., The war [MASK] many people → The war *uprooted* many people). The major issue with such masking strategy is that it ignores the semantic mapping between the literal verb and the metaphorical verb. Moreover, there are only 11,593 such parallel instances, still too small to train a neural model. The lack of semantic mapping between the artificial parallel training data samples, coupled with limited size thus affects the lexical diversity and meaning preservation of generated metaphors at test time. In light of these challenges, we propose to compose a large-scale parallel corpora with literal and metaphorical sentence pairs to learn the semantic mappings. We start with collecting a large-scale corpora of metaphorical sentences (Section 2.1) and leverage masked language model and symbolism-relevant common sense knowledge to create literal version for each metaphorical sentence (Section 2.2).

### 2.1 Metaphor Dataset Collection

Metaphors are frequently used in *Poetry* to explain and elucidate emotions, feelings, relationships and

<sup>1</sup>Our code, data and models are available at <https://github.com/tuhinjubcse/MetaphorGenNAACL2021>

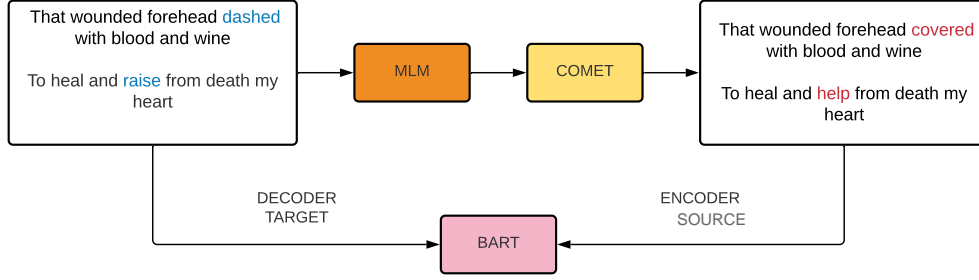


Figure 1: A schematic illustration of our system, which shows the **data creation** and **training** process where we use MLM along with COMET to transform an original metaphorical input to a literal output evoking similar symbolic meaning and use them to fine-tune BART.

other elements that could not be described in ordinary language. We use this intuition to identify a naturally occurring poetry corpus that contains metaphors called Gutenberg Poetry Corpus (Jacobs, 2018).<sup>2</sup> The corpus contains 3,085,117 lines of poetry extracted from hundreds of books. Not every sentence in the corpus contains a metaphorical verb. So as a first step, we identify and filter sentences containing a metaphorical verb.

We build a classifier by fine-tuning BERT (Devlin et al., 2018) on a metaphor detection corpus VU AMSTERDAM (Steen, 2010). Since our work is focused on verbs, we only do token classification and calculate loss for verbs. Figure 2 illustrates the BERT-based token-level classifier. The classification accuracy on test set is 74.7%, which is on par with most state of art methods.

Using the metaphor detection model, we identify 622,248 (20.2%) sentences predicted by our model as containing a metaphoric verb. Considering the classifier can introduce noise as the accuracy of the metaphor detection model is far from oracle 100%, we only retain sentences which are predicted by our model with a confidence score of 95% (i.e., prediction probability 0.95). This results in a total number of 518,865 (16.8%) metaphorical sentences.

## 2.2 Metaphoric to Literal Transformation with Symbolism

After identifying high quality metaphorical sentences, we want to obtain their literal counterparts to create a parallel training data. Masked language models like BERT (Devlin et al., 2018), or roBERTa (Liu et al., 2019) can be used for fill-in-the-blank tasks, where the model uses the context words surrounding a masked token to predict the masked word. We borrow this framework to mask

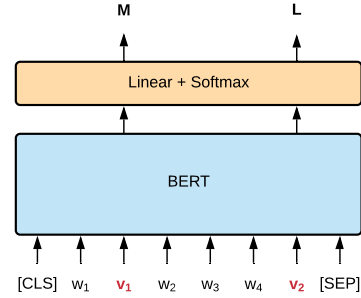


Figure 2: BERT-base-based model to identify metaphoric verbs, where  $v_1$  and  $v_2$  represent the verbs in a sentence. (M) denotes softmax probability of a verb being metaphorical, while (L) denotes it literal softmax probability.

the metaphorical verb (Table 2 Row1 vs Row2) from a sentence and use BERT-base-based model to obtain the top 200 candidate verbs to replace the metaphorical one to generate literal sentences (Table 2 Row3). There are two main issues in solely relying on MLM predicted verbs: 1) they are not necessarily literal in nature; 2) after replacing the default MLM predicted verb, the metaphorical sentence and the new sentence with the replaced verb might be semantically dissimilar.

### 2.2.1 Ensuring Literal Sense

Even though our inductive biases tell us that the chance of a predicted token having a literal sense is higher than having a metaphorical one, this cannot be assumed. To filter only literal candidate verbs we re-rank the MLM predicted mask tokens based on literal scores obtained from 2.1 since the model can predict the softmax probability of a verb in a sentence being either literal or metaphorical (Table 2 Row 4).

### 2.2.2 Ensuring Meaning Preservation

While we can potentially pair the sentence with the top most literal ranked verb with the input

<sup>2</sup><https://github.com/aparrish/gutenberg-poetry-corpus>

Input	The turbulent feelings that <b>surged</b> through his soul .
Masked	The turbulent feelings that [MASK] through his soul .
Ranked by MLM Prob	(‘tore’, 0.11), (‘ran’, 0.10), (‘ripped’, 0.09), (‘flowed’, 0.03), (‘rushed’, 0.01), ..... , (‘eased’, 0.01),..... , (‘continued’, 0.0005),...
Ranked by Meta Prob	(‘eased’, 0.12), (‘continued’, 0.0008), (‘spread’, 0.0004), (‘kicked’, <b>0.99</b> ), (‘punched’, <b>0.99</b> ),....., (‘screamed’, <b>0.99</b> ),.....

Table 2: Table showing a metaphorical sentence (Row1) where the metaphorical verb *surge* is masked (Row2). Row3 shows predicted tokens ranked by default LM probability. Row4 shows predicted tokens ranked by metaphoricity scores obtain from model described in 2.1. Lower scores means more literal.

Meta Input	The turbulent feelings that <b>surged</b> through his soul .	
Inp Symbol	love, loss, despair, sorrow, loneliness	
Lit Output1	The turbulent feelings that <i>eased</i> through his soul .	✗
Symbol	peace, love, happiness, joy, hope	
Lit Output2	The turbulent feelings that <i>continued</i> through his soul .	✓
Symbol	love, loss, despair, sorrow, loneliness	

Table 3: Table showing input metaphorical sentence and literal outputs along with the associated symbolic meaning obtained from COMET (Bosselut et al., 2019). Lit Output1 is an incorrect candidate since the symbolic meanings are divergent.

sentence containing the metaphorical verb, they might symbolically or semantically represent different abstract concepts. For example, in Table 3, after replacing the metaphorical verb “surge” with the top most literal verb “eased”, the sentence “*The turbulent feelings that eased through his soul*” evoke a different symbolic meaning of *peace, love, happiness, joy & hope* in comparison to the input containing the metaphorical verb, which evokes a symbolic meaning of *love, loss, despair, sorrow & loneliness*. To tackle this problem we ensure that the transformed literal output represents the same symbolic meaning as the metaphorical input.

To generate the common sense SYMBOL that is implied by the literal or metaphorical sentences, we feed the sentences as input to COMET (Bosselut et al., 2019) and restrict it to return top-5 beams. COMET is an adapted knowledge model pre-trained on ConceptNet.<sup>3</sup> Our work only leverages the **SymbolOf** relation from COMET.

<sup>3</sup>[https://mosaickg.apps.allenai.org/comet\\_conceptnet](https://mosaickg.apps.allenai.org/comet_conceptnet)

We now need a method to combine information from MLM and symbolic knowledge obtained from COMET described above. To do this, we filter candidates from MLM token predictions based on the symbolic meaning overlap between the metaphorical input and literal output first. To ensure that the quality is high, we put a strict requirement that all the 5 symbolic beams (typically words or short phrases) for the input metaphorical sentence should match all the 5 symbolic beams for the output literal sentence. Between multiple literal candidates all having beam overlap of 5, they are further ranked by reverse metaphoricity (i.e., literal) scores. The top most candidate is returned thereafter. We finally end up with 90,000 pairs for training and 3,498 pairs for validation.

### 3 Metaphor Generation

Our goal of generating metaphors can be broken down into two primary tasks: 1) generating the appropriate substitutions for the literal verb while being pertinent to the context; 2) ensuring that the generated utterances are actually metaphorical.

#### 3.1 Transfer Learning from BART

To achieve the first goal, we fine-tune BART (Lewis et al., 2019), a pre-trained conditional language model that combines bidirectional and autoregressive transformers, on the collected parallel corpora. Specifically, we fine-tune BART by treating the literal input as encoder source and the metaphorical output as the the decoder target (Figure 1). One issue of the pre-trained language models is that they have a tendency to generate literal tokens over metaphorical ones. To overcome this, we introduce a rescoring model during the decoding process to favor more metaphorical verbs. The rescoring model is inspired by Holtzman et al. (2018); Goldfarb-Tarrant et al. (2020) and detailed in the next section.

#### 3.2 Discriminative Decoding

We have a base metaphor generation model  $p(\mathbf{z}|\mathbf{x})$  which is learned by fine-tuning BART (Lewis et al., 2019) on pairs of literal ( $x$ ) and metaphorical ( $z$ ) sentences. We propose to modify the decoding objective to incorporate a Metaphor detection rescoring model  $a$  and re-rank the base, or “naive” BART generated hypotheses, bringing the metaphoric representation closer to the rescoring model’s specialty



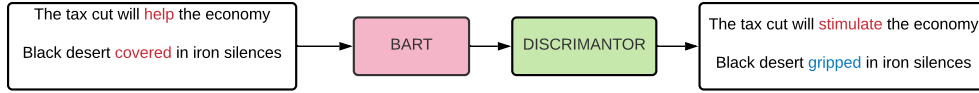


Figure 3: Schematic showing the **decoding** step where we use fine-tuned BART along with a metaphor detecting discriminator to generate a metaphorical sentence conditioned on a literal input

and desirable attribute. The modified decoding objective becomes:

$$f_{\lambda}(\mathbf{x}, \mathbf{z}) = \sum_i^m -\log p(z|z < i, \mathbf{x}) + \lambda a(\mathbf{x}, z_{i...m}) \quad (1)$$

where  $\lambda$  is a weight of the score given by  $a$ .

**Implementation Details** We use top-k sampling strategy (Fan et al., 2018) ( $k=5$ ) to generate metaphors conditioned on a literal input. Our rescoring model  $a$  is a RoBERTa model fine-tuned on a combined dataset of (Steen, 2010; Beigman Klebanov et al., 2018) to classify sentences as literal or metaphorical based on whether there exists a metaphorical verb. It is a sentence level task where the model predicts a sentence as literal or metaphorical. We down-sample the data to maintain a ratio of (1 : 1) between two classes and use 90% of the data to train and 10% for validation. We achieve a considerably decent validation accuracy of 83%. We manually tune  $\lambda$  using grid search on a small subset of 3,498 validation samples from our parallel automatic data and choose the best value.

Figure 3 shows the process of re-ranking BART hypothesis using the discriminator described above to generate novel metaphorical replacements for literal verbs. All the hyper-parameters for data creation, fine-tuning and discriminative decoding are exactly the same as mentioned in **Appendix A**.

The reason to use a separate discriminator for decoding instead of using the same BERT based classifier used for parallel data creation, was to avoid introducing dataset biases or spurious correlations. The BERT-based classifier used for automatically creating the parallel dataset ideally has already picked up salient metaphorical phenomena in the VUA dataset. To further guide the decoding process, we hypothesize that a model trained on datasets not seen during training would lead to better generalization. We experimented with using the BERT model trained on VUA for rescoring, but the results were not better.

## 4 Experimental Setup

To compare the quality of the generated metaphors, we benchmark our MERMAID model against human

performance (i.e., the two creative writing experts HUMAN1 (a novelist) & HUMAN2 (a poet) who are not the authors of the paper) (Section 4.2) and three baseline systems described below.

### 4.1 Baseline Systems

**Lexical Replacement (LEXREP):** We use the same idea as our data creation process (Section 2.2). We use our model described in Section 2.1 to re-rank the predicted tokens from a mask language model based on metaphoricity scores. We filter the top 25 ranked metaphorical candidates and further rerank them based on symbolic meaning overlap with the literal meaning using COMET (Bosselut et al., 2019) and replace the literal verb with the top scoring candidate.

**Metaphor Masking (META\_M):** We use the metaphor masking model proposed by Stowe et al. (2020) where the language model learns to replace a masked verb with a metaphor. They train a seq2seq model with the encoder input of the format (The tax cut [MASK] the economy) and the decoder output being the actual metaphorical sentence (The tax cut lifted the economy). During inference, they mask the literal verb and expect the language model to infill a metaphorical verb.

**BART:** We use generations from a BART model fine-tuned on our automatically created data without the discriminative decoding. This helps us gauge the effect of transfer learning from a large generative pre-trained model, which also accounts for context unlike the retrieval based methods.

### 4.2 Test Data

To measure the effectiveness of our approach, we need to evaluate our model on a dataset that is independent of our automatically created parallel data and that is diverse across various domains, genres and types. Hence we rely on test data from multiple sources. As our first source, we randomly sample literal and metaphorical sentences with high confidence ( $> 0.7$ ) and unique verbs from the existing dataset introduced by Mohammad et al. (2016). For the metaphorical sentences from Mohammad et al. (2016) we convert them to their literal equivalent

the same way as discussed in Section 2.2 without the use of COMET as we do not need it. To ensure diversity in genre, as our second source we scrape WRITINGPROMPT and OCPOMETRY subreddits for sentences with length up to 12 words, which are *literal* in nature based on prediction from our model described in Section 2.1. We collate 500 such sentences combined from all sources and randomly sample 150 literal utterance for evaluation.

We use two literary experts (not authors of this paper) — a student in computer science who is also a poet, and a student in comparative literature who is the author of a novel — to write corresponding metaphors for each of these 150 inputs for evaluation and comparison.

### 4.3 Evaluation Criteria

**Automatic evaluation.** One important aspect in evaluating the quality of the generated metaphors is whether they are faithful to the input: while we change literal sentences to metaphorical ones, it should still maintain the same denotation as the input. To this end, we calculate the *Semantic Similarity* between the metaphorical output and the input using sentence-BERT (SBERT) (Reimers and Gurevych, 2019). We also calculate corpus-level BLEU-2 (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) with human written references.

**Human evaluation.** Since automatic evaluation is known to have significant limitations for creative generation (Novikova et al., 2017), we further conduct human evaluation on a total of 900 utterances, 600 generated from 4 systems and 300 generated by the two human experts. We propose a set of four criteria to evaluate the generated output: (1) *Fluency (Flu)* (“How fluent, grammatical, well formed and easy to understand are the generated utterances?”), (2) *Meaning (Mea)* (“Are the input and the output referring or meaning the same thing?”) (3) *Creativity (Crea)* (“How creative are the generated utterances?”), and (4) *Metaphoricity (Meta)* (“How metaphoric are the generated utterances?”). The human evaluation is done on the Amazon Mechanical Turk platform. Each Turker was given a literal input and 6 metaphorical outputs (4 from system outputs – 3 baselines and our proposed system MERMAID, and 2 from humans) at a time, with the metaphorical outputs randomly shuffled to avoid potential biases. Turkers were instructed to evaluate the quality of the metaphorical sentences with respect to the input and not in isolation. As we

System	Similarity ↑	BLEU-2↑	BertScore↑
LEXREP	79.6	<b>68.7</b>	0.56
META_M	73.2	61.0	0.62
BART	83.6	65.0	<u>0.65</u>
MERMAID	<u>85.0</u>	<u>66.7</u>	<b>0.71</b>
HUMAN1	<b>86.6</b>	-	-
HUMAN2	84.2	-	-

Table 4: Automatic evaluation results on test set where MERMAID significantly outperforms other automatic methods for 2 out of 3 metrics ( $p < .001$ ) according to approximate randomization test). BLEU-2 and BertScore is calculated w.r.t to Human references (HUMAN1 & HUMAN2). Corpus level BLEU-2 and Semantic Similarity are in range of (0-100) while BertScore is in range (0-1)

System	Flu	Mea	Crea	Meta
HUMAN1	<b>3.83</b>	<b>3.77</b>	<b>4.02</b>	<b>3.52</b>
HUMAN2	3.29	3.43	3.58	3.16
LEXREP	2.21	2.59	2.16	1.98
META_M	2.10	1.91	2.00	1.89
BART	3.33	3.08	3.16	2.85
MERMAID	<u>3.46</u>	<u>3.35</u>	<u>3.50</u>	<u>3.07</u>

Table 5: Human evaluation on four criteria of metaphors quality for systems and humans generated metaphors. We show average scores on a likert scale of 1-5 where 1 denotes the worst and 5 be the best. Bold-face denotes the best results overall and underscore denotes the best among computational models.

evaluate on four dimensions for 900 utterances, we have a total of 3600 evaluations. Each criteria was rated on a likert scale from 1 (not at all) to 5 (very). Each group of utterances was rated by three separate Turkers, resulted in 42, 48, 44 and 53 Turkers for the four evaluation tasks respectively. We pay them at a rate of \$15 per hour.

## 5 Results

Based on the semantic similarity metric shown in column 1 of Table 4, our system MERMAID is better in preserving the meaning of the input than the other baselines. As mentioned, we calculate BLEU-2 and BERTScore between system outputs and human references. MERMAID is better than the other baselines according to BERTScore. In terms of BLEU-2, MERMAID is second best.

Table 5 shows the average scores for the human evaluation on four metaphor quality criteria for MERMAID, the baselines, and human written metaphors on the test set. The inter-annotator agreements computed using Krippendorff’s alpha for Creativity, Meaning, Fluency and Metaphoricity are 0.44, 0.42, 0.68, 0.52 respectively. The results demonstrate that MERMAID is significantly better

Literal	System	Metaphor	Flu	Mea	Crea	Meta
The scream <i>filled</i> the night	HUMAN1	The scream <i>pierced</i> the night	<b>4.3</b>	<b>5.0</b>	<b>3.7</b>	<b>4.0</b>
	HUMAN2	The scream <i>covered</i> the night	2.7	4.0	3.0	3.0
	LEXREP	The scream <i>held</i> the night	1.7	3.7	2.0	1.7
	META_M	The scream <i>opened</i> the night	1.0	1.0	1.0	1.0
	BART	The scream <i>filled</i> the night	2.3	1.0	2.3	1.0
	MERMAID	The scream <i>pierced</i> the night	<b>4.3</b>	<b>5.0</b>	<b>3.7</b>	<b>4.0</b>
The wildfire <i>spread</i> through the forest at an amazing speed	HUMAN1	The wildfire <i>ravaged</i> through the forest at an amazing speed	<b>4.7</b>	<b>4.3</b>	4.0	3.0
	HUMAN2	The wildfire <i>leapt</i> through the forest at an amazing speed	3.7	3.0	<b>5.0</b>	3.7
	LEXREP	The wildfire <i>saw</i> through the forest at an amazing speed	1.3	1.0	2.7	3.3
	META_M	The wildfire <i>grows</i> through the forest at an amazing speed	3.7	2.7	2.7	<b>4.0</b>
	BART	The wildfire <i>swept</i> through the forest at an amazing speed	4.0	3.7	4.7	<b>4.0</b>
	MERMAID	The wildfire <i>danced</i> through the forest at an amazing speed	3.0	4.0	4.0	3.7
My heart <i>beats</i> when he walks in the room	HUMAN1	My heart <i>skips</i> when he walks in the room	4.7	<b>5.0</b>	4.0	<b>4.3</b>
	HUMAN2	My heart <i>sings</i> when he walks in the room	<b>5.0</b>	4.3	3.7	3.3
	LEXREP	My heart <i>made</i> when he walks in the room	1.0	1.0	1.0	1.0
	META_M	My heart <i>came</i> when he walks in the room	1.7	1.0	1.3	1.3
	BART	My heart <i>sings</i> when he walks in the room	<b>5.0</b>	4.3	3.7	3.7
	MERMAID	My heart <i>jumps</i> when he walks in the room	4.7	4.7	<b>4.3</b>	4.0
After a glass of wine, he <i>relaxed</i> up a bit	HUMAN1	After a glass of wine, he <i>loosened</i> up a bit	<b>4.7</b>	<b>5.0</b>	<b>5.0</b>	<b>4.0</b>
	HUMAN2	After a glass of wine, he <i>unfurled</i> up a bit	2.0	<b>5.0</b>	2.0	3.7
	LEXREP	After a glass of wine, he <i>followed</i> up a bit	3.7	1.0	2.7	1.7
	META_M	After a glass of wine, he <i>touched</i> up a bit	1.3	1.0	1.7	2.0
	BART	After a glass of wine, he <i>dried</i> up a bit	2.7	1.0	2.3	2.0
	MERMAID	After a glass of wine, he <i>loosened</i> up a bit	4.3	<b>5.0</b>	<b>5.0</b>	3.7
The tax cut will <i>help</i> the economy	HUMAN1	The tax cut will <i>uplift</i> the economy	4.7	<b>5.0</b>	<b>4.7</b>	<b>4.0</b>
	HUMAN2	The tax cut will <i>fertilize</i> the economy	4.0	4.3	4.3	3.7
	LEXREP	The tax cut will <i>bring</i> the economy	1.7	3.0	2.7	1.7
	META_M	The tax cut will <i>prevent</i> the economy	1.7	1.0	2.0	1.0
	BART	The tax cut will <i>strengthen</i> the economy	<b>5.0</b>	<b>5.0</b>	4.3	3.7
	MERMAID	The tax cut will <i>stimulate</i> the economy	<b>5.0</b>	4.7	3.7	<b>4.0</b>
I tried to <i>resolve</i> things over between them	HUMAN1	I tried to <i>tide</i> things over between them	4.3	3.0	3.7	<b>4.3</b>
	HUMAN2	I tried to <i>patch</i> things over between them	<b>4.7</b>	<b>4.7</b>	<b>5.0</b>	2.0
	LEXREP	I tried to <i>push</i> things over between them	3.3	1.0	2.3	2.0
	META_M	I tried to <i>make</i> things over between them	4.0	1.0	2.7	2.7
	BART	I tried to <i>put</i> things over between them	<b>4.7</b>	2.0	3.0	2.7
	MERMAID	I tried to <i>smooth</i> things over between them	<b>4.7</b>	<b>4.7</b>	<b>5.0</b>	4.0

Table 6: Examples of generated outputs from different systems (with human written metaphors as references). We show average scores (over three annotators) on a 1-5 scale with 1 denotes the worst and 5 be the best. The italics texts in the literal column represent the *verb* while those in Metaphor column represents the generated metaphorical *verb*. Boldface indicates the best results.

than the baselines on all four criteria ( $p < .001$  according to approximate randomization test).

Table 6 presents several generation outputs from different systems along with human judgements on individual criteria. We observe that incorporating a discriminator often guides our model to generate better metaphors than the already strong baseline using BART. Finally, incorporating symbolic meaning in data creation step helps our model to maintain the same meaning as the input.

## 6 Task Based Evaluation

Metaphors are frequently used by creative writing practitioners, in particular poets, to embellish their work. We posit that MERMAID can be used to edit literal sentences in poems to further enhance creativity. To test this hypothesis, we first crawl origi-

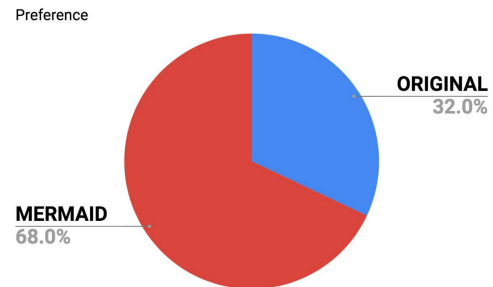


Figure 4: Percentage of Preference of Original Quatrains vs Quatrains rewritten by MERMAID

nal poems submitted by authors from the sub-reddit OCPoetry. The poems are of variable lengths, so to ensure parity we break them into *Quatrains* (four sentence stanza). We randomly sample 50 such Quatrains containing at least one sentence with a

And the hills have a shimmer of light between, And the valleys are <i>covered</i> with misty veils, And ....., ....
And the hills have a shimmer of light between, And the valleys are <i>wrapped</i> with misty veils, And ....., ....
Leaves on a maple, <i>burst</i> red with the shorter days; Falling to the ground. ....
Leaves on a maple, <i>burgeoned</i> red with the shorter days; Falling to the ground. ....

Table 7: Example Quatrains from reddit where MERMAID rewrites a sentence containing a literal verb to make it metaphorical.

literal verb in it. We use our metaphor detection model (Section 2.1) to detect literal verbs.

We then select a sentence containing a literal verb from each Quatrain and use MERMAID to re-write it so that the resulting output is metaphorical. We ignore common verbs like *is, was, are, were, have, had*. If there are more than one sentence in Quatrain with literal verbs, we choose the sentence with a literal verb that has the highest probability for being literal. For sentences with multiple literal verbs, we choose the verb with highest literal probability.

Our goal is to see if re-written poems are qualitatively better than the original forms. To do this, we hire Turkers from Amazon Mechanical Turk and present them with hits where the task is to choose the better version between the original Quatrain and the re-written version. 15 Turkers were recruited for the task. Each Quatrain was evaluated by 3 distinct Turkers. Table 7 shows metaphorical transformations by a MERMAID Figure 4 shows that poems rewritten by MERMAID were considered better by the Turkers.

## 7 Related Work

Most researchers focused on identification and interpretation of metaphor, while metaphor generation is relatively under-studied.

### 7.1 Metaphor Detection

For metaphor detection, researchers focused on variety of features, including unigrams, imageability, sensory features, WordNet, bag-of-words features (Klebanov et al., 2014; Tsvetkov et al., 2014; Shutova et al., 2016; Tekiroğlu et al., 2015; Hovy et al., 2013; Köper and im Walde, 2016).

With advent of deep learning approaches, Gao et al. (2018) used BiLSTM models based on GloVe (Pennington et al., 2014) and ELMo word vectors (Peters et al., 2018) to detect metaphoric verbs. Inspired by the linguistic theories, MIP (Semino et al., 2007; Steen, 2010) and SPV (Wilks, 1975, 1978), Mao et al. (2019) proposed two detection models consisting of BiLSTM with attention mechanisms that relied on GloVe and ELMo embeddings. Recent work on metaphor detection have also used pretrained language models (Su et al., 2020; Gong et al., 2020). While we focus on metaphor generation, we use (Devlin et al., 2018) to detect metaphoric verbs to create parallel data and (Liu et al., 2019) to rescore our generated hypothesis during decoding.

### 7.2 Metaphor Generation

Some early works made contributions to use template and heuristic-based methods (Abe et al., 2006; Terai and Nakagawa, 2010) to generate “A is like B” sentences, more popularly referred to as *similes*. Chakrabarty et al. (2020) concentrated on simile generation, applying seq2seq model to paraphrase a literal sentence into a simile. Other attempts learned from the mappings of different domains and generated conceptual metaphors of pattern “A is B” (Hervás et al., 2007; Mason, 2004; Gero and Chilton, 2019). These works paid attention to the relationship between nouns and concepts to create elementary figurative expressions.

Recent metaphor generation works focus mainly on verbs. Yu and Wan (2019) proposed an unsupervised metaphor extraction method, and developed a neural generation model to generate metaphorical sentences from literal-metaphorical verb pairs. They however do not focus on literal to metaphorical sentence transfer, but generate a sentence given a metaphorical fit word. The closest to our work is that of Stowe et al. (2020), who focus on building a seq2seq model, using a special mask token to mask the metaphorical verbs as input, and the original metaphorical sentences as output. However, this model face challenges in transferring the literal sentences to metaphorical ones, while maintaining the same meaning. We, on the contrary, focus on maintaining the same meaning through parallel data creation focusing on symbolism. Additionally, we incorporate a metaphor detection model as a discriminator to improve decoding during generation.



## 8 Conclusion

We show how to transform literal sentences to metaphorical ones. We propose a novel way of creating parallel corpora and an approach for generating metaphors that benefits from transfer learning and discriminative decoding. Human and automatic evaluations show that our best model is successful at generating metaphors. We further show that leveraging symbolic meanings helps us learn better abstract representations and better preservation of the denotative meaning of the input. Future directions include learning diverse conceptual metaphoric mapping using our parallel data and constraining our metaphoric generations based on particular mapping.

## 9 Ethics

Our data is collected from Reddit and we understand and respect user privacy. Our models are fine-tuned on sentence level data obtained from user posts. These do not contain any explicit detail which leaks information about a users name, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, sexual orientation, trade union membership, alleged or actual commission of crime.

Second, although we use language models trained on data collected from the Web, which have been shown to have issues with bias and abusive language (Sheng et al., 2019; Wallace et al., 2019), the inductive bias of our models should limit inadvertent negative impacts. Unlike model variants such as GPT, BART is a conditional language model, which provides more control of the generated output. Furthermore, we specifically encode writing style from a poetic corpus in our models and train on parallel data in the direction of literal to metaphorical style. Open-sourcing this technology will help to generate metaphoric text assisting creative writing practitioners or non native language speakers to improve their writing. We do not envision any dual-use that can cause harm for the use of our the metaphor generation system.

## Acknowledgments

This work was supported in part by the MCS program under Cooperative Agreement N66001-19-2-4032, and the CwC program under Contract W911NF-15-1-0543 with the US Defense Advanced Research Projects Agency (DARPA). The

views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors would like to thank the members of PLUS-Lab at the University of California Los Angeles and University of Southern California and the anonymous reviewers for helpful comments.

## 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.
- Lawrence W Barsalou et al. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.
- Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. [A corpus of non-native written English annotated for metaphor](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: Commonsense transformers for automatic knowledge graph construction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. [Generating similes effortlessly like a pro: A style transfer approach for simile generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.
- Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870*.
- Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153.
- Raquel Hervás, Rui P Costa, Hugo Costa, Pablo Gervás, and Francisco C Pereira. 2007. Enrichment of automatically generated texts using metaphor. In *Mexican International Conference on Artificial Intelligence*, pages 944–954. Springer.
- Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*.
- Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.
- Arthur M Jacobs. 2018. The gutenber english poetry corpus: exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:5.
- Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17.
- Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of german particle verbs. In *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies*, pages 353–362.
- 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*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898.
- James H Martin. 2006. A corpus-based analysis of context effects on metaphor comprehension.
- Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.
- 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, Berlin, Germany. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, Amanda Ceras Curry, and Verena Rieser. 2017. *Why we need new evaluation metrics for NLG*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.
- 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.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Elena Semino et al. 2007. Mip: a method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language 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 3398–3403.
- Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of*

- the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.
- Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.
- Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic analysis improves neural metaphor detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 362–371.
- Kevin Stowe and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26.
- Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych. 2020. Metaphoric paraphrase generation. *arXiv preprint arXiv:2002.12854*.
- Chuangdong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiquan Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39.
- Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39.
- 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.
- Yulia Tsvetkov, Leonid Boytsov, Anatole Gershan, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.
- Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.
- Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.
- Yorick Wilks. 1978. Making preferences more active. *Artificial intelligence*, 11(3):197–223.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

## A Appendix

For retrieving commonsense symbolism of the sentences, we use the pre-trained COMET model <sup>4</sup> and retrieve top 5 candidates for each input.

1. **No of Parameters:** For metaphor detection at token level we use BERT-base-cased model (110M). For generation we use the BART large checkpoint (400M parameters) and use the implementation by FAIRSEQ (Ott et al., 2019) <sup>5</sup>. For discriminative decoding we use RoBERTa large model (355M)
2. **No of Epochs:** For metaphor detection at token level for parallel data creation we fine-tune it for 3 epochs. We fine-tune pre-trained BART for 70 epochs for MERMAID model and save best model based on validation perplexity. For discriminator we fine-tune RoBERTa-large model for 10 epoch and save the checkpoint for best validation accuracy
3. **Training Time:** For metaphor detection training time is 40 minutes. Our training time is 280 minutes for BART. For discriminator we train it for 60 minutes
4. **Hardware Configuration:** We use 4 RTX 2080 GPU
5. **Training Hyper parameters:** We use the same parameters mentioned in the github repo

<sup>4</sup><https://github.com/atcbosselut/comet-commonsense>

<sup>5</sup><https://github.com/pytorch/fairseq/tree/master/examples/bart>

where BART was fine-tuned for CNN-DM summarization task with the exception of MAX-TOKENS (size of each mini-batch, in terms of the number of tokens.) being 1024 for us. For discriminator finetuning of roberta we use same parameters as RTE task <sup>6</sup>

6. **Decoding Strategy & Hyper Parameters:**For decoding we generate metaphors from our models using a top-k random sampling scheme (Fan et al., 2018). At each timestep, the model generates the probability of each word in the vocabulary being the likely next word. We randomly sample from the k = 5 most likely candidates from this distribution.

---

<sup>6</sup><https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md>