Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation

Yuyan Chen[1], Yichen Yuan[2], Panjun Liu[3], Dayiheng Liu[4], Qinghao Guan[5], Mengfei Guo[6], Haiming Peng[1], Bang Liu[7*], Zhixu Li[1*], Yanghua Xiao[1,8*†]

[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Computer Science, Beijing Institute of Technology
[4]Alibaba DAMO Academy
[5]University of Zurich
[6] Beijing Jiaotong University,
[7] RALI & Mila, Université de Montréal
[8]Fudan-Aishu Cognitive Intelligence Joint Research Center, Shanghai, China
{chenyuyan21@m., hmpeng21@m., zhixuli@, shawyh@}fudan.edu.cn, yuanyichen2024@ia.ac.cn, panjunliu@outlook.com
liudayiheng.ldyh@alibaba-inc.com, qinghao.guan@uzh.ch, guomengfei@bjtu.edu.cn, bang.liu@umontreal.ca

Abstract

Humor is a crucial part of human communication. Understanding humor and generating humorous responses in dialogue can provide natural and empathic human-computer interactions. However, most existing pre-trained language models (PLMs) perform unsatisfactorily in humor generation. On one hand, the serious shortage of humor corpus and datasets pose challenges for constructing models that can understand and generate humorous expressions. On the other hand, humor generation relies on rich knowledge and commonsense, which is often tacit and unspoken. In this paper, we construct the largest Chinese Explainable Humor Response Dataset to date with chain-of-humor and humor mind map annotations, which can be used to comprehensively evaluate as well as improve the humorous response ability of PLMs. We further design humor-related auxiliary tasks to further enhance PLMs' humorous response performance. Extensive evaluations demonstrate that our proposed dataset and auxiliary tasks effectively help PLMs to generate humorous responses, laying the groundwork for future humor research.

Introduction

Humor is an important skill in human interaction. Possessing a sense of humor requires a comprehensive and deep understanding of external knowledge, including semantics and cultural background, etc. Therefore, it is a big challenge for machines to understand what the "funny" is and enable them with a sense of humor (Yang et al. 2022; Bechade, Duplessis, and Devillers 2016; Binsted 1996). However, existing NLP research on humor mainly focuses on humor recognition (Chauhan et al. 2022; Xu et al. 2022) and humor rewriting (Petrović and Matthews 2013; Valitutti et al. 2016). Humor recognition aims to determine whether a sentence is humorous from a human perspective, and humor rewriting aims to rewrite an ordinary sentence into a humorous one. In comparison, humor response generation is a more challenging task that aims to generate humor text in dialogues (Bertero and Fung 2016b,a). Although existing Pre-trained Language Models (PLMs) can achieve superior performance in human-machine conversations, their ability to humorous response generation is still poor and there is little research on this. For example in Fig. 1, the human-generated humorous response to the context "Why do I get acne?" is "Because you're so cute that you're bubbling." which gives an unexpected answer that a person is so cute with metaphor. In contrast, the response given by a PLM is "Because you stay up late." which directly present the common reason.

The first reason for the unsatisfactory performance of PLMs is existing humor corpus and datasets are limited (Engelthaler and Hills 2018; Hossain, Krumm, and Gamon 2019) and of low quality, and the second is that humor generation requires abundant knowledge and commonsense.

Therefore, in this paper, we propose **`TalkFunny`**, a large-scale Chinese explainable humorous response dataset. `TalkFunny` consists of 4,116 high-quality context-response pairs crawled from various platforms such as RED and Zhihu. Specially, each context-response pair in `TalkFunny` features a manually created *chain-of-humor* field that interprets how the humorous text is generated, as well as a corresponding *humor mind map* field that unfolds the underlying knowledge and logic backbone for generating the humor response. As shown in Fig. 1, the chain-of-humor field explains the inference process from the input context to the humorous response. The corresponding humor mind map reveals the implicit knowledge and logic that can be

*Corresponding author.
†Yanghua Xiao is also a member of Research Group of Computational and AI Communication at Institute for Global Communications and Integrated Media, Fudan University.

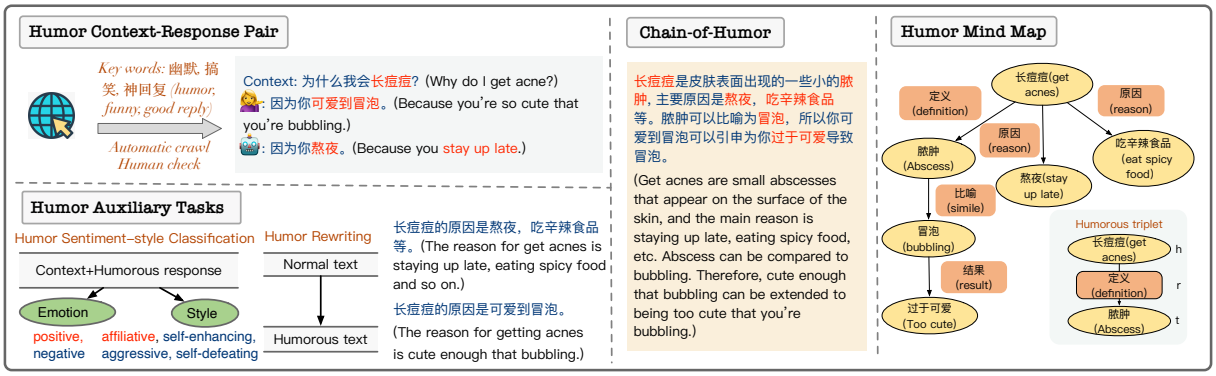

Figure 1: Responses from humans and PLMs respectively to the given context, with the necessary *chain-of-humor*, *humor mind map* and auxiliary tasks for inferring humorous response.

helpful for NLP models (e.g., PLMs) to better understand how humor arises. To obtain a chain-of-humor and humor mind map for each humor example, we design a four-step annotating mechanism inspired by humor theory such as incongruity (Forabosco 1992).

To further help improve the humor generation ability of NLP models, we augment `TalkFunny` with auxiliary tasks which can be trained together with the humorous response generation task. Specifically, we design humor sentiment-style classification and humor rewriting tasks for this purpose, and annotate the corresponding task labels. As shown in Fig. 1, the humor sentiment-style classification task aims to classify the sentiment and style of a given context-response pair, and the humor rewriting task transforms a normal text into a humorous one. Considering these tasks also require understanding humor elements, we suppose they can aid humor generation in a multitasking manner. The experimental results prove that the chain-of-humor and humor mind maps can improve the quality of humor responses for PLMs. It also justifies the benefits brought by the humor sentiment-style classification and humor rewriting tasks in humorous response generation.

Dataset Construction

In this section, we collect and annotate 4,116 pieces of explainable humorous context-response pairs with statistics shown in Table 1.

Collecting Humor Context-Response Pairs

We initiate the process by automatically extracting context-response pairs from web pages, specifically those labeled as "humor," "funny," or "good reply," while confirming that each has under 50 tokens. If a context has several responses, we select the one with the most thumbs-up. For content presented as images, we leverage OCR techniques to capture humorous context-response pairs. After collecting context-response pairs, we recruit three human volunteers to evaluate whether a piece of response is actually humorous to the context. Responses with controversial labels given by the three volunteers will be discarded. For the humorous context-response pair, we also remove privacy information to prevent any potential harm that may arise from the disclosure of personal information and discard humorless parts. All volunteers offer to help without being compensated and they are professionally trained. For quality control purposes, each piece of data is annotated and reviewed by different persons.

Chain-of-Humor Annotation

The chain-of-humor annotation in our dataset is a piece of text that explains how humorous responses arise in the underlying thought process, which contains four steps:

Step 1 - Restate contexts: we transform an interrogative context into a declarative sentence to make the annotation related to the interrogative context.

Step 2 - Generate normal responses: we manually create a normal response to the context or retrieve one through a search engine. This response is in line with the intuition of most people.

Step 3 - Tag humor anchors: we find out the *turning point* between the normal response and the final response, such as rhetoric, etc.

Step 4 - Generate humorous responses: we write humorous response based on the turning point.

Figure 2 gives an example about the above four-step annotating process. Finally, human volunteers will write a complete natural language sentence with the guidance of the above four steps. The resulting sentence is considered a piece of chain-of-humor, which represents an inference path from the context to the humorous response.

We also utilize some common patterns shared by many chain-of-humors to accelerate and unify the annotating process. Taking unity of opposites as an example: "X: hobby" has the attribute "x_1: the thing people like doing", which can generally be explained as "y_1: the thing that has a positive effect for life". Then "y_1: the thing people like doing" can be "skewed" as the attribute "z_1: the thing that has no positive effect for life" (note: this is a turning point), which is an attribute included by event "Z: dreaming". Therefore, the event "X: hobby" can be inferred to obtain the event "Z: dreaming". The pattern can also be formally represented as:

$$\{x_1, ..., x_n\} \in X; x_1 \rightarrow y_1 \rightarrow z_1; z_1 \in Z; \therefore X \rightarrow Z. \quad (1)$$

To minimize manpower, human annotators are asked to first determine the pattern to which a given context-response pair belongs. Keywords in the context-response pairs are then automatically identified and injected into the corresponding patterns to generate the Chain-of-the-Humor using GPT-3.5 [1]. The prompt used to identify and inject keywords is "Please identify the keywords in the following context-response pair X and inject them into the patterns Y." Human annotators are then responsible for proofreading the generated Chain-of-the-Humor. For example, in a context-response pair "Is the advantages of college students' 'juan' outweigh the disadvantages or vice versa? The ancients said: 'kai juan you yi'.", "juan" is an internet buzzword, which means "working hard", and "kai juan you yi" is an idiom written in Chinese Pinyin, which means " reading books is benificial", but is used to represent "working hard is benificial" here.

Content	Amount	Content	Amount
Avg. # in context	16.7	Avg. # in HMM	55.8
Max # in context	32	Max # in HMM	76
Min # in context	8	Min # in HMM	20
Avg. # in responses	15.3	Avg. # in ESC's input	29.5
Max # in responses	22	Max # in ESC's input	49
Min # in responses	5	Min # in ESC's input	18
Avg. # in COH	78.6	Avg. # in NR input	18.8
Max # in COH	137	Max # in NR input	26
Min # in COH	37	Min # in NR input	8

Table 1: Statistics of the proposed `TalkFunny` dataset. #: tokens; COH: Chain-of-Humor; HMM: Humor Mind Map; ESC: Humor sentiment-style classification task; NR: Normal responses in humor rewriting task.

Humor Mind Map Annotation

To help PLMs capture both explicit and implicit semantic knowledge contained in humorous responses and the underlying logical backbone, we build a comprehensive humor mind map based on each chain-of-humor annotation with an annotating process illustrated in Fig. 2. We first automatically extracts keywords from the Chain-of-the-Humor with GPT-3.5, allowing annotators to efficiently select suitable sets for generating Humor mind maps. Next, annotators are expected to manually spot humor-related entities $\{h_i, t_i\}$ in the extracted keywords and establish relationships r_i between each entity pair with pre-defined relation types. During this process, human annotators are also allowed to use GPT-3.5 to annotate humor mind map, which they must then proofread. After that, we generate a humor mind map from the resulting humor triplets $\{h_i, r_i, t_i\}$.

Specifically, a humor-related entity represents a word or phrase which acts as a component of the humor response inference path. It can be an event, a sort of status or description, etc. Pre-defined relations are divided into general relations and humor-related relations. General relations include "reason/result", "definition", etc. Humor-related relations include "metaphor", "anthropomorphic", etc. For example, in the Fig 2, we first select "get acnes", "abscess", "stay up late", "eat spicy food", "bubbling", "too cute" and so on as humor-related entities. Next, we establish a general relation to form a triplet "(get acnes, definition, abscess)", and establish a humor-related relation to form another triplet "(abscess, metaphor, bubbling)".

In order to control the quality of human annotation, we utilize an open-sourced annotation tool [2] to visualize the annotating process and conduct double checks by other human annotators. After that, we connect the triplets to generate the final humor mind map.

Auxiliary Tasks Annotation

Humor sentiment-style classification task. We hypothesize that PLMs can generate more reasonable humorous responses with a better understanding of the sentiment and style of the context-response pairs. Therefore, we further annotate the sentiment and style labels for context-response pairs in `TalkFunny`. For sentiment, we classify the pairs into *positive* or *negative* classes. For style, we classify them as *affiliative, self-enhancing, aggressive, self-defeating* according to humor theory (Martin et al. 2003). According to the statistics, affiliative, self-enhancing, aggressive, and self-defeating context-response pairs occupy 19.1%, 23.6%, 27.9%, and 29.4%, respectively, with 42.7% positive pairs and 56.3% negative pairs.

Humor rewriting task. PLMs generally generate normal responses instead humorous ones. Therefore, we construct normal-humorous text pairs to help PLMs learn the semantic incongruity between normal-humorous pairs and improve their ability of humor generation. We first input contexts from `TalkFunny` into a search engine (i.e., Baidu Zhidao [3]) or use GPT-3.5 to retrieve common responses to them. Next, based on grammar rules, we combine contexts and common responses to generate 4,116 normal context-response pairs, which correspond to 4,116 humorous context-response pairs.

Methods

We now investigate the humor response ability of PLMs, and evaluate whether our annotated information (e.g., chain-of-humor, humor mind maps, auxiliary tasks) can help with improving the ability. The overall process is shown in Fig. 3.

Humor Response With PLMs

It is necessary to first evaluate the humor response generation ability of existing PLMs. Therefore, we select several SOTA generation models, such as T5 (Raffel et al. 2019), BART (Lewis et al. 2020), and CPT (Shao et al. 2021), to perform humorous response generation by taking only the contexts in `TalkFunny` as the input and generate the responses R_{PLM}.

[1] https://openai.com/blog/chatgpt

[2] http://www.jinglingbiaozhu.com/

[3] https://zhidao.baidu.com/

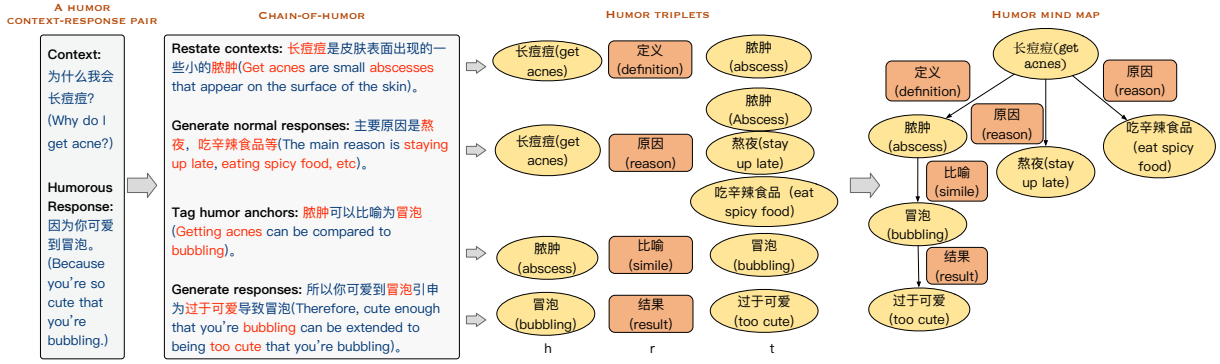

Figure 2: The process of constructing a humor mind map given the chain-of-humor of a context-response pair.

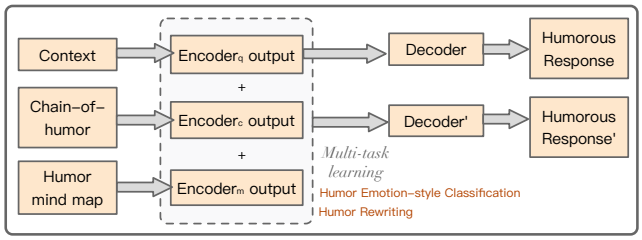

Figure 3: The multitask-enhanced humor response generation framework for evaluating PLMs.

Annotation-Enhanced Humor Response

Besides the dialogue context, we also take chain-of-humor and humor mind maps as part of the inputs and feed them into PLMs to evaluate whether they can help improve the performance of the humor response generation. First, the humor mind maps are linearized to be encoded. Second, we use breadth-first search and construct the linear sequence by appending the visited nodes to a list, and then make encoding. Specifically, we encode each type of input with a corresponding encoder of PLMs and then concatenate their outputs as follows:

$$h_Q = \text{Enc}(Q), \quad h_M = \text{Enc}(M), \tag{2}$$

$$R_{\text{annotate}} = \text{Dec}([h_Q; h_C; h_M]), \tag{3}$$

where Q and M denote the context and humor mind map, respectively, R_{annotate} denotes the generated humor response, and Enc and Dec are the encoder and decoder of a PLM. By evaluating the quality of R_{annotate} and comparing it with the generated responses R_{PLM} that are generated without extra annotations in `TalkFunny`, we can estimate how helpful our annotations can be.

Multitask-Enhanced Humor Response

We further evaluate whether our auxiliary humor tasks, i.e., humor sentiment-style classification and humor rewriting, can further improve PLMs' humor response performance. For humor sentiment-style classification, we utilize CPT (Shao et al. 2021) as the backbone for classifying the sentiment and style label for an input context-response pair. For humor rewriting, we utilize T5 (Raffel et al. 2019) as the backbone for rewriting an input normal text into a humorous version.

After that, we utilize multi-task learning to jointly learn the humor response generation task with the auxiliary tasks. Specifically, we minimize the weighted sum of the loss functions for the three tasks, where the weights are hyperparameters. By training the response generation model in such a multitask manner, we can obtain the generated responses $R_{\text{multitask}}$. By evaluating $R_{\text{multitask}}$ and comparing it with R_{annotate}, we can evaluate how helpful the auxiliary tasks can be.

Moreover, we find that the PLMs may generate toxic or very undesired responses. To address potential concerns, we've implemented a filtering mechanism for the outputs. For instance, responses containing words or phrases that promote hate, discrimination, violence, racial slurs, derogatory terms, and explicit language or any form of harm will be flagged and removed. This ensures that the content generated remains not only relevant but also respectful and safe for users.

Experiments

Experiment Setup and Baselines

The experiment is carried out on one Tesla V100 GPUs with Pytorch in Python. For the annotation-enhanced humor response, the maximum source length and targeted length is set to 512 and 128, respectively. We initialize the learning rate from 2e-5 to 4e-5 and batch size to 8 according to the memory of the machine, and use early stopping with 50 epochs. We choose some SOTA PLMs, such as T5 (Raffel et al. 2019), BART (Lewis et al. 2020), CPT (Shao et al. 2021), which achieve satisfying performance on other downstream tasks, as baselines, for both original and multitask annotation-enhanced humor response tasks in the fine-tuning setting. For some current powerful LLMs which spend a huge amount of computational cost, such as GPT-3, GPT-3.5, we only evaluate their corresponding humor response ability by inference, that is combining contexts and the chain-of-humor as the input of the PLMs, and output the

humorous responses. We also compare the performance of GPT-3.5 with prompt instructions and that without prompt instructions in humor response generation. The prompt instruction used is "Humor degree is determined based on unexpectedness, personification capabilities, sentiment correlation, informativeness, and coherence. Please answer question q_i with a touch of humor. ", which is demonstrated the best in the preliminary experiments.

Metrics

Automatic metrics contain Distinct-N (Dist-1, Dist-2) (Li et al. 2016), Greedy Match (GM), Embedding Average (EA) (Liu et al. 2016), BARTScore (BS) (Yuan, Neubig, and Liu 2021). Specifically, Dist-1, Dist-2 are used to evaluate the degree of diversity, including the distinct unigrams and bigrams in generated responses. GM and EA are used to evaluate the semantic correlation of generated responses and ground-truth responses at word-level and sentence-level, respectively. BS is to evaluate a generated text from different perspectives (e.g. informativeness, fluency, or factuality) based on BART, an encoder-decoder based pre-trained model. Due to the subpar performance of perplexity (Jelinek et al. 1977) and BLEU (Papineni et al. 2002), we omit these records in the experimental report.

Manual metrics contain unexpectedness, personification capabilities, sentiment correlation, informativeness, coherence, and humor degree. Specifically, unexpectedness (Unexp.) denotes the incongruity of the response to the context. Personification capabilities (Per.) denotes the human-like expression ability/fluency of the response. Sentiment correlation (Sen.) denotes the emotional relevance and rationality of the response. Informativeness (Info.) denotes the amount of information of the response. Coherence (Coh.) denotes the coherence and relevance to the context of the response. Humorous degree is the weighed average of the above five metrics, which is a comprehensive metric that indicating the degree of humor of a response. The rating scale of human metrics are all from 1 to 5, where 1 means the worst and 5 means the best. The final scores will be scaled to 1-100. Most importantly, we allocate the weights of the above five metrics as 4:1:1:2:2 based on the correlation coefficient in experiments.

We enroll three volunteers, and each of them is required to give scores for the randomly selected 1000 generated humorous responses by each PLM. We also calculate Inter-rater agreement of Krippendorff's Alpha (IRA) to ensure the confidence of human ratings. For the controversial ratings which have low agreements ($<$0.7), we discard this humorous response.

Main Results

The overall results are shown in Table 2, Table 3, and Table 4. From the results, we observe that both automatic and manual results on original PLMs are unsatisfying. It represents that PLMs have little ability to respond humorously without fine-tuning on `TalkFunny`. After we introduce chain-of-humor and humor mind maps to PLMs, both results are improved to a large degree. It represents that chain-of-humor and humor mind maps make great contributions on PLMs to enhance their humor response ability. Next, when we design auxiliary tasks for annotation-enhanced PLMs, their performance is further improved, especially the humor rewriting task, which will be proved in the ablation study (Sec.). It represents that some humor-related tasks have positive effect on PLMs' humor response performance. Specifically, there is not much difference in performance between different models or sizes in evaluating PLMs in three settings. This result suggests that our dataset is relatively robust and is suitable for evaluating the humor generation ability of various models. Moreover, the performance of powerful PLMs like GPT-3 and GPT-3.5 without prompt instructions is unsatisfactory for humor response generation without fine-tuning. However, we have observed a significant improvement in both automatic and manual metrics after presenting prompt instructions to GPT-3.5, emphasizing the importance of prompt instructions in this task for PLMs.

Ablation Study

We carry out ablation study to analyze the function of respective component in the overall humor response evaluation framework as shown in Table 5. In addition to the already presented components, we also designed a special humor mind map completion module to deal with the situation that the humor mind map is incomplete due to the difficulty of extraction and construction. To mimic the real situation, we randomly drop 15% of nodes in the humor mind map and then predict the missing nodes based on T5. The predicted complete humor brain map is then used in the follow-up humorous response generation task. Specifically, the input of the humor mind map completion task is chain-of-humor and masked humor mind maps, and the output is a completed humor mind maps. We adopt Kullback-Leibler divergence (Kullback and Leibler 1951) as the loss function.

From the results, we observe that chain-of-humor plays more important role than humor mind map in humor response performance of PLMs (see the row a vs. b). PLMs also have a certain of humor understanding ability of the chain-of-humor after fine-tuning so as to predict the missing information in the humor mind maps (see the row c vs. b). After we design auxiliary tasks to help PLMs in humor response, we observe that the performance increase a lot (hijklm vs. fg), especially the humor rewriting task (i vs. h; l vs. k). It represents that humor-related tasks have co-operating function.

Case Study and Error Analysis

In this section, we present some good examples generated by multitask-based annotation-enhanced humor response framework as shown in Fig. 4 and make analysis for some bad cases. For example, generated response of Q_2 adopts metaphorical rhetoric which compares "Chongming Island" to "unsinkable aircraft carrier" to echo "catamaran aircraft carrier" in the context "Is there a catamaran aircraft carrier?". Moreeover, generated response of Q_3 restate some information such as "eliminate the potential harm of genetically modified crops". In addition, some generated responses are not fluent or can not be understood by humans. We consider that although chain-of-humor and humor mind

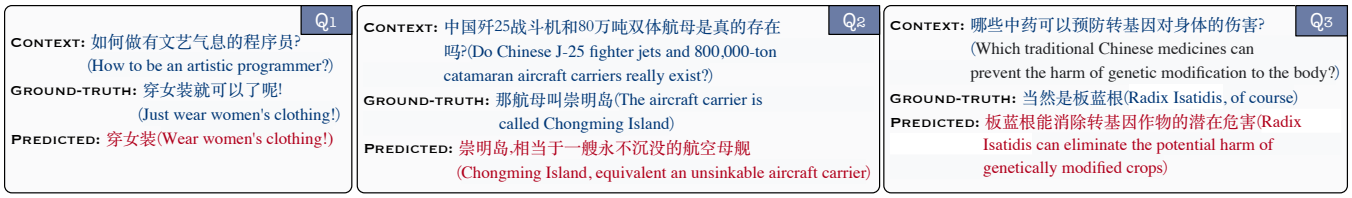

Figure 4: Some good cases generated by PLMs based on chain-of-humor, humor mind maps and auxiliary tasks.

Models	Original					Annotation-enhanced					Multitask-based				
	Dist-1	Dist-2	GM	EA	BS	Dist-1	Dist-2	GM	EA	BS	Dist-1	Dist-2	GM	EA	BS
BART-base	43.2	61.8	60.7	72.5	-9.1	45.3	74.8	70.8	81.8	-8.3	47.3	75.3	71.8	81.3	-7.8
BART-large	43.6	61.9	61.0	73.0	-8.5	45.9	74.9	71.2	81.9	-8.1	47.7	75.5	72.0	82.2	-7.6
T5-base	43.8	62.2	61.2	73.1	-7.7	46.8	75.3	71.6	82.4	-7.1	48.3	76.9	73.2	83.1	-6.6
T5-large	44.2	62.8	62.0	73.5	-7.3	46.9	75.4	72.1	82.5	-7.0	48.4	77.0	73.3	83.2	-6.3
CPT-base	44.4	63.2	62.2	73.6	-6.8	46.6	75.6	72.2	82.1	-6.2	49.4	77.1	73.8	83.0	-5.7
CPT-large	44.6	63.4	62.5	73.8	-6.3	47.0	75.9	72.8	82.4	-5.7	50.9	77.5	74.5	83.8	-4.9
GPT-3	43.7	62.0	61.1	73.3	-6.2	43.9	62.2	61.2	73.4	-5.8	-	-	-	-	-
GPT-3.5(w/o p)	44.1	62.3	62.5	74.3	-5.9	44.3	62.4	62.6	74.5	-5.5	-	-	-	-	-
GPT-3.5(w/ p)	62.5	75.8	62.8	74.5	-5.7	69.9	80.2	79.4	87.8	-5.3	-	-	-	-	-
Average	46.0	63.9	61.8	73.5	-7.1	48.5	73.0	70.4	81.0	-6.6	48.7	76.6	73.1	82.8	-6.5
Improve	-	-	-	-	-	5.4	14.2	13.9	10.2	7.0	0.3	4.9	3.8	2.2	1.5

Table 2: The automatic evaluation for PLMs' original, annotation-enhanced and multitask-based humor response performance. The average value and the improve rate are to compare with the corresponding metrics of the original PLMs. w/o p: without prompt instructions. w/ p: with prompt instructions. The results of three settings show statistical significance ($p<0.01$).

Models	Original					
	Unexp.	Per.	Sen.	Info.	Coh.	Hum.
BART-base	47.6	49.8	52.4	51.0	43.3	48.1
BART-large	47.7	50.1	52.7	51.3	44.0	48.4
T5-base	47.9	51.1	52.8	51.5	44.3	48.7
T5-large	48.0	51.3	53.1	51.7	44.4	48.9
CPT-base	48.2	51.6	53.2	52.0	44.7	49.1
CPT-large	48.7	52.1	54.0	52.3	45.0	49.6
GPT-3	76.4	47.6	86.5	52.1	63.7	67.1
GPT-3.5(w/o p)	82.5	49.8	94.6	53.6	65.0	71.2
GPT-3.5(w/ p)	92.0	95.4	94.8	87.4	88.9	91.1
Average	59.9	55.4	66.0	55.9	53.7	58.0

Table 3: The manual evaluation for PLMs' original humor response performance.

maps are introduced to PLMs, they are too difficult to be learned by PLMs. Therefore, some additional knowledge to further explain the chain-of-humor and humor mind maps is necessary for generating humorous responses of given contexts. Moreover, some examples get low scores in the automatic evaluation but the manual evaluation is satisfying. There are also examples where the automatic evaluation gets low scores but the manual evaluation is ok. For example, the question is "Does anyone use scary wallpapers on their phone?", the ground-truth answer is "Your account balance", and the predicted answer is "A wallet with no money". Although the predicted answer is not similar with the reference answer, it can answer the question and still humorous.

Related Work

Humorous datasets. Some researchers are dedicated to constructing computational humor datasets and corpora. For example, Engelthaler and Hills (2018) construct humor norms of English words with 4,997 English words; Hossain, Krumm, and Gamon (2019) construct pairs of newspaper headlines with a humorous counterpart, including making analysis of creative text editing for humorous headlines; Stock (1996) construct verbal humor in the interface to finish password swordfish; Chen et al. (2023) constructs a comprehensive Chinese humor evaluation dataset to evaluate PLMs' humor understanding ability and whether it will be improved with the exsiting knowledge. Moreover, multimodal data is also released by Kamrul et al. (2019), which includes not only linguistic, but also acoustic and facial expression features as reaction to humorous expressions in TED talks. Li et al. (2023) build a Chinese Comical Crosstalk dataset, which is for a popular Chinese performing art called "Xiangsheng". However, these study does not investigate how to make the chatbots talk funny in the conversation.

Humor generation. Humor generation aims to generate humorous texts, including question-answer jokes (Hong and Ong 2009; Labutov and Lipson 2012) and narrative jokes (Sjöbergh and Araki 2009; Yu, Tan, and Wan 2018). Main approaches are divided into template-based and neural model-based methods. The template-based methods adopt lexical replacement, such as using synonymy, meronymy, hyponymy to replace the original texts based on WordNet, ConceptNet. Some research adopts ontologies for variable selection. For example, Sjöbergh and Araki (2008) con-

Models	Annotation-enhanced						Multitask-based					
	Unexp.	Per.	Sen.	Info.	Coh.	Hum.	Unexp.	Per.	Sen.	Info.	Coh.	Hum.
BART-base	63.1	70.8	65.9	63.7	70.5	65.8	69.1	78.8	73.4	72.1	81.0	73.5
BART-large	63.4	71.2	67.2	66.8	73.1	67.2	69.8	79.2	73.9	72.4	81.5	74.0
T5-base	64.4	74.8	70.4	67.8	76.8	69.2	70.8	80.3	75.6	73.3	83.0	75.2
T5-large	66.7	75.7	71.0	67.8	77.2	70.4	70.9	80.4	75.7	73.9	83.1	75.4
CPT-base	68.5	77.3	72.2	69.1	78.9	72.0	71.1	80.6	76.3	74.7	83.6	75.8
CPT-large	68.9	78.1	72.9	70.5	79.9	72.7	71.3	80.8	76.8	75.2	84.1	76.1
GPT-3	76.6	47.9	86.8	52.3	63.9	67.4	-	-	-	-	-	-
GPT-3.5(w/o p)	82.6	49.9	95.1	53.7	65.3	71.3	-	-	-	-	-	-
GPT-3.5(w/ p)	92.9	95.7	95.3	87.9	89.0	91.6	-	-	-	-	-	-
Average	71.9	71.3	77.4	66.6	75.0	72.0	70.5	80.0	75.3	73.6	82.7	75.0
Improvement(%)	20.0	28.7	17.3	19.1	39.7	24.0	17.7	44.4	14.1	31.7	54.0	4.2

Table 4: The manual evaluation for PLMs' annotation-enhanced and multitask-based humor response performance. The results of three settings show statistical significance ($p<0.01$).

Models	Automatic					Manual					
	Dist-1	Dist-2	GM	EA	BS	Unexp.	Per.	Sen.	Info.	Coh.	Hum.
a. [Q;C]	35.5	62.8	67.8	77.4	-9.3	55.1	45.6	40.9	44.7	48.9	46.8
b. [Q;M]	30.3	61.0	63.4	75.9	-9.9	53.8	43.7	39.8	42.1	46.6	44.8
c. [Q;M']	29.5	60.1	62.5	74.4	-10.3	51.7	40.9	38.4	41.1	45.2	43.2
d. [Q;C;M]	38.1	64.8	69.6	80.0	-9.2	59.8	47.5	43.9	48.7	52.1	50.1
e. [Q;C;M']	34.7	62.7	67.8	76.5	-9.7	56.5	42.6	42.1	42.5	50.3	47.0
f. [h_Q;h_C;h_M]	**46.8**	**75.3**	**71.6**	**82.4**	**-7.7**	**74.8**	**70.4**	**67.8**	**76.8**	**64.4**	**69.2**
g. $[h_Q; h_C; h_{M'}]$	46.2	75.0	71.3	82.0	-8.8	74.6	70.1	67.5	76.4	64.2	68.9
h. $[h_Q; h_C; h_M] + C$	46.8	75.7	72.0	82.2	-7.9	75.2	71.0	68.0	76.8	64.7	69.5
i. $[h_Q; h_C; h_M] + R$	47.8	75.5	72.3	82.7	-6.3	79.1	76.6	73.2	82.5	70.3	74.8
j. [h_Q;h_C;h_M]+C+R	**48.3**	**76.9**	**73.2**	**83.1**	**-5.9**	**80.3**	**75.6**	**73.3**	**83.0**	**70.8**	**75.2**
k. $[h_Q; h_C; h_{M'}] + C$	46.4	75.3	71.7	82.1	-8.4	74.8	70.5	67.8	76.6	64.4	69.2
l. $[h_Q; h_C; h_{M'}] + R$	47.5	75.4	72.1	82.6	-7.1	78.4	74.6	72.3	82.0	69.8	74.1
m. $[h_Q; h_C; h_{M'}] + C + R$	47.9	75.5	72.4	82.7	-6.6	80.1	75.2	72.8	82.3	70.5	74.8

Table 5: The function of respective component in the humor response evaluation framework, including the annotation-enhanced evaluation module and multitask-based evidence-enhanced evaluation module.

struct a complete and modestly funny system for generating and performing Japanese stand-Up comedy, Hong and Ong (2009) automatically extract word relationships for pun generation. Others adopt n-gram co-occurrence as quantitative measures for variable selection. For example, Labutov and Lipson (2012) take humor as circuits in semantic networks, Petrović and Matthews (2013) use unsupervised methods to generate jokes from big data based on witty analogies. However, the generated humor with templates is less creative and not in line with human intuition. Neural model-based methods adopt neural networks, such as sequence-to-sequence models, language models, to generate humorous texts. For example, Li, Liu, and Wang (2022) and Yu, Tan, and Wan (2018) use neural approaches for pun generation. Although these methods can generate high level of creative output, the generated text may not be an appropriate humorous sentence from the perspective of human evaluation.

Conclusions and Future Work

Humor response generation is a challenge task in NLP. In this paper, we propose **`TalkFunny`**, a large-scale Chinese explainable humorous response dataset. Based on the constructed dataset, we comprehensively evaluate PLMs' humor response ability and enhance PLMs' humor response performance based on chain-of-humor and humor mind maps. Moreover, we design two humor-related auxiliary tasks, including humor sentiment-style classification and humor rewriting, to further improve the performance. Experimental results demonstrate our methods can effectively help PLMs to generate humorous response. In the future, we would like to construct a large-scale humor-related knowledge graph to generate humorous responses.

Acknowledgements

This work is supported by Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), the National Natural Science Foundation of China (No.62072323), Shanghai Science and Technology Innovation Action Plan (No. 22511104700), and the Zhejiang Lab Open Research Project (NO. K2022NB0AB04).

References

Bechade, L.; Duplessis, G. D.; and Devillers, L. 2016. Empirical study of humor support in social human-robot interaction. In *International Conference on Distributed, Ambient, and Pervasive Interactions*, 305–316. Springer.

Bertero, D.; and Fung, P. 2016a. A Long Short-Term Memory Framework for Predicting Humor in Dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 130–135. San Diego, California: Association for Computational Linguistics.

Bertero, D.; and Fung, P. 2016b. Predicting humor response in dialogues from TV sitcoms. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 5780–5784. IEEE.

Binsted, K. 1996. Machine humour: An implemented model of puns.

Chauhan, D. S.; Singh, G. V.; Arora, A.; Ekbal, A.; and Bhattacharyya, P. 2022. A Sentiment and Emotion aware Multimodal Multiparty Humor Recognition in Multilingual Conversational Setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, 6752–6761.

Chen, Y.; Li, Z.; Liang, J.; Xiao, Y.; Liu, B.; and Chen, Y. 2023. Can Pre-trained Language Models Understand Chinese Humor?

Engelthaler, T.; and Hills, T. T. 2018. Humor norms for 4,997 English words. *Behavior research methods*, 50(3): 1116–1124.

Forabosco, G. 1992. Cognitive aspects of the humor process: The concept of incongruity.

Hong, B. A.; and Ong, E. 2009. Automatically Extracting Word Relationships as Templates for Pun Generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 24–31. Boulder, Colorado: Association for Computational Linguistics.

Hossain, N.; Krumm, J.; and Gamon, M. 2019. " President Vows to Cut¡ Taxes¿ Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. *arXiv preprint arXiv:1906.00274*.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63.

Kamrul, M.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. URFUNNY: A Multimodal Language Dataset for Understanding Humor. 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)*, 2046–2056. Hong Kong, China: Association for Computational Linguistics.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Labutov, I.; and Lipson, H. 2012. Humor as Circuits in Semantic Networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 150–155. Jeju Island, Korea: Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 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*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, J.; Wu, X.; Liu, X.; Xie, Q.; Tiwari, P.; and Wang, B. 2023. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7581–7596.

Li, Z.; Liu, J.; and Wang, Y. 2022. Performance Analysis on Deep Learning Models in Humor Detection Task. In *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, 93–97. IEEE.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Martin, R. A.; Puhlik-Doris, P.; Larsen, G.; Gray, J.; and Weir, K. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of research in personality*, 37(1): 48–75.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Petrović, S.; and Matthews, D. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 228–232. Sofia, Bulgaria: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Shao, Y.; Geng, Z.; Liu, Y.; Dai, J.; Yang, F.; Zhe, L.; Bao, H.; and Qiu, X. 2021. CPT: A Pre-Trained Unbalanced

Transformer for Both Chinese Language Understanding and Generation. *arXiv preprint arXiv:2109.05729*.

Sjöbergh, J.; and Araki, K. 2008. A Complete and Modestly Funny System for Generating and Performing Japanese Stand-Up Comedy. In *Coling 2008: Companion volume: Posters*, 111–114. Manchester, UK: Coling 2008 Organizing Committee.

Sjöbergh, J.; and Araki, K. 2009. A Measure of Funniness, Applied to Finding Funny Things in WordNet. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics 2009*, 236–241.

Stock, O. 1996. Password Swordfish: Verbal humour in the interface. In *Proc. Intern. Workshop on Computational Humor*. Citeseer.

Valitutti, A.; Doucet, A.; Toivanen, J. M.; and Toivonen, H. 2016. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5): 727–749.

Xu, H.; Liu, W.; Liu, J.; Li, M.; Feng, Y.; Peng, Y.; Shi, Y.; Sun, X.; and Wang, M. 2022. Hybrid Multimodal Fusion for Humor Detection. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 15–21.

Yang, H.; Xu, H.; Zhang, Y.; Liang, Y.; and Lyu, T. 2022. Exploring the effect of humor in robot failure. *Annals of Tourism Research*, 95: 103425.

Yu, Z.; Tan, J.; and Wan, X. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1660. Melbourne, Australia: Association for Computational Linguistics.

Yuan, W.; Neubig, G.; and Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34: 27263–27277.