I run as fast as a rabbit, can you? A Multilingual Simile Dialogue Dataset

Longxuan Ma[1] and **Weinan Zhang**[1*] and **Shuhan Zhou**[1,2]
and **Churui Sun**[3] and **Changxin Ke**[3] and **Ting Liu**[1]
[1] Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology
[1] lxma,wnzhang,shzhou,tliu@ir.hit.edu.cn
[2] School of Information Science, Beijing Language and Culture University
[3] School of Computer Science, Harbin Institute of Technology
[3] sunchurui@hit.edu.cn, cxke@stu.hit.edu.cn

Abstract

A simile is a figure of speech that compares two different things (called the tenor and the vehicle) via shared properties. The tenor and the vehicle are usually connected with comparator words such as "like" or "as". The simile phenomena are unique and complex in a real-life dialogue scene where the tenor and the vehicle can be verbal phrases or sentences, mentioned by different speakers, exist in different sentences, or occur in reversed order. However, the current simile research usually focuses on similes in a triplet tuple (tenor, property, vehicle) or a single sentence where the tenor and vehicle are usually entities or noun phrases, which could not reflect complex simile phenomena in real scenarios. In this paper, we propose a novel and high-quality multilingual simile dialogue (MSD) dataset to facilitate the study of complex simile phenomena. The MSD is the largest manually annotated simile data (∼20K) and it contains both English and Chinese data. Meanwhile, the MSD data can also be used on dialogue tasks to test the ability of dialogue systems when using similes. We design 3 simile tasks (recognition, interpretation, and generation) and 2 dialogue tasks (retrieval and generation) with MSD. For each task, we provide experimental results from strong pretrained or state-of-the-art models. The experiments demonstrate the challenge of MSD and we will release the data/code on GitHub.

1 Introduction

Simile plays an important role in human language to make utterances more vivid, interesting, and graspable (Zhang et al., 2021; He et al., 2022) and is an increasingly studied phenomenon in computational linguistics (Song et al., 2021; He et al., 2022). A simile is a figure of speech that compares two things from different categories (called the tenor and the vehicle) via shared properties (Paul, 1970). A tenor and a vehicle are usually connected with comparator words such as "like" or "as". For example, in the first example of Table 1, the tenor is "The boy", the vehicle is "a rabbit", the event is "run", the comparator is "as ... as" and the shared property is "fast".

	Examples	Simile
1	The boy runs as fast as *a rabbit*.	Yes
2	The girl looks like her mother.	No
3	A: Look that fireman over the street. B: Wow, he is so strong. A: I agree, strong as *a bull*.	Yes
4	A: Like *a monster*, right? B: Yes, that man is really rude.	Yes
5	A: Arguing with parents is not wise. B: It is like *throwing an egg at a rock*.	Yes
6	A: He walks into the crowd and disappears. B: It is like *a fish swims into the ocean*.	Yes

Table 1: Examples to illustrate simile. The underline font represents **tenors**. The italic font means *vehicles*. A and B are different Speakers.

The current simile research usually focuses on the simile in a triplet (tenor, shared property, vehicle) (Song et al., 2021) or a single sentence (Bizzoni and Lappin, 2018; Liu et al., 2018; Li et al., 2022). For example, the simile recognition (Birke and Sarkar, 2006; Liu et al., 2018) task is judging whether a sentence contains a simile, such as distinguishing which of the first and second examples in Table 1 contains a simile. However, a simile in a triplet or a single sentence is not enough to reflect the complex simile phenomena in the real scenario. In this paper, we study similes in real-life dialogue where a tenor and a vehicle can be mentioned by different speakers, exist in different sentences, or occur in reversed order. The third example in Table 1 shows a simile dialogue where the tenor "That fireman" and the vehicle "a bull" are in different utterances. The fourth example in Table 1 shows a simile where the tenor and the vehicle are mentioned by different speakers and the vehicle occurs before the tenor. What is more, different from previous research where the tenor and vehicle are usually single entities (Song et al., 2021) or simple

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 7223–7237
July 9-14, 2023

nominal phrases (Bizzoni and Lappin, 2018), *the tenor and vehicle in a real-life dialogue may be a verbal phrase or a long sentence*. A verbal phrase can function as the subject or object of a verb, such as the fifth example in Table 1. A sentence is a set of words expressing a statement, a question, or an order, usually containing a subject and a verb. The sixth example in Table 1 shows sentences as the tenor and vehicle. The verbal phrase and sentences can convey richer content and emotions, making the real-life dialogue more vivid and interesting. Studying these complex simile phenomena in a dialogue scenario needs to consider both the dialogue context and the various forms of the tenor and vehicle, and will lead the simile research to a brand new level. However, similes in real-life dialogue scenarios have not been studied by previous research so there are no public benchmarks available nowadays.

To facilitate the simile study, we release a human-annotated, high-quality simile dialogue dataset, which contains both English and Chinese data. The complex simile phenomena in real-life dialogue scenarios not only bring more difficulties to traditional simile tasks such as recognition, interpretation (Su et al., 2016), and generation (Li et al., 2022) but also raise challenges for dialogue research, e.g. generation and retrieval tasks. To address the simile dialogue tasks, dialogue models need to understand the simile relations between entities/phrases/sentences. Our contributions are:

- To the best of our knowledge, we are the first to study the simile phenomenon in dialogue and propose a high-quality multi-lingual simile dialogue (MSD) dataset to assist both the simile and dialogue research.
- There are 5 tasks with the proposed MSD dataset. For simile research, we design the dialogue simile recognition/interpretation/generation tasks. For dialogue research, we design the response retrieval and generation tasks.
- We verify how strong pre-trained models and the state-of-the-art simile models perform on the 5 tasks we designed. Experimental results reveal that simile in dialogue is a difficult task and requires further study. Our code and data will be released on GitHub[1].

[1]https://github.com/malongxuan/MSD

Metaphor Category	Example
Noun phrase	The nurse is *an angel*.
Adjective	These words are cold. The soldier had a warm heart.
Verbal	The process was killed. They plant the seeds of change.
Adverb-Verb	He speak fluidly.
Verbal phrase	Taking care of pets is like *raising children*.
Sentence	I rushed to the terminal like *a cheetah chasing its prey*.

Table 2: Different metaphor categories. The underline font represents **tenors**. The italic font means *vehicles*. The similes in our MSD data cover Noun phrases, Verbal phrases, and Sentence categories. The two examples in Adjective show two different Adjective-Noun modes. The two examples in Verbal are Subject-Verb and Subject-Verb-Object modes.

2 Related Work

2.1 Simile and Metaphor

The simile is a kind of metaphor that is frequently used in human languages to make utterances more vivid and graspable (Niculae and Danescu-Niculescu-Mizil, 2014) and expresses human sentiments (Li et al., 2012; Mohammad et al., 2016). Previous researchers defined different metaphor categories. We present examples for these categories in the first four lines of Table 2. For example, Bizzoni and Lappin (2018) categorized metaphor into Noun phrases, Adjectives, Verbs, and Multi-word; Li et al. (2022) categorized metaphor into Nominal, Verbal (Subject-Verb-Object), Adjective-Noun, and Adverb-Verb. Previous work usually denoted the Noun phrase metaphor as a simile (Li et al., 2022; He et al., 2022; Chen et al., 2022). *Following previous work, we also categorize Noun phrase metaphor as a simile. Meanwhile, we extend the tenor and vehicle to verbal phrases and sentences according to the simile phenomena in dialogue.* The examples of verbal phrases and sentences in simile are shown in the last two lines of Table 2.

2.2 Tasks in Metaphor/Simile

The tasks in metaphor are also suitable for simile, such as recognition (Birke and Sarkar, 2006; Liu et al., 2018), interpretation (Su et al., 2016), and generation (Li et al., 2022). The recognition task is also called identification (Steen, 2010; Li et al., 2022) or detection (Tsvetkov et al., 2014; Mohler et al., 2016), which aims to identify whether a given phrase or sentence contains a metaphor/simile. The interpretation is also called explanation (Liu et al., 2018) which usually assigns an appropriate inter-

Dataset	Lan.	Form	Task	Size	Man.
CM	Ch	sentence	I	85	Yes
SRC	Ch	sentence	R	11,337	Yes
CMC	Ch	sentence	G	11,581	Yes
MCP	En	sentence	I	1,633	Yes
SLS	En	sentence	G	87K	No
WPS	Ch	sentence	G	5M	No
Ours	Ch/En	Dialogue	R/I/G	19,565	Yes

Table 3: Survey of existing simile datasets. "Lan."/ "Ch"/ "En"/ "R"/ "I"/ "G"/ "Man." is short for "Language"/ "Chinese"/ "English"/ "recognition"/ "interpretation"/ "generation"/ "manual", respectively.

pretation to a metaphorical expression (Bizzoni and Lappin, 2018) or infers the shared properties of the tenor and the vehicle (Song et al., 2021; He et al., 2022; Chen et al., 2022). The generation task also has different forms. For example, when giving an input tenor, it can generate a simile sentence conditioned on the input tenor (Li et al., 2022); when giving both the tenor and the shared property in simile, it can generate the vehicle (Song et al., 2021; Chen et al., 2022); when providing a literal sentence, it can generate a metaphoric sentence which paraphrases that input (Chakrabarty et al., 2020; Stowe et al., 2021), or generating a specific simile according to the location where the simile interpolation should happen (Zhang et al., 2021). In this paper, we also define recognition, interpretation, and generation tasks. *However, different from previous work that only focused on similes in a triplet tuple or a sentence, we investigate a more challenging scenario where the simile happens in a multi-turn dialogue.*

2.3 Survey of Simile Datasets

Table 3 shows the comparison between our MSD dataset with the existing simile datasets. Su et al. (2016) constructed a small Chinese Metaphor (CM) data with 85 nominal and 35 verbal metaphors for the interpretation task. Liu et al. (2018) introduced Simile Recognition in Chinese (SRC) data containing sentences with a special comparator 像 (like). The Chinese Nominal Metaphor Corpus (CMC) (Li et al., 2022) data merges other Chinese metaphor datasets (Liu et al., 2018) for simile generation. He et al. (2022) proposed a simile property probing task and constructed Multi-choice Probing (MCP) datasets. Chakrabarty et al. (2020) collected Reddit comments containing similes and then auto-constructed a parallel simile corpus with a pre-trained model powered by commonsense knowledge (Bosselut et al., 2019). However, their Self-labeled Similes (SLS) dataset is limited to a "like a" pattern which appears only at the end of a sentence. Zhang et al. (2021) introduced the Writing Polishment with Similes (WPS) dataset where models need to locate the simile position in a sentence and then generate a simile in that position. The SLS and WPS are much larger than other existing data but they are not manually annotated. *Our MSD data is extracted from more than 166M dialogue data (shown in Table 4). It is the first multi-lingual simile dialogue data and the largest* ***manually annotated simile data*** *so far. What's more, benefiting from the strict annotation schedule, the MSD contains necessary simile components so that it can be used for simile recognition/interpretation/generation simultaneously.*

Dataset	Dialogue examples			
	Original	Coarse	Fine	Final
LCCC	12M	20K	4K	1,214
PchatbotW	139M	1M	82K	12,830
Reddit-dialogue	15M	71K	32K	8,510

Table 4: Statistics of the dialogue datasets we collected.

3 Multilingual Simile Dialogue Dataset

In this section, we introduce the collection, annotation, and statistics of our MSD data.

3.1 Data Collection

Since we aim to extract the simile in a real-life dialogue, we adopt the existing open-domain dialogue corpus collected from social platforms such as Reddit.com and Weibo.com. For English similes, we use the 3 turns version Reddit Dialogue dataset (Dziri et al., 2018) which contains more than 15 million dialogues. For Chinese similes, we use two datasets: PchatbotW and LCCC. The PchatbotW (Qian et al., 2021) is the largest dialogue data we can find and contains 139 million 2 turns dialogues from Weibo. The LCCC (Wang et al., 2020) is also from Weibo and contains 12 million 2 or 3 turns dialogues. We treat the last utterance in a dialogue as a response and the utterances in front of the response as a dialogue context. We extract dialogues from these large-scale datasets with a rigorous data collection pipeline, which is built based on a set of rules we will introduce in this section. Notice that we do not make any changes to the original dialogue data and only extract those dialogues with comparators in the response.

In the first step, we select the dialogue examples where the responses contain comparators such as

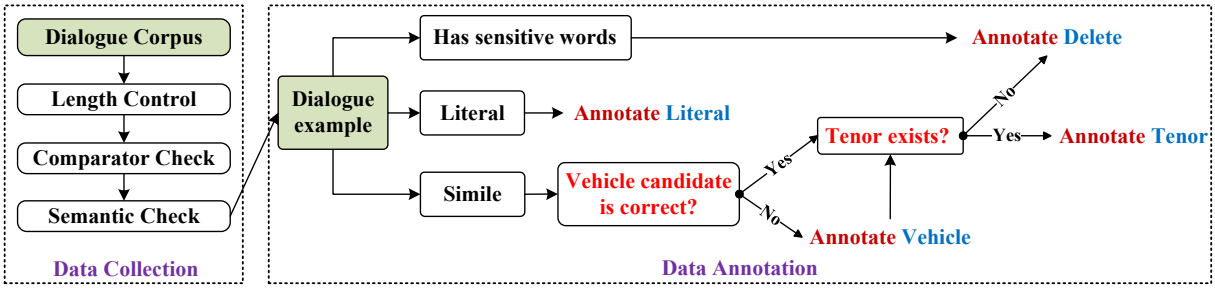

Figure 1: The data collection and annotation process.

"像...一样"/"like"/"as...as"[2]. We only select dialogue examples with context lengths between 15 and 30 words so that the dialogue context is both informative and not too long for the annotators to read. These examples are denoted by *the coarse version* of the simile dialogue data and the statistics are shown in Table 4.

In the second step, we use machine translation[3] to ensure that a sentence contains a comparator. We only reserve the dialogue examples that still contain comparator when they are translated into another language. For example, an English simile candidate sentence "I run as fast as a rabbit" contains a comparator "as...as". When translating it into Chinese, this sentence is "我跑得像兔子一样快" and still contains a comparator "像"(like). After the machine translation checking, we got *the fine version* of the simile dialogue candidates. The fine version needs further improvement since the candidate tenor/vehicle connected by the comparator is not always a simile. For example, the sentence "The Poodle is as tall as a Corgi" is not a simile since the sentence compares the height of two different kinds of dogs. So we conduct a third step to further remove examples that are not similes.

In the third step, we adopt a semantic dependency tool[4] to locate the candidate tenor/vehicle, then we compute the similarity between them to retain the examples with low similarity so that the remaining candidate tenor/vehicle are from different categories. The similarity is computed with dense representations of the candidate tenor/vehicle from BERT (Devlin et al., 2019). After the above pipeline, we obtain *the final version* of simile dialogue data for annotation. The statistics of the fine/final version we obtained are also shown in Table 4.

[2]The full Chinese comparators are listed in Appendix 10
[3]We use https://ai.youdao.com/ to conduct the translation.
[4]https://stanfordnlp.github.io/CoreNLP

3.2 Data Annotation

We recruited 7 students majoring in English for annotating the English data and recruit other 6 well-educated native speakers (graduate students) for annotating the Chinese data. We randomly select 100 examples in the final version, finding that the vehicle candidates we extracted have an acceptable accuracy (above 80%). However, the accuracy of the tenor candidate is not good (below 60%). Hence, we provided annotators with "dialogue context", "response", "comparator", and "vehicle candidate" for each dialogue. We use the annotation tool proposed by Yang et al. (2018) to simplify the operation so that the annotators can use a mouse and a few shortcuts on the keyboard to annotate.

There are some difficulties when annotating similes in the dialogue scenario apart from the fact that the tenor may exist in different sentences or occur after the vehicle. For example, the tenor may not exist in the dialogue even if the response is a simile. We ask the annotators to delete these examples. There are other situations that a dialogue that contains commonly used phrases or slang that makes the dialogue seem like a simile but not. For instance, "make like a tree" is not a simile but slang means "leave". Besides, English words usually have different meanings. For example, according to the Oxford Dictionary, the word "body" means "the whole physical structure of a human or an animal" as well as "a group of people who work or act together, often for an official purpose". So the sentence "This association is like the body that represents its members." is not a simile. Furthermore, there are many abbreviations used on social platforms such as FTW (for the win) and OP (original poster). These difficult linguistic phenomena require the annotators to have a good understanding of the dialogue context so that they could determine whether a response contains a simile.

We conduct preliminary training for the recruited annotators so that they are aware of the professional

Category	Ch	En
Simile	5,515	3,576
Literal	5,904	4,570
Tenor in context	32.8%	48.9%
Tenor in response	67.2%	51.1%
Vehicle before **Tenor**	5.7%	0.9%
Tenor before *Vehicle*	94.3%	99.1%
Ave. context words in simile	20.76	22.22
Ave. response words in simile	18.86	17.83

Table 5: The statistics of the MSD dataset. "diff." means "different". "Ave." is short for "Average".

standards. We ask the annotators to first check whether the response in this dialogue example contains a simile. The example will be annotated "Literal" if the response is not a simile. Otherwise, they should check whether the vehicle candidate in the response is correct. They need to annotate the correct vehicle (can be word/phrase/sentence) if the candidate is not accurate. If the candidate vehicle is correct, they can annotate the tenor (can be word/phrase/sentence) if it exists. We present the annotation schedule in Figure 1. Our annotation schedule ensures that the tenor and vehicle are in the data.

Quality Evaluation. During the annotation, each time we send a small "*.txt" file containing hundreds of dialogue examples to the annotators and conduct a random sampling test after they return the annotated data[5]. The annotator who returns a low-quality file will be asked to check their annotation again before we send the next file. The whole annotation takes 35 days, and each dialogue is annotated by 3 annotators. When determining the final result, the majority will be adopted when there is a disagreement among the three annotators[6]. The overall inter-rater agreement measured by Fliess' Kappa is 0.61, indicating a substantial agreement among the annotators.

3.3 Data Statistics

After the annotation, we get a total of 19,565 (8,146 English and 11,419 Chinese) dialogues. The MSD has multiple comparators for both English and Chinese data. In MSD English data, the "like" mode is around 52.4% and the "as" mode is around 47.6%. In MSD Chinese data, "像...一样" accounts for the most[7]. The proportion of each comparator is similar in simile and literal data. Table 5 shows some of the statistics of the MSD data. Please refer to the data link for more details.

[5]During annotation, we randomly selected 5% of the examples from one annotated file and checked if the annotator made accurate annotations for these random examples. The annotators were preliminary trained so that they were expected to make as few errors as possible. We expected no more than 1 error per 20 examples in the random sampling test. Otherwise, the file will be sent back for revision.

[6]There are a few cases where the three annotators disagree with each other, we decide these cases by ourselves.

Model	Precision	Recall	F1
MSD-En			
ChatGLM(zero-shot)	0.4793	0.8441	0.6114
BERT(fine-tuned)	0.7154	0.6759	0.6951
MSD-Ch			
ChatGLM(zero-shot)	0.4992	0.8772	0.6363
BERT(fine-tuned)	0.7754	0.7519	0.7635

Table 6: Simile recognition results.

4 Tasks and Results

In this section, we introduce the 5 tasks defined with our MSD dataset. Including the definition of the task, the baselines, evaluation metrics, experimental results, and analysis. The implementation details are shown in the Appendix A.

4.1 Simile Recognition Task

Following previous work (Liu et al., 2018; Li et al., 2022), we define simile recognition as a binary classification task where the model needs to distinguish whether an input sequence contains a simile. The input is a multi-turn dialogue and the output is True (simile) or False (literal).

4.1.1 Baselines and Evaluation Metrics

We use two baselines: 1) BERT is widely used and proven to be effective in classification tasks. We randomly split our MSD-En/Ch data into train/validation/test (8:1:1) sets and use the train set to fine-tune BERT. We use the output vector of the first input token <cls> of BERT to calculate the classification score for the input dialogue (see Appendix A); 2) a large language model (ChatGLM[8]). The input to ChatGLM is a concatenation of three parts: the definition of simile "A simile is a figure of speech that compares two different things via their shared properties."; a requirement "answer yes or no to this question: is the following dialogue example contains a simile?"; a simile dialogue examples such as in Table 1. Then we calculate the results according to the prediction of the baselines. Following previous work (Liu et al., 2018), we use Precision/Recall/F1 to measure the results.

[7]There are total of 11 comparators in Chinese data. Please refer to Appendix B for more details.

[8]https://www.datalearner.com/ai-models/pretrained-models/ChatGLM-6B

4.1.2 Results and Analysis

Table 6 shows the simile recognition results. We can see that BERT(fine-tuned) performs much better on Precision and F1 than ChatGLM on both MSD-En and MSD-Ch[9]. It is reasonable since the BERT models are fine-tuned on our training set. On the other hand, the ChatGLM is much better on Recall with a zero-shot setting. Overall, the classification results on both BERT and ChatGLM still have a lot of room to improve. Using syntactic structure information to locate simile components may help this task.

4.2 Simile Interpretation/Generation Tasks

Following the previous simile interpretation task (Song et al., 2021; He et al., 2022) and simile generation task (Song et al., 2021), we define Simile Interpretation/Generation (SI/SG) as a Multi-choice task with the "as...as" mode in our MSD-En[10] data (we test with 450 examples) since the shared property naturally exists in the comparator.

For **interpretation task**, we have a simile dialogue where the shared property between two "as"s is removed and replaced with a blank. The model needs to select a property from 4 choices (one correct answer and three distractors) for the blank. We construct the distractors with ConceptNet (Speer et al., 2017). In particular, we first use the tenor and some relations to find the related concept to the tenor and then use the HasProperty relation to find the distractors. Notice that for the examples where the tenor is a phrase of a sentence we could not find in ConceptNet, we use keywords (e.g. the subject of the sentence, the noun in the phrase) as the tenor to search ConceptNet.

Similar to the simile interpretation task, we remove the vehicle in a simile dialogue and leave a blank for the **simile generation task**. The model needs to select a proper vehicle for this blank from 4 candidates (one correct answer and three distractors). We also construct the distractors with ConceptNet. We use the vehicle and certain relations in the ConceptNet to find the related concepts to the vehicle as the distractors. Notice that for the examples where the vehicle is a phrase or sentence that we could not find in ConceptNet, we use the vehicles from other dialogues in MSD dataset as the distractors.

To ensure the distractors are true-negative, we randomly select 50 dialogue examples and manually check the quality of the distractors. We find that 92% of the distractors are well selected and the rest 8% are not as ideal as we expected but can still serve as distractors. More details about using ConceptNets are shown in Appendix C.

Model	Interpretation	Generation
BERT-large	0.5603	0.2967
BERT-Probe	0.5804	0.3375
BERT-ANT	0.4621	0.3337

Table 7: Simile interpretation and generation results (Hit@1) on MSD-En.

4.2.1 Baselines and Evaluation Metrics

The first baseline is a BERT-large model which takes the whole dialogue with the shared property or the vehicle masked and predicts the masked words. The second baseline is the BERT-Probe (He et al., 2022) that fine-tunes BERT with the simile interpretation task. To compare both SI and SG tasks with this baseline. We further fine-tune the BERT-Probe model with the SG task using the data proposed by He et al. (2022). The third baseline is BERT-ANT (Chen et al., 2022) which is trained with masked word prediction with metaphor data and can solve the Simile Interpretation and Generation tasks in a unified framework of simile triple completion. For example, when giving tenor=fireman and vehicle=bull, BERT-ANT can generate a list of words including the shared property like "strong" or "brave". All baselines are based on a BERT-large-uncased model. Since there are multiple masked words in our SI/SG experiments. We encode the predicted words and the candidates into dense vectors with a sentence-transformer (huggingface.co/sentence-transformers/all-MiniLM-L6-v2). Then we compute the cosine similarity between the predicted words and each of the candidates. The candidate with the highest similarity is chosen as the answer. We use Hit@1 to measure the accuracy.

4.2.2 Results and Analysis

Table 7 shows the results of simile interpretation/generation tasks. We can see that BERT-Probe performs better than BERT-large in this task, showing that a model pre-trained on simile data can better align the simile components in an input sequence and predict the missing component, even though the training data is much different from our

[9]For Chinese, we use https://huggingface.co/bert-base-chinese

[10]We did not conduct simile interpretation/generation on MSD-Ch in this paper since we did not annotate the shared property in Chinese data and we leave it for future work.

proposed data. The BERT-ANT performs similarly to the other two models on SG tasks but not as well at SI. It is because the training data of BERT-ANT is more of a metaphor data rather than simile data, a large portion of the metaphor data does not have shared properties. Hence, BERT-ANT is more powerful in connecting tenor and vehicle but is less powerful when predicting shared properties. Overall, the results on both simile interpretations/generations still have a lot of room to improve. How to exploit the semantic information in context to help these tasks requires further study.

4.3 Response Retrieval Task

Following previous work in retrieval (Guo et al., 2016), we define Response Retrieval as a ranking task. The input is a multi-turn dialogue context and multiple response candidates (including the correct one) and the model needs to rank all the candidates so that the correct one has the highest score. In particular, for each "dialogue context" in MSD simile data (both English and Chinese), we randomly select 19 responses from other dialogue as the negative examples.

4.3.1 Baselines and Evaluation Metrics

We use BERT-base for our baseline in response retrieval since it is widely used and proven to be effective in retrieval tasks. We concatenate dialogue context and each of the response candidates as the input sequence to the pre-trained model. Then we use the output of the first input token <cls> to compute the score for the input sequence as in Appendix A. Finally, the response candidate with the highest score will be chosen as the answer.

We first randomly split the Reddit dialogue data into train/validation/test (14.99M/5K/5K) sets. Then we used the BERT model to train an English dialogue retrieval model with this train/validation data. The model is denoted by BERT(Reddit). We choose a checkpoint with the best performance on the validation set. Then we use this checkpoint to compare its performance on both the Reddit Test set and the MSD-En set. Similarly, we combine LCCC and PchatbotW and randomly select 12M/5K/5K from the combined data as train/validation/test sets and train a Chinese dialogue retrieval model. The trained BERT[11] model is denoted by BERT(Ch) and used to do the comparison of the performance on the LCCC+PchatbotW Test set and the MSD-Ch set. We measure the accuracy of the retrieval with Recall@1/2/5.

[11] https://huggingface.co/bert-base-chinese

Model	R_{20}@1	R_{20}@2	R_{20}@5
MSD-En simile data			
BERT(Reddit)	0.4212	0.4960	0.6391
Reddit Test set (5K)			
BERT(Reddit)	0.8012	0.9066	0.9319
MSD-Ch simile data			
BERT(Ch)	0.3706	0.4632	0.6191
LCCC+PchatbotW Test set (5K)			
BERT(Ch)	0.4221	0.5217	0.8024

Table 8: Response retrieval results.

4.3.2 Results and Analysis

Table 8 shows the results of the response retrieval task. The performance of BERT(Reddit) and BERT(LCCC) on MSD is lower than their performance on Reddit and LCCC+PchatbotW Test sets, respectively. The results show that the data distribution in MSD is different from the data used to extract it and selecting a simile response is much harder than selecting a proper response. The low Recall results show that the dialogue retrieval task on MSD simile data needs further study. This requires a model that judges not only the relevance between context and response but also the plausibility of similes.

4.4 Response Generation Task

The traditional response generation task uses dialogue context as input and outputs the response of the context. In this section, we also introduce a new generation task that completes the response sentence behind the comparator. Taking the fifth simile dialogue "Arguing with parents is not wise. It is like throwing an egg at a rock." as an example, we give the model "Arguing with parents is not wise. It is like" as input and ask the model to generate the rest "throwing an egg at a rock.". This is different from the Writing Polishment with Similes Zhang et al. (2021) task since our task is a dialogue scene. The model needs to understand the difference between different speakers and complete the simile sentence. We use the simile data in MSD for the generation experiments. We conduct comparative experiments on the Reddit-dialogue Test set and the LCCC+PchatbotW Test set we used in the response retrieval task to show the difference between datasets.

4.4.1 Baselines and Evaluation Metrics

For the traditional response generation task, we use the DialoGPT (Zhang et al., 2020) and GODEL

Model	PPL	BLEU(1/2/3/4)(%)	ROUGE(1/2/L)(%)	METEOR(%)	Distinct(1/2)(%)
Reddit-dialogue Test set (En)					
DialoGPT	236.74	0.01 / 0.00 / 0.00 / 0.00	2.05 / 0.00 /1.79	1.24	6.67 / 23.84
GODEL	3.70	0.53 / 0.02 / 0.00 / 0.00	2.80 / 0.00 / 1.98	2.41	6.54 / 36.01
MSD-En (simile data)					
DialoGPT	329.55	11.29 / 3.58 / 1.45 / 0.70	7.53 / 0.57 / 6.39	8.48	8.39 / 28.16
GODEL	6.10	17.10 / 5.99 / 2.61 / 1.37	10.91 / 0.87 / 8.94	11.78	7.00 / 23.37
MSD-En (simile data) on Response Completion					
DialoGPT	-	17.29 / 8.50 / 5.24 / 3.35	23.71 / 5.13 / 23.04	12.85	14.64 / 43.51
LCCC+PchatbotW Test set (Ch)					
CDialGPT(Ch)	102.00	3.01 / 0.64 / 0.16 / 0.05	5.42 / 0.21 / 4.77	2.24	11.10 / 40.41
GPT-2(Ch)	129.28	5.20 / 1.50 / 0.59 / 0.26	7.09 / 0.87 / 6.14	3.04	23.23 / 66.14
MSD-Ch (simile data)					
CDialGPT(Ch)	113.75	3.07 / 0.72 / 0.26 / 0.09	5.46 / 0.24 / 4.85	2.30	11.36 / 40.58
GPT-2(Ch)	101.24	5.89 / 1.11 / 0.27 / 0.10	6.35 / 0.19 / 5.47	2.98	12.15 / 48.18
T5-base(Ch)	118.60	7.61 / 2.57 / 1.40 / 0.94	8.66 / 0.94 / 7.66	4.25	22.15 / 66.59
BART-large(Ch)	44.28	10.16 / 3.34 / 1.64 / 1.00	11.13 / 1.09 / 8.82	6.56	15.26 / 51.91

Table 9: Dialogue generation and completion results.

(Peng et al., 2022) for English data; use T5-base[12], BART-large[13] (Lewis et al., 2020), GPT-2[14] (Radford et al., 2019), and CDialGPT[15] (Wang et al., 2020) for Chinese data. We choose these baselines since 1) they are widely used and proven to be effective in dialogue generation tasks. For example, GODEL (Grounded Open Dialogue Language Model) is pre-trained for dialogue and is initiated from T5 (Raffel et al., 2020). CDialGPT and BART-large are pre-trained with LCCC-large; 2) the different size models can provide more insight into the experiments. For our proposed response generation (completion) task, we conduct the experiment on English data with DialoGPT.

We use the following automatic evaluation metrics employed by dialogue research. Perplexity (PPL), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007), and Distinct (Li et al., 2016). PPL measures the probability of the model predicting the real response. BLEU measures the n-gram overlap between the generated response and the reference one. ROUGE is based on the calculation of the recall rate of the common sub-sequence of generating response and the real one. METEOR further considers the alignment between the generated and the real responses to improve BLEU. Distinct measures the diversity of responses by calculating the proportion of distinct n-grams in the total number of n-grams. Higher BLEU/ROUGE/METEOR/Distinct means better performance. The PPL is provided for comparing models with the same vocabularies, and the results are also useful for future research.

[12]huggingface.co/shibing624/prompt-t5-base-chinese
[13]huggingface.co/HIT-TMG/dialogue-bart-large-chinese
[14]huggingface.co/shibing624/gpt2-dialogbot-base-chinese
[15]huggingface.co/thu-coai/CDial-GPT_LCCC-large

4.4.2 Results and Analysis

Table 9 shows the generation and completion results. On most metrics of English data, DialoGPT and GODEL perform better on MSD-En than on Reddit-dialogue. CDialoGPT and GPT-2 have comparable performance on the LCCC+PchatbotW Test set and MSD-Ch. This is different from the response retrieval tasks where the MSD data is more difficult than the original data used to extract MSD. The reason may be the dialogue context in MSD provides more information than the context in the original data, so the generation models could leverage the rich context information to construct an informative response. Experiments also verify that larger models (GODEL/T5/BART) have a better performance. However, even the performance of the best baseline can still be improved. We analyze the generation results. Although there are some interesting cases, most of the results are not similes. It means the simile dialogue generation task requires a specific model design to capture the simile relations in context. We provide a case study in Appendix D.

For the response completion task, when giving the comparator, DialoGPT has a big performance gain. It proves that the simile generation can benefit from the guide. Please refer to our code/data link for more experimental results about this simile dialogue completion task.

5 Conclusion

We propose manually annotated multilingual simile dialogue (MSD) data for both simile and dialogue research. We design 3 simile tasks (recognition, interpretation, and generation) and 2 dialogue tasks (retrieval and generation) with MSD. Experiments

with strong baselines show the challenge of each task. Future works include but are not limited to **1)** Dataset enlargement (e.g., more annotated examples with more kinds of comparators); **2)** Model designing (e.g., models with a specific structure to address the proposed tasks); **3)** New task designing (e.g., detecting tenor in the coarse/fine data). We encourage using the MSD in future simile and dialogue research.

Limitations

Due to time constraints, we were unable to implement some unreleased models as baselines for the proposed tasks. We did not conduct simile interpretation/generation on MSD-Ch in this paper since we could not automatically annotate the shared property in Chinese data like the "as...as" mode in English. We are currently working on this annotation and plan to release the Chinese simile interpretation/generation results on the data link. The coarse/fine version data we introduced in this paper can still be used for enlarging the MSD data. We will study to utilize them for more simile data and richer language phenomena.

Ethics Statement

We provide and emphasize some details of our work to address potential ethical concerns. First, all the data sources used in the data collection process are publicly available. We did not make any changes to the data sources and only extracted dialogue examples from these data. We carried out strict quality control during the extraction and annotation process. We made sure that there are no sensitive words even though the original data sources have already conducted this kind of checking. However, using our data to train or fine-tune a pre-trained generation model may still generate semantic errors or unpleasant similes or responses. One reason is that simile is a difficult task that compares two different things, mistakes could happen even when humans use similes. The other reason is that the knowledge stored in the original parameters of the pre-trained models may dominate the generation. We protect the privacy rights of annotators and paid 0.55 Chinese Yuan for annotating each dialogue data. The income of each annotator was above 100 Chinese Yuan per hour (On January 20, 2023, 100 yuan can be converted into 14.73 dollars).

Acknowledgements

This paper is supported by the Science and Technology Innovation 2030 Major Project of China (No. 2021ZD0113302) and the National Natural Science Foundation of China (No. 62076081, No. 61772153, and No. 61936010).

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing, Fig-Lang@NAACL-HLT 2018, New Orleans, Louisiana, 6 June 2018*, pages 45–55. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6455–6469. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5875–5887. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*,

pages 4171–4186. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory W. Mathewson, and Osmar R. Zaïane. 2018. Augmenting neural response generation with context-aware topical attention. *CoRR*, abs/1811.01063.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 55–64. ACM.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7875–7887. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *WMT@ACL*, pages 228–231. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bin Li, Haibo Kuang, Yingjie Zhang, Jiajun Chen, and Xuri Tang. 2012. Using similes to extract basic sentiments across languages. In *Web Information Systems and Mining - International Conference, WISM 2012, Chengdu, China, October 26-28, 2012. Proceedings*, volume 7529 of *Lecture Notes in Computer Science*, pages 536–542. Springer.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119. The Association for Computational Linguistics.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6468–6479. International Committee on Computational Linguistics.

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

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1543–1553. Association for Computational Linguistics.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, *SEM@ACL 2016, Berlin, Germany, 11-12 August 2016*. The *SEM 2016 Organizing Committee.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc T. Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 2008–2018. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

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

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. GODEL: large-scale pre-training for goal-directed dialog. *CoRR*, abs/2206.11309.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2470–2477. ACM.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. A knowledge graph embedding approach for metaphor processing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:406–420.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Gerard Steen. 2010. A method for linguistic metaphor identification: From mip to mipvu. volume 14. John Benjamins Publishing.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 323–336. Association for Computational Linguistics.

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Eng. Appl. Artif. Intell.*, 48:188–203.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 248–258. The Association for Computer Linguistics.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 91–103. Springer.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A lightweight collaborative text span annotation tool. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 31–36. 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14383–14392. AAAI Press.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

A Implementation Appendix

The implementations of the pre-trained models in this paper are all based on the public Pytorch implementation [16]. The hyper-parameters follow the default settings. We did not truncate any of the dialogue because the dialogue length in MSD data is much smaller than the maximum input length of the pre-trained models. We use a single Tesla v100s GPU with 32gb memory to conduct experiments, the batch size is 8 for all experiments. Checkpoints are chosen with the best performance on the corresponding validation set. In simile recognition and dialogue retrieval tasks, the first input position of the model $\mathcal{G}$ is a special token "<cls>", and the corresponding output vector E_{cls} is fed into a non-linear layer to compute the final score of the input sequence:

$$\mathcal{G}(input) = \sigma(W_2 \cdot \mu(W_1 \cdot E_{cls} + b_1) + b_2), \quad (1)$$

where $W_{1,2}$ and $b_{1,2}$ are training parameters; σ/μ is the sigmoid/tanh function, respectively. When training the simile recognition model, the loss is cross-entropy between predicted labels y_i and ground-truth label $\bar{y}_i$:

$$\mathcal{L}_{simile} = -\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i log P(y_i)) \quad (2)$$

Where N is the number of simile examples. When training the dialogue retrieval model, the loss is calculated as follows:

$$\mathcal{L}_{dr} = -\sum_{i=1}^{N} \log\left(\frac{e^{\mathcal{G}(\mathrm{C}_i,\mathrm{R}_i^+)}}{e^{\mathcal{G}(\mathrm{C}_i,\mathrm{R}_i^+)} + \sum_{j=1}^{\alpha} e^{\mathcal{G}(\mathrm{C}_i,\mathrm{R}_j^-)}}\right), \quad (3)$$

where C is the dialogue context, R is the response, and α is a hyper-parameter meaning the number of different negative samples for a positive one. We set α = 9 in our training.

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

Comparators	Proportion (%)
像...一样	49.5
跟...一样	34.8
跟...似的	11.6
像...似的	2.7
像	0.3
仿佛	0.3
简直是	0.3
如...般	0.2
像...般	0.1
如...一样	0.1
仿佛...一样	0.1

Table 10: Comparators in the Chinese MSD data.

Relation: *Definition*
RelatedTo: *The most general relation. There is some positive relationship between A and B, but ConceptNet can't determine what that relationship is based on the data. Symmetric. learn <-> erudition*
Causes: *A and B are events, and it is typical for A to cause B. exercise -> sweat*
Desires: *A is a conscious entity that typically wants B. Many assertions of this type use the appropriate language's word for "person" as A. person -> love*
DistinctFrom: *A and B are distinct member of a set; something that is A is not B. Symmetric. red <-> blue; August <-> September*
SymbolOf: *A symbolically represents B. red -> fervor*
MannerOf: *A is a specific way to do B. Similar to "IsA", but for verbs. auction -> sale*
LocatedNear: *A and B are typically found near each other. Symmetric. chair <-> table*
CausesDesire: *A makes someone want B. having no food -> go to a store*
MadeOf: *A is made of B. bottle -> plastic*

Table 11: Relations in ConceptNet we used to find distractors. "<->" means Symmetric relation for A and B. "->" means Asymmetric relation that A entails B.

B Statistic Appendix

In Table 10, we present all the comparators and their proportions in MSD-Chinese.

C ConceptNet Appendix

We use ConceptNet to construct the distractors in simile interpretation and generation tasks. ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges (Speer et al., 2017). Two concepts (A and B) are connected with relations such as "IsA" or "PartOf". In the *simile interpretation* task, we need to find three distractors for the shared property of tenor and vehicle. We use the "Antonym" relation to extract the antonym of the property as the first distractor. We adopt the nine relations in Table 11 to find the related concepts to the tenor and then use the "HasProperty" relation to find the properties of these related concepts. Finally, we randomly choose two of the nine properties as the distractors. For example, if the (tenor, property, vehicle) is (fireman, strong, bull). We first have "weak" as the Antonym and the first distractor of "strong". Then we find that "fireman" is related to (RelateTo) "fire" and "fire" has a property (HasProperty) "hot". So "hot" is the second distractor for "strong". We can get up to nine distractors and choose two of them along with the Antonym as the final three distractors. Notice that for the examples where the tenor is a phrase or sentence that we could not find in ConceptNet, we use keywords (e.g. the subject of the sentence, the noun in the phrase) as the tenor to do the search in ConceptNet. In the *simile generation* task, we also use the relations in Table 11 to find the related concepts to the vehicle as the distractors. Different from the interpretation task, for the examples where the vehicle is a phrase or sentence that we could not find in ConceptNet, we use the vehicles from other dialogues in MSD dataset as the distractors.

English example in MSD
Dialogue context: *You'll never love someone exactly the same as her, because there will never be another person exactly like her. But, that doesn't mean you won't ever be as in love with someone else.*
Ground truth response: *It will just be a different kind of love, as unique to that person as a fingerprint, yet just as strong and deep as before, or maybe even stronger and deeper.*
Response from DialoGPT: *I know, I know. I just don't want to be like her.*
Completion from DialoGPT (additionally input "it will just be a different kind of love, as unique to that person as"): *a gift which will never be forgotten.*
Response from GODEL: *I'm not sure if this is true, but I'm pretty sure I've been in love with someone exactly the same as her.*
Chinese example in MSD
Dialogue context: 鲜艳的月季花，伴你快乐的生活。 *(Bright Chinese rose flowers accompany your happy life.)*
Ground truth response: 愿我们的生活像花一样美好！ *(May our life be as beautiful as flowers!)*
Response from CDialGPT: 谢谢*(Thanks!)*
Response from GPT-2: 谢谢姐姐*(Thanks. Sister!)*
Response from T5-base: 花攒锦簇*(The flowers gather in abundance.)*
Response from BART-large: 月季花开的时候，我们的心情也会像花一样美丽！*(When the Chinese rose opens, our mood will be as beautiful as the flower!)*

Table 12: Case study of simile response generation task.

D Generation Cases Appendix

As we introduced in the simile response generation section, most of the generated results are not similes since the baselines are not designed for this task.

In Table 12, we provide two cases to show the dialogue cases in MSD and the generation results from different models. In the first English example, both DialoGPT and GODEL generate fluent responses and contain the comparator "like" or "as". However, both models fail to generate a simile response like the ground truth one. The Chinese example is extracted from the LCCC data, we can see BART-large performs the best and gives an informative response with a simile in it. The GPT-2 gives a general response and T5-base gives an informative response. The CDialGPT also gives a general response even if it is trained with the LCCC dataset. The two cases in Table 12 further verify that simile dialogue generation is challenging. However, in the response completion task, when adding the comparator in the input, we can see the DialoGPT outputs a simile and makes the dialogue more vivid and interesting.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ 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)?
Left blank.

☐ 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?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ 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.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

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?
Left blank.

☐ 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?
Left blank.

☐ 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.)?
Left blank.

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

Left blank.

☐ 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.?
Left blank.

☐ 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)?
Left blank.

☐ 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?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.