HAUSER: Towards Holistic and Automatic Evaluation of Simile Generation

Qianyu He[1], **Yikai Zhang**[1], **Jiaqing Liang**[2]*,
Yuncheng Huang[1], **Yanghua Xiao**[1]*, **Yunwen Chen**[3]
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]School of Data Science, Fudan University [3]DataGrand Inc., Shanghai, China
{qyhe21, ykzhang22, yunchenghuang22}@m.fudan.edu.cn,
{liangjiaqing, shawyh}@fudan.edu.cn, chenyunwen@datagrand.com

Abstract

Similes play an imperative role in creative writing such as story and dialogue generation. Proper evaluation metrics are like a beacon guiding the research of simile generation (SG). However, it remains under-explored as to what criteria should be considered, how to quantify each criterion into metrics, and whether the metrics are effective for comprehensive, efficient, and reliable SG evaluation. To address the issues, we establish HAUSER, a holistic and automatic evaluation system for the SG task, which consists of five criteria from three perspectives and automatic metrics for each criterion. Through extensive experiments, we verify that our metrics are significantly more correlated with human ratings from each perspective compared with prior automatic metrics. Resources of HAUSER are publicly available at https://github.com/Abbey4799/HAUSER.

1 Introduction

Similes play a vital role in human expression, making literal sentences imaginative and graspable. For example, Robert Burns famously wrote "*My Luve is like a red, red rose*" to metaphorically depict the beloved as being beautiful. In this simile, "*Luve*" (a.k.a. topic) is compared with "*red rose*" (a.k.a. vehicle) via the implicit property "*beautiful*" and the event "*is*". Here, topic, vehicle, property, and event are four main *simile components* (Hanks, 2013). As a figure of speech, similes have been widely used in literature and conversations (Zheng et al., 2019; Chakrabarty et al., 2022).

Simile generation (SG) is a crucial task in natural language processing (Chakrabarty et al., 2020; Zhang et al., 2021; Lai and Nissim, 2022), with the aim of polishing literal sentences into similes. In Fig. 1, the literal sentence "*He yelps and howls.*" is polished into a simile by inserting the phrase "*like a wolf*", resulting in "*He yelps and howls like a wolf*". The ability to generate similes can assist various downstream tasks, such as making the generations more imaginative in story or poet generation task (Tartakovsky and Shen, 2018; Chakrabarty et al., 2022) and the generated response more human-like in dialogue generation task (Zheng et al., 2019).

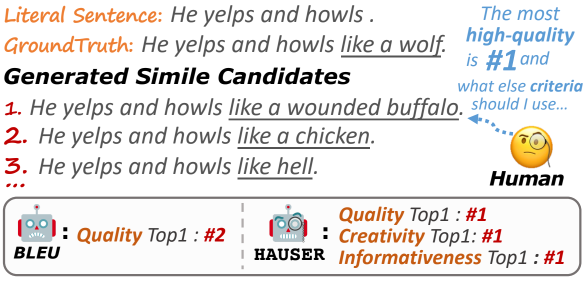
Literal Sentence: He yelps and howls .
GroundTruth: He yelps and howls like a wolf.
Generated Simile Candidates
1. He yelps and howls like a wounded buffalo.
2. He yelps and howls like a chicken.
3. He yelps and howls like hell.
...
The most high-quality is #1 and what else criteria should I use...

Figure 1: An example of Simile Generation (SG) Evaluation. The commonly used automatic metric BLEU deems the second candidate as the most *high-quality* one among all the generated similes, while our proposed metrics HAUSER deem the first candidate as the best one regarding its *quality*, *creativity* and *informativeness*, which better correlates with human ratings and also provides more criteria for SG evaluation.

Automatic evaluation is critical for the SG task since it enables efficient, systematic, and scalable comparisons between models in general (Celikyilmaz et al., 2020). However, existing studies are inadequate for effective SG evaluation. Task-agnostic automatic metrics (Papineni et al., 2002; Zhang et al., 2019; Li et al., 2016) are widely adopted for SG evaluation (Zhang et al., 2021; Lai and Nissim, 2022), which have several limitations: (1) The simile components should receive more attention than other words during SG evaluation (e.g. "*he*" and "*wolf*" in Fig. 1), while there are no automatic metrics that consider the key components. (2) The SG task is open-ended, allowing for multiple plausible generations for the same input (Chakrabarty et al., 2020) (e.g. the howling man can be compared to "*wolf*", "*buffalo*", or "*tiger*" in Fig. 1). Hence, the metrics based on word overlap with a few references are inadequate to accurately mea-

[*] Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12557–12572
July 9-14, 2023 ©2023 Association for Computational Linguistics

Criterion		Literal Sentence	Example Simile Candidates
Quality	Relevance	Some raindrops struck the roof, window and ran down its panes.	Some raindrops struck the roof, window and ran down its panes (**like tears** | like arrows).
	Logical Consistency	Stefan moved, every movement easy and precisely controlled.	Stefan moved (like lightning | **like a dancer**), every movement easy and precisely controlled.
	Sentiment Consistency	The idea resounded throughout the land.	The idea resounded (like an earthquake | **like a thunderous wave**) throughout the land.
Creativity		He possessed a power of sarcasm which could scorch.	He possessed a power of sarcasm which could scorch (**like vitriol** | like fire).
Informativeness		They gleamed.	They gleamed (like the eyes of a cat | **like the eyes of an angry cat**).

Table 1: Examples of our criteria for Simile Generation (SG) Evaluation. We design five criteria from three perspectives. The vehicles of the better simile candidates given by each criterion are highlighted in bold.

sure the overall quality of generated similes. As shown in Fig. 1, the commonly used metric BLEU deems the second candidate as the highest quality, as it has more overlapped words with the only referenced groundtruth, while human deems the first candidate as the most coherent one. (3) The existing metrics are inadequate to provide fine-grained and comprehensive SG evaluation, considering that the creative generation tasks have distinct criteria for desired generations (Celikyilmaz et al., 2020), such as novelty and complexity for story generation (Chhun et al., 2022) and logical consistency for dialogue generation (Pang et al., 2020).

However, establishing a comprehensive, efficient, and reliable evaluation system for SG is non-trivial, which raises three main concerns: (1) What criteria should be adopted to evaluate the SG task in a comprehensive and non-redundant fashion? (2) How to quantify each criterion into a metric thus enabling efficient and objective SG evaluation, given that the human evaluation of creative generation task is not only time-consuming but also subjective and blurred (Niculae and Danescu-Niculescu-Mizil, 2014; Celikyilmaz et al., 2020)? (3) Whether the proposed metrics are effective in providing useful scores to guide actual improvements in the real-world application of the SG model?

In this paper, we establish HAUSER, a **H**olistic and **AU**tomatic evaluation system for **S**imile g**E**ne**R**ation task, consisting of five criteria (Tab. 1): (1) The *relevance* between topic and vehicle, as the foundation of a simile is to compare the two via their shared properties (Paul, 1970). (2) The *logical consistency* between the literal sentence and generated simile, since the aim of SG task is to polish the original sentence without altering its semantics (Tversky, 1977). (3) The *sentiment consistency* between the literal sentence and generated simile, since similes generally transmit certain sentiment polarity (Qadir et al., 2015). (4,5) The *creativity* and *informativeness* of the simile, since novel similes or those with richer content can enhance the literary experience (Jones and Estes, 2006; Roncero and de Almeida, 2015; Addison, 2001). Overall, these five criteria can be categorized into three perspectives: *quality* (which considers relevance, logical, and sentiment consistency jointly), *creativity*, and *informativeness*. We further quantify each criterion into automatic metrics (Fig. 2) and prove their effectiveness through extensive experiments.

To the best of our knowledge, we are the first to systematically investigate the automatic evaluation of the SG task. To summarize, our contributions are mainly three-fold: (1) We establish a holistic and automatic evaluation system for the SG task, consisting of five criteria based on linguistic theories, facilitating both human and automatic evaluation of this task. (2) We design automatic metrics for each criterion, facilitating efficient and objective comparisons between SG models. (3) We conduct extensive experiments to verify that our metrics are significantly more correlated with human ratings than prior metrics.

2 Related Work

2.1 Simile Generation Task

There are two primary forms of the simile generation (SG) task: simile triplet completion and literal sentence polishing. For simile triplet completion, a model receives simile components, topic and property, and is required to generate the vehicle (Roncero and de Almeida, 2015; Zheng et al., 2019; Chen et al., 2022; He et al., 2022). For literal sentence polishing, a model receives a literal sentence and is expected to convert it into similes (Zhang

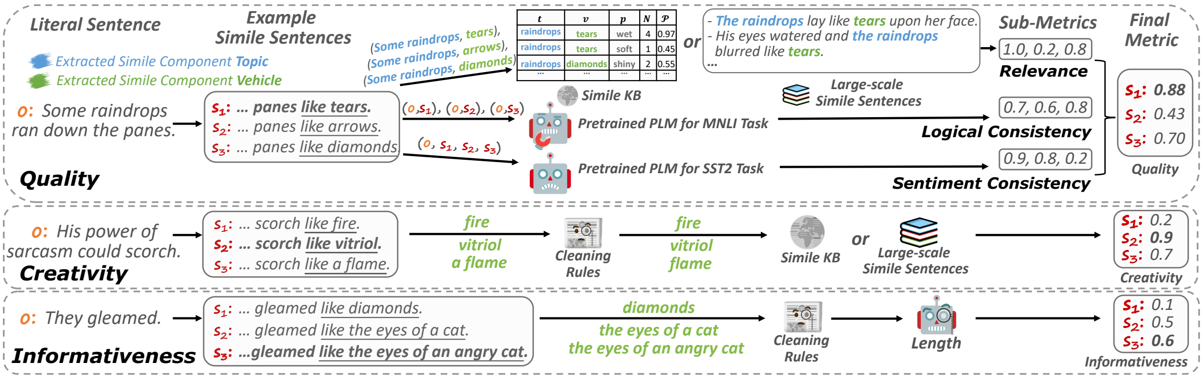
Literal Sentence
Example Simile Sentences
Extracted Simile Component Topic
Extracted Simile Component Vehicle
o: Some raindrops ran down the panes.
s1: ... panes like tears.
s2: ... panes like arrows.
s3: ... panes like diamonds.
(Some raindrops, tears),
(Some raindrops, arrows),
(Some raindrops, diamonds)
t | v | p | N | 𝒫
raindrops | tears | wet | 4 | 0.97
raindrops | tears | soft | 1 | 0.45
raindrops | diamonds | shiny | 2 | 0.55
... | ... | ... | ... | ...
or
- The raindrops lay like tears upon her face.
- His eyes watered and the raindrops blurred like tears.
...
Simile KB
Large-scale Simile Sentences
(o,s1), (o,s2), (o,s3)
(o, s1, s2, s3)
Pretrained PLM for MNLI Task
Pretrained PLM for SST2 Task
Sub-Metrics
1.0, 0.2, 0.8
Relevance
0.7, 0.6, 0.8
Logical Consistency
0.9, 0.8, 0.2
Sentiment Consistency
Final Metric
s1: 0.88
s2: 0.43
s3: 0.70
Quality
Quality
o: His power of sarcasm could scorch.
s1: ... scorch like fire.
s2: ... scorch like vitriol.
s3: ... scorch like a flame.
fire
vitriol
a flame
Cleaning Rules
fire
vitriol
flame
Simile KB
or
Large-scale Simile Sentences
s1: 0.2
s2: 0.9
s3: 0.7
Creativity
Creativity
o: They gleamed.
s1: ... gleamed like diamonds.
s2: ... gleamed like the eyes of a cat.
s3: ...gleamed like the eyes of an angry cat.
diamonds
the eyes of a cat
the eyes of an angry cat
Cleaning Rules
Length
s1: 0.1
s2: 0.5
s3: 0.6
Informativeness
Informativeness

Figure 2: The framework of our automatic metrics design. We design the automatic metric for each criterion in Tab. 1.

et al., 2021; Stowe et al., 2020; Chakrabarty et al., 2020; Lai and Nissim, 2022). We focus on the latter. However, prior works mainly adopt task-agnostic automatic metrics to evaluate the SG task, raising concern as to whether the claimed improvements are comprehensive and reliable.

2.2 Automatic Evaluation for NLG Systems

Existing automatic metrics for Natural Language Generation (NLG) evaluation can be categorized into task-agnostic and task-specific metrics. Task-agnostic metrics can be applied to various NLG tasks, which generally focus on the coherence of generations (Papineni et al., 2002; Zhang et al., 2019), including n-gram-based metrics (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2014) and embedding-based metrics (Zhang et al., 2019; Zhao et al., 2019). There are also many metrics for evaluating the diversity of generations (Li et al., 2016; Zhu et al., 2018; Tevet and Berant, 2021). Task-specific metrics are proposed to evaluate NLG systems on specific tasks (Tao et al., 2018; Dhingra et al., 2019; Ren et al., 2020). Specifically, various works systematically study the evaluation of the creative generation task (Pang et al., 2020; Tevet and Berant, 2021; Chhun et al., 2022). Different from these works, we revisit SG evaluation, propose holistic criteria based on linguistic theories, and design effective automatic metrics for it.

3 HAUSER for SG evaluation

We establish HAUSER, a holistic and automatic evaluation system for SG evaluation, containing five criteria from three perspectives, and further design automatic metrics for each criterion (Fig. 2).

3.1 Quality

We measure the overall quality of generated similes using three criteria: *relevance*, *logical consistency*, *sentiment consistency*. The key simile components - topic and vehicle - should be relevant, as the foundation of a simile is to compare the two via their shared properties (*relevance*) (Paul, 1970). In Tab. 1, comparing "*raindrops*" to "*tears*" is more coherent than to "*arrows*". Additionally, the generated simile should remain logically consistent with the original sentence (*logical consistency*), as the SG task aims to polish the plain text without changing its semantics (Tversky, 1977). In Tab. 1, comparing "*Stefan*" to "*dancer*" better depicts his controlled and easy movement than to "*lightning*". Furthermore, as similes generally transmit certain sentiment polarity (Qadir et al., 2015), the generated simile should enhance the sentiment polarity of the original sentence (*sentiment consistency*). In Tab. 1, the vehicle "*thunderous wave*" enhances the positive polarity of the original sentence, while the vehicle "*earthquake*" brings a negative sentiment polarity in opposition to the original sentence.

3.1.1 Relevance

For the *relevance* score, if the components of one simile are relevant, they tend to co-occur in simile sentences (Xiao et al., 2016; He et al., 2022) and possess shared properties (Paul, 1970; Tversky, 1977). Hence, obtaining the relevance score requires large-scale simile sentences as references, as well as knowledge about the properties (adjectives) of each simile component. For a simile s, the relevance score is defined as follows:

$$r = \frac{1}{m_p} \sum_{(t,v) \in s} \sum_{e \in \Gamma(t,v)} P_e(t,v), \qquad (1)$$

where there are m_p topic-vehicle pairs extracted from simile s, each denoted as (t, v)[1]. $\Gamma(t, v)$ is the set of similes containing (t, v) as simile components, each denoted as e. $P_e(t, v)$ is the probability that the simile components (t, v) share properties in the context of the simile sentence e.

An effective way to obtain the frequency information $\Gamma(t, v)$ and property knowledge $P_e(t, v)$ is to utilize the large-scale probabilistic simile knowledge base MAPS-KB (He et al., 2022), which contains millions of simile triplets in the form of (*topic*, *property*, *vehicle*), along with frequency and two probabilistic metrics to model each triplet[2]. Specifically, the probabilistic metric *Plausibility* is calculated based on the confidence score of the simile instance (*topic*, *property*, *vehicle*, *simile sentence*) supporting the triplet, indicating the probability that the topic and vehicle share the property. The relevance score r can be calculated as follow:

$$r = \frac{1}{m_p} \sum_{(t,v)\in s} \sum_{(t,p,v)\in \mathcal{G}_{(t,v)}} n(t,p,v) \cdot \mathcal{P}(t,p,v), \tag{2}$$

where $\mathcal{G}_{(t,v)}$ is the set of triplets (t, p ,v) containing the (t, v) pair in MAPS-KB, with p referring to the property. n and $\mathcal{P}$ are the metrics provided by MAPS-KB, where n and $\mathcal{P}$ denote the frequency and the plausibility of the triplet respectively.

It is noticed that the metric is not coupled with MAPS-KB, as the frequency information can be obtained by referencing a large set of simile sentences and the property knowledge can be contained via other knowledge bases. More methods are beyond the scope of this paper. However, we additionally provide a method to approximate the relevance score. If we assume the probability that the simile components (t, v) share properties in each sentence is 1, the relevance score can be approximated as:

$$r \approx \frac{1}{m_p} \sum_{(t,v)\in s} n(t,v), \tag{3}$$

where $n(t, v)$ denotes the number of samples that contain the simile components (t, v) in large-scale simile sentences. We discuss the effects of the referenced dataset size in Sec. 4.2.1.

3.1.2 Logical Consistency

The literal sentence and the generated simile that are logically inconsistent generally exhibit contradictory logic. Hence, for a generated simile, we input the <literal text(l), simile(s)> sentence pair into existing pre-trained Multi-Genre Natural Language Inference (MNLI) model[3], which determines the relation between them is *entailment*, *neutral*, or *contradiction*. The logical consistency score c_l of this simile is defined as follows (Pang et al., 2020):

$$c_l = 1 - P(h_{<l,s>} = c), \tag{4}$$

where $P(h_{<l,s>} = c)$ represents the probability that the model predicts the relation of the sentence pair $< l, s >$ to be *contradiction* (denoted as c).

3.1.3 Sentiment Consistency

Better similes tend to enhance the sentiment polarity of the original sentence (Qadir et al., 2015). Hence, we first apply the model fine-tuned on the GLUE SST-2 dataset[4] to classify each simile as being either *positive* or *negative*. Then, the sentiment consistency score c_s is defined as follows:

$$c_s = P(h_s = a) - P(h_l = a), \tag{5}$$

where a is the sentiment polarity of the literal sentence (*positive* or *negative*) predicted by the model. $P(h_s = a)$ and $P(h_l = a)$ denote the probabilities that the model predicts the sentiment polarity of the simile s and the literal sentence l to be a, respectively.

It is noticed that different <topic, vehicle> pairs within a sentence may have distinct sentiment polarities, such as <*She*, *scared rabbit*> and <*I*, *bird*> in the simile "*If she escapes like a scared rabbit, I will fly like a bird to catch her.*". Directly inputting text containing multiple topic-vehicle pairs into the sentiment classification model will result in inferior performance. Therefore, for each simile, only the text from the beginning up to the first *vehicle* is input into the model (i.e. "*If she escapes like a scared rabbit*" in the given example), and for each literal sentence, the text from the beginning up to the first *event* (i.e. "*If she escapes*" in the given example) is input into the model.

[1] All the simile components in our work are extracted and cleaned using rules from (He et al., 2022) which determines the optimal semantics a component should carry, e.g., "a kid in a candy store" instead of just "a kid".

[2] More details of MAPS-KB is provided in Appx. D

[3] We use the checkpoint of the model (roberta-base_mnli_bc) that achieves the SOTA performance on the GLUE (Wang et al., 2018) MNLI dataset at the time of submission, according to https://paperswithcode.com/sota/text-classification-on-glue-mnli.

[4] We apply the checkpoint of the model (distilbert-base-uncased-finetuned-sst-2-english) with the most download times on the GLUE SST-2 dataset at the time of submission, according to https://huggingface.co/models.

3.1.4 Combination

Since the aim of the SG task is to polish the plain text, the quality of similes generated from different texts can not be compared. Therefore, the normalized score among the simile candidates for each original text is utilized. Suppose there are m simile candidates $\mathcal{S} = \{s_1, s_2, ..., s_m\}$ for the literal text l, the original relevance scores of $\mathcal{R}$ is $\mathcal{R} = \{r_1, r_2, ..., r_m\}$ respectively. The normalized relevance score r'_i of s_i is formulated as follows:

$$r'_i = \frac{r_i - min(\mathcal{R})}{max(\mathcal{R}) - min(\mathcal{R})}, \tag{6}$$

which ranges from 0 to 1. Then, the normalized logical and sentiment consistency score c'_{li}, c'_{si} for each simile s_i are obtained in the same manner[5].

Finally, the *quality* for simile s_i is defined as the weighted combination of three parts as follows:

$$\mathcal{Q}_i = \alpha \cdot r'_i + \beta \cdot c'_{li} + \gamma \cdot c'_{si}, \tag{7}$$

where α, β, and γ are hyperparameters.

3.2 Creativity

Creative similes can provide a better literary experience (Jones and Estes, 2006). In Tab. 1, comparing "*sarcasm*" to "*vitriol*" is less common than to "*fire*", yet it better conveys the intensity of a person's sarcasm. Hence, we design *creativity* score.

Previous studies mainly evaluate the creativity of text generation tasks via human evaluation (Sai et al., 2022), since measuring the creativity of open-ended text is a relatively difficult task (Celikyilmaz et al., 2020). Although there have been many works evaluating the diversity of open-ended text generation (Li et al., 2016; Zhu et al., 2018; Tevet and Berant, 2021), these metrics are not suitable for measuring the creativity of the text. Because the *diversity* metrics take a set of generated text as input and output one score, while a *creativity* metric is required to measure each text individually and output a set of corresponding scores.

Different from other open-ended generation tasks, the components of the generated similes enable us to evaluate creativity automatically. According to linguists, the creativity of a simile is determined by vehicles (Pierce and Chiappe, 2008; Roncero and de Almeida, 2015). Intuitively, the generated simile may be less creative if its extracted topic-vehicle pair co-occurs frequently, or if many topics are compared to its vehicle in the corpus. Therefore, we adopt large-scale corpora as references when designing our creativity metric. The creativity score of s is calculated as follows:

$$\mathcal{C}_i = -log(\frac{1}{m_v}\sum_{v \in s} N_v + 1), \tag{8}$$

where there are m_v vehicles extracted from the simile s, each denoted as v. N_v denotes the frequency of the vehicles appearing in the similes in the corpora. The log transformation aims to reduce the influence of extreme values.

An effective way to obtain the adequate frequency information N_v is to utilize the million-scale simile knowledge base MAPS-KB, where the N_v can be defined as follows:

$$N_v = \sum_{(t,p,v) \in \mathcal{G}_v} n(t, p, v), \tag{9}$$

$\mathcal{G}_v$ is the set of triplets containing the vehicle v in MAPS-KB, n denotes the frequency of the triplet.

It is noticed that the metric is not coupled with MAPS-KB, as N_v can also be obtained by counting the samples containing the vehicle v in large-scale simile sentences. The method of obtaining the simile sentences is beyond the scope of this paper. Nevertheless, we discuss the effects of the referenced dataset size in Sec. 4.2.2.

3.3 Informativeness

The vehicle with richer content can create a more impact and vivid impression(Addison, 2001). In the example from Tab. 1, the addition of the word "*angry*" makes the similes more expressive. Therefore, we design the metric *informativeness* to measure the content richness of the vehicles.

Intuitively, the more words a vehicle contains, the richer its content will be. Hence, for a given simile s, we adopt the average length of the extracted vehicles to be the *informativeness* score[6] (Chakrabarty et al., 2020; Zhang et al., 2021), defined as $\mathcal{I}_i = \frac{1}{m_v}\sum_{v \in s} \text{len}(v)$, where there are m_v vehicles extracted from simile s.

4 HAUSER Analysis

In this section, we conduct experiments to verify the effectiveness of our automatic metrics.

[5] If all the relevance scores r_i in $\mathcal{R}$ are the same, the normalized relevance scores r'_i in $\mathcal{R}'$ are set to 0.5 uniformly.

[6] Different from the *quality* metric, we do not use a normalized score for *creativity* and *informativeness*, since they mainly depend on the generated vehicles, rather than the original text.

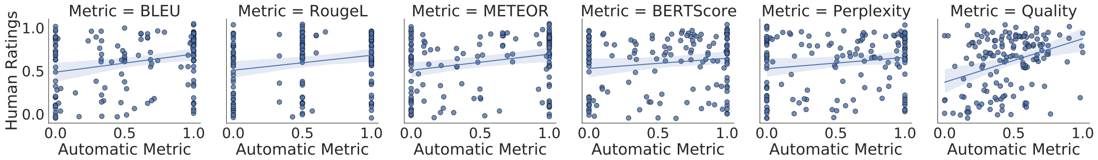

Figure 3: Correlation between automatic metrics and human ratings when evaluating quality. Here, BLEU2, Rouge2, and BERTScore$_{\text{large}}$ are presented since they perform the best in their respective category. To avoid overlapping points, random jitters sampled from $\mathcal{N}(0, 0.05^2)$ were added to human ratings after fitting the regression.

4.1 Experiment Setup

4.1.1 Simile Generation

The existing datasets for the SG task are either Chinese (Zhang et al., 2021), limited to the simile triplet completion (Roncero and de Almeida, 2015; Chen et al., 2022), or having all vehicles located at end of the sentence (Chakrabarty et al., 2022; Lai and Nissim, 2022), which are not practical for English simile generation in a real-world application. To bridge the gap, we construct a large-scale English dataset for SG task based on simile sentences from (He et al., 2022), which contains 524k simile sentences labeled with topic and vehicle. The output decoder target is the simile sentence s and the input encoder source is s rewritten to drop the comparator "*like*" and the vehicle. For example, given s = "*The idea resounded like a thunderclap throughout the land.*", the encoder source would be "*The idea resounded throughout the land.*". In particular, we remove the simile sentences whose event is a linking verb (e.g. *be*, *seem*, *turn*) as they would be meaningless after the vehicle is removed. The final train, validation and test sets contain 139k, 2.5k, and 2.5k sentence pairs, respectively.

Based on our constructed dataset, we finetune a pre-trained sequence-to-sequence model, BART (Lewis et al., 2020), for the SG task, which has been demonstrated to be an effective framework for various figurative language generation (Zhang and Wan, 2021; Chakrabarty et al., 2022; He et al., 2022; Lai and Nissim, 2022). The experiments are run on RTX3090 GPU and the implementation of BART is based on the HuggingFace Transformers[7]. The experiments are run with a batch size of 16, a max sequence length of 128, and a learning rate of 4e-5 for 10 epochs.

4.1.2 Evaluation Dataset Construction

Firstly, we randomly sample 50 literal sentences from the test set and adopt the trained SG model to generate five candidates for each one. Then, for each perspective, three raters are asked to rate each simile from 1 to 5, where 1 denotes the worst and 5 denotes the best[8]. Since evaluating the quality of generated similes is subjective and blurred (Niculae and Danescu-Niculescu-Mizil, 2014), we remove the simile-literal sentence pairs if (1) raters argue that the pairs lack context and are difficult to rate (e.g. "*Nobody can shoot.*") or (2) some raters rate them as low quality (quality score of 1-2), while others rate them as high quality (scores of 4-5) (Niculae and Danescu-Niculescu-Mizil, 2014). Moreover, we measure the inter-rater agreement by holding out the ratings of one rater at a time, calculating the correlations with the average of the other rater's ratings, and finally calculating the average or maximum of all the held-out correlations (denoted as "*Mean*" and "*Max*", respectively). The inter-rater agreement before and after applying the filtering strategies is shown in Tab. 2. Overall, the final inter-rater agreement ensures the reliability of our evaluation of automatic metrics and the filtering strategies improve the inter-rater agreement generally. We finally get 150 simile candidates generated from 44 literal sentences.

Setting	Metric	Pearson		Spearman	
		Mean	Max	Mean	Max
Before	Quality	0.573	**0.626**	**0.542**	**0.595**
	Creativity	**0.537**	0.671	0.550	0.678
	Informativeness	0.833	0.857	0.799	0.816
After	Quality	0.812	0.833	0.735	0.759
	Creativity	0.551	0.643	0.568	0.650
	Informativeness	0.848	0.893	0.817	0.841

Table 2: The inter-rater agreement before and after applying the removal strategies. Bold numbers are the worst results, indicating that the raters are quite divided on this metric.

4.2 Results

4.2.1 Quality

We compare our *quality* metric with the following automatic metrics[9]: (1) **BLEU** (Papineni et al.,

[7] https://github.com/huggingface/transformers/

[8] The details about human ratings, including the instructions provided to raters and examples of human ratings are provided in Appx. A.

[9] These metrics are normalized among simile candidates for a literal sentence, since the quality score between the similes

Metrics	Pearson	Spearman
N-gram-level Metrics		
BLEU1	0.229	0.218
BLEU2	0.255	0.208
BLEU3	0.193	0.172
BLEU4	*0.159*	*0.140*
Rouge1	0.185	0.176
Rouge2	0.210	0.190
RougeL	0.173	*0.152*
METEOR	0.234	0.233
Sentence-level Metrics		
BERTS$_{\text{base}}$	*0.107*	*0.075*
BERTS$_{\text{large}}$	*0.143*	*0.120*
Perplexity	*0.157*	*0.120*
HAUSER		
Quality	**0.320**(+6.5%)	**0.292**(+5.9%)
−relevance	0.206	0.194
−consistency$_l$	0.259	0.217
−consistency$_s$	0.307	0.265

Table 3: Correlation between automatic metrics and human ratings when evaluating quality. All measures with p-value > 0.05 are italicized. Bold numbers are the best results. The second best results are marked by "___". "−" denotes the removal of the sub-metric.

2002) calculates the precision of n-gram matches, (2) **RougeL** (Lin, 2004) is a recall-oriented metric, (3) **METEOR** (Denkowski and Lavie, 2014) proposes a set of linguistic rules to compare the hypothesis with the reference, (4) **BERTScore** (Zhang et al., 2019) calculates the cosine similarity between the BERT embeddings, (5) **Perplexity** (Pang et al., 2020) measures the proximity of a language model, the inverse of which is utilized.

Correlations with Human Ratings. Tab. 3 shows the correlation coefficients between automatic metrics and human ratings. Firstly, our metrics are significantly more correlated with human ratings than prior automatic metrics. Moreover, all the sentence-level metrics, which consider the semantics of the entire sentence, perform worse than almost all the n-gram-level metrics, which compare the n-grams between the hypothesis and the reference, which reveals that simile components need to be specifically considered during SG evaluation.

According to the visualized correlation result in Fig. 3, datapoints from prior automatic metrics tend to scatter at 0 or 1, while the datapoints from our metric are distributed closer to the fitter line, proving that our metric can better measure the quality.

Recommendation Task. We compare the rankings given by automatic metrics with human rankings[10]. We adopt the following metrics: Hit Ratio at rank K (**HR@K**(K=1,3)), Normalized Discounted Cumulative Gain at rank K (**NDCG@K**(K=1,3))[11], and Mean Reciprocal Rank (**MRR**). From Tab. 4, our metric achieves significant improvement compared to other metrics, indicating that our metric can yield more accurate rankings for quality. Also, the n-gram-level metrics generally outperform sentence-level metrics, which is consistent with the result in Tab. 3.

Metrics	HR@1	HR@3	nDCG@1	nDCG@3	MRR
N-gram-level Metrics					
BLEU1	0.429	**0.857**	0.893	**0.945**	0.662
BLEU2	0.314	0.838	0.892	0.936	0.600
BLEU3	0.286	0.838	0.859	0.924	0.648
BLEU4	0.286	0.838	0.882	0.929	0.581
Rouge1	0.400	0.848	0.907	0.941	0.655
Rouge2	0.400	0.848	0.905	0.937	0.650
RougeL	0.429	0.848	0.901	0.937	0.670
METEOR	0.286	**0.857**	0.884	0.936	0.589
Sentence-level Metrics					
BERTS$_{\text{base}}$	0.314	0.829	0.870	0.934	0.585
BERTS$_{\text{large}}$	0.257	0.838	0.895	0.939	0.570
Perplexity	0.257	0.810	0.898	0.940	0.549
HAUSER					
Quality	**0.457**	0.848	**0.915**	0.937	**0.688**

Table 4: Comparison of automatic metrics ranking and human ranking when evaluating quality.

Ablation Study. To investigate the importance of different sub-metrics in *quality* metric, we compare the correlation between *quality* metric and human ratings after removing each sub-metric individually. From Tab. 3, the removal of any sub-metric leads to a decline in performance, which proves the effectiveness of each sub-metric. Among three components, the removal of the *relevance* results in the largest performance drop, which reveals that *relevance* is the most important sub-metric.

The Effects of Hyperparameters. Since different sub-metrics have varying levels of importance, we study the correlation results when gradually increasing the weight of *relevance* component and decreasing the weight of *sentiment consistency* component (as in Tab. 5). From Fig. 4 (left), increasing the weight of the *relevance* component consistently results in improved performance, peaking at the combination [7]($\alpha, \beta, \gamma = 3/6, 2/6, 1/6$), before eventually causing a decline in performance. This reveals that although *relevance* is the most important sub-metric, too much weight on it can be detrimental.

The Effects of Referenced Dataset Size. We sample different numbers of simile sentences from (He et al., 2022) as references for relevance

generated from different literal sentences can not be compared. Please refer to Appx. C for the implementation of them.

[10] We remove the literal sentences with fewer than three valid simile candidates in this task, as they are too simple to rank. We finally get 134 sentences from 35 literal sentences.

[11] The formulated NDCG@K in our setting is provided in Appx. B, with the optimal ranking being human rankings.

Combination	α, β, γ
[1]	1/12, 1/12, 5/6
[2]	1/6, 1/6, 4/6
[3]	1/6, 2/6, 3/6
[4]	1/6, 3/6, 2/6
[5]	2/6, 2/6, 2/6
[6]	2/6, 3/6, 1/6
[7]	3/6, 2/6, 1/6
[8]	4/6, 1/6, 1/6
[9]	5/6, 1/12, 1/12

Table 5: The setting of each hyperparameters combination for the *quality* metric. The result is shown in Fig. 4 (left).

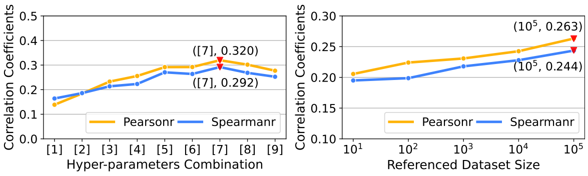

Figure 4: Correlation between *quality* metric and human ratings with different hyperparameters (left) and different referenced corpus size (right).

score and study the correlation between the *quality* metric and human ratings[12]. From Fig. 4 (right)[13], correlations grow linearly with exponential growth in referenced dataset size, indicating that using datasets larger than 100k will improve the correlation coefficients. Moreover, the performance at the peak surpasses the prior automatic metrics, proving the effectiveness of our approximation method.

4.2.2 Creativity

We compare our *creativity* metric with the following automatic metrics: (1) **Perplexity** which is often utilized to measure diversity as well (Tevet and Berant, 2021), (2) **Self-BLEU** (Zhu et al., 2018) calculates the BLEU score of each generation against all other generations as references, (3) **Distinct n-grams(Dist)** (Tevet and Berant, 2021), which is the fraction of distinct n-grams from all possible n-grams across all generations.

Correlations with Human Ratings. From Tab. 6, our metric *creativity* is significantly more correlated with human evaluation scores compared with prior diversity metrics. According to the visualized correlation result in Fig. 5, the prior diversity metrics have either wide confidence intervals (Perplexity, Dist) or scattered datapoints (self-BLEU), whereas our creativity metrics exhibit stronger linear correlation and narrower confidence intervals (Creativty w/ Log), implying higher reliability.

Recommendation Task. We compare the rankings given by automatic metrics with human rankings. According to Tab. 7, our creativity metric outperforms prior automatic metrics, which proves our metric can better measure the creativity of simile candidates given a literal sentence, which is consistent with the results in Tab. 6.

Metrics	Pearson	Spearman
Prior Diversity Metrics		
Perplexity	*0.088*	*0.041*
Self-BLEU3	*0.118*	*0.076*
Self-BLEU4	0.196	*0.175*
Self-BLEU5	0.128	*0.077*
Dist1	0.278	0.311
Dist2	0.319	0.369
Dist3	0.299	0.379
HAUSER		
Creativty	**0.592**(+27.3%)	**0.645**(+26.6%)
−log	0.394	0.571

Table 6: Correlation between metrics and human ratings when evaluating creativity. All measures with p-value > 0.05 are italicized. "−log" denotes the removal of log transformation.

Metrics	HR@1	HR@3	nDCG@1	nDCG@3	MRR
Prior Diversity Metrics					
Perplexity	0.314	0.800	0.800	0.903	0.566
Self-BLEU3	0.257	0.771	0.765	0.892	0.520
Self-BLEU4	0.257	0.762	0.756	0.889	0.518
Self-BLEU5	0.229	0.762	0.751	0.882	0.504
Dist1	0.486	0.800	0.862	0.927	0.671
Dist2	0.571	0.810	0.893	0.939	0.737
Dist3	0.543	0.838	0.877	0.938	0.725
HAUSER					
Creativty	**0.629**	**0.914**	**0.944**	**0.976**	**0.784**

Table 7: Comparison of automatic metrics ranking and human ranking when evaluating creativity.

Ablation Study. According to Tab. 6, removing the log transformation leads to significant performance drops. According to the visualized correlation result in Fig. 5, the datapoints are distributed closer to the fitter line and exhibit narrower confidence intervals after applying the log transformation, which further proves that log transformation is essential for our creativity metric.

The Effects of Referenced Dataset Size. According to Fig. 6 (left), the correlation coefficients increase continuously and eventually converge as the number of referenced sentences increases. Moreover, the performance after convergence is comparable to that given by the *creativity* metric based on the simile KB. The trend reveals that our metric referencing 10k similes can achieve a promising correlation with human ratings.

4.2.3 Informativeness

The Pearson and Spearman correlation coefficients between our *informativeness* metric and human ratings are 0.798 and 0.882, respectively. According to Fig. 6 (right), the strong linear correlation between the metric and human ratings proves that our

[12] The results are averaged over three random seeds.

[13] The best hyper-parameter combination is applied.

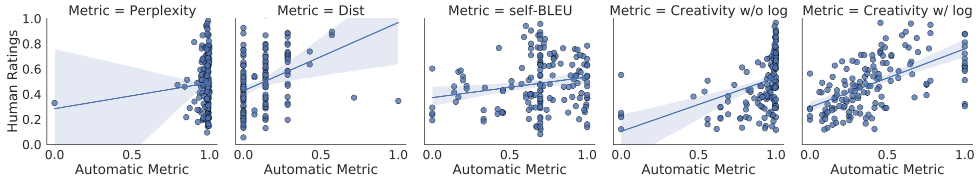

Figure 5: Correlation between automatic metrics and human ratings when evaluating creativity. Here, Self-BLEU4 and Dist2, which perform the best in their respective category in Tab. 6, are presented. "w/o log" and "w/ log" denotes whether the log transformation is applied or not.

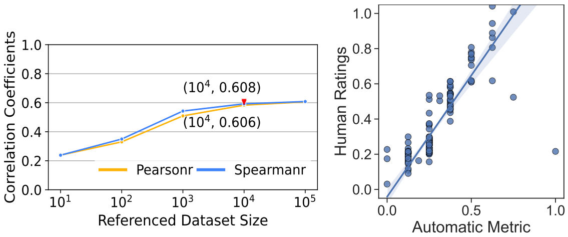

Figure 6: Correlation between *creativity* metric and human ratings with varying referenced corpus size (left), and correlation between *informativeness* metric and human ratings (right).

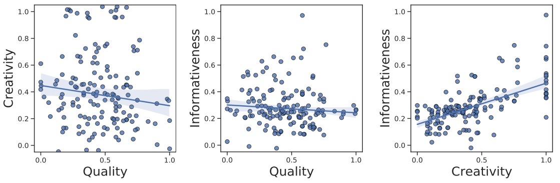

Figure 7: The pair-wise correlations between the metrics.

informativeness metric is simple yet quite effective.

4.2.4 Relation between Metrics

We present pair-wise correlations between the three automatic metrics in Tab. 8 and also visualize them in Fig. 7. Among the three metrics, creativity correlates with informativeness moderately, mainly because shorter vehicles tend to be less creative than longer ones. The correlations of all other pairwise metrics are relatively weak. Thus, it is evident that the three metrics are independent of each other and it is necessary to measure each one of them to obtain a holistic view of SG evaluation.

Metrics	Pearson	Spearman
Quality & Creativity	*-0.116*	*-0.130*
Quality & Informativeness	*-0.040*	*-0.118*
Creativity & Informativeness	0.652	0.635

Table 8: The pair-wise correlations between our automatic metrics. All measures with p-value $>$ 0.05 are italicized.

5 HAUSER Application

We perform a case study to prove that our designed automatic metrics are effective for various methods. Here, we apply our metrics to a retrieval method (Zhang et al., 2021) (denoted as **BM25**), which utilizes the 20 context words around the insertion position given by groundtruth to retrieve the 5 most similar samples based on the BM25 ranking score from the training set, and adopts the vehicles from these samples to be those of simile candidates. This method ensures the diversity of generated similes. The method introduced in Sec. 4.1 is denoted as **Ours**. Given the candidates generated by each method, we rerank them using a weighted combination of quality, creativity, and informativeness rankings obtained by HAUSER, with a ratio of 2:2:1.

From Tab. 11 in Appendix, the candidates generated by various methods can be more correlated with human rankings after being ranked by our metrics, thus proving the generality of our metrics. It is noticed that the insertion position for **BM25** is provided by the groundtruth, while the insertion position for **Ours** is predicted by the model, thus proving the effectiveness of our generation method.

6 Conclusion

In this work, we systematically investigate the evaluation of the Simile Generation (SG) task. We establish a holistic and automatic evaluation system for the SG task, containing five criteria from three perspectives, and propose holistic automatic metrics for each criterion. Extensive experiments verify the effectiveness of our metrics.

Acknowledgements

This research is funded by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0103), National Natural Science Foundation of China (No. 62102095).

Limitations

We analyze the limitations of our work as follows. Firstly, although applying a million-scale simile knowledge base or large-scale simile sentences as reference makes our designed metric significantly more correlated with humans than prior reference-based metrics (e.g. BLEU, Rouge, BERTScore), our metrics are still reference-based and rely on the quality and scale of referenced data. We have discussed the effect of referenced dataset size in our paper and will design reference-free metrics to further complement our metrics in future work. Additionally, since our metrics utilize a million-scale simile knowledge base or large-scale simile sentences as references, the efficiency of our method is slightly lower than the automatic metrics based on a few references. Nevertheless, this limitation does not prevent our metrics from performing systematic and scalable comparisons between SG models.

Ethical Considerations

We provide details of our work to address potential ethical considerations. In our work, we propose holistic and automatic metrics for SG evaluation and construct an evaluation dataset to verify their effectiveness (Sec. 4.1). All the data sources used in our evaluation dataset are publicly available. The details about human ratings, such as the instructions provided to raters, are provided in Appx. A. In our case study (Sec. 5), the human rankings are discussed by three raters. We protect the privacy rights of raters. All raters have been paid above the local minimum wage and consented to use the evaluation dataset for research purposes covered in our paper. Our work does not raise any ethical considerations regarding potential risks and does not involve the research of human subjects.

References

Catherine Addison. 2001. "so stretched out huge in length": Reading the extended simile. *Style*, 35(3):498–516.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

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.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5875–5887.

Cyril Chhun, Pierre Colombo, Fabian M Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *29th International Conference on Computational Linguistics (COLING 2022)*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895.

Patrick Hanks. 2013. *Lexical analysis: Norms and exploitations*. Mit Press.

Qianyu He, Xintao Wang, Jiaqing Liang, and Yanghua Xiao. 2022. Maps-kb: A million-scale probabilistic simile knowledge base. *arXiv preprint arXiv:2212.05254*.

Lara L Jones and Zachary Estes. 2006. Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55(1):18–32.

Huiyuan Lai and Malvina Nissim. 2022. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5939–5954.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629.

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 of the Association for Computational Linguistics*, pages 311–318.

Anthony M Paul. 1970. Figurative language. *Philosophy & Rhetoric*, pages 225–248.

Russell S Pierce and Dan L Chiappe. 2008. The roles of aptness, conventionality, and working memory in the production of metaphors and similes. *Metaphor and symbol*, 24(1):1–19.

Ashequl Qadir, Ellen Riloff, and Marilyn Walker. 2015. Learning to recognize affective polarity in similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 190–200.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *arXiv preprint arXiv:2009.10297*.

Carlos Roncero and Roberto G de Almeida. 2015. Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior research methods*, 47(3):800–812.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

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

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Roi Tartakovsky and Yeshayahu Shen. 2018. 'simple as a fire': Making sense of the non-standard poetic simile. *Journal of Literary Semantics*, 47(2):103–119.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kat Agres, Hannu Toivonen, et al. 2016. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In *Proceedings of the Seventh International Conference on Computational Creativity*. Sony CSL Paris.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yunxiang Zhang and Xiaojun Wan. 2021. Mover: Mask, over-generate and rank for hyperbole generation. *arXiv preprint arXiv:2109.07726*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2019. "love is as complex as math": Metaphor generation system for social chatbot. In *Workshop on Chinese Lexical Semantics*, pages 337–347. Springer.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A Human Ratings

The instructions given to raters are detailed as follows:

1. All raters are provided with the necessary background information on similes and the simile generation task, including the definition of similes, the main simile components, and the motivation of our proposed criteria.

2. To ensure the quality of ratings, all the raters label a small set of 20 samples to reach an agreement on the labeling criteria for each metric before the formal labeling.

3. For each perspective (i.e. quality, creativity, informativeness), three raters are asked to rate each simile from 1 to 5, where 1 denotes the worst and 5 denotes the best. **The examples of our human ratings** are provided in Tab. 10.

4. During the rating, raters are asked to specifically label the simile-literal sentence pairs which lack context and are thus difficult to rate (e.g. "*Nobody can shoot.*").

B NDCG Formulation

In our setting, the optimal rankings are human rankings. Hence, given m simile candidates $\mathcal{S} = \{s_1, s_2, ..., s_m\}$, the NDCG@k given by each automatic metric is defined as follows:

$$\text{NDCG}(k) = \frac{\text{DCG}(\mathcal{O}_{\text{hypo}}, k)}{\text{DCG}(\mathcal{O}_{\text{ref}}, k)} \tag{10}$$

$$\text{DCG}(\mathcal{O}, k) = \sum_{i=1}^{k} \frac{\mathcal{O}[\mathcal{I}(i)]}{log_2(1+i)} \tag{11}$$

where O_{ref} and O_{hypo} represent the score list given by humans and each automatic metric respectively, $\mathcal{O}[j]$ denote the score of s_j, $\mathcal{I}(i)$ denotes the index of the i-th largest score in $\mathcal{O}$.

C The Implementation of Prior Metrics

We report the packages used to implement prior automatic metrics in Tab. 9. For the metric denoted with an asterisk(*), we apply the corresponding package to implement the key parts, based on the definition from the cited papers. The formulation of NDCG in our setting is provided in Appx. B. The rest of the metrics are entirely implemented by us according to the cited papers.

Metric	**Packages**
BLEU, METEOR	NLTK
Rouge	rouge
BERTScore	bert_score
Self-BLEU*	NLTK
Distinct n-grams*	NLTK

Table 9: The packages used to implement the metrics.

D The Details of MAPS-KB

MAPS-KB (He et al., 2022) is a million-scale probabilistic simile knowledge, containing 4.3 million simile triplets from 70 GB corpora, along with frequency and two probabilistic metrics, *plausibility* and *typicality*, to model each triplet. The simile triplet is in the form of (topic, property, vehicle)(t, p, v).

In our paper, we specifically adopt the *frequency* and *plausibility* information from MAPS-KB to implement our relevance metric. With regard to *plausibility*, it evaluates the quality of simile triplets based on the confidence score of their supporting simile instances (simile sentence, topic, property, vehicle)(s_i, t, p, v). In each simile instance, the topic and vehicle are extracted from the simile sentence, while the property is generated via generative commonsense model COMET (Bosselut et al., 2019) and prompting the PLMs. MAPS-KB adopt the *noisy-or* model to measure the plausibility of the triplet (t, p, v), which is defined as follows:

$$\mathcal{P}(t,p,v) = 1 - \prod_{i=1}^{\eta}(1 - S(s_i, t, p, v)),$$

where $S(s_i, t, p, v) = P(p|s_i, t, v)$ is the confidence score of each simile instance during generation and η is the number of simile instances supporting the simile triplet (t, p, v).

#	Literal Sentence	Vehicles in the Generated Similes	Q	C	I
1	Some raindrops struck the roof, window and ran down its panes [insert].	like diamonds	2.3	3.3	2.0
		like tears	3.3	3.3	2.0
		like arrows	1.0	3.0	2.0
		like a stream	**4.0**	2.7	2.3
		like a stream of diamonds	**4.0**	**4.7**	**4.0**
2	As suddenly as she'd jumped up from the sofa, Jaklin collapsed [insert].	like a rag doll	3.0	3.3	2.7
		like a deflated balloon	**4.7**	4.0	**3.3**
		like a pricked bladder	3.0	**4.7**	**3.3**
		like a pricked balloon	4.3	4.3	**3.3**
	As suddenly as [insert] she'd jumped up from the sofa, Jaklin collapsed.	like a flash	3.3	2.0	2.3
3	In the other direction the Empire State Building loomed [insert].	like a dark shadow	4.0	2.3	3.3
		like a huge black monster	**4.7**	3.3	4.0
		like a giant black monster	**4.7**	3.7	4.0
		like a huge black shadow	4.3	3.0	4.0
		like a huge black monster of destruction	**4.7**	**4.3**	**5.0**
4	His hormones boiled and steamed [insert] and yet he did not reach for the succulent young flesh there beside him.	like a boiling caldron	3.0	4.3	3.0
		like a volcano	4.3	2.7	2.0
		like a boiling cauldron	4.0	**4.7**	3.0
		like a cauldron of boiling water	**4.7**	**4.7**	**4.7**
		like a cauldron of boiling water*	**4.7**	**4.7**	**4.7**
5	The coil whistled through the air. It fell right over the mate's shoulder. He clutched at it as the fore, topmast crosstrees, with the full force of the surge, struck him from behind, and he sank [insert].	like a stone	**4.7**	1.7	2.0
		like a log	3.0	1.7	2.0
		like lead	4.3	2.3	1.7
		like an empty sack	1.3	**3.7**	**3.0**
		like an empty barrel	1.3	**3.7**	**3.0**

Table 10: Examples of human ratings for each perspective (Q, C, I denoting *Quality*, *Creativity*, *Informativeness*, respectively). The indicators "[insert]" denotes the insertion positions of vehicles within the generated similes given by models, which do not exist in the literal sentences. Bold numbers indicate the highest ranking among the simile candidates generated from a literal sentence. An asterisk (*) indicates that the generated simile introduces noise to the context word through additions, deletions, or changes within two words.

#	Method	Literal Sentence	Vehicles in the Generated Similes		
			Original Rank	HAUSER Rank	Human Rank
1	BM25	Stefan moved [Insert], every movement easy and precisely controlled.	like water	**like a ballerina**	**like a ballerina**
			like hell	like a predator	like a predator
			like a ballerina	like a drum	like a drum
			like a drum	like water	like water
			like a predator	like hell	like hell
	Ours	Stefan moved [Insert], every movement easy and precisely controlled.	like a cat	**like a dancer**	**like a dancer**
			like a dancer	like an automaton	like an automaton
			like lightning	like lightning	like a cat
			like an automaton	like a cat	like a cat*
			like a cat*	like a cat*	like lightning
2	BM25	But his next line called for him to howl [Insert].	like a fiend	**like a wounded buffalo**	**like a wounded buffalo**
			like a drug	like a fiend	like a fiend
			like a chicken	like a trail	like a chicken
			like a trail	like a chicken	like a trail
			like a wounded buffalo	like a drug	like a drug
	Ours	But his next line called for him to howl [Insert].	like a wolf	**like a wounded animal**	**like a wounded animal**
			like a dog	like a dog	like a coyote
			like a coyote	like a coyote	like a coyote*
			like a wounded animal.	like a coyote*	like a wolf
			like a coyote*	like a wolf	like a dog
3	BM25	She wondered absently if those soldiers would survive the coming war, if they would earn glory or run [Insert].	like a rabbit	**like a very coward**	**like a very coward**
			like bees about their friend	like bees about their friend	like a pack of wolves
			like wildfire	like a pack of wolves	like a rabbit
			like a very coward	like wildfire	like wildfire
			like a pack of wolves	like a rabbit	like bees about their friend
	Ours	She wondered absently if those soldiers would survive the coming war, if they would earn glory or run [Insert].	like cowards	like scared rabbits	**like frightened sheep**
			like scared rabbits	like hares	like scared rabbits
			like frightened sheep	**like frightened sheep**	like cowards
			like hares	like cowards	like cowards*
			like cowards*	like cowards*	like hares
4	BM25	As suddenly as she'd jumped up from the sofa, Jaklin collapsed [Insert].	**like a pricked bubble**	like a grocery bag	**like a pricked bubble**
			like a boy	**like a pricked bubble**	like a ragdoll
			like a panther	like a ragdoll	like a grocery bag
			like a ragdoll	like a boy	like a panther
			like a grocery bag	like a panther	like a boy
	Ours	As suddenly as she'd jumped up from the sofa, Jaklin collapsed [Insert].	like a rag doll	like a sack of potatoes*	**like a deflated balloon**
			like a deflated balloon	**like a deflated balloon**	like a pricked balloon
			like a sack of potatoes	like a pricked balloon	like a rag doll
			like a pricked balloon	like a sack of potatoes	like a sack of potatoes
			like a sack of potatoes*	like a rag doll	like a sack of potatoes*
5	BM25	They gleamed [Insert].	like golden fire	**like the eyes of great cats**	**like the eyes of great cats**
			like silver	like golden fire	like golden fire
			like the eyes of great cats	like silver	like sparks of fire
			like a second skin	like sparks of fire	like silver
			like sparks of fire	like a second skin	like a second skin
	Ours	They gleamed [Insert].	like polished ebony	like the eyes of a cat	**like the eyes of a wild beast**
			like polished steel	like the eyes of a wild animal	like the eyes of a wild animal
			like the eyes of a cat	**like the eyes of a wild beast**	like the eyes of a cat
			like the eyes of a wild animal	like polished ebony	like polished ebony
			like the eyes of a wild beast	like polished steel	like polished steel
6	BM25	The idea resounded [Insert] throughout the land.	like a gong	like the beating of a bass drum	**like the crack of a whip in the silence of the hall**
			like an agonized lament	**like the crack of a whip in the silence of the hall**	like prolonged theater applause
			like the beating of a bass drum	like prolonged theater applause	like a gong
			like the crack of a whip in the silence of the hall	like an agonized lament	like the beating of a bass drum
			like prolonged theater applause	like a gong	like an agonized lament
	Ours	The idea resounded [Insert] throughout the land.	like thunder	like a trumpet	**like a thunderclap**
			like a thunderclap	like a thunderclap*	like a thunderclap*
			like an earthquake	**like a thunderclap**	like thunder
			like a trumpet	like an earthquake	like a trumpet
			like a thunderclap*	like thunder	like an earthquake

Table 11: The examples of simile candidates reranked via HAUSER, which are generated by various methods. The indicators "[insert]" denotes the insertion positions of vehicles within the generated similes given by models, which do not exist in the literal sentences. An asterisk (*) indicates that the generated simile introduces noise to the context word through additions, deletions, or changes within two words. A darker shade of green indicates a higher rank bestowed by humans.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
the "Limitations" Section.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
the "Abstract" Section and the Section 1.

☑ A4. Have you used AI writing assistants when working on this paper?
I adopt the free api from text-davinci-003 to polish the language of the whole paper.

B ☑ Did you use or create scientific artifacts?

Section 4,5

☑ B1. Did you cite the creators of artifacts you used?
Section 4,5 and Appendix A, C, D

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4, 5, the "Ethical Consideration" Section, and Appendix C, D

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3, 4, 5, and Appendix D

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
the "Ethical Consideration" Section

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4, Appendix A

C ☑ Did you run computational experiments?

Section 4, 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4, 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, 4, Appendix C, D

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
the "Ethical Consideration" Section.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
the "Ethical Consideration" Section.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.