Can Pre-trained Language Models Understand Chinese Humor?

Yuyan Chen
chenyuyan21@m.fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
Shanghai, China

Zhixu Li*
zhixuli@fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
Shanghai, China

Jiaqing Liang
liangjiaqing@fudan.edu.cn
School of Data Science, Fudan University
Shanghai, China

Yanghua Xiao*
shawyh@fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University & Fudan-Aishu Cognitive Intelligence Joint Research Center
Shanghai, China

Bang Liu
bang.liu@umontreal.ca
RALI & Mila, Université de Montréal
Montréal Québec, Canada

Yunwen Chen
chenyunwen@datagrand.com
DataGrand Inc.
Shanghai, China

Abstract

Humor understanding is an important and challenging research in natural language processing. As the popularity of pre-trained language models (PLMs), some recent work makes preliminary attempts to adopt PLMs for humor recognition and generation. However, these simple attempts do not substantially answer the question: *whether PLMs are capable of humor understanding?* This paper is the first work that systematically investigates the humor understanding ability of PLMs. For this purpose, a comprehensive framework with three evaluation steps and four evaluation tasks is designed. We also construct a comprehensive Chinese humor dataset, which can fully meet all the data requirements of the proposed evaluation framework. Our empirical study on the Chinese humor dataset yields some valuable observations, which are of great guiding value for future optimization of PLMs in humor understanding and generation.

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Humor evaluation framework, Humor understanding, Chinese humor dataset, Pre-trained language models

ACM Reference Format:
Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. 2023. Can Pre-trained Language Models Understand Chinese Humor?. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570431

*The corresponding authors.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570431

1 Introduction

Humor is an advanced language art prevalently used in human languages. However, it is very challenging to let machines possess a sense of humor as humans, since it requires a deep understanding of semantics as well as cultural background. Nowadays, as the development of human-machine interaction systems and applications, how to let machines have a sense of humor has become an increasingly important topic in Natural Language Processing (NLP). Its success or failure may potentially forecast whether a Babel of human-machine interaction could finally be built.

Due to its importance, great efforts has been made on humor-relevant tasks in the NLP community, which mainly focuses on Humor Recognition and Humor Generation. Early work mainly relies on shallow linguistic features and templates to recognize or generate humors. For instance, Cattle and Ma [4] and Yang et al. [31] recognize humor with words associations and the latent semantic structures, while Aggarwal and Mamidi [1] and He et al. [14] generate poetic three liner jokes and puns through analyzing the structure and retrieve-and-edit approach, respectively. However, these methods rely on unaffordable human cost to design features or templates for different datasets, which can only recognize or generate a very limited range of humorous expressions.

As the popularity of pre-trained language models (PLMs), some recent work, such as Yu et al. [32] and Rodriguez et al. [23], make preliminary attempts to finetune PLMs for humor recognition and generation. Thanks to the powerful understanding and generation capabilities that PLMs have learned from massive amounts of data, they significantly reduce human cost and enable the recognition (or generation) on more types of humorous expressions. However, these simple endeavors do not substantially answer an important question: *whether PLMs are capable of humor understanding?*

This is a deep question worth exploring, which should be answered firstly when we utilize PLMs for various humor understanding and generation tasks. To answer this question, we would like

to investigate the humor understanding ability of PLMs in the following several aspects: i) Whether PLMs can understand humor before or after fine-tuning? ii) Whether existing external knowledge can help improve PLMs' humor understanding ability? iii) Whether PLMs can detect interpretable clue words that fit human intuitive understanding of humor? To this end, we need a well-designed evaluation framework and corresponding comprehensive dataset, both of which cannot be directly obtained from the existing humor-relevant tasks [1, 4, 14, 31].

In this paper, we first propose a three-step evaluation framework, each step of which is responsible for answering one of the above questions. Next, within this framework, we employ four representative humor-relevant tasks to conduct the systematically evaluation on PLMs, including humor recognition, humor type classification, humor level classification and punchline detection. Meanwhile, we construct a comprehensive Chinese humor dataset, which fully meets all data requirements of the four tasks and three steps evaluation framework. We choose to construct the Chinese humor dataset since Chinese humor is as worthy of study as English humor and more challenging. However, the existing Chinese humor datasets[1,2] are far less abundant than the English humor datasets, thus we want to fill in the gap for Chinese humor research.

Our empirical study based on this Chinese humor dataset suggests that: 1) By fine-tuning on the constructed humor dataset, the humor understanding ability of PLMs has been greatly improved. 2) Some external knowledge, such as Chinese pinyin information, has a positive effect on improving the PLMs' performance on humor-related tasks. 3) Moreover, a portion of the detected clue words are considered being in line with human perception of humor, but there is much room for improvement in PLMs' ability to understand humor. To summarize, our contributions in this paper are threefold:

- We are the first work to systematically evaluate the ability of PLMs in understanding Chinese humor. For this purpose, a comprehensive framework with three evaluation steps and four evaluation tasks is designed.
- We construct a more comprehensive Chinese humor dataset compared with the prior research, which fully meets all the data requirements of the proposed evaluation framework.
- Our empirical study justifies the positive effect of fine-tuning and external knowledge for PLMs on humor-relevant tasks, which has important guiding value for future optimization of PLMs in humor understanding and generation.

2 Evaluation Framework

In this paper, we focus on evaluating the ability of PLMs in understanding humor. Only when PLMs are capable of understanding humor can they generate more reasonable humorous texts, so we leave the investigation of the humor generation ability of PLMs for future work. As depicted in Fig. 1, we propose a comprehensive evaluation framework with three evaluation steps based on four evaluation tasks to investigate the capability of PLMs in understanding humor. In the following, we introduce them in detail.

2.1 Evaluation Tasks

We evaluate four representative humor-relevant tasks as follows:

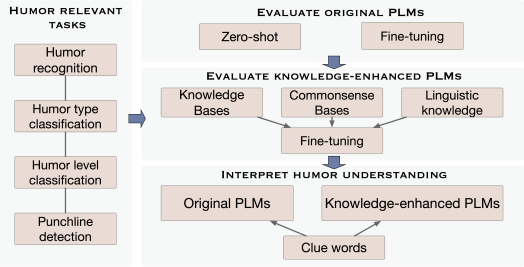

Figure 1: The evaluation framework of PLMs' humor understanding, including four tasks and three steps.

Humor Recognition. This task aims to distinguish humorous texts from humorless ones Taylor and Mazlack [29]. For each input text, the task outputs *humorous* or *humorless* as shown in Fig 2(a).
Humor Type Classification. This task first appears in the CCL2019 competitions [1]. Given a piece of humor text as input, it classifies humorous texts into several predefined humorous types and outputs *harmonic*, *ambiguous* or *incongruous* as shown in Fig 2(b).
Humor Level Classification. This task works on judging the level of humor for input texts, also first proposed in the CCL2019 competitions, where the humor can be divided into five levels from the weakest to the strongest. Here we modify the five continual levels into three discrete levels. Given a piece of humor text, the task outputs its corresponding humor level *(strong, medium or weak)* as shown in Fig 2(c).
Punchline Detection. According to humor theory [27, 30], this task determines whether there is a semantic incongruity between the previous context and its punchline (or laugh-point) ending, which originates but is slightly different from the research by Chen and Lee [5]. Specifically, the input of this task is a pair of texts: 1) the context of a humorous text before the punchline ending sentence and 2) its corresponding punchline ending sentence or a non-punchline normal ending sentence, and it will determine whether this ending sentence is a punchline one as shown in Fig 2(d).

The above four tasks are adopted for different evaluation purposes: Initially, humor recognition is the most basic task to examine a model's ability in discriminating humor. Further, humor type classification evaluates a model's ability in distinguishing between different types of humorous text, while humor level classification indicates whether the model is sensitive to different levels of humor. Last but not the least, punchline detection can reflect whether a model has a fine-grained understanding on a humorous text and thus judges whether the ending sentence is a punchline or normal one.

2.2 Evaluation Steps

To perform a thorough and insightful evaluation based on the above four humor-relevant tasks, three evaluation steps are designed for investigating the ability of PLMs in humor understanding:
Evaluate Original PLMs. The first step is responsible for evaluating the humor understanding ability of the original PLMs, which is expected to tell us: 1) Whether the original PLMs have humor understanding ability; and 2) What are the improvements and shortcomings of PLMs after simple fine-tuning on the humor dataset?

[1] http://www.cips-cl.org/static/CCL2019/call-evaluation.html

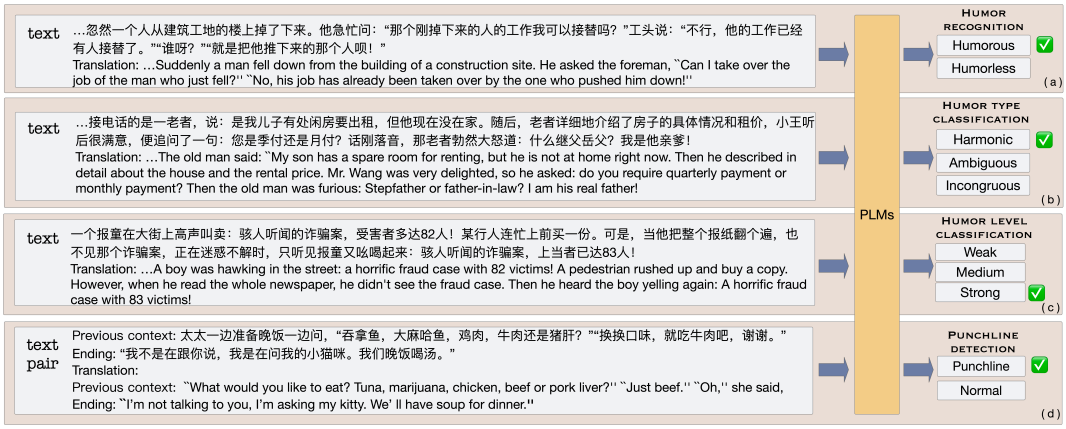

Figure 2: The process for PLMs to perform on four humor-relevant tasks.

Table 1: The overview of the constructed Chinese humor dataset used for four humor-relevant tasks.

Humor Recognition		Punchline Detection	
Property	**Amount**	**Property**	**Amount**
Humorous	18709	Punchline	18709
Humorless	7709	Normal	18709
Humor Type Classification		**Humor Level Classification**	
Property	**Amount**	**Property**	**Amount**
Harmonic	557	Weak	3722
Ambiguous	972	Medium	5764
Incongruous	2977	Strong	3009

Evaluate Knowledge-enhanced PLMs. The second evaluation step tries to inject different kinds of external knowledge into PLMs, which is to let us know: What kind of knowledge can help improve PLMs in humor understanding and to what extent?

Interpret Humor Understanding. The third step explores whether PLMs' humor understanding ability is in line with human perception. For this purpose, it investigates whether the PLMs (including the original PLMs, fine-tuned PLMs, and knowledge-enhanced PLMs) can detect appropriate clue words from the humorous text in the concerned humor tasks.

3 The Chinese Humor Dataset

To fully meet all data requirements of the evaluation framework, we construct a large-scale Chinese humor dataset, which consists of four sub-datasets as depicted in Table 1.

Humor Recognition Sub-dataset. The humor recognition sub-dataset contains humorous texts mainly from released data[1,2] and humorless text crawled from various platforms. On another hand, to construct negative examples, we mainly crawl short monologues or dialogues, such as hint fictions, fables, celebrity stories, or monologue-based diaries as the humorless texts. Such texts have similar language styles with the humorous ones. The lengths of the texts are ranged from 20 to 200 characters. Each humorless text is tagged by three recruited human volunteers to evaluate whether this piece of text is actually humorless. Texts with controversial labels given by the three volunteers will be discarded.

Humor Type Classification Sub-dataset. The humor type classification sub-dataset contains three types of humorous texts:

- *Harmonic humor*: it means a word-pair has a similar pronunciation but different meanings in a piece of humorous text, such as "季付 (meaning: pay quarterly, Chinese pinyin: jifu)" and "继父 (meaning: stepfather, Chinese pinyin: jifu)", "月付 (meaning: pay monthly, Chinese pinyin: yuefu)" and "岳父 (meaning: father-in-law, Chinese pinyin: yuefu)" in the first row of Table 3 ;
- *Ambiguous humor*: it means at least two definitions of a word are simultaneously used in a piece of humorous text. For example, "十分" can mean "very" and "ten scores" at the same time. in the second row of Table 3 ;
- *Incongruous humor*: it means there is a semantic incongruity in a humorous text, which doesn't follow humans' expectation. For example, in the third row of Table 3, "你看这个姑娘就很有素质一直很冷静嘛" which means "the girl is calm without dispute", and the response is "我住30楼，跑下来累了歇会儿再骂你" which means "the girl keeps calm because she is too tired to quarrel". Such response has semantic incongruity which doesn't follow humans' normal expectation.

We first collect the data from the release data[1,2] which already have labels. In order to guarantee the accuracy of the type labels, we proofread and then discard the wrong ones with the help of the recruited three human raters.

Humor Level Classification Sub-dataset. The humor level classification sub-dataset classifies humorous texts into three levels of humor: weak, medium, and strong. We collect the data from the same places as the humor type classification sub-dataset, and also perform similar manual proofreading operations to guarantee the accuracy of the labels.

Punchline Detection Sub-dataset. Based on the humorous text in humor recognition sub-dataset, we construct humorous and humorless context-ending pairs.

The humorous ending is extracted by dividing each humorous text into two parts: previous context and punchline. According

[2]https://github.com/liuhuanyong/ChineseHumorSentiment

Table 2: Human evaluation on normal endings generated by CPM based on semantics, correctness and readability.

Criteria	(Scores) Contents
Semantics	(5) Link very closely to the previous context. (4) Link highly closely to the previous context. (3) A majority part links to the previous context. (2) A minority part of links to the previous context. (1) Can't link to the previous context completely.
Correctness	(5) Completely factually correct. (4) Highly factually correct. (3) A majority part is factually correct. (2) A minority part is factually correct. (1) Completely factually wrong.
Readability	(5) Extremely readable without grammar mistakes. (4) Highly readable 1 grammar mistakes. (3) A majority part is fluent with a few grammar mistakes. (2) A minority part is fluent with many grammar mistakes. (1) Not readable at all.

to the theories of humor [3, 9, 21, 27, 30], the reason why humor introduces laughter is that a piece of text presents an unexpected sentence which is incongruous with the previous context. Thus, we extract the unexpected sentence in a humorous text as the punchline with the help of human annotation. Specifically, we first enroll another three volunteers, and each of them is required to vote punchline sentences for all humorous texts. We then choose the sentence which has the highest vote as the final punchline sentence for each humorous text and discard the following content in this piece of text. If more than one sentences have equal high votes, we discard this piece of humorous text.

The normal ending is generated by a text generator. We input the previous context into a large language model such as CPM [35, 36], and generate a normal ending which has similar length with the punchline. We also carry out human evaluation and machine evaluation to guarantee the quality of the generated normal endings. The steps of quality evaluation are as follows: i) We first enroll another three volunteers and randomly select 3,000 normal endings. Each of the volunteers needs to give a rating for the overall 3,000 normal endings based on a scoresheet shown in Table 2. We 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) or a normal ending is rated below 0.85, we re-generate a new normal ending. ii) Next, we use simCSE [11] to calculate similarity scores to guarantee the generated normal endings have similar semantics with the corresponding punchlines and the previous contexts. The similarity score between the normal ending and the punchline ending is s_1, and the similarity score between context-punchline pair and context-normal-ending pair is s_2. The final similarity score s is the average of s_1 and s_2. For some normal endings whose s are below 0.85, we also re-generate new ones.

4 Methodology

Fig 3 illustrates our framework for evaluating PLMs' ability in understanding humor, which consists of three parts: evaluate original/fine-tuned PLMs, evaluate knowledge-enhanced PLMs, and interpret humor understanding in PLMs.

4.1 Evaluate Original/Fine-tuned PLMs

In this module, we adopt several SOTA PLMs to model humor understanding through four representative tasks: humor recognition, humor type classification, humor level classification, and punchline detection (see Fig 3(a)). The first three are text classification tasks that take a piece of text as input and output text properties (i.e., humorous or humorless, humor types, and humor levels). The last task is performing text matching of a context-ending pair, and outputs the similarity of the pair to indicate whether the ending is a punchline or not. We calculate the similarity scores based on sentence-level embeddings, and the loss function we adopt is online contrastive loss [3], which is proved better than contrastive loss [12] in our experiments. If the similarity of the pair is over 0.5, we regard the ending as a normal ending. Otherwise, we regard the ending as a punchline.

4.2 Evaluate Knowledge-enhanced PLMs

In this module, we consider several types of external knowledge, such as general knowledge bases, commonsense bases, and linguistic knowledge. For each type of external knowledge as shown in Fig 3(b), we further use two ways of knowledge enhancement, i.e., *implicit embedding* and *explicit fusion*, to enhance PLMs.

Knowledge Embedding Construction. We utilize open-sourced Tencent AI Lab Embedding Corpus [4] and ConceptNet [5] as knowledge embeddings from general knowledge bases and commonsense bases, respectively. For the linguistic knowledge, we learn a pinyin embedding for each character as the knowledge embedding to detect different semantic meanings for characters with the same or similar pronunciations. We first utilize the *pypinyin* [6] package to generate pinyin with one of four tones for each Chinese character in a given text. For polyphonic characters, we select the first pronunciation. Inspired by Sun et al. [28], we use special tokens to denote tones. The maximum length of input pinyin sequence is set as 8 and we use a special letter "-" for padding short pinyin sequences. We adopt a Convolution Neural Network model [16] to make pinyin embedding as shown below:

$$Emb_{pinyin} = Maxpool(CNN(pypinyin(Seq_{in}))) \quad (1)$$

Fusion Layer. After constructing three types of knowledge embeddings, we normalize each of the embeddings by aggregating along the token dimension of each word or letter dimension of each pinyin (as each word is a sequence of tokens and each pinyin is a sequence of letters), respectively, with a fully connected layer to get normalized knowledge embedding(s) $\overline{Emb}_k$. Then we use two ways of knowledge enhancement for PLMs, which are Implicit embedding and Explicit fusion. Specifically, if we use BERT as the PLM, implicit embedding is to add word embedding Emb_{word}, segment embedding Emb_{seg}, position embedding Emb_{pos} and normalized knowledge embedding $\overline{Emb}_k$ together as the input embeddings for PLMs. The output embeddings for PLMs are used to make new predictions. The process is shown as follows:

$$Emb_{in} = Emb_{word} + Emb_{seg} + Emb_{pos} + \overline{Emb}_k \quad (2)$$

$$Emb_{out} = PLM(Emb_{in}) \quad (3)$$

$$y = \boldsymbol{W} * Emb_{out} + b \quad (4)$$

[3] https://www.sbert.net/
[4] https://ai.tencent.com/ailab/nlp/zh/embedding.html
[5] https://github.com/commonsense/conceptnet-numberbatch
[6] https://pypi.org/project/pypinyin/

Table 3: An example of the humor type classification sub-dataset, including harmonic humor, ambiguous humor and incongruous humor. The underlined part is the clues indicating a specific type.

Property	Content (Chinese)	Content (English)	Length
Harmonic	...您是季付还是月付？什么继父岳父？我是他亲爹！	...do you require quarterly payment or monthly payment? Stepfather or father-in-law? I am his real father!	162
Ambiguous	...十分简单。...十分简单，剩下九十分很难！	...Very easy. ...ten points are easy, the ninety points are difficult!	116
Incongruous	..."你看这个姑娘就很有素质一直很冷静嘛。""我住30楼，跑下来累了歌会儿再骂你。"	..."Look at this girl, she is very calm." "I need a rest after running down from the 30^{th} floor and then to scold you."	125

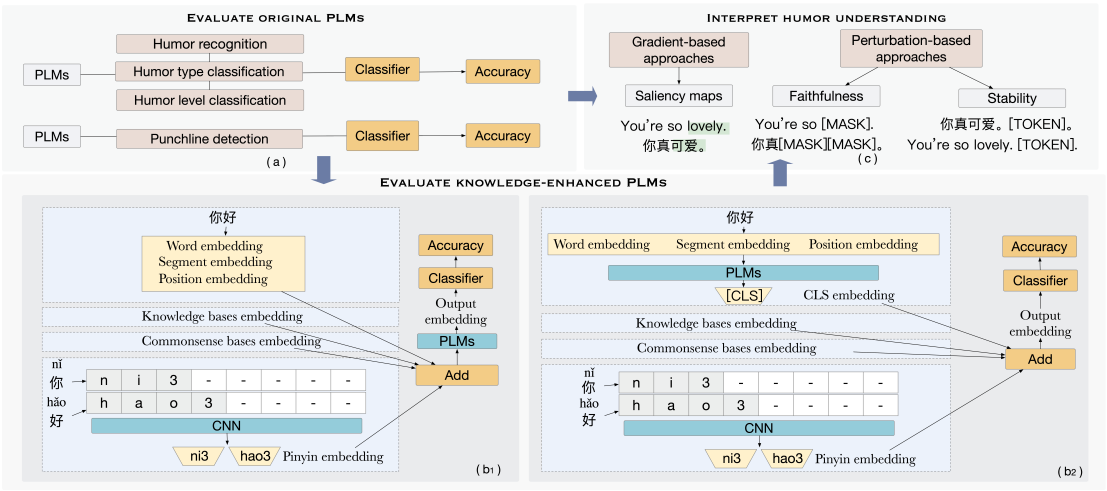

Figure 3: The evaluation framework of PLMs' humor understanding, including three steps: evaluate original PLMs (a), evaluate knowledge-enhanced PLMs (b_1,b_2), and interpret humor understanding (c).

Besides, explicit fusion is to enhance the sentence embedding given by a PLM with a fusion layer. It's well-known that the embedding of the [CLS] token in BERT-style PLMs represents the information of a whole sentence. We add the normalized knowledge embedding $\overline{Emb_k}$ to the PLMs' sentence embedding Emb_{sen} before the head of the output to make new predictions. The process is as follows:

$$Emb_{in} = Emb_{word} + Emb_{seg} + Emb_{pos} \tag{5}$$

$$Emb_{sen} = PLM(Emb_{in}) \tag{6}$$

$$Emb_{out} = Emb_{sen} + \overline{Emb_k} \tag{7}$$

$$y = \boldsymbol{W} * Emb_{out} + b \tag{8}$$

where $\boldsymbol{W}$ and b are trainable parameter and bias, respectively.

4.3 Interpret Humor Understanding

In this module, given an input text, we aim at figuring out which input words are critical for a PLM to make it correctly perform humor-relevant tasks and whether these words better interpret PLMs' humor understanding ability. We utilize gradient-based and perturbation-based techniques with the Captum package [7] for this purpose (see Fig. 3(c)).

Gradient-based approaches [26] compute saliency map based on the gradient of the input with respect to the output. We first take differentiable embeddings of tokens as the input for PLMs. Next, we aggregate the embeddings gradients with L2 normalization. Then we use Input X Gradient [25], which multiplies the gradient with the normalized embeddings, to improve the sharpness of the saliency scores, thus to compare them better. After that, we use a visualizer to present saliency maps for saliency scores to find clue words for PLMs' correct predictions on humor-relevant tasks.

Perturbation-based techniques perturb the input to find which input regions have a significant impact on the prediction. Given an input text, we randomly perturb it by adding or removing a few tokens and further check the new prediction of the PLMs. Based on perturbation-based techniques, we first investigate the faithfulness of the saliency maps to detect whether PLMs' correct prediction are not based on arbitrary choices. We replace the top-N (we set $N = 3$) most salient words with a mask token and then measure the drop of the PLMs' performance. Next, we investigate the stability of the saliency maps to detect whether insignificant words affect saliency maps. We add some random words at the end of the texts and then measure the correlation between the change of the prediction and the change of the saliency scores based on the Pearson correlation coefficient and Spearman correlation coefficient.

5 Experiments

Following the proposed evaluation framework, we carry out experiments to investigate whether the pre-trained language models (PLMs) have the ability of humor understanding (see Sec. 5.1), whether external knowledge can improve their humor understanding ability (see Sec. 5.2), and the interpretability of the detected

[7] https://captum.ai/

clue words which lead to PLMs' correct prediction on humor understanding tasks (see Sec. 5.3).

Experiment Setup. Our experiments are carried on GeForce RTX 3090 GPU (on our machine) and TPU (on Google Colab) with Pytorch in Python. The sequence length is set to 200. We initialize the learning rate to 2e-5 and batch size from 4 to 32 according to the memory of the machine, and use early stopping with 20 epochs.

Baselines, Datasets and Metrics. We adopt some representative PLMs, including base and large versions of BERT [7], RoBERTa [18], BART [17], T5 [22], CPT [24] for humor recognition, humor type classification, and humor level classification tasks. For the punchline detection task, we adopt base and large versions of simCSE-BERT [11] and simCSE-RoBERTa [11], which are proved to perform better in text matching. We divide each dataset into a training set and a dev set at a ratio of 7 to 3, and make down-sampling for all training and dev set to balance sample numbers in different classes. We use accuracy with percentage as the metric.

5.1 Results on Original/Fine-tuned PLMs

The evaluation results on original and fine-tuned PLMs based on four humor-relevant tasks are shown in Table 4 (see the column "zs" and "ft"). From the results, we observe that the original PLMs have weak ability on humor understanding with the average value 54.98, 33.77, 33.32, and 49.96 in the zero-shot learning on the humor recognition, humor type classification, humor level classification, and punchline detection task, respectively. After fine-tuning on the corresponding sub-datasets, the performance is improved by 68.67%, 72.64%, 40.87%, and 91.89%, respectively. The accuracy on humor recognition and punchline detection are both over 90%. It suggests that *PLMs have a certain degree of ability in humor recognition and punchline detection after fine-tuning on the humor dataset.*

5.2 Results on Knowledge-enhanced PLMs

In this part, we only present the experimental results by injecting Chinese pinyin into PLMs in the way of explicit fusion as shown in Table 4. That is because we observe in our experiments that either injecting Chinese pinyin in the way of implicit embedding, or injecting another one or two types of knowledge embeddings in any ways do not improve PLMs' performance in all the four humor-relevant tasks. Due to space limitation, we omit these results and will make analysis later.

See the column "K-ft" in the above Tables, we find that PLMs perform better, improving by 70.78%, 77.47%, 44.07%, and 94.63% on the humor recognition, humor type classification, humor level classification, and punchline detection task respectively. This group of experiments demonstrates that *external linguistic knowledge such as Chinese pinyin has a positive effect for PLMs in humor-relevant tasks*, and injecting external knowledge by explicit fusion is more possible to maintain important information in the knowledge than by implicit embedding.

However, for another one or two types of knowledge, which do not improve the performance of PLMs in all the four evaluation tasks, we give the possible reasons as follows: 1) The huge amount of data used for training PLMs may already contain most of the factual knowledge and commonsense knowledge, thus the existing knowledge bases or commonsense bases can not contribute more for PLMs in humor understanding. 2) Some humorous texts need complicated specific knowledge or inference paths to understand, which can not be provided by existing knowledge. For example, in the following harmonic humorous text "...一来闻，二来闻，三来闻..", the Chinese phrase "一来闻" has a similar pronunciation with the English word "eleven", where "e" and "leven" correspond the pronunciation of Chinese character "一" and "来闻", respectively. Therefore, "二来闻(two-leven)" and "三来闻(three-leven)", are analogous to "一来闻(one-leven)", , which produce harmonic humor. It's a much difficult inference process for PLMs to understand and make correct humor type classification.

Moreover, when we inject other existing knowledge from knowledge bases and commonsense bases except Pinyin knowledge in the way of implicit embedding into PLMs, the performance for PLMs in the humor-relevant tasks do not have any improvement. Due to space limitation, we also omit these results in our paper and give some possible analysis as follows: 1) Fusing Knowledge from different sources will do harm to separate feature of each type of knowledge. It's difficult for PLMs to learn effective information from the fused knowledge. 2) Humor is a much tough issue. The existing knowledge is not powerful enough for PLMs in humor understanding. Thus, *besides linguistic knowledge, PLMs also need humor-relevant background knowledge for better performance in humor-relevant tasks and better humor understanding ability.*

5.3 Results on Interpretability Analysis

To visualize the interpretability of PLMs' humor understanding ability, we draw saliency maps for sentences to show the detected clue words. We first investigate the stability of the clue words detection results to verify whether these clue words are faithful to PLMs. We add some random characters such as "我(I)" at the ending position of samples, and then measure the correlation between the change in the prediction and the change in the saliency scores. The p-values of Pearson correlation coefficient and Spearman correlation coefficient are 0.0087 and 0.0010, respectively in the zero-shot learning, 0.0086 and 0.0060, respectively in the fine-tuning, and 0.0080 and 0.0064, respectively in the knowledge-enhanced fine-tuning. The changes are statistically different due to p-values all below 0.05, which suggests that saliency scores highly relate to predictions, thus the saliency scores are stable and the clue words detection results are faithful to PLMs, which can be trusted by humans for interpreting PLMs' humor understanding ability.

Fig 4 gives the saliency maps of several samples got from BERT-base humor-relevant tasks, and the faithfulness of the saliency map on four humor-relevant tasks is shown in Table 5. In the humor recognition task, we observe that the original PLMs only take some special tokens, such as `[CLS]`, `[UNK]`, or punctuations (see Fig 4 (a_1)) as the clue words (which have deeper color). We mask the top three salient words except `[SEP]` of each instance, and the average performance has a slight drop from 54.98 to 53.61. In another three humor-relevant classification tasks, the results in the zero-shot learning as shown in Fig 4 (b_1) are similar with those on the humor recognition task. These results prove that the original PLMs (without fine-tuning) can hardly understand humor, and their predictions on humor are nearly from arbitrary choices.

After that, we fine-tune PLMs on four humor-relevant tasks, respectively. We observe that on the humor recognition sub-dataset, the saliency maps for fine-tune PLMs show that the models focus

Table 4: The evaluation results on original PLMs and knowledge-enhanced PLMs based on humor recognition (Column 2 to 4), humor type classification (Column 5 to 7), humor level classification (Column 8 to 10), and punchline detection (Column 11 to 13). zs: zero-shot learning, ft: fine-tuning, K-ft: knowledge-enhanced fine-tuning. The improvement rate is to compare with column "zs".

PLMs	Humor recognition			Humor type classification			Humor level classification			PLMs	Punchline detection		
	zs	ft	K-ft	zs	ft	K-ft	zs	ft	K-ft		zs	ft	K-ft
BERT-base	51.08	92.21	**93.17**	33.50	60.14	**62.36**	34.98	46.35	**47.24**	S-BERT-base	49.27	95.21	**96.30**
BERT-large	51.69	92.87	**94.32**	34.35	61.04	**64.31**	35.03	47.36	**48.49**	S-BERT-large	50.01	95.24	**96.7**
RoBERTa-base	52.83	91.79	**93.46**	32.51	61.42	**63.50**	36.13	45.45	**46.37**	S-RoBERTa-base	52.43	96.51	**97.88**
RoBERTa-large	52.96	92.30	**93.24**	33.16	62.58	**64.01**	37.11	46.26	**47.02**	S-RoBERTa-large	48.13	96.52	**97.98**
BART-base	49.35	91.46	**92.58**	29.40	46.23	**47.01**	31.09	46.22	**46.80**	-	-	-	-
BART-large	50.06	93.21	**93.89**	33.73	48.32	**49.10**	32.07	47.25	**47.97**	-	-	-	-
T5-base	52.05	91.47	**92.45**	33.32	55.10	**56.22**	31.28	46.82	**47.96**	-	-	-	-
T5-large	54.26	93.19	**93.99**	35.28	57.87	**58.86**	32.09	47.51	**48.76**	-	-	-	-
CPT-base	66.21	94.35	**95.85**	34.37	64.91	**66.30**	31.24	47.88	**49.24**	-	-	-	-
CPT-large	69.27	94.45	**95.91**	38.04	65.33	**67.60**	32.20	48.30	**50.21**	-	-	-	-
Average	54.98	92.73	**93.89**	33.77	58.29	**59.93**	33.32	46.94	**48.01**	Average	49.96	95.87	**97.24**
Improve rate	-	68.67%	**70.78%**	-	72.64%	**77.47%**	-	40.87%	**44.07%**	Improve rate	-	91.89%	**94.63%**

Table 5: The results of masking the top three salient words except `[SEP]` of each instance based on the zero-shot learning, fine-tuning in humor recognition (Column 2 to 3), humor type classification (Column 4 to 5), humor level classification (Column 6 to 7), and punchline detection (Column 9 to 10). mzs: mask words in the zero-shot learning, mft: mask words in the fine-tuning, mkft: mask words in the knowledge-enhanced fine-tuning. The decline is to compare with the corresponding columns in Table 4.

PLMs	humor recognition		Humor type classification		Humor level classification		PLMs	Punchline detection	
	mzs	mft	mzs	mft	mzs	mft		mzs	mft
BERT-base	50.67	52.35	32.19	30.02	33.21	32.48	S-BERT-base	48.21	47.35
BERT-large	49.42	49.38	33.11	34.37	33.22	34.57	S-BERT-large	48.29	50.52
RoBERTa-base	51.92	50.66	30.98	30.99	34.75	34.21	S-RoBERTa-base	51.33	52.01
RoBERTa-large	50.27	51.33	32.67	33.01	34.55	33.17	S-RoBERTa-large	47.09	46.50
BART-base	48.65	47.36	27.58	28.02	30.23	30.53	-	-	-
BART-large	49.25	46.98	32.37	33.21	30.59	32.04	-	-	-
T5-base	50.22	51.34	32.38	31.72	30.69	31.11	-	-	-
T5-large	53.21	52.89	34.33	33.09	31.19	30.08	-	-	-
CPT-base	64.30	64.17	32.88	32.74	30.58	29.29	-	-	-
CPT-large	68.17	69.32	36.59	37.21	31.07	32.51	-	-	-
Average	53.61	53.58	32.51	32.44	32.01	32.00	Average	48.73	49.10
Decline rate	2.55%	73.07%	3.87%	79.71%	4.11%	46.69%	Decline rate	2.52%	96.27%

more on the sentiment words, such as "嚷(shout)", "魔(demon)", when making correct predictions (see Fig 4 (a_2)). We also mask the top three salient words of each instance, and the average performance has a dramatic drop from 92.73 to 53.58. When we fine-tune PLMs on humor type classification sub-dataset, PLMs focus more on the significant words, such as "税(meaning: taxes, pinyin: shui)", "睡(meaning: sleep, pinyin: shui)", in the harmonic humorous text as shown in Fig 4 (c_2), "药(drug)", "吊(use)", in the ambiguous humorous text, and "考(have an examination)", "英(English)", in the incongruous humorous text when making correct predictions. We conjecture that *PLMs find deep semantic correlations among these clue words, which help them make correct predictions.* Moreover, when we fine-tune PLMs on the punchline detection sub-dataset, PLMs extract some significant words such as "校(meaning: school, pinyin: xiao)", "孝(meaning: filial, pinyin: xiao)", in the previous context and ending, respectively, which lead to correct predictions for a punchline ending. However, these clue words are still not very apparent and some punctuations are also regarded as salient words , which may interpret bad performance with serious threshold (0.8/0.2). Therefore, *fine-tuned PLMs have a certain degree of humor understanding ability after being fine-tuned on humor datasets. They focus on some significant words, such as sentiment words, which are partly in line with human perception on humor.*

6 Further Evaluation on Downstream Tasks

Similar to happiness, sadness, and anger, humor is one of the essential emotions of humans. Therefore, we carry out further evaluations to investigate whether PLMs which have been fine-tuned on the humor dataset can achieve better performance on the downstream tasks. We choose four Chinese sentiment classification datasets [8]: 1) ChnSentiCorp-htl-all (Chn), which is a hotel review dataset with more than 5,000 positive reviews and more than 2,000 negative reviews; 2) Waimai-10k (Wai), which contains 4,000 positive and 8,000 negative user reviews; 3) Online-shopping-10-cats (Shop), which has 30000 positive and 30000 negative user comments with online shopping; 4) weibo-senti-100k (Wei), which has about 50000 positive and 50000 negative comments.

We fine-tune BERT, which has been fine-tuned on our constructed Chinese humor dataset, including all the four sub-datasets. The results are shown in Table 6. The baseline results are based on BERT, which are derived from their published research [10, 33, 34, 37].We observe that the performance increases a little after

[8]https://github.com/SophonPlus/ChineseNlpCorpus/raw/master/datasets/

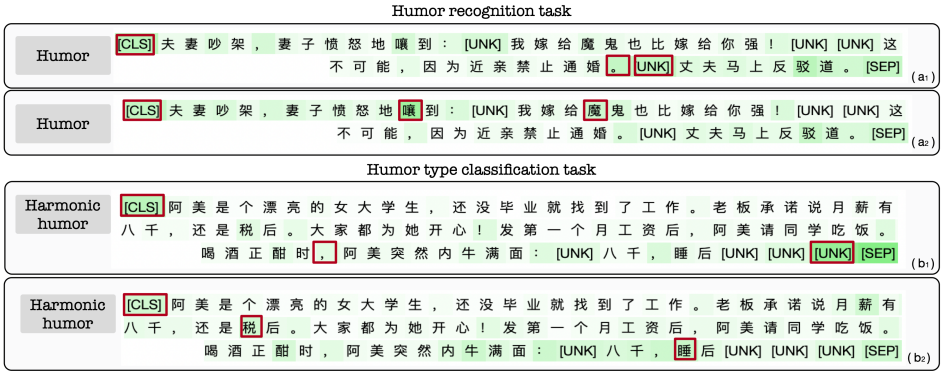

Figure 4: Saliency maps of some samples on the humor recognition sub-dataset based on the zero-shot learning (a_1) and fine-tuning (a_2), humor type classification sub-dataset based on the zero-shot learning (b_1) and fine-tuning (b_2). The top three salient words except `[SEP]` are highlighted with red boxes. We use BERT-base in these samples.

Table 6: The accuracy of baselines, fine-tuned PLMs, and knowledge-enhanced fine-tuned PLMs on four Chinese sentiment classification datasets. The improvement is to compare with the baselines.

Methods	Chn	Wai	Shop	Wei	Average	Improve rate
baseline	93.32	92.42	93.20	97.90	94.21	-
ft	**94.21**	**93.87**	**93.66**	**97.91**	**94.91**	**0.74%**
K-ft	94.20	92.98	93.76	97.88	94.70	0.52%

fine-tuning on the humor dataset (see the row "ft"). The results suggest that *the sense of humor has common characteristics with other emotions, which thus humor-fine-tuned PLMs have a positive effect on other sentiment analysis tasks.* We also find that there is no further improvement after injecting external knowledge in any above-mentioned ways (see the row "K-ft"). It suggests that *Chinese pinyin is a unique and important characteristic for humor understanding, which is not very useful for other relevant downstream tasks.*

7 Related Work

Humor Datasets and Corpora. Some research on humor dedicate themselves to construct a large-scale humor datasets and corpora. For example, Engelthaler and Hills [8] design a humor dataset which provides researchers with a list of humor ratings with 4,997 English words. Chiruzzo et al. [6] present the development of a corpus of 30,000 Spanish tweets that were crowd-annotated with humor value and funniness score. Hossain et al. [15] introduce a new dataset called Humicroedit that design simple replacement edits to make English news headlines funny. Hasan et al. [13] introduce a multimodal English humor dataset to detect humorous expressions in TED talks. Different from the above research, we construct a Chinese humor dataset, which includes four sub-datasets, each of which can be used for one representative humor-relevant task.

Humor Recognition. Other research on humor focus on humor recognition which is to decide whether a given sentence expresses a certain degree of humor. For example, Mihalcea and Strapparava [19] report text classification techniques are a viable approach to recognize humorous one-liners. Barbieri and Saggion [2] design several linguistic features to automatically detect irony and humor in twitter. Cattle and Ma [4] adopt the minimum, maximum, and average Word2Vec similarity between ordered word pairs to recognize humor and extract humor anchor. Yang et al. [31] investigate the latent semantic structures behind humor in four aspects. Morales and Zhai [20] propose a generative language model and design some key component to identify humor in reviews. Chen and Lee [5] use semantic structural features and semantic distance features to predict audience's laughter in TED Talk Data based on convolutional Neural Network. However, most of them design linguistic features to recognize humor. They ignore the powerful learning abilities of pre-trained language models (PLMs). Different from them, we design a comprehensive evaluation framework to make a research on PLMs' humor understanding ability.

8 Conclusions and future work

Humor understanding for PLMs is a challenging research in Natural Language Processing. In this work, we systematically investigate the humor understanding ability of PLMs with a designed comprehensive framework with three evaluation steps and four evaluation tasks. We also construct a comprehensive Chinese humor dataset, and our empirical study on it yields some valuable observations : 1) While the original PLMs can hardly understand humor, they could gain a certain degree of humor understanding ability through fine-tuning. 2) Linguistic knowledge such as Chinese Pinyin has a positive effect for PLMs in humor-relevant tasks. 3) The existing knowledge bases and commonsense bases can not provide much required knowledge for humor understanding. As a future work, we would like to find ways to collect more humor-relevant background and cultural knowledge for the optimization of PLMs in humor understanding.

9 Acknowledgement

This work was supported by Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), National Natural Science Foundation of China (No.62072323), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Shanghai Science and Technology Innovation Action Plan (No. 22511104700), and National Natural Science Foundation of China (No. 62102095). Yanghua Xiao is also a member of Research Group of Computational and AI Communication at Institute for Global Communications and Integrated Media, Fudan University.

References

[1] Srishti Aggarwal and Radhika Mamidi. 2017. Automatic Generation of Jokes in Hindi. In *Proceedings of ACL 2017, Student Research Workshop*. Association for Computational Linguistics, Vancouver, Canada, 69–74. https://www.aclweb.org/anthology/P17-3012

[2] Francesco Barbieri and Horacio Saggion. 2014. Automatic Detection of Irony and Humour in Twitter.. In *ICCC*. 155–162.

[3] Arthur Asa Berger. 1987. Humor: an introduction. *American Behavioral Scientist* 30, 3 (1987), 6–15.

[4] Andrew Cattle and Xiaojuan Ma. 2018. Recognizing Humour using Word Associations and Humour Anchor Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1849–1858. https://www.aclweb.org/anthology/C18-1157

[5] Lei Chen and Chong MIn Lee. 2017. Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584* (2017).

[6] Luis Chiruzzo, Santiago Castro, and Aiala Rosá. 2020. HAHA 2019 Dataset: A Corpus for Humor Analysis in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5106–5112. https://www.aclweb.org/anthology/2020.lrec-1.628

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Tomas Engelthaler and Thomas T Hills. 2018. Humor norms for 4,997 English words. *Behavior research methods* 50, 3 (2018), 1116–1124.

[9] Giovannantonio Forabosco. 1992. Cognitive aspects of the humor process: The concept of incongruity. (1992).

[10] Chenquan Gan, Qingdong Feng, and Zufan Zhang. 2021. Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis. *Future Generation Computer Systems* 118 (2021), 297–309.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[13] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: 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)*. Association for Computational Linguistics, Hong Kong, China, 2046–2056. https://doi.org/10.18653/v1/D19-1211

[14] He He, Nanyun Peng, and Percy Liang. 2019. Pun Generation with Surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1734–1744. https://doi.org/10.18653/v1/N19-1172

[15] Nabil Hossain, John Krumm, and Michael Gamon. 2019. " President Vows to Cut< Taxes> Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. *arXiv preprint arXiv:1906.00274* (2019).

[16] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[17] 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*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] Rada Mihalcea and Carlo Strapparava. 2005. Computational laughing: Automatic recognition of humorous one-liners. In *Proceedings of Cognitive Science Conference*. Citeseer, 1513–1518.

[20] Alex Morales and Chengxiang Zhai. 2017. Identifying Humor in Reviews using Background Text Sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 492–501. https://doi.org/10.18653/v1/D17-1051

[21] Lisa Glebatis Perks. 2012. The ancient roots of humor theory. *Humor* 25, 2 (2012), 119–132.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[23] Mariano Rodriguez, Reynier Ortega-Bueno, and Paolo Rosso. 2021. RoMa at HAHA-2021: Deep Reinforcement Learning to Improve a Transformed-based Model for Humor Detection. (2021).

[24] Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A Pre-Trained Unbalanced Transformer for Both Chinese Language Understanding and Generation. *arXiv preprint arXiv:2109.05729* (2021).

[25] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[27] Mary F Spence. 2006. What s SO Bloody Funny? (2006).

[28] Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038* (2021).

[29] Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 26.

[30] Yang Wen et al. 2021. A Study of American Verbal Humor in The Big Bang Theory from the Perspective of Cooperative Principle. *Academic Journal of Humanities & Social Sciences* 4, 7 (2021).

[31] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2367–2376. https://doi.org/10.18653/v1/D15-1284

[32] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1650–1660. https://doi.org/10.18653/v1/P18-1153

[33] Linkun Zhang, Yuxia Lei, and Zhengyan Wang. 2020. IAS-BERT: An Information Gain Association Vector Semi-supervised BERT Model for Sentiment Analysis. In *International Conference on Cloud Computing*. Springer, 31–42.

[34] Linkun Zhang, Yuxia Lei, and Zhengyan Wang. 2020. Long-Text Sentiment Analysis Based on Semantic Graph. In *2020 IEEE International Conference on Embedded Software and Systems (ICESS)*. IEEE, 1–6.

[35] Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. 2021. CPM-2: Large-scale Cost-efficient Pre-trained Language Models.

[36] Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021. CPM: A large-scale generative Chinese pre-trained language model. *AI Open* 2 (2021), 93–99.

[37] Jinlin Zhou, Haixin Song, Wendong Wang, Yao Niu, and Wenhao Rao. 2021. Takeaway Comments Sentiment Analysis Based on BERT. (2021).