MUMOR: A Multimodal Dataset for Humor Detection in Conversations

Jiaming Wu, Hongfei Lin, Liang Yang(✉), and Bo Xu

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

Abstract. Humor detection attracts increased attention in natural language processing for its potential applications. Prior work focus on analyzing humor on isolated, textual data, but humor usually comes from the interaction among speakers in a multimodal way. In this paper, we proposed a novel dataset named MUMOR, which consists of multimodal dialogues in both English and Chinese. It contains a total of 29,585 utterances belonging to 1,298 dialogues from two TV-sitcoms. We manually annotated each utterance with humor, emotion, and sentiment labels. To our best knowledge, this is the first corpus containing Chinese conversations for humor detection. This dataset could be used for research on humor detection, humor generation, and multi-task learning on emotion and humor analysis. We released this dataset publicly.

Keywords: Mulimodal · Sentiment analysis · Humor detection

1 Introduction

Humor plays an important role in human communication. It not only creates an entertaining atmosphere, but also helps regulate conversations, reduce stress, and build trust between partners [1]. Humor recognition attracts increased attention for its potential application in human-computer interaction, which can be used in advertising, healthcare and education area [2].

Recent years, there are many papers about construction of humor detection dataset. Mihalcea and Strapparava [3] introduced a dataset containing 16,000 humorous and 16,000 non-humorous text in English. The humorous data are one-liners collected from Internet, and the non-humorous sentences were collected from one-liners, Reuters titles, BNC sentences, and proverbs. Zhang and Liu [4] constructed an English tweets dataset for recognizing humor on Twitter. The corpus contains 3,000 tweets with annotation of 3 categories: humorous tweets, non-humorous tweets, and humorous non-tweets. Castro et al. [5] established a Spanish corpus for humor detection. They collected 39,363 tweets from both humorous accounts and non-humours accounts. After filtering and manually labeling, they finally got a corpus containing 33,531 Spanish tweets with humorous and non-humorous labels. Khandelwal et al. [6] proposed a corpus with 3,543 English-Hindi code-mixed tweets. Each tweet is annotated with humorous or non-humorous label. Blinov et al. [7] constructed a large size dataset for humor

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13028, pp. 619–627, 2021.
https://doi.org/10.1007/978-3-030-88480-2_49

recognition in Russian. They collected jokes and funny dialogues from various online resources. This dataset contains more than 300,000 short tests, which is significantly larger than any previous humor-related corpus.

Most of the previous work in the field of humor detection focusing on isolated, textual data, few works have been devoted to detecting humor in conversations. Bertero and Fung [8] proposed a LSTM based network to detect humor in dialogues. They constructed a corpus with 43,672 utterances from 1,589 scenes. The corpus is collected from the subtitles and scripts of the TV-sitcom *"The Big Bang Theory"*. In their later work [9], they combined acoustic and language features to detect humor in conversations. Experiment results on three English sitcom corpus showed that combining the acoustic features brought improvement on humor detection.

It is important and difficult to detect humor in conversations. The formation of humor has a setup process. Sometimes one sentence itself is not humorous, but it becomes humorous when combined with the context. And the interaction between speakers in a dialogue often produces humorous effect. In addition, multimodal information also helps detect humor. Sometimes the reason makes an utterance funny is the vocal tonality, facial expressions, and body gesture of the speaker but not the meaning of the utterance text. And a multimodal dialogue scene is a common scenario in reality. Detecting humor in multimodal dialogue is a very challenging task. It requires the model to obtain the contextual information and integrate the features of different modalities.

Figure 1 shows an example of humor in multimodal dialogue. The topic of this dialogue is quitting smoke. Each utterance in this conversation, if being treated separately, is not humorous. However, considering the contextual information the emotional changes of the characters in the dialogue, utterance 3, 4, 6 become humorous.

In this work, we constructed *Multimodal Utterance-level Humor Dataset* (MUMOR), a dataset for humor detection in multimodal conversations. It contains two language corpora: English and Chinese. Both corpora contain textual dialogues with their corresponding video and audio segments, which means each utterance in this dataset has three modal sources. MUMOR provides humor label for recognizing humor in dialogues. Furthermore, each utterance is annotated with emotion and sentiment labels, which can be used in multi-task learning on emotion and humor analysis as research has shown that modeling sentiment is effective for humor detection [10]. Our contributions are as follows.

- We proposed a dataset, MUMOR, contains both English and Chinese corpus. To our knowledge, this is the first Chinese conversational corpus for humor detection.
- We introduced the data processing and labeling process of this dataset, which has reference significance for the construction of multi-modal data set. This paper shows the data distribution and statistical information of MUMOR.
- MUMOR provides audio, visual, and textual modal sources. It is a multilingual, multi-label dialogue dataset. It can be used for multi-modal sentiment analysis, humor recognition and dialogue generation research.

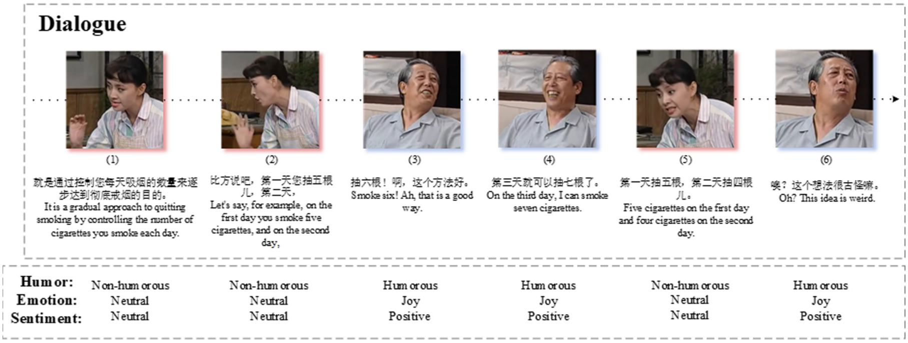

Dialogue
(1)
就是通过控制您每天吸烟的数量来逐步达到彻底戒烟的目的。
It is a gradual approach to quitting smoking by controlling the number of cigarettes you smoke each day.
(2)
比方说吧，第一天您抽五根儿，第二天，
Let's say, for example, on the first day you smoke five cigarettes, and on the second day,
(3)
抽六根！啊，这个方法好。
Smoke six! Ah, that is a good way.
(4)
第三天就可以抽七根了。
On the third day, I can smoke seven cigarettes.
(5)
第一天抽五根，第二天抽四根儿。
Five cigarettes on the first day and four cigarettes on the second day.
(6)
唉？这个想法很古怪嘛。
Oh? This idea is weird.
Humor: Non-humorous Non-humorous Humorous Humorous Non-humorous Humorous
Emotion: Neutral Neutral Joy Joy Neutral Joy
Sentiment: Neutral Neutral Positive Positive Neutral Positive

Fig. 1. A example of humor in multimodal dialogue.

2 Dataset

2.1 Data Source

The MUMOR dataset contains English and Chinese corpus which we name as MUMOR-EN and MUMOR-ZH, respectively.

MUMOR-EN is constructed based on the MELD dataset [11]. MELD is a multimodal dataset used for emotion recognition task. It contains 1,433 dialogues from TV-sitcom *"Friends"*. Each dialogue contains several utterances belonging to the same scene. And each utterance encompasses audio, visual and textual modalities. Since the purpose of our dataset is to recognize humor in long conversations rather than short texts, we discarded the conversations with a small number of utterance in the original dataset. We filtered out dialogues with less than 3 utterances and made humor annotation on the rest data.

For MUMOR-ZH, we choose a popular Chinese sitcom "我爱我家" (I Love My Family) as our data source. We collected 81 episodes of this TV series video and extracted utterance text and its timestamps from subtitle files. We cut the video into clips according to the timestamps of each utterance. The utterances are grouped into dialogues following the constraint that all the utterances in a dialogue comes from the same episode and scene. Finally, we got 19,103 utterances belonging to 519 dialogues.

2.2 Data Format

Each utterance is identified by a dialogue ID and an utterance ID, which also name the corresponded video clip file saved in *.mp4* format.

In Table 1, we show the format of our dataset, which contains the information of the utterance, the speaker, the humor label, the emotion label, the sentiment label, the dialogue ID and the utterance ID.

Table 1. Dataset format.

Utterance	Speaker	Humor	Emotion	Sentiment	D_ID	U_ID
All right, there you go!	Ross	Non-humorous	Joy	Positive	439	12
Yeah, you hang in there Teddy!	Joey	Non-humorous	Anger	Negative	439	13
I'm Andrew, and I didn't pay for this pear.	Older Scientist	Humorous	Neutral	Neutral	439	14
Okay, good-good for you.	Ross	Non-humorous	joy	Positive	439	15
I'm Rhonda, and these aren't real!	Tour Guide	Humorous	Neutral	Neutral	439	16

2.3 Data Annotation

The MUMOR dataset contains humor, emotion, and sentiment labels for each utterance. We ask three annotators to watch the video clips with subtitles of each utterance, and let them decide whether this utterance is humorous or not and which kind of emotion it belongs to.

The annotators are Chinese postgraduate students with at least 10 years of English learning experience. In addition, we displayed both English and translated Chinese subtitles while annotating English data. Before the formal annotation, all annotators did some pre-annotation test to ensure the quality of the annotation.

For humor label, we provide two categories: *humorous* and *non-humorous.* The overall Fleiss' kappa score of humor annotation process is 0.81, which indicates a substantial agreement among annotators.

For emotion label, We keep the original emotion labels in MELD dataset for MUMOR-EN, and make our emotion annotation by 3 annotators on MUMOR-ZH. The emotion label contains six universal emotions *Joy*, *Sadness*, *Fear*, *Anger*, *Surprise*, and *Disgust* [12] in addition with *Neutral.* For utterances that 3 annotators can not reach agreement, we label it with a *None* label. There exists 410 *None* in totally 19,103 utterances. The Fleiss' kappa score of emotion task is 0.45 (kappa of MELD emotion annotation process is 0.43).

For sentiment label, we apply the scheme proposed by Poria et al. [11]. It considered *Anger*, *Disgust*, *Fear*, *Sadness* as *Negative*, *Joy* as *Positive* and made further annotation in class *Surprise* to decide whether is *Positive* or *Negative.* Our 3 annotators labeled the utterances in MUMOR with the tactic mentioned above. The Fleiss' kappa score of sentiment annotation process is 0.84.

3 Dataset Analysis

First, we display the main statistical information of the two language data in Table 2.

It can be seen that the scale of the Chinese data is about twice that of the English data, showing a roughly 2:1 ratio between the total duration of the video and the total number of utterance. The length of an utterance is the

Table 2. Data statistics.

	MUMOR-EN	MUMOR-ZH
Total video duration (h)	9.03	18.12
Avg duration of utterance (s)	3.11	3.42
# Dialogues	779	519
# Utterances	10,482	19,103
D-length	13.46	36.81
U-length	10.37	10.04
Humorous percentage (%)	24.59	28.36
# Speakers	259	91
Total number of words	108746	191770
Number of unique words	6869	17804

number of words in it, this number on the Chinese and English data is very close, being 10.04 and 10.37 respectively. The average duration of one utterance is also relatively close, 3.42 s and 3.11 s respectively. It can be seen that the narrative length and speech speed of the actors in the two sitcoms are relatively close. The big difference between the two languages is the length of the conversation. Among them, the value of Chinese data is 36.81, which is close to three times of 13.64 on English data. It indicates that a paragraph in the Chinese data has a longer sequence and contains more context information. In addition, the proportion of humor in the data of different languages is similar, and there is not much difference. The percentage of humor in the Chinese data is 28.36%, and the median value of the English data is slightly lower, at 24.58%, and the ratio of positive and negative cases is about 3:1. The Chinese data contains a total of 191,770 words, of which the size of the non-repeated vocabulary is 17,804. In contrast, there are only 6,869 different words in the English data. It can be seen that the vocabulary variety of the Chinese data is higher than that of the English data.

We split our dataset into training, development, and testing set on two corpus, respectively. Table 3 shows the data statistics on the 3 sets. It can be seen that the main statistical information on the training, development, and testing set is very close.

Figure 2 shows the distribution of humor percentage in one dialogue on two corpora. There are 9 Chinese dialogues that do not contain humorous utterances, while the number is 20 in English corpus. The proportion of humorous utterances in most conversations is 10% to 40%. While the 20%-30% range had the largest number of dialogues, with 252 dialogues in the Chinese corpus and 192 dialogues in the English corpus.

We also calculate the utterance proportion for all sitcom characters. For those with less than 2% utterances, we group them as *Other*. The result is shown at Fig. 3. We can see both corpus contain 6 main characters, and the utterance proportion of main characters in the English corpus is more balanced.

Table 3. Dataset division.

	MUMOR-EN			MUMOR-ZH		
	Train	Dev	Test	Train	Dev	Test
# Dialogues	551	70	158	348	43	128
# Utterances	7,472	914	2,096	12,677	1,632	4,794
D-length	13.56	13.06	13.27	36.43	37.95	37.45
U-length	10.34	10.20	10.59	9.99	10.49	10.02
Humorous percentage %	23.73	27.02	26.62	28.76	26.23	28.04
# Speakers	215	36	75	76	24	51

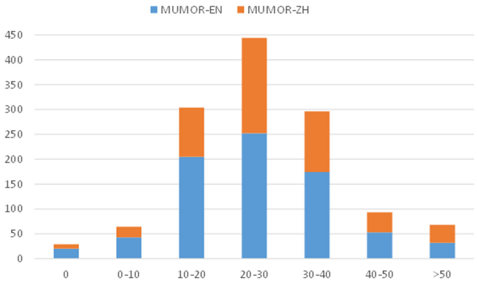


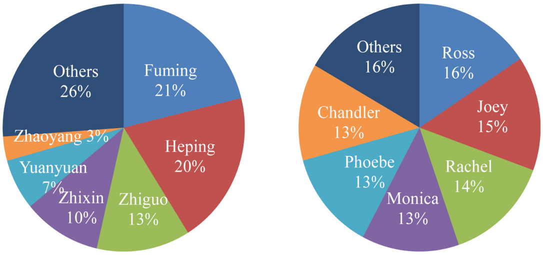
Fig. 2. Humor distribution.

Figure 4 shows the humorous percentage of each character in two corpora, respectively. We can see some characters with lower utterance proportion but have higher humor percentage, they played an important role in pleasing the audience, like *Yuanyuan* and *Chandler* in their respective sitcoms. Both of them have a humorous percentage of over 33%.

4 Comparison with Existing Dataset

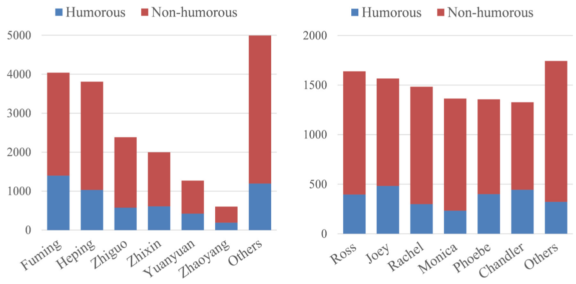
In this section, we will compare MUMOR dataset with two related multimodal datasets and introduce the potential applications of our dataset.

UR-FUNNY [13] is a multimodal dataset for humor detection. It contains 16,514 speech data extracted from TED speech. Each speech segment contains several utterances and labeled with humours or non-humours label. The positive instances end with a punchline utterance, the negative instances sampled from sentences in the same distribution but not end with a punchline. UR-FUNNY dataset is making classification on dialogue level while MUMOR dataset works on utterance level. From the example in Fig. 1, we can see that compared with the punchline in the speech, the humor in the dialogue does not only appear in the ending utterance, but is distributed in the whole dialogue. Furthermore, the average length of context in UR-FUNNY dataset is 2.86 which is much shorter



(a) MUMOR-ZH (b) MUMOR-EN

Fig. 3. Character distribution.



(a) MUMOR-ZH (b) MUMOR-EN

Fig. 4. Humor distribution.

than that in MUMOR dataset, which means models need powerful context modeling capabilities to achieve good results on MUMOR dataset.

MUStARD [14] is a multimodal dataset for sarcasm detection. Its data are extracted from three English sitcoms. Sarcasm utterance in this dataset is accompanied with several historical utterances as its context which is much shorter than the length of context in MUMOR dataset. Unlike MUStARD dataset, MUMOR dataset focus on humor detection. Although there is a strong relationship between sarcasm and humor, humor does not only come from sarcasm. Table 4 shows the comparison between MUMOR dataset and the existing dataset.

Table 4. Comparison with existing dataset.

	MUMOR	UR-FUNNY	MUStARD
Number of videos	1298	16514	690
Avg duration of utterance(s)	3.31	4.64	5.22
Annotation granularity	Utterance-level	Dialogue-level	Dialogue-level
Labels	Humor, sentiment, emotion	Humor	Sarcasm
Languages	English & Chinese	English	English

MUMOR dataset can be used in the research of detecting humor in conversations involving multiple speakers. It can also be used to study humor differences in multiple languages as it provides corpus in both Chinese and English. In addition, the emotion and sentiment labels can be used to analysis the relationship between emotion and humor through multi-task learning.

5 Conclusion and Future Work

In this work, we constructed MUMOR, a multimodal dialogue dataset. It provides two language corpus: English and Chinese. It totally contains 29,585 utterances from 1,298 dialogues from two sitcoms. Each utterance in MUMOR has textual, audio, and visual modal sources. We introduced the process of building this dataset and the kappa score indicated a high quality of our dataset.

Our dataset provided emotion, sentiment and humor label. Moreover, it can be used for emotion recognition, humor response generation, and multi-task learning on emotion and humor analysis. In addition, research about multimodal feature extraction and fusion can be explored on our dataset.

References

1. Morse, D.: Use of humor to reduce stress and pain and enhance healing in the dental setting. J. N.J. Dent. Assoc. **78**(4), 32–36 (2007)
2. Nijholt, A., Niculescu, A.I., Alessandro, V., Banchs, R.E.: Humor in human-computer interaction: a short survey (2017)
3. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Vancouver, Canada, 6–8 October 2005, pp. 531–538 (2005). https://www.aclweb.org/anthology/H05-1067/
4. Zhang, R., Liu, N.: Recognizing humor on twitter. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, 3–7, November 2014, pp. 889–898 (2014). https://doi.org/10.1145/2661829.2661997, https://doi.org/10.1145/2661829.2661997
5. 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
6. 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, Miyazaki, Japan, 7–12, May 2018 (2018). http://www.lrec-conf.org/proceedings/lrec2018/summaries/363.html
7. Blinov, V., Bolotova-Baranova, V., Braslavski, P.: Large dataset and language model fun-tuning for humor recognition. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Long Papers, vol. 1, pp. 4027–4032. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1394

8. Bertero, D., Fung, P.: A long short-term memory framework for predicting humor in dialogues. In: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, 12–17, June 2016, pp. 130–135 (2016). https://www.aclweb.org/anthology/N16-1016/
9. Bertero, D., Fung, P.: Multimodal deep neural nets for detecting humor in TV sitcoms. In: 2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, 13–16, December 2016, pp. 383–390 (2016). https://doi.org/10.1109/SLT.2016.7846293
10. Liu, L., Zhang, D., Song, W.: Modeling sentiment association in discourse for humor recognition. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20, July 2018, Short Papers, vol. 2, pp. 586–591 (2018). https://doi.org/10.18653/v1/P18-2093, https://www.aclweb.org/anthology/P18-2093/
11. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Long Papers, vol. 1, pp. 527–536 (2019). https://www.aclweb.org/anthology/P19-1050/
12. Ekman, P., Friesen, W.V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. **53**(4), 712 (1987)
13. Hasan, M.K., et al.: UR-FUNNY: a multimodal language dataset for understanding humor. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, 3–7, November 2019, pp. 2046–2056. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1211
14. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S.: Towards multimodal sarcasm detection (an _obviously_ perfect paper). In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Long Papers, vol. 1, pp. 4619–4629. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1455