Memeplate: A Chinese Multimodal Dataset for Humor Understanding in Meme Templates

Zefeng Li, Hongfei Lin, Liang Yang(✉), Bo Xu, and Shaowu Zhang

Dalian University of Technology, Dalian 116024, China
chinese_lzf@mail.dlut.edu.cn, {hflin,liang,xubo,zhangsw}@dlut.edu.cn

Abstract. Humor plays an important role in human communication. Besides language, multimodal information is also of great significance in humor expression and understanding, which promotes the development of multimodal humor research. However, in existing datasets, images and text often have a one-to-one relationship, making it difficult to control image modality variables. It causes the low correlation and low enhancement between the two modalities in humor recognition tasks. Moreover, with the development of Vision Transformers (ViTs), the generalization ability of visual models has been greatly enhanced. Using ViTs alone can achieve impressive performance, but is difficult to explain. In this paper, we introduce Memeplate (Our dataset is available at https://github.com/chineselzf/memeplate.), a novel multimodal humor dataset containing 203 templates, 5,184 memes and manually annotated humor levels. The template transfers images and text into a one-to-many relationship, which can make it easier for researchers to cut through the linguistic lens to multimodal humor. And it provides examples closer to human behavior for generation research. In addition, we provide multiple baseline results on the humor recognition task, which demonstrate the effectiveness of our control over image modality and the importance of introducing multimodal cues.

Keywords: Multimodality · Sentiment analysis · Humor recognition

1 Introduction

As a linguistic phenomenon, humor plays an important role in human communication. Making the AI systems enable to recognize and understand humor will greatly improve the level of linguistic intelligence and bring it closer to human mind. Early research on computational humor is mainly in the NLP community. Mihalcea and Strapparava [18] created a corpus of 16,000 one-liner jokes, and proposed humor-specific features including alliteration, antonymy, and adult slang. After that, they [19] explored several computational models for incongruity resolution and introduced a dataset containing set-ups followed by coherent continuations. In addition, much research [7,24,26] emphasized incongruity in the

W. Lu et al. (Eds.): NLPCC 2022, LNAI 13551, pp. 527–538, 2022.
https://doi.org/10.1007/978-3-031-17120-8_41

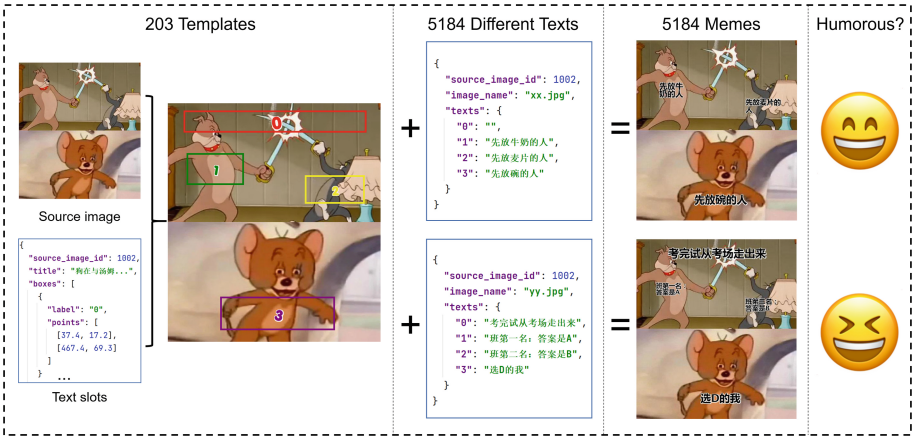

Fig. 1. An example of the Memeplate dataset. A template consists of a source image and several corresponding text slots described by a specific structure. One template can generate memes with different text which may have different humor levels. We offer 5,184 instances with annotated humor levels.

humor. In brief, humor production needs preparation and a sudden twist using a punchline. For example, in the joke "The god promised him a wish, and he said he would become a millionaire. The next day, he became a Zimbabwean.", the first sentence is the context, and the second sentence is the punchline.

However, humor can be expressed not only through language, but visual, acoustic and other modal information as well. As the example shown in Fig. 1, the image provides information not mentioned in the text, i.e., the dog is fighting with Tom, and Jerry is smiling wryly. Each object in the image is associated with a piece of text, which is impossible to obtain the complete logical relationship from the text individually, let alone express humor. After correctly combining the image content with the text (for the second meme), we can get the context "After the exam, the first and second students in the class are arguing over the answer" and the punchline "But I choose a different answer from them".

Therefore, it is important to study humor from a multimodal perspective. Multimodal humor research involves trimodal form (acoustics, vision and text) like videos, and bimodal form (image and text) like memes. For the former, Hasan et al. [9] introduced the UR-FUNNY dataset. And for the latter, Sharma et al. [21] organized SemEval-2020 Task8. Their baseline results show that the combination of two modalities performs better than single modality while Bonheme et al. [5] found that using the single modality led to better results with their models, and the two modalities were uncorrelated. To verify their conclusions, we conducted experiments using the latest models.

Table 1 shows the experimental results. The XCit and BEit show a strong generalization ability, achieving results close to the SOTA. Although image models may learn humorous information, existing research has few conclusions about

Table 1. In the table, the results of SemEval-2020 Task8 show the metrics of four-categories classification for the humorous sub-label of Task C, and the results of CCL-2021 Task4 show the metrics of three-categories classification of Task2.

Modality	Model	SemEval-2020 Task8		CCL-2021 Task4	
		ACC	F1	ACC	F1
Text	BERT [13]	32.20	24.86	45.30	41.15
	RoBERTa [16]	32.96	25.57	44.90	41.52
Image	Swin [17]	30.16	24.59	52.90	48.56
	XCiT [2]	**34.94**	**25.61**	55.90	54.34
	BEiT [3]	30.59	24.90	**63.00**	59.56
Text+Image	RoBERTa+Swin	29.89	24.62	54.30	51.95
	RoBERTa+XCiT	32.10	25.35	49.60	47.17
	RoBERTa+BEiT	30.32	25.43	62.10	**60.80**

humor features in image modality. And we cannot determine whether the patterns learned by image models are relevant to humor. It is more feasible to explain the mechanism of multimodal humor from a textual perspective. Therefore, we intend to create a dataset that constrains image modality but has no effects on multimodal properties.

We are inspired by the meme secondary-creations on the Internet, where a source image is filled with different text to produce different humorous effects. In this case, the information provided by image modality is almost identical for the memes with the same source image, and the humorous effect is mainly determined by the text filled in. The secondary-creation can be considered as using a template approach, by which we create Memeplate, a novel multimodal humor dataset. We aim to make researchers easier to apply the linguistic findings to multimodal humor, and provide examples of meme generation closer to human behavior for humor generation research. Overall, our main contributions are as follows:

- We design a novel annotation scheme to focus on humor in memes from the textual perspective. The main idea is to use templates to constrain the image modality.
- We create a multimodal dataset containing 203 templates and 5,184 memes with manually annotated humor levels based on the scheme. And we offer detailed annotation process and statistical information on the dataset.
- We conduct experiments with multiple baseline models for the humor recognition task on the dataset. The experimental results demonstrate the effectiveness of our control over image modality and the importance of combining multimodal cues.

2 Related Work

The humor dataset is fundamental in computational humor research. Mihalcea and Strapparava [18] created a corpus of 16,000 one-liner jokes. Yang et al. [24] introduced Pun of the Days for pun recognition. Both of them are binary classification tasks. Recently, the humor dataset evolves in a diverse direction, which reflects trends in computational humor.

The first trend is from the classification of humor levels to the research of humor mechanisms. Ahuja et al. [1] collected jokes from Twitter and Reddit, and recognized three major characteristics reflected across all types of jokes including modes, theme and topics. Zhang et al. [25] developed a Chinese humor corpus containing 9,123 jokes with reference to the General Theory of Verbal Humor (GTVH). Their annotations of linguistic humor not only contain the degree of funniness, but contain keywords that trigger humor as well as character relationship, scene, and humor categories as well. Tseng et al. [22] developed a Chinese humor corpus containing 3,365 jokes with five levels of funniness, eight skill sets of humor, and six dimensions of intent.

The second trend is from English to other languages or multi-languages. Castro et al. [7] introduced a humor corpus containing 33,531 Spanish Tweets. The dataset involves binary annotation and humor level annotation with five levels. Then they [6] revised it with crowd notes and presented a 27,000 tweets dataset in total. Instead of using the five-point annotation, they used five different emojis to represent the levels of humor. Blinov et al. [4] constructed a Russian language dataset with over 300,000 short jokes using an automated approach. Khandelwal et al. [14] create a corpus containing English-Hindi code-mixed tweets annotated with humorous or non-humorous tags.

The third trend is from text-based to multimodal. For the trimodal form (acoustics, vision and text), Hasan et al. [9] collected data from TED talks and published the UR-FUNNY multimodal humor dataset with 8,257 humor and 8,257 non-humor instances. Wu et al. [23] used a similar approach to construct the first Chinese multimodal humor dataset, MUMOR. Kayatani et al. [12] study facial expressions in the visual modality and constructed a dataset from The Big Bang Theory. For the bimodal form (image and text), Sharma et al. [21] released approx 10K annotated memes with manually annotated labels, including humorous, sarcasm, offensive, and motivation. Ziser et al. [27] combine humor recognition tasks with the domain of Product Question Answer (PQA). Annotators were presented with a question and the associated product image with caption, and were asked to classify whether the question is humorous or not.

3 Dataset

3.1 Data Collection

To make the dataset objective and comprehensive, data was collected from a range of sources, including social media (Weibo and Tieba) and image recognition websites (Baidu Image and Yandex Image). We searched with keywords,

including “迷因图(meme)”, “表情包(sticker)”, “搞笑图片(funny images)” and “幽默图片(humorous images)” to obtain data from both content and blogger perspectives. For the content results, we directly downloaded them. And for the blogger results, we selected 20 bloggers who were most followed and downloaded the image contents posted by them. To obtain textual information, we extracted the text in memes using PaddleOCR[1] for reference in the data filter and annotation process. In terms of genre, the data we collected included memes, stickers and comics. In general, the language in memes is more standardized and completed, while the language in stickers and comics is more verbalized and fragmented.

3.2 Data Filter

Data filter could be divided into two stages, the first stage was the preliminary filter. Images with unqualified content (e.g. meaningless, politically sensitive, emoji without text, non-Chinese and non-English), text-based content (e.g. a WeChat conversation between two people), and duplicate text were dropped. However, low-resolution images were reserved, as long as the text could be recognized. Because high-quality images could be generated using high-resolution source images and recognized text after template creation. The second stage was the counting filter for further template creation and one-to-many text examples. We counted the memes group by source images manually, and those with less than four repetitions were dropped.

3.3 Image Recognition

Image recognition was conducted with Baidu Image and Yandex Image for two purposes. The first was to obtain high-quality source images, either original images or ones with text boxes but not filled yet (a template in a sense, but requiring manual edit). The second was to expand the existed data using the similar images search function. Since such websites crawl images across the whole Internet, search results can be plentiful. We conducted the similar image search on memes in terms of source images, downloaded the results, and repeated the data filter steps described above.

3.4 Data Annotation

After collection, we obtained **a)** filtered memes, and **b)** source images (one corresponding to several memes without text). Then the dataset was annotated in the following steps:

– **Text slot annotation.** As the example shown in Fig. 1, a template has several text slots labeled with different colors. We specified that the text in memes must be filled in the slots. Each template could have more than

[1] https://github.com/PaddlePaddle/PaddleOCR.

one rectangular text slot, but should be able to reasonably accommodate all the text in the filtered memes. One annotator completed this part of the annotation using Labelme[2].

- **Image caption annotation.** As mentioned earlier, humor production requires context and punchlines. Since the text modality rarely contained complete information about context and punchlines, we provided captions to supply image information to text modality from another perspective. We specified that captions needed to describe the objects (required) and actions (optional) in the images. For example, the caption of the image in Fig. 1 is "狗在与汤姆持剑打斗，杰瑞苦笑 (The dog is fighting with Tom with swords and Jerry is smiling wryly)". One annotator completed this part of the annotation.
- **Text extraction and filling.** In the process of data collection, we used OCR tools to extract the text in the memes, however, the quality was poor and manual revision was required. In addition, since we manually annotated text slots in the templates, different paragraphs of text needed to be filled into the corresponding slots manually. The data was split into 40 pieces, and the annotation work was performed by 40 people, each of whom annotated one piece of the data and revised another one.
- **Humor level annotation.** The humor level reflects whether a meme is funny and how funny it is. We classified memes as not funny, slightly funny (tending to be funny), funny (somewhere between slightly and very funny) and very funny (making people laugh) ones. Forty people participated in the annotation and the data was split equally into eight groups, each group was annotated by five people, and each meme was annotated five times.

An important aspect to analyze is the disagreement of the annotation. Different from binary annotation, we annotated no humor and three humor levels, so it is necessary to consider the disagreement between them. For example, a disagreement between strong humor and medium humor should be considered different from a disagreement between strong humor and no humor. We chose Krippendorff's alpha measure [15], which considers this into the formula by using a generic distance function.

We calculated the alpha value for each group (eight in total) and the mean value was 0.348, which was not good enough. After removing the memes with both not funny and very funny annotations, the alpha value went up to 0.399. The results showed the subjectivity of humor, especially when the humor intensity was subdivided. Finally, five annotations of each meme were comprehensively considered.

4 Data Analysis

Our dataset consists of 203 templates (source images) and 5,184 memes (paragraphs of text). According to the characteristics of the dataset, we analyze the data on the meme scale and template scale.

[2] https://github.com/wkentaro/labelme.

Table 2. Data Analysis in the meme scale. Here, "#" denotes number, and "avg" denotes average. The train, development and test folds share no template with each other.

Meme	Total	Train	Dev	Test
$\#\text{instances}_{\text{all}}$	5184	3746	700	738
$\#\text{instances}_{\text{not funny}}$	402	263	72	67
$\#\text{instances}_{\text{slightly funny}}$	1945	1447	244	254
$\#\text{instances}_{\text{funny}}$	2364	1710	313	341
$\#\text{instances}_{\text{very funny}}$	473	326	71	76
#characters	105169	76121	14261	14787
#distinct characters	2880	2685	1644	1727
Avg $\#\text{characters}_{\text{all}}$	20.29	20.32	20.37	20.03
Avg $\#\text{characters}_{\text{not funny}}$	19.12	19.25	19.14	18.57
Avg $\#\text{characters}_{\text{slightly funny}}$	19.27	19.29	19.05	19.35
Avg $\#\text{characters}_{\text{funny}}$	20.83	20.94	21.00	20.13
Avg $\#\text{characters}_{\text{very funny}}$	22.77	22.53	23.41	23.21
#distinct templates	203	128	34	41
Avg #text slots used	2.59	2.61	2.52	2.54

Table 2 presents high-level statistics of the dataset on the meme scale. There are 5,184 memes in the whole dataset. In terms of categories, not funny category and very funny category account for a relatively small proportion. It also shows the standard train, development and test folds of the dataset. We try to make the data in each fold have the same distribution. And they share no template with each other, hence standard folds have source image independence which avoids the label leakage.

Figure 2 shows an overview of some important statistics of the dataset. Figure 2(a) shows the distribution of the text length. Figure 2(b) shows the word cloud. The frequent words include "我(I)", "妈妈(mum)", "老师(teacher)", and "朋友(friend)", which shows that we like to tell jokes about our daily life. For example, complaining about mums nagging at home and teachers assigning endless homework at school. Besides, there are some frequent sentences such as "最开始的我(The very beginning of me)" and "时间穿越成功了(Time travel has been successful)", which are the fixed topic sentences in some templates. For the content of the source image, animals are the most frequent, followed by anime characters such as Tom and SpongeBob, indicating that people like to express their humor with cute and funny things. Figure 2(c) demonstrates the distribution of class. Figure 2(d) shows the distribution of the number of corresponding memes for the templates and Fig. 2(e) shows the distribution of the number of text slots for the templates.

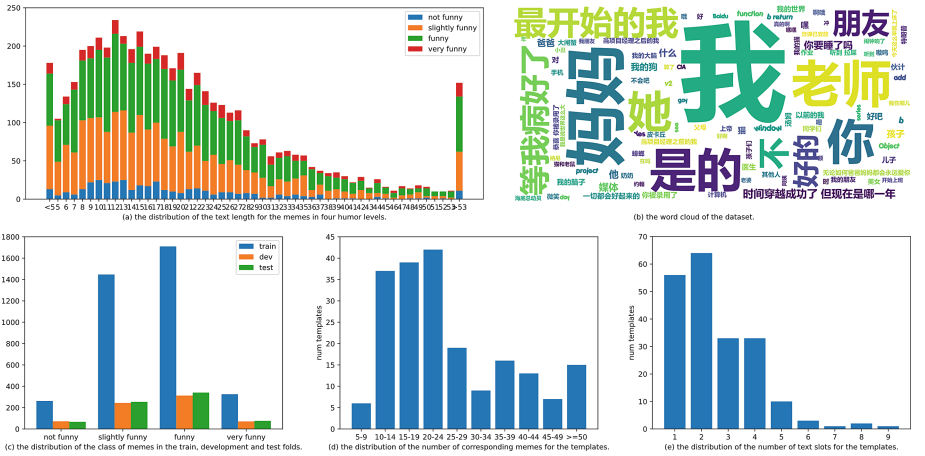

Fig. 2. Overview of the dataset statistics. (a) the distribution of the text length for the memes in four humor levels. (b) the word cloud of the dataset. (c) the distribution of the class of memes in the train, development and test folds. (d) the distribution of the number of corresponding memes for the templates. (e) the distribution of the number of text slots for the templates.

5 Experiments

In this section, our goal is to establish a performance baseline on the humor recognition task for our dataset. We define the task as a four-categories classification task: to classify the given meme as a not funny, slightly funny, funny or very funny one. Accuracy and macro-F1 score are used as evaluation metrics, for the evaluation of overall classes and individual class performance on the models. We also aim to demonstrate the effectiveness of our scheme in dataset construction and figure out the following questions:

- **Q1**: Can baseline models learn some knowledge related to humor from our dataset?
- **Q2**: What is the performance of using single modality, and does our dataset constrain the interference of irrelevant features in image modality?
- **Q3**: Does the information provided by image modality contribute to humor recognition even if many memes have the same source image?
- **Q4**: If the answer to **Q3** is yes, which is the most effective way?
- **Q5**: Is using cross-modal pre-trained models better than using two modal pre-trained models separately?

5.1 Baseline Models

We used the model shown in Fig. 3 to recognize humor. For text input, we encoded the text using Transformers encoder including BERT, RoBERTa and MacBERT [8], while for image input, we extracted both the overall features

using image classification models and the ROIs features using object detection models. The image classification models included ResNet-50 [10], XCit and BEit, and the object detection model was Faster-RCNN [20]. We also extracted the features using a Chinese cross-modal pre-trained model namely WenLan [11], in which both modalities interact in the pre-training stage and the cross-modal relationship can be better modeled.

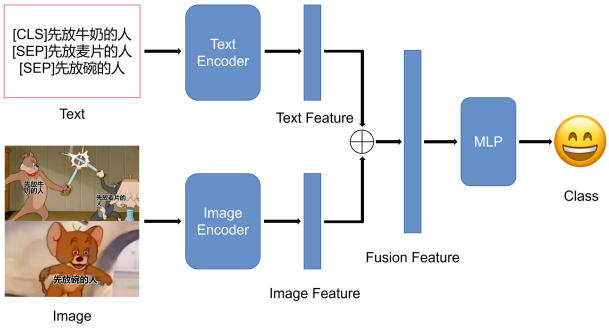

Fig. 3. The structure of our baseline model.

After obtaining both modal feature vectors, we applied the concatenation method as the modality fusion strategy. Subsequently, we added a two-layer MLP after the fusion vector with dropout and ReLU activation function. We used cross-entropy loss function, Adam optimizer and trained models with the early-stop strategy. The train, development and test sets were the standard ones introduced in Sect. 4. We also tested the performance of single modal models by directly connecting the MLP layer with the single modal feature vector. In addition, we introduced image modal information using a different approach, i.e., joining the image caption with the text in memes. This approach only changes the input of the text with caption rather than using the image encoder.

5.2 Results and Analysis

Table 3 shows the experimental results. The results of using normal labels and random labels demonstrate that various models can learn certain knowledge in humor, excluding using image only, which answers **Q1**. Experimental results for using text only show that text modality has a relatively good ability to recognize humor individually. In contrast, the performance of using single image modality is much lower than other models, proving that our annotation scheme, to a large extent, avoids the interference of irrelevant features of image modality, which answers **Q2**. However, it still brings improvement compared to the random cases, indicating that image models have learned some patterns beyond our expectation, which is worth further study.

Table 3. The performance of the baseline models.

Type	Model		Normal		Random	
	Text	Image	ACC	F1	ACC	F1
Text	BERT	–	48.51	40.65	–	–
	RoBERTa	–	50.41	42.91	37.40	24.55
	MacBERT	–	49.45	42.48	–	–
Image	–	ResNet-50	40.38	27.37	–	–
	–	XCiT	46.07	32.61	43.22	19.98
	–	BEiT	43.22	30.80	–	–
	–	Faster-RCNN	41.73	22.26	–	–
Text+Caption	BERT	–	49.59	43.14	–	–
	RoBERTa	–	49.32	43.83	38.48	24.77
	MacBERT	–	50.00	41.46	–	–
Text+Image	RoBERTa	ResNet-50	51.08	45.82	–	–
	RoBERTa	XCiT	49.32	**46.18**	42.82	27.11
	RoBERTa	BEiT	**52.30**	45.33	–	–
	RoBERTa	Faster-RCNN	50.54	43.31	–	–
Text&Image	WenLan	WenLan	50.81	43.02	45.25	16.26

The introduction of image modality brings improvement in most cases, either by introducing image captions in a text form or by introducing image features, which answers **Q3**. In terms of image features, the overall features extracted by CNNs or ViTs work better than the ROIs features extracted by Faster-RCNN, probably because the ROIs features do not interact well with the text features, or the Faster-RCNN weights are frozen in the training process, which answers **Q4**. Furthermore, the improvement of the cross-modal pre-trained model is not significant for our humor recognition task, probably because of its lack of humor-related domain knowledge, which answers **Q5**.

In summary, the baseline models perform not as well as we expected, showing that multimodal humor recognition is a challenging task. Understanding humor needs to match the context to the punchline. However, the context or punchline in a meme may be in the image or text, and combining them correctly can be difficult. In addition, text corresponds to the adjacent object in many memes, but existing methods are difficult to establish this relationship, which brings incoherence to the semantics.

6 Conclusion

This paper presents a novel multimodal humor dataset, Memeplate, which contains 203 templates and 5,184 memes with manually labeled humor levels. It can be used for humor recognition and humor generation research. Unlike previous

datasets, we introduce templates that change the relationship between image and text from one-to-one to one-to-many. Since the template controls the variables of image modality, it is more convenient to extend linguistic findings on humor mechanisms to multimodal humor research. And it provides examples closer to human behavior for humor generation tasks. We offer multiple baseline results for the humor recognition task, which confirm the validity of our dataset and the importance of combining multimodal cues. We hope Memeplate will provide future researchers with valuable multimodal training data and contribute to the development of automatic humor understanding systems.

Acknowledgements. This work is supported by National Natural Science Foundation of China (NSFC) Program (No. 62076046). And we would like to thank the anonymous reviewers for their insightful and valuable comments.

References

1. Ahuja, V., Bali, T., Singh, N.: What makes us laugh? Investigations into automatic humor classification. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 1–9 (2018)
2. Ali, A., et al.: XCiT: cross-covariance image transformers. Adv. Neural Inf. Process. Syst. **34**, 20014–20027 (2021)
3. Bao, H., Dong, L., Wei, F.: BEiT: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
4. Blinov, V., Bolotova-Baranova, V., Braslavski, P.: Large dataset and language model fun-tuning for humor recognition. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4027–4032 (2019)
5. Bonheme, L., Grześ, M.: SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 804–816 (2020)
6. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A crowd-annotated Spanish corpus for humor analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, pp. 7–11 (2018)
7. Castro, S., Cubero, M., Garat, D., Moncecchi, G.: Is this a joke? Detecting humor in Spanish tweets. In: Montes-y-Gómez, M., Escalante, H.J., Segura, A., Murillo, J.D. (eds.) IBERAMIA 2016. LNCS (LNAI), vol. 10022, pp. 139–150. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47955-2_12
8. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 657–668 (2020)
9. Hasan, M.K., et al.: 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), pp. 2046–2056 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Huo, Y., et al.: WenLan: bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561 (2021)

12. Kayatani, Y., et al.: The laughing machine: predicting humor in video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2073–2082 (2021)
13. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
14. Khandelwal, A., Swami, S., Akhtar, S.S., Shrivastava, M.: Humor detection in English-Hindi code-mixed social media content: Corpus and baseline system. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
15. Krippendorff, K.: Computing Krippendorff's alpha-reliability (2011)
16. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
18. Mihalcea, R., Strapparava, C.: Making computers laugh: investigations in automatic humor recognition. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 531–538 (2005)
19. Mihalcea, R., Strapparava, C., Pulman, S.: Computational models for incongruity detection in humour. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 364–374. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12116-6_30
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28** (2015)
21. Sharma, C., et al.: Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 759–773 (2020)
22. Tseng, Y.H., Wu, W.S., Chang, C.Y., Chen, H.C., Hsu, W.L.: Development and validation of a corpus for machine humor comprehension. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1346–1352 (2020)
23. Wu, J., Lin, H., Yang, L., Xu, B.: MUMOR: a multimodal dataset for humor detection in conversations. In: Wang, L., Feng, Y., Hong, Yu., He, R. (eds.) NLPCC 2021. LNCS (LNAI), vol. 13028, pp. 619–627. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88480-2_49
24. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2367–2376 (2015)
25. Zhang, D., Zhang, H., Liu, X., Lin, H., Xia, F.: Telling the whole story: a manually annotated Chinese dataset for the analysis of humor in jokes. 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), pp. 6402–6407 (2019)
26. Zhang, R., Liu, N.: Recognizing humor on twitter. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 889–898 (2014)
27. Ziser, Y., Kravi, E., Carmel, D.: Humor detection in product question answering systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 519–528 (2020)