Multi-Figurative Language Generation

Lai, Huiyuan; Nissim, Malvina

Multi-Figurative Language Generation

Huiyuan Lai and **Malvina Nissim**
Center for Language and Cognition (CLCG)
University of Groningen / The Netherlands
{h.lai, m.nissim}@rug.nl

Abstract

Figurative language generation is the task of reformulating a given text in the desired figure of speech while still being faithful to the original context. We take the first step towards multi-figurative language modelling by providing a benchmark for the automatic generation of five common figurative forms in English. We train **mFLAG** employing a scheme for multi-figurative language pre-training on top of BART, and a mechanism for injecting the target figurative information into the encoder; this enables the generation of text with the target figurative form from another figurative form without parallel figurative-figurative sentence pairs. Our approach outperforms all strong baselines. We also offer some qualitative analysis and reflections on the relationship between the different figures of speech.

1 Introduction

Figurative language is commonly used in speaking and writing to accomplish a constellation of communicative goals (Roberts and Kreuz, 1994). Figures of speech, such as metaphors, or idiomatic expressions, can make an expression stand out by making it more interesting and captivating, and can evoke stronger emotions than more factual, literal phrases thereby making the text more engaging.

Automatic figurative language generation has received growing attention with the progress of neural networks, especially the emergence of large pre-trained models (Raffel et al., 2020; Lewis et al., 2020). We see there are two core values for this task: (i) computational approaches can be employed to provide a better understanding of linguistic phenomena and more specifically in this case different figures of speech; (ii) we can explore how much models can handle creativity and devise ways to employ them in the support of creative writing, so as to yield more varied and human-like generated text, including in the context of machine translation (Guerberof-Arenas and Toral, 2022).

Forms	Sentences
Literal	Old Mr. Smith has been teaching here for a very long time.
Hyperbole	Old Mr. Smith has been teaching here since the Stone Age.
Literal	My niece will babysit for you for a little bit of money.
Idiom	My niece will babysit for you for pin money.
Literal	I hate it when they run the same commercial twice in a row.
Sarcasm	I love when they run the same commercial twice in a row.
Literal	He remembers a road of my broken works.
Metaphor	He made a road of my broken works.
Literal	You can publish the whole thing old.
Simile	You can publish the whole thing like a diary.

Table 1: Examples of figurative language generation from literal texts.

There are many related tasks that have been proposed and studied by NLP researchers, including the generation of hyperbole (Tian et al., 2021; Zhang and Wan, 2022), idiom (Zhou et al., 2021), sarcasm (Zhu et al., 2019; Chakrabarty et al., 2020a), metaphor (Abe et al., 2006; Stowe et al., 2021b), and simile (Chakrabarty et al., 2020b; Zhang et al., 2021). Table 1 shows examples of figurative language generation from literal texts.

Previous works focus on modelling single figurative forms, generally rewriting a literal sentence into one with a specific figure of speech. This results in having to train separate models, one for each figure of speech, and in not exploiting knowledge transfer across figurative forms. However, since different figures of speech can share some features related to non-literality, and a text may also contain and combine multiple figures of speech at the same time, it is possible that substantial knowledge gains can be transferred from one figure to another. Moreover, the generation between different figures of speech (e.g. generating an idiomatic text from the hyperbolic one) is under-explored.

In this work we suggest to model multiple figures of speech jointly, with the ultimate goal of having a single model that can handle the generation of multiple figurative forms from both literal and figurative inputs.

Intuitively, multi-task learning (Collobert and Weston, 2008) and the usage of a domain la-

bel (Kobus et al., 2017) could be a good method for multi-figurative language modelling, adding a special token to the beginning of the sentence to guide text generation. Such a method requires parallel data (i.e. aligned texts with the same context but different figures of speech) for training; this is usually unavailable, especially between different figures of speech, and costly to produce.

We rely on existing parallel data between literal sentences and single figures of speech and propose mFLAG (**M**ulti-**F**igurative **La**nguage **G**eneration), an approach which is applicable to the generation between different forms, both literal and figurative. In a nutshell, mFLAG is trained in two stages, in both of which we also exploit the contribution of generic paraphrase data: (i) a specifically designed pre-training for multi-figurative language, where a special label is added at the beginning of each sentence to indicate its figure of speech; (ii) a supervised training where the parallel literal-figurative sentence pairs for all figurative languages are combined to achieve multi-figurative language generation. For (ii), we introduce an innovative mechanism that allows the form labels to leak their own figurative information into the input embedding, thus guiding the encoder to represent the source sentence. This mechanism makes it possible to generate between different figures of speech without parallel figurative-figurative data. For comparison, and to allow for wider flexibility in generation choices as well as linguistic analysis, we also use the literal form corresponding to each figure of speech, which is available through the separate parallel datasets, as pivot to run figurative-to-figurative transformation. We expect that with the direct figurative-figurative transformation the source figurative form might still be maintained in the generated sentence, with the addition of the target figurative form, while this should not be the case when using the literal form as pivot.

Contributions Considering five common figures of speech in English, (i) we propose a novel task of multi-figurative language generation, and explore the potential of its computational modelling; (ii) we introduce a pre-training scheme for multi-figurative language modelling, which boosts performance substantially by leveraging paraphrase data and cross-figurative language knowledge transfer; (iii) we design a mechanism for injecting the desired figurative information into the encoder to achieve the generation between different figures of speech with-

out parallel figurative-figurative sentence pairs; this mechanisms could be applied to other tasks, too; (iv) we compare figurative-figurative and figurative-literal-figurative generation, thereby assessing the feasibility, the limits, and the characteristics of direct multi-figurative language generation; and (v) we provide a benchmark for multi-figurative language generation, which can hopefully foster the progress of figurative language processing. [1]

2   Background

Transforming text involving a figure of speech, either in source or in target or both, is closely related to three other NLP tasks, namely paraphrasing, text style transfer, and figurative language detection. We discuss relevant background on such tasks, and why and how they play a role in our work.

Paraphrasing Paraphrasing is the task of generating a text semantically (almost) identical to a given input, but with variations in wording or syntax (Prakash et al., 2016; Cao et al., 2017). The large amount of parallel paraphrase data available can be used to teach models a general rewriting task in the context of various downstream NLP tasks, such as semantic parsing (Berant and Liang, 2014), machine translation (Callison-Burch et al., 2006), question answering (Dong et al., 2017), and text style transfer (Lai et al., 2021). As figurative generation can be viewed as a special paraphrasing task, where texts are expected to include specific figurative forms, we also leverage paraphrase data for figurative generation modelling.

Text Style Transfer The goal of text style transfer is to transform a given text of one style into another while preserving the style-independent content. A common task, for example, is formality transfer, where an informal sentence is turned into formal, or viceversa (Rao and Tetreault, 2018). Generally speaking, both text style transfer and figurative language generation aim to achieve the generation of text with specific attributes. Regarding sentence changes, for text style transfer, often multiple parts of the sentence might be modified at the same time, such as capitalization at the beginning of the sentence, punctuation at the end, and some phrasing in the middle. Figurative language generation, instead, often concerns the rewriting of some specific expressions, while other (possi-

---

[1]Data, code, and model are available at `https://github.com/laihuiyuan/mflag`.

bly large) portions of the input sentence could be retained (Zhou et al., 2021). Also, in figurative language generation, the original figurative form could be still present in the transformed sentence, while text style transfer aims to alter the original style fully.

It should also be pointed out that addressing multi-figurative language generation is particularly challenging since not all figures of speech considered require the same kinds of alterations in text.

Figurative Language Detection Most past work on figurative language processing focuses on detection rather than generation. The detection of figurative language generally involves two levels: sentence-level and word-level. At sentence-level, the task is usually formulated as a binary classification problem, namely automatically detecting whether a given sentence is literal or non-literal (Troiano et al., 2018). At word-level, the task is concerned with identifying the exact words within a sentence which trigger the figurative reading (Beigman Klebanov et al., 2016; Mao et al., 2018). This task is a crucial component in retrieval-based approaches to figurative language generation, which usually require first the identification of triggering words in a sentence, followed then by other operations such as replacement and generation (see next paragraph.)

Figurative Language Generation Early work on figurative language generation is mainly template-based. Abe et al. (2006) employ simple expressions "A is like B" for metaphor generation. Veale (2016) use template-like structures to generate metaphoric tweets. These methods usually lack the flexibility to cope with the variability intrinsic to (creative) natural language. In recent years, figurative language modelling has mostly shifted to neural-based end-to-end approaches, showing good degrees of creativity, for example in the generation of puns and metaphors (Yu et al., 2018; Yu and Wan, 2019). To provide better explainability, Zhou et al. (2021) propose a neural-based pipeline for idiom generation that contains three explicit steps: retrieve, extract, and generate. Most recently, and as in most NLP tasks, impressive results for figurative language generation have been achieved leveraging pre-trained models. For example, Stowe et al. (2021a) and Chakrabarty et al. (2021) successfully generate metaphors fine-tuning T5 (Raffel et al., 2020) and BART (Lewis et al., 2020),

Forms	Task	Train	Valid	Test
Hyperbole	Literal Form↔Hyperbole	509(+668)	50	150
Idiom	Literal Form↔Idiom	3,784	876	876
Sarcasm	Literal Form↔Sarcasm	16,762	1,500	1,470
Metaphor	Literal Form↔Metaphor	118,807	6,254	150
Simile	Literal Form↔Simile	82,687	5,145	150

Table 2: Dataset statistics.

respectively. Fine-tuning BART is successful for the generation of simile (Chakrabarty et al., 2020b), and hyperbole (Zhang and Wan, 2022), too. Stowe et al. (2021b) also propose to control the metaphor generation process by encoding conceptual mappings in the form of FrameNet frames. All these works focus on single figurative forms, modelling generation between literal and figurative. Instead, while still leveraging parallel literal-figurative data for single forms, we aim to model multiple figures of speech jointly thereby also generative between different figurative forms.

3 Task and Dataset

We define the task of figurative language generation as the transformation of a text written in (or with) a given form (literal or figurative) to a text in (or containing) another form, while preserving the original general context.

We use five existing datasets for the figures of speech we consider in this paper; Table 2 shows sizes and splits.

- **Hyperbole** Troiano et al. (2018) introduce HYPO, a corpus of 709 hyperbolic sentences with their non-hyperbolic formulations. We boost this small dataset with some automatically obtained pairs. We fine-tune BART with HYPO, and use this model to transform into literal the hyperbolic texts contained in the non-parallel dataset HYPO-Red (Tian et al., 2021). We then select literal generations with a low hyperbolic score σ as predicted by a binary classifier based on BERT (Devlin et al., 2019) trained on HYPO, for an additional 668 training pairs.[2]

- **Idiom** Zhou et al. (2021) use the existing MAGPIE corpus (Haagsma et al., 2020) to create a parallel dataset of literal and idiomatic pairs.

- **Sarcasm** Peled and Reichart (2017) release a dataset of 3,000 pairs of sarcastic tweets each augmented with five interpretations. We complement this by adding to the training set 4,762

---

[2]Generated literal texts with $\sigma < 0.5$ are selected.
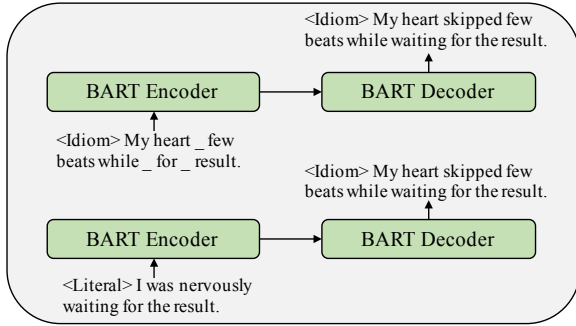
(a) Multi-figurative language denoising pre-training and fine-tuning.
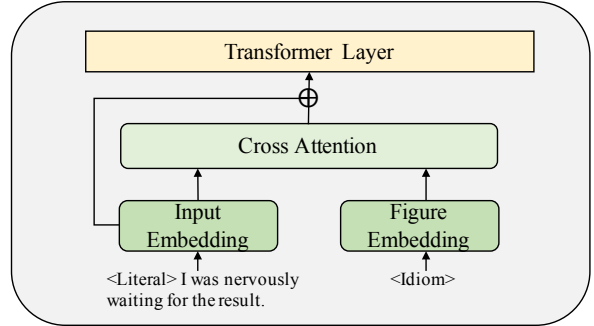
(b) An overview of the mechanism for injecting the figurative information into the Encoder.

Figure 1: Overview of multi-figurative language modelling. In 1(a), there is the framework for our multi-figurative language denoising pre-training (top) where word masking as the injected noise, and fine-tuning on downstream task of figurative language generation (down); in 1(b), the figurative information is injected into the encoder using cross-attention and residual learning.

sentence pairs from a sarcasm dataset (Ghosh et al., 2020).

- **Metaphor** Stowe et al. (2021b) build a literal-metaphor dataset exploiting the Gutenberg Poetry corpus (Jacobs, 2018): metaphoric verbs are identified, masked, and eventually replaced with infilling from a language model.

- **Simile** Chakrabarty et al. (2020b) automatically collect a set of self-labelled similes via distant supervision, using the phrase *like a*; similes are converted into their literal versions leveraging the structured common sense knowledge obtained from COMET (Bosselut et al., 2019).

Pre-Training Data Given that figurative generation is a special paraphrasing task, we use the available paraphrase data from PARABANK 2 (Hu et al., 2019) for multi-figurative language modelling, but only selecting more relevant pairs for the pre-training phase. To do so, we fine-tune BERT with the above figurative data to obtain five binary classifiers (each one literal vs figurative). With them, we do figurative language detection on paraphrase data, and only retain pairs where the probability that the source and target sentences are in literal form and figurative form, respectively, is greater than a threshold σ.[3]

4 Multi-figurative Language Modelling

We propose an approach to model multi-figurative language on top of the large pre-trained sequence-to-sequence model BART (Lewis et al., 2020),

by performing further, figurative language-specific pre-training, and then fine-tuning.

BART is a seq2seq model trained as a denoising autoencoder, and to reconstruct the original text T given $g(T)$ where g is a noising function that is used to corrupt text:

$$L_\theta = - \sum \log(T \mid g(T); \theta) \qquad (1)$$

with θ being the parameters of BART.

4.1 Multi-figurative Language Pre-training

We further pre-train BART for multi-figurative language modelling with a procedure that creates one model capable of modelling multiple figurative languages at once, so that (i) only one model needs to be maintained, and (ii) the model can benefit from cross-figurative knowledge transfer.

Inspired by Tang et al. (2020), we use a special token as a prefix in both the source and target text. That is, the text format is [form code] T [eos] with T being the text and the [form code] represents the form of the sentence. In the pre-training stage, we incorporate all the pre-training data of five figures of speech (Section 3) by concatenating data: $D = \{D_1, ..., D_i\}$ where each D_i is a collection of texts in a figurative form. Following Liu et al. (2020), our model is trained on a denoising task, where it is asked to reconstruct text from a version corrupted with a noise function that randomly masks 35% of the words in the sentence. The [form code] is used as the initial token to predict the sentence (Figure 1(a) (top)).

---

[3]More details about the pre-training data for each figure of speech are in Appendix A.1.

4.2 Literal↔Figurative Form Generation

In *Literal↔Figurative* generation, the model generates a text with the desired figure of speech given a literal text, or viceversa. First, following Lai et al. (2021), we use the parallel paraphrase pre-training data to make the model learn the basic task of rewriting. In practice, we incorporate all the data and add the corresponding form code at the beginning of each sentence to train the model in a supervised regime. Second, we fine-tune the model with the literal↔figurative parallel data (Table 2) in the same way (PT-to-FT; Figure 1(a) (down)). Since hyperbole and idiom datasets are too small, we upsample them by replication obtaining training sets of 10,000 sentence pairs.

4.3 Figurative↔Figurative Form Generation

In *Figurative↔Figurative* generation, the model takes a text with a given figurative form, and generates a text with the target figurative form. It is important to note that this procedure can have two outcomes: the target figure of speech *substitutes* the original one, or it is *added* to it, yielding a text that contains both the original and the target figurative forms.

Specifically, given a sentence of tokens $x = \{x_1, \cdots, x_n\}$ with the figure of speech s, the model is asked to generate the corresponding sequence $y = \{y_1, \cdots, y_m\}$ with the target figure of speech t. To overcome the lack of parallel data in different figures of speech which would be necessary to train such a model, we design a mechanism which can leak the information of the desired figure of speech to the encoder with a figurative embedding as additional input. Formally, we employ cross attention to inject the figurative information into word embedding of the input in the fine-tuning process (mFLAG; Figure 1(b)).

$$\text{CrossAttn}(\mathbf{W}, \mathbf{F}) = \text{softmax}(\frac{\mathbf{W}\mathbf{F}^{\mathrm{T}}}{\sqrt{d}})\mathbf{F} \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{m \times d}$ represents the embedding of the source sentence. $\mathbf{F} \in \mathbb{R}^{1 \times d}$ is the embedding of the target form code T. To avoid introducing new parameters and catastrophic forgetting, we do not use the commonly used feed-forward block here. We also employ a residual connection (He et al., 2016) for the word embedding:

$$\mathbf{C} = \text{CrossAttn}(\mathbf{W}, \mathbf{F}) + \mathbf{W} \quad (3)$$

Forms	Precision Score	Recall Score	F1 Score
Hyperbole	0.858	0.967	0.909
Idiom	0.897	0.961	0.928
Sarcasm	0.763	0.847	0.803
Metaphor	0.716	0.707	0.711
Simile	1.000	0.700	0.824

Table 3: Accuracy of classifiers for different forms.

The probability of the output can be computed conditioned both on the input sentence x and the target form code T. It can be formulated as:

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}, T) = \prod_{t=1}^{m} p_\theta(y_t|y_{1,\ldots,t-1}; \mathbf{C})) \quad (4)$$

We also first use the pre-training data to enhance model's rewriting ability, and employ upsampling to augment the gold training data for hyperbole and idiom. We use two settings for generation: (i) the model generates text in the target form directly from the source form (mFLAG-DR), meaning that direct figurative-figurative transformation is achieved; (ii) the model uses literal forms as pivot: it first transforms the source text back into its literal form, and then uses this obtained literal form to generate in the target form (mFLAG-BT). Comparing these two models will contribute to better understand the benefits of modelling multi-figurative language generation directly.

5 Experiments

All experiments are implemented atop Transformers (Wolf et al., 2020) using BART-large (Lewis et al., 2020). We train models with batch size 32, accumulating gradients over 8 update steps, using the Adam optimiser (Kingma and Ba, 2015) with learning rate 1e-5. We use early stopping (patience 5) if validation performance does not improve.

5.1 Evaluation Method

To assess the model performance we use automatic metrics commonly used in figurative language generation and text style transfer, which focus on form strength and context preservation.

Form Strength To evaluate the form accuracy of the generated text, we reuse the binary classifiers trained for selecting pre-training data. High confidence for the target figurative form, suggests high accuracy in the generation. The performance of the classifiers on the test set (Table 3), suggests that they are very reliable for Simile, Idiom, and Hyperbole, and slightly less for Metaphor and Sarcasm.

	TGT	BLEU	BERT	BLEURT	COMET	HM	TGT	BLEU	BERT	BLEURT	COMET	HM
		Literal Form→Hyperbole						Literal Form→Idiom				
BART-Single	0.627	0.513	0.693	0.280	0.461	0.564	0.711	**0.791**	**0.855**	**0.595**	**0.808**	0.749
BART-Multi	0.707	0.541	0.698	0.260	0.352	0.613	0.637	0.747	0.829	0.498	0.706	0.688
PT-to-FT	0.833	**0.582**	**0.733**	**0.379**	**0.490**	**0.686**	**0.769**	0.765	0.841	0.536	0.738	**0.767**
mFLAG	**0.844**	0.556	0.726	0.349	0.463	0.670	0.764	0.761	0.839	0.539	0.735	0.762
		Literal Form→Sarcasm						Literal Form→Metaphor				
BART-Single	0.679	**0.491**	**0.611**	**0.052**	**0.188**	0.570	0.720	0.595	0.771	0.364	0.720	0.652
BART-Multi	0.743	0.483	0.598	0.011	0.137	0.585	0.767	0.577	0.780	0.434	0.785	0.659
PT-to-FT	**0.765**	0.485	0.609	0.040	0.162	**0.594**	0.867	**0.643**	**0.812**	**0.493**	0.842	**0.738**
mFLAG	0.762	0.487	0.609	0.043	0.169	**0.594**	**0.880**	0.628	0.809	0.490	**0.844**	0.733
		Literal Form→Simile						Figurative→Literal Form				
BART-Single	0.647	0.724	0.720	**0.017**	**0.321**	0.683	0.733	0.606	0.742	0.284	0.455	0.663
BART-Multi	0.420	0.658	0.681	-0.025	0.178	0.513	0.725	0.622	0.762	0.364	0.522	0.670
PT-to-FT	0.907	0.729	0.722	-0.021	0.219	0.808	**0.801**	0.634	0.766	**0.542**	0.544	**0.708**
mFLAG	**0.953**	**0.745**	**0.727**	-0.021	0.220	**0.836**	0.796	**0.637**	**0.769**	0.375	**0.681**	0.707

Table 4: Results of literal↔figurative form generation. TGT represents the accuracy of output labeled as the target form by the classifier; the results of figurative→literal form generation are averaged across all figures of speech.

Context Preservation To assess this aspect, we adopt BLEU and BERTScore (F1-Score) (Zhang et al., 2020) following previous work (Chakrabarty et al., 2020b; Zhang and Wan, 2022; Zhou et al., 2021; Tian et al., 2021). In addition, we employ BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020), two learnable metrics that have shown promising results in the evaluation of formality transfer (Lai et al., 2022). For all metrics, we calculate scores between model outputs and references for the literal↔figurative generation, and between outputs and source sentences (and literal sentences) for figurative↔figurative generation as the latter has no parallel data available. [4]

Overall Score We compute the harmonic mean (HM) of figurative accuracy and BLEU score for a direct comparison to baselines.

5.2 Baselines

We compare our systems to two strong baselines.

BART-Single For each figure of speech, we fine-tune BART on the corresponding parallel data. For figurative→figurative generation, we use each figurative-to-literal model to generate the literal text, and then feed it into the model of the target form to generate the output.

BART-Multi We concatenate the five parallel training sets and fine-tune BART for multi-figurative language modelling, thereby enabling the generation between different forms.

5.3 Literal↔Figurative Generation

Table 4 presents the results of literal↔figurative form generation. BART-Multi outperforms BART-Single on most generation directions, except literal-to-idiom and literal-to-simile. This suggests that the model does benefit from multi-figurative language modelling with cross-figurative knowledge transfer. Compared to BART-Single and BART-Multi, both of our proposed models PT-to-FT and mFLAG have consistently stronger results. Specifically, we observe that BART-Single has the best performance only on context preservation for literal-to-idiom and literal-to-sarcasm generation, while our models are better for the rest, especially with a good balance between form strength and context preservation. The results confirm that our pre-training scheme and strategies significantly improve performances for multi-figurative language modelling. When looking at PT-to-FT and mFLAG, we see that these two models' performances are very close on all tasks and do not show a clear and consistent trend. The main reason for this is most likely that the settings of the two models are almost identical except that mFLAG has a figurative injection mechanism, and they are both trained with parallel literal↔figurative sentence pairs.

5.4 Figurative↔Figurative Generation

Table 5 reports results of figurative↔figurative form generation.[5] We see that both BART-Multi and PT-to-FT perform poorly on the form strength

---

[4] In our evaluation, we take `multi-bleu.perl` to calculate BLEU score, and models `bleurt-large-512` and `wmt-large-da-estimator-1719` for BLEURT and COMET, respectively.

[5] Complete results are in Appendix A.2.

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
					Hyperbole→Others							
BART-Single	0.470	0.425	0.665	0.782	0.459	0.472	0.519	0.488	0.700	0.294	0.248	0.454
BART-Multi	0.328	0.242	0.602	0.761	0.455	0.443	0.345	0.505	0.731	0.427	0.385	0.327
PT-to-FT	0.252	0.258	0.590	0.749	0.437	0.420	0.359	**0.507**	**0.732**	**0.438**	**0.407**	0.342
mFLAG-DR	**0.922**	0.608	**0.815**	**0.893**	**0.753**	**0.836**	**0.696**	0.411	0.633	0.036	-0.105	0.490
mFLAG-BT	0.482	**0.644**	0.539	0.702	0.253	0.246	0.586	0.421	0.662	0.169	0.093	**0.509**
					Idiom→Others							
BART-Single	0.290	0.309	0.783	0.864	0.575	0.646	0.443	0.749	0.844	0.578	0.659	0.438
BART-Multi	0.273	0.204	0.785	0.873	0.602	0.674	0.324	0.758	0.859	0.630	0.701	0.408
PT-to-FT	0.204	0.207	0.771	0.867	0.594	0.662	0.326	**0.760**	**0.860**	**0.646**	**0.715**	0.325
mFLAG-DR	**0.910**	0.400	**0.901**	**0.940**	**0.822**	**0.869**	0.554	0.694	0.799	0.328	0.375	0.507
mFLAG-BT	0.328	**0.409**	0.724	0.831	0.491	0.566	0.523	0.703	0.816	0.490	0.569	**0.517**
					Sarcasm→Others							
BART-Single	0.577	0.370	0.877	0.899	0.650	0.792	0.520	0.454	0.579	-0.088	-0.051	0.408
BART-Multi	0.569	0.247	0.903	0.923	0.701	0.838	0.388	0.471	**0.593**	-0.049	-0.014	0.324
PT-to-FT	0.464	0.252	0.863	0.891	0.613	0.774	0.390	**0.468**	0.592	**-0.031**	**0.000**	0.328
mFLAG-DR	**0.840**	0.438	**0.907**	**0.928**	**0.813**	**0.872**	0.591	0.442	0.563	-0.198	-0.143	0.440
mFLAG-BT	0.583	**0.481**	0.808	0.831	0.460	0.604	**0.605**	0.430	0.554	-0.164	-0.133	**0.454**
					Metaphor→Others							
BART-Single	0.163	0.314	0.603	0.776	0.412	0.555	0.413	0.575	0.773	0.381	0.486	0.406
BART-Multi	0.255	0.249	0.647	0.825	0.554	0.723	0.360	0.632	0.820	0.550	0.689	0.357
PT-to-FT	0.147	0.254	0.671	0.832	0.599	**0.763**	0.369	**0.648**	**0.824**	**0.507**	**0.665**	0.365
mFLAG-DR	**0.795**	0.518	**0.697**	**0.846**	**0.614**	0.706	**0.594**	0.516	0.758	0.320	0.410	0.517
mFLAG-BT	0.387	**0.557**	0.502	0.734	0.329	0.434	0.528	0.496	0.743	0.317	0.417	**0.525**
					Simile→Others							
BART-Single	0.057	0.607	0.469	0.559	-0.406	-0.429	0.529	0.588	0.667	0.160	-0.102	0.597
BART-Multi	0.007	0.272	0.629	0.686	-0.043	-0.051	0.380	**0.765**	**0.818**	**0.262**	**0.415**	0.401
PT-to-FT	0.000	0.314	0.622	0.671	-0.031	-0.067	0.417	0.754	0.804	0.244	0.394	0.443
mFLAG-DR	**0.440**	0.685	**0.849**	**0.884**	**0.637**	**0.690**	0.758	0.589	0.698	-0.016	-0.057	0.633
mFLAG-BT	0.132	**0.687**	0.606	0.670	-0.069	-0.064	0.644	0.672	0.766	0.163	0.250	**0.679**

Table 5: Results of figurative↔figurative form generation. Notes: (i) SRC (TGT) represents the accuracy of output labeled as the source (target) form by the classifier of the source (target) form; (ii) results for each block are averaged for all generations from one figurative language to others.

and the context preservation computed against the source text. The low form strength (SRC and TGT, see table's caption) and high scores of context preservation (using literal text) suggest that these two models transform the source text into the literal form. BART-Single, interestingly, shows a better performance on both form strength and context preservation. For mFLAG-DR and mFLAG-BT, we see that they show the best performance across the board: (i) mFLAG-DR shows a significant improvement in target figurative form (TGT) while maintaining the original form (SRC) very much; it also achieves the best performance on context preservation; (ii) mFLAG-BT achieves the highest form accuracy in the target figure of speech while reducing the original form strength.

It is interesting to note that the direct generation method might allow for the source figure of speech to be retained in the generated sentence, as we do not explicitly remove it by transforming the sentence to its literal form first. For example, with

hyperbolic input "*I am not happy that he urged me to finish all the hard task in the world*", one of our sarcastic transformations reads "*Thank you for encouraging me to finish all the hardest tasks in the world*", where the hyperbolic part ("*all the hardest tasks in the world*") is preserved unchanged (see Table 6).

Overall, the results show that mFLAG with the mechanism for injecting the figurative information into the encoder can generate from one figure of speech to another even without task-specific parallel data.

6 Analysis and Discussion

Case Study Table 6 shows a group of example outputs for hyperbole→others generated by various models.[6] From the results of hyperbole→literal generation, we see that mFLAG generates the literal sentence from the hyperbolic one very well,

---

[6]More example outputs of mFLAG are in Appendix A.3.

Forms	Models	Sentences	Suc.
Hyperbole	Input	I am not happy that he urged me to finish all the hardest tasks in the world.	-
Literal	BART-Single	I am not happy that he urged me to finish all the hard tasks in the world.	✗
	BART-Multi	I am glad that he urged me to finish all the hardest tasks in the world.	✗
	PT-to-FT	I am not happy that he urged me to complete all the difficult tasks in the world.	✗
	mFLAG	I am not happy that he urged me to finish all the difficult tasks.	✔
Idiom	BART-Single	I am not happy that he urged me to finish all the hard tasks in the world.	✗
	BART-Multi	I am so glad that he urged me to finish all the hardest tasks in the world.	✗
	PT-to-FT	I am not happy that he urged me to complete all the difficult tasks.	✗
	mFLAG-DR	I am not happy that he urged me to tick off all the hardest tasks in the world.	✔
	mFLAG-BT	I am not happy that he urged me to get cracking on all the difficult tasks.	✔
Sarcasm	BART-Single	I'm so glad my dad encouraged me to finish all the hard tasks in the world.	✔
	BART-Multi	Thank you for encouraging me to finish all the hardest tasks in the world.	✔
	PT-to-FT	I am happy that he urged me to finish all the hardest tasks in the world.	✔
	mFLAG-DR	Glad he urged me to finish all the hardest tasks in the world.	✔
	mFLAG-BT	Glad he urged me to finish all the difficult tasks.	✔
Metaphor	BART-Single	I am not happy that he urged me to bear all the difficult tasks.	✔
	BART-Multi	I am so glad that he urged me to finish all the hardest tasks in the world.	✗
	PT-to-FT	I am not happy that he urged me to complete all the difficult tasks in the world.	✗
	mFLAG-DR	I am not happy that he urged me to bear all the hardest tasks in the world.	✔
	mFLAG-BT	I am not happy that he pressed me to finish all the difficult tasks.	✗
Simile	BART-Single	I am not happy that he urged me to finish all the difficult tasks.	✗
	BART-Multi	I am so glad that he urged me to finish all the hardest tasks in the world.	✗
	PT-to-FT	I am not happy that he urged me to complete all the difficult tasks in the world.	✗
	mFLAG-DR	I am not happy that he urged me to finish all the like a million things.	✔
	mFLAG-BT	I am not happy that he urged me to finish all the difficult tasks.	✗

Table 6: Examples outputs generated by various models from hyperbolic text, where red denotes appropriate words/phrases for desired forms. Suc.==Successful.

confirming that texts generated by mFLAG-BT tend to contain less the source form by substituting it with the target form. In figurative↔figurative generation, all models nicely generate sarcastic text while all baselines usually fail at generating the other forms. Since the metaphor generation dataset we used focuses on metaphorical verb aspect, we consider the outputs of BART-Single and mFLAG-DR to be successful. Overall, our proposed mFLAG based models perform better on all generation directions.

Probing Figurative Information for Encoder To measure the distribution of source and target sentences encoded by the Encoder with/without the mechanism of injecting figurative information, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the Encoder outputs and visualise relations between tokens in a two-dimensional space. Fig. 2(a) and 2(b) show the results of a source literal text "*He was nervous waiting for the result.*" and a target hyperbolic text "*He was on pins and needles waiting for the result.*". We see the word "He" and 'was' of the two sentences are not in the same cluster in 2(a) while it is interesting to see that all distances between token pairs of 2(b) are closer, especially the phrase "on pins and needles", and "nervous" are almost in the same cluster in 2(b). Fig. 2(c) and 2(d) show the results of a source idiomatic text "*I felt like I had a feather in my cap after I aced that exam.*" and a target hyperbolic text "*I felt like I was a star after I aced that exam.*". We observe that the token pairs like "I", "like" and 'felt' of mFLAG are closer than those of PT-to-FT. It is also interesting to see that the phrase "a feather in my cap" and the token "star" make more of a cluster in 2(d). We believe this benefits the decoder, especially decoding into the target figurative form.

How similar are different forms? To analyze the connection between literal and figurative forms,
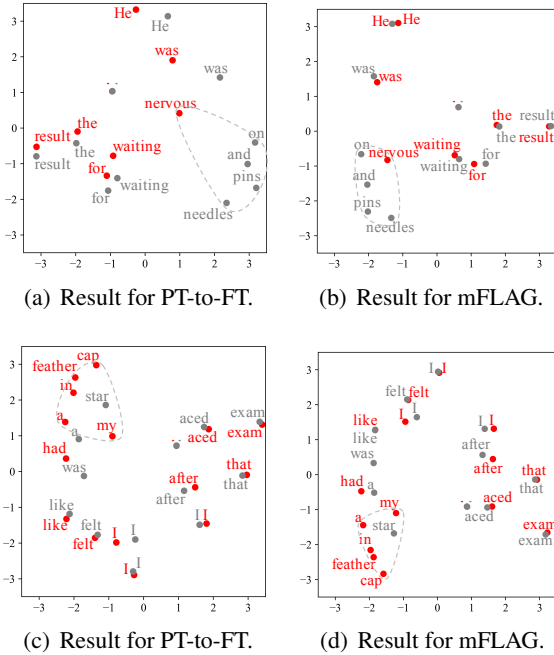
(a) Result for PT-to-FT. (b) Result for mFLAG.

(c) Result for PT-to-FT. (d) Result for mFLAG.

Figure 2: PCA token representations of encoder outputs for literal→hyperbole (top) and idiom→hyperbole (down).

Figure 3: Performances (F1 score) of classifiers on different figurative forms. Each row represents results of a classifier tested on each/all figurative form(s).

and between different figures of speech, we evaluate each figurative classifier on the test sets of the other figurative forms (Figure 3). We first see that the overall model (literal vs figurative) achieves F1 scores of over 0.69 for each figure of speech, confirming the feasibility of multi-figurative modelling. For each figure of speech, we observe: (i) classifiers for hyperbole and idiom have high F1 scores on the test set of simile (0.79 and 0.84), suggesting that sentences with similes may also be hyperbolic or idiomatic; (ii) for sarcasm and metaphor, classifiers have medium scores on other forms; (iii) the classifier of simile achieves F1 scores of less than 0.11 on other figures of speech; this is due to the fact that the simile dataset was created using the format *like a*, which is easy for the model to learn. Different figurative forms are related to each other, confirming that models can benefit from cross-figurative knowledge transfer. Further (computational) analysis of similarities and differences will help to even better leverage such transfer.

7 Conclusion and Outlook

We have proposed a novel task of multi-figurative language generation, and shown that our models do benefit from cross-figurative knowledge transfer. Paraphrasing data can be leveraged in further pre-training to enhance both form strength and context

preservation in figurative language generation. We have also proposed a mechanism for injecting the target figurative information into the encoder, so that we can achieve generation between different figures of speech even without parallel figurative-figurative pairs.

While we innovatively explore multi-figurative language generation across literal and five figurative forms, and our model achieves the best performances compared to baselines, there is still substantial room for improvement and further extensions.

The current lack of human references for automating the evaluation of figurative-to-figurative generation is surely a limitation in terms of better understanding of the models' behaviour and potential improvements. More in general, figurative language generation is a relatively new task, which still lacks standardised evaluation methods, both in terms of automatic metrics and human-based evaluation.

Also, we introduce for the first time generation across literal expressions and five figurative forms, but there are many more forms of creative writing that could be modelled. Moreover, we only limited our attention to English, due to data availability, but are convinced that datasets in other languages would greatly benefit research in this area. Indeed, multilingual modelling would make it possible to make connections across different languages, thus shedding more light on cross-lingual regularities in figurative language use, and thus also open up potential avenues to tackle this task better.

References

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

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106, Berlin, Germany. Association for Computational Linguistics.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland. 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.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3152–3158. AAAI Press.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020a. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020b. 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.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, page 160–167, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*.

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, Marseille, France. European Language Resources Association.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Arthur M. Jacobs. 2018. The gutenberg english poetry corpus: Exemplary quantitative narrative analyses. *Frontiers Digit. Humanit.*, 5:5.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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, Melbourne, Australia. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021a. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, Online. Association for Computational Linguistics.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021b. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 6724–6736, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint, arXiv: 2008.00401.*

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.

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.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14383–14392.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yunxiang Zhang and Xiaojun Wan. 2022. MOVER: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030, Seattle, United States. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, Srihari Nanniyur, and Suma Bhat. 2021. From solving a problem boldly to cutting the gordian knot: Idiomatic text generation. *arXiv preprint, arXiv: 2104.06541.*

Mengdi Zhu, Zhiwei Yu, and Xiaojun Wan. 2019. A neural approach to irony generation. *arXiv preprint, arXiv: 1909.06200.*

A Appendices:

This appendices include: (i) Dataset statistics of pre-training data (A.1); (ii) Detailed results for figurative↔figurative generation (A.2); (iii) Example outputs of mFLAG (A.3) .

A.1 Pre-Training Data

Forms	Task	σ	Train	Valid
Hyperbole	Literal text↔Hyperbole	0.94	102,887	5,000
Idiom	Literal text↔idiom	0.95	133,285	5,000
Sarcasm	Literal text↔Sarcasm	0.70	22,550	5,000
Metaphor	Literal text↔Metaphor	0.95	206,554	5,000
Simile	Literal text↔Simile	0.76	57,566	5,000

Table A.1: Dataset statistics for generic pre-training data. Note that σ is the threshold used to select sentence pairs.

A.2 Detailed Results for Figurative↔Figurative Generation

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
Hyperbole→Idiom												
BART-Single	0.513	0.513	0.653	0.781	0.469	0.466	0.575	0.471	0.692	0.294	0.240	0.491
BART-Multi	0.313	0.233	0.595	0.755	0.439	0.425	0.335	0.505	0.730	0.429	0.385	0.386
PT-to-FT	0.240	0.200	0.587	0.747	0.445	0.422	0.298	0.506	0.729	0.442	0.402	0.287
mFLAG-DR	0.900	0.733	0.766	0.876	0.729	0.758	0.749	0.401	0.637	0.063	-0.089	0.518
mFLAG-BT	0.653	0.707	0.599	0.743	0.368	0.380	0.649	0.409	0.650	0.136	-0.011	0.518
Hyperbole→Sarcasm												
BART-Single	0.407	0.387	0.673	0.785	0.464	0.595	0.491	0.499	0.710	0.300	0.298	0.436
BART-Multi	0.333	0.313	0.601	0.760	0.464	0.447	0.412	0.500	0.730	0.427	0.386	0.385
PT-to-FT	0.267	0.373	0.587	0.744	0.400	0.399	0.456	0.505	0.728	0.392	0.385	0.429
mFLAG-DR	0.900	0.447	0.873	0.922	0.883	0.947	0.591	0.431	0.645	0.073	-0.006	0.439
mFLAG-BT	0.373	0.507	0.545	0.699	0.283	0.265	0.525	0.442	0.678	0.233	0.233	0.472
Hyperbole→Metaphor												
BART-Single	0.407	0.533	0.653	0.784	0.501	0.509	0.587	0.499	0.712	0.369	0.331	0.515
BART-Multi	0.320	0.407	0.597	0.758	0.439	0.432	0.484	0.505	0.730	0.422	0.383	0.451
PT-to-FT	0.253	0.447	0.592	0.756	0.450	0.432	0.509	0.513	0.736	0.451	0.423	0.478
mFLAG-DR	0.927	0.773	0.823	0.902	0.762	0.870	0.797	0.412	0.634	0.033	-0.081	0.538
mFLAG-BT	0.300	0.753	0.495	0.692	0.227	0.235	0.597	0.433	0.686	0.252	0.226	0.550
Hyperbole→Simile												
BART-Single	0.553	0.267	0.680	0.779	0.402	0.416	0.383	0.481	0.687	0.214	0.123	0.342
BART-Multi	0.347	0.013	0.616	0.771	0.476	0.467	0.025	0.511	0.733	0.431	0.387	0.025
PT-to-FT	0.247	0.013	0.595	0.747	0.451	0.424	0.025	0.505	0.732	0.465	0.418	0.332
mFLAG-DR	0.960	0.480	0.798	0.873	0.639	0.709	0.599	0.400	0.616	-0.026	-0.242	0.436
mFLAG-BT	0.600	0.607	0.525	0.674	0.135	0.105	0.551	0.401	0.634	0.055	-0.077	0.563

Table A.2: Results of hyperbole→others generation.

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
					Idiom→Hyperbole							
BART-Single	0.311	0.103	0.788	0.867	0.585	0.653	0.182	0.751	0.844	0.575	0.651	0.181
BART-Multi	0.269	0.031	0.784	0.872	0.600	0.671	0.059	0.758	0.859	0.632	0.702	0.059
PT-to-FT	0.232	0.041	0.782	0.874	0.614	0.681	0.078	0.763	0.862	0.647	0.717	0.078
mFLAG-DR	0.929	0.232	0.847	0.908	0.716	0.769	0.364	0.667	0.783	0.286	0.317	0.344
mFLAG-BT	0.564	0.172	0.728	0.836	0.523	0.574	0.278	0.679	0.799	0.415	0.477	0.274
					Idiom→Sarcasm							
BART-Single	0.277	0.335	0.795	0.872	0.602	0.671	0.471	0.761	0.853	0.609	0.692	0.465
BART-Multi	0.281	0.292	0.785	0.875	0.608	0.679	0.426	0.756	0.857	0.623	0.693	0.421
PT-to-FT	0.230	0.319	0.773	0.866	0.587	0.657	0.452	0.755	0.854	0.620	0.690	0.449
mFLAG-DR	0.924	0.376	0.927	0.955	0.871	0.919	0.535	0.711	0.804	0.345	0.395	0.492
mFLAG-BT	0.233	0.405	0.721	0.828	0.485	0.570	0.519	0.710	0.821	0.515	0.613	0.516
					Idiom→Metaphor							
BART-Single	0.280	0.692	0.768	0.858	0.561	0.643	0.728	0.734	0.840	0.571	0.667	0.728
BART-Multi	0.268	0.485	0.784	0.872	0.600	0.671	0.599	0.759	0.859	0.633	0.703	0.592
PT-to-FT	0.170	0.467	0.762	0.862	0.581	0.656	0.579	0.760	0.862	0.656	0.728	0.579
mFLAG-DR	0.866	0.798	0.879	0.938	0.821	0.876	0.837	0.687	0.803	0.359	0.420	0.739
mFLAG-BT	0.247	0.798	0.703	0.828	0.482	0.580	0.747	0.688	0.820	0.515	0.620	0.739
					Idiom→Simile							
BART-Single	0.293	0.106	0.782	0.859	0.550	0.616	0.187	0.748	0.839	0.557	0.627	0.186
BART-Multi	0.274	0.007	0.786	0.874	0.601	0.673	0.014	0.759	0.860	0.633	0.704	0.014
PT-to-FT	0.184	0.000	0.766	0.864	0.592	0.655	0.000	0.762	0.862	0.662	0.726	0.000
mFLAG-DR	0.920	0.193	0.949	0.959	0.878	0.909	0.321	0.712	0.805	0.322	0.368	0.304
mFLAG-BT	0.266	0.259	0.744	0.832	0.475	0.539	0.384	0.736	0.825	0.514	0.566	0.383

Table A.3: Results of idiom→others generation.

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
					Sarcasm→Hyperbole							
BART-Single	0.568	0.405	0.907	0.921	0.727	0.855	0.560	0.470	0.590	-0.050	-0.010	0.435
BART-Multi	0.558	0.347	0.898	0.918	0.690	0.828	0.501	0.471	0.592	-0.048	-0.013	0.400
PT-to-FT	0.459	0.384	0.878	0.901	0.635	0.799	0.534	0.473	0.595	-0.022	0.010	0.349
mFLAG-DR	0.823	0.466	0.914	0.936	0.862	0.904	0.617	0.449	0.569	-0.169	-0.114	0.457
mFLAG-BT	0.612	0.473	0.821	0.849	0.548	0.675	0.595	0.438	0.562	-0.123	-0.095	0.455
					Sarcasm→Idiom							
BART-Single	0.582	0.429	0.853	0.889	0.615	0.730	0.571	0.441	0.575	-0.098	-0.090	0.435
BART-Multi	0.568	0.299	0.901	0.921	0.697	0.836	0.449	0.472	0.593	-0.051	-0.017	0.366
PT-to-FT	0.422	0.276	0.862	0.886	0.599	0.700	0.418	0.462	0.594	-0.024	0.006	0.394
mFLAG-DR	0.847	0.517	0.875	0.911	0.749	0.808	0.650	0.426	0.554	-0.229	-0.193	0.467
mFLAG-BT	0.599	0.527	0.791	0.825	0.442	0.570	0.633	0.417	0.550	-0.176	-0.166	0.466
					Sarcasm→Metaphor							
BART-Single	0.571	0.483	0.851	0.881	0.591	0.788	0.616	0.445	0.571	-0.112	-0.049	0.463
BART-Multi	0.561	0.337	0.900	0.919	0.693	0.830	0.490	0.471	0.592	-0.046	-0.014	0.393
PT-to-FT	0.514	0.344	0.870	0.901	0.654	0.796	0.493	0.472	0.592	-0.037	-0.007	0.398
mFLAG-DR	0.833	0.534	0.907	0.928	0.805	0.906	0.672	0.439	0.563	-0.203	-0.119	0.482
mFLAG-BT	0.520	0.578	0.790	0.827	0.424	0.627	0.668	0.431	0.556	-0.166	-0.100	0.494
					Sarcasm→Simile							
BART-Single	0.585	0.163	0.897	0.906	0.666	0.793	0.276	0.460	0.581	-0.091	-0.056	0.241
BART-Multi	0.588	0.003	0.911	0.932	0.725	0.857	0.006	0.471	0.594	-0.050	-0.013	0.005
PT-to-FT	0.459	0.003	0.842	0.874	0.565	0.730	0.006	0.465	0.587	-0.042	-0.008	0.006
mFLAG-DR	0.857	0.235	0.932	0.937	0.835	0.870	0.375	0.452	0.566	-0.191	-0.144	0.309
mFLAG-BT	0.599	0.344	0.821	0.822	0.424	0.544	0.485	0.433	0.547	-0.189	-0.171	0.383

Table A.4: Results of sarcasm→others generation.

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
					Metaphor→Hyperbole							
BART-Single	0.173	0.480	0.617	0.786	0.446	0.582	0.540	0.588	0.779	0.399	0.511	0.529
BART-Multi	0.260	0.427	0.643	0.826	0.562	0.722	0.513	0.635	0.825	0.561	0.700	0.511
PT-to-FT	0.233	0.480	0.711	0.870	0.709	0.832	0.573	0.639	0.827	0.508	0.667	0.548
mFLAG-DR	0.827	0.653	0.662	0.846	0.634	0.717	0.657	0.516	0.769	0.359	0.450	0.576
mFLAG-BT	0.453	0.620	0.511	0.755	0.438	0.511	0.560	0.496	0.762	0.404	0.492	0.551
					Metaphor→Idiom							
BART-Single	0.240	0.447	0.542	0.744	0.361	0.459	0.490	0.518	0.748	0.350	0.411	0.480
BART-Multi	0.253	0.280	0.643	0.825	0.559	0.724	0.390	0.633	0.822	0.550	0.694	0.388
PT-to-FT	0.113	0.260	0.646	0.819	0.573	0.748	0.371	0.657	0.834	0.554	0.683	0.373
mFLAG-DR	0.887	0.547	0.640	0.829	0.582	0.708	0.590	0.542	0.787	0.444	0.561	0.544
mFLAG-BT	0.653	0.547	0.557	0.771	0.453	0.586	0.552	0.524	0.774	0.416	0.541	0.536
					Metaphor→Sarcasm							
BART-Single	0.133	0.240	0.623	0.788	0.424	0.604	0.347	0.597	0.782	0.391	0.532	0.347
BART-Multi	0.233	0.280	0.654	0.820	0.527	0.712	0.392	0.621	0.807	0.510	0.652	0.386
PT-to-FT	0.153	0.267	0.683	0.832	0.574	0.761	0.384	0.645	0.812	0.462	0.650	0.378
mFLAG-DR	0.720	0.347	0.788	0.883	0.760	0.843	0.482	0.557	0.767	0.377	0.486	0.428
mFLAG-BT	0.273	0.427	0.511	0.732	0.322	0.496	0.465	0.516	0.742	0.334	0.500	0.467
					Metaphor→Simile							
BART-Single	0.107	0.087	0.631	0.785	0.418	0.574	0.153	0.598	0.775	0.384	0.489	0.152
BART-Multi	0.273	0.007	0.647	0.828	0.569	0.733	0.014	0.637	0.826	0.579	0.710	0.014
PT-to-FT	0.087	0.007	0.643	0.808	0.540	0.711	0.014	0.650	0.822	0.503	0.661	0.014
mFLAG-DR	0.747	0.500	0.696	0.827	0.479	0.554	0.581	0.450	0.710	0.099	0.142	0.474
mFLAG-BT	0.167	0.633	0.428	0.679	0.102	0.142	0.511	0.447	0.695	0.115	0.135	0.524

Table A.5: Results of metaphor→others generation.

	Form Strength		Source Text					Literal Text				
	SRC	TGT	BLEU	BERT	BLEURT	COMET	HM	BLEU	BERT	BLEURT	COMET	HM
					Simile→Hyperbole							
BART-Single	0.093	0.713	0.492	0.575	-0.358	-0.358	0.582	0.603	0.656	-0.135	-0.127	0.653
BART-Multi	0.007	0.293	0.634	0.689	-0.040	-0.045	0.401	0.770	0.821	0.261	0.418	0.424
PT-to-FT	0.000	0.327	0.649	0.692	0.003	-0.012	0.435	0.777	0.818	0.261	0.417	0.460
mFLAG-DR	0.527	0.893	0.895	0.918	0.772	0.811	0.894	0.583	0.685	-0.041	-0.090	0.705
mFLAG-BT	0.240	0.820	0.640	0.687	-0.035	-0.022	0.719	0.657	0.756	0.162	0.171	0.730
					Simile→Idiom							
BART-Single	0.127	0.627	0.488	0.554	-0.367	-0.440	0.549	0.589	0.646	-0.169	-0.204	0.607
BART-Multi	0.007	0.207	0.634	0.689	-0.040	-0.045	0.273	0.770	0.821	0.261	0.418	0.326
PT-to-FT	0.000	0.173	0.644	0.684	0.007	-0.038	0.273	0.781	0.830	0.307	0.470	0.283
mFLAG-DR	0.420	0.800	0.810	0.848	0.508	0.554	0.805	0.600	0.710	0.013	-0.025	0.686
mFLAG-BT	0.200	0.773	0.617	0.683	-0.018	-0.009	0.686	0.636	0.761	0.170	0.212	0.698
					Simile→Sarcasm							
BART-Single	0.007	0.440	0.479	0.572	-0.402	-0.420	0.459	0.618	0.704	-0.113	0.001	0.514
BART-Multi	0.007	0.233	0.611	0.671	-0.070	-0.086	0.337	0.748	0.806	0.252	0.396	0.355
PT-to-FT	0.000	0.387	0.551	0.623	-0.128	-0.178	0.455	0.677	0.743	0.117	0.242	0.492
mFLAG-DR	0.373	0.367	0.877	0.892	0.671	0.692	0.517	0.619	0.714	0.001	-0.014	0.598
mFLAG-BT	0.073	0.380	0.618	0.672	-0.057	-0.057	0.471	0.726	0.792	0.241	0.362	0.499
					Simile→Metaphor							
BART-Single	0.000	0.647	0.418	0.536	-0.497	-0.499	0.508	0.541	0.660	-0.222	-0.083	0.589
BART-Multi	0.007	0.353	0.638	0.694	-0.022	-0.026	0.455	0.772	0.824	0.273	0.429	0.484
PT-to-FT	0.000	0.367	0.643	0.685	-0.007	-0.041	0.467	0.782	0.825	0.289	0.445	0.500
mFLAG-DR	0.440	0.680	0.815	0.878	0.595	0.702	0.741	0.552	0.681	-0.036	-0.099	0.609
mFLAG-BT	0.013	0.773	0.550	0.638	-0.167	-0.166	0.643	0.668	0.757	0.079	0.256	0.717

Table A.6: Results of simile→others generation.

A.3 Example Outputs of mFLAG

Forms	Sentences	Suc.
Literal [Input]	Old mr. smith has been teaching here for a very long time.	-
Hyperbole	Old mr. smith has been teaching here since the stone age.	✔
Literal [Input]	I like your new haircut!	-
Hyperbole	I love your new haircut!	✗
Hyperbole [Input]	I am not happy that he urged me to finish all the hardest tasks in the world.	-
Idiom	I am not happy that he urged me to tick off all the hardest tasks in the world.	✔
Sarcasm	Glad he urged me to finish all the hardest tasks in the world.	✔
Metaphor	I am not happy that he urged me to bear all the hardest tasks in the world.	✔
Simile	I am not happy that he urged me to finish all the like a million things.	✔
Idiom [Input]	Your exam results this term were not up to the mark.	-
Hyperbole	Your exam results this term were way over the top.	✔
Sarcasm	Your exam results this term were up to the mark.	✔
Metaphor	Your exam results this term were not up to the mark.	✗
Simile	Your exam results this term were not up to the mark.	✗
Sarcasm [Input]	Being stuck in an airport is fun.	-
Hyperbole	Being stuck in an airport is the best.	✔
Idiom	Being stuck in an airport is always fun.	✗
Metaphor	Being stuck in an airport is no fun.	✗
Simile	Being stuck in an airport is like a rollercoaster.	✔
Metaphor	The bad news rocked her hopes.	-
Hyperbole	The bad news crushed her hopes.	✔
Idiom	The bad news crushed her hopes.	✗
Sarcasm	The good news just keeps on getting better.	✔
Simile	The bad news rocked her like a rock.	✔
Simile [Input]	One day you decide to test it out and what you experience is like a magic trick.	-
Hyperbole	One day you decide to test it out and what you experience is magic.	✔
Sarcasm	One day you decide to test it out and what you experience is awesome.	✔
Idiom	One day you decide to test it out and what you experience is dangerous.	✗
Metaphor	One day you decide to test it out and what you experience is dangerous.	✗

Table A.7: Example outputs generated by mFLAG-DR, where red denotes appropriate words for desired forms. Suc.==Successful.