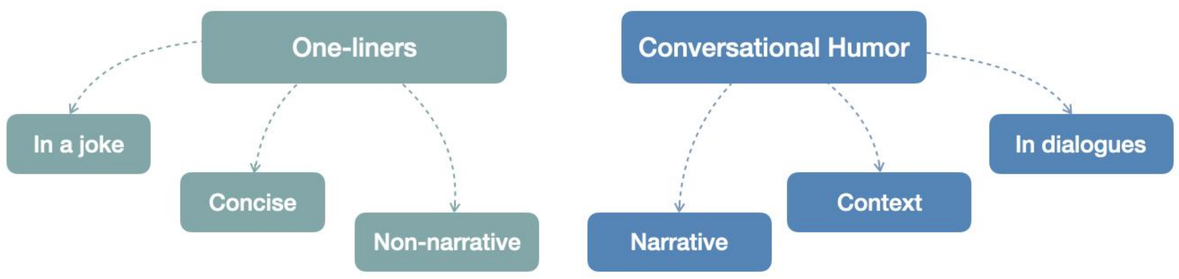
MUCH: A Multimodal Corpus Construction for Conversational Humor Recognition Based on Chinese Sitcom

Hongyu Guo[1], Wenbo Shang[2], Xueyao Zhang[1], Shubo Zhang[1], Xu Han[3,*], Binyang Li[1,*]
[1]University of International Relations, Beijing, China
[2]Hong Kong Baptist University, Hong Kong, China
[3]Capital Normal University Information Engineering College, Beijing, China

1. Introduction

Humor, an essential element in interpersonal communication, serves as a vital medium for expressing emotions in humans.

There are two main forms of humorous expression:

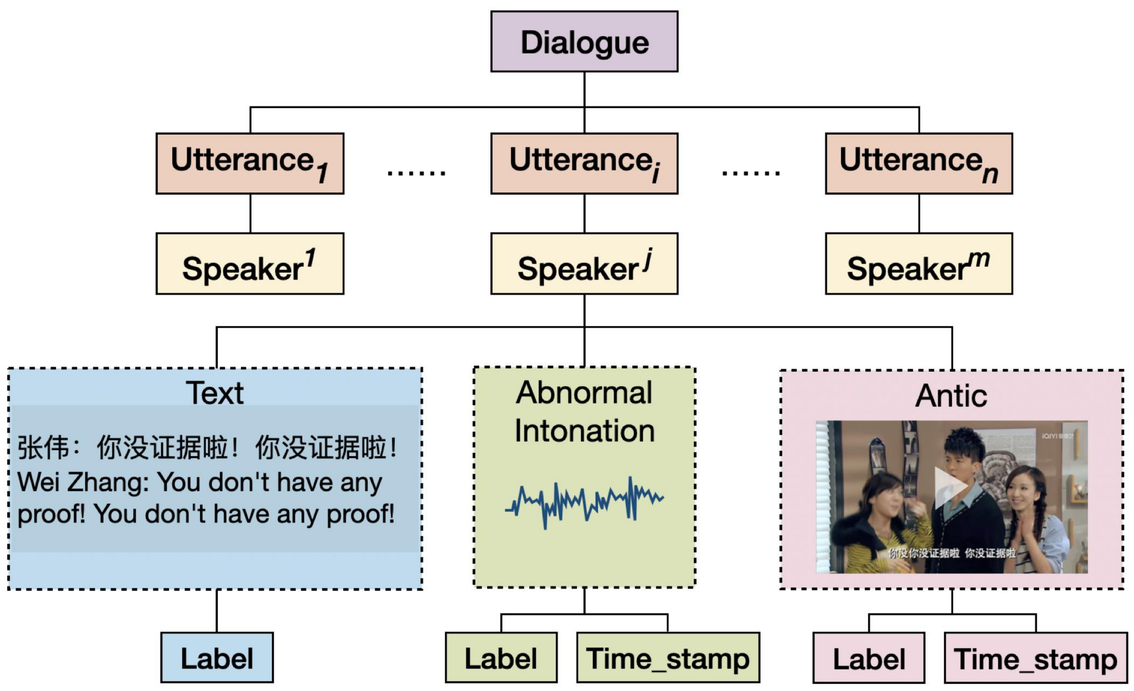


Different from one-liners, conversational humor is generated based on the context of the dialogue and expressed more flexibly in conversations. Therefore, conversational humor recognition is significant for capturing the humorous semantics and dialogue comprehension.

Most of the existing research mainly used textual modality to recognize conversational humor, which made it difficult to identify conversational humor from unimodal. In the real world, a conversation between individuals is usually conducted face-to-face. Therefore, in addition to text, other modalities will be used to generate humor, such as funny facial expressions and quirky intonations. However, existing multimodal datasets did not cover all modalities, they are coarse-grained in capturing the differentiation of multimodalities and perform suboptimal in conversational humor recognition.

Therefore, in order to better recognize conversational humor, this paper constructed a multimodal humor corpus annotation scheme and manually annotated a fine-grained conversational humor corpus (**MUCH**) based on a Chinese sitcom, *iPartment*.

2. Annotation Scheme



Speaker plays a particular role in the generation of humor, and different speakers have different styles of humor.

Text represents the textual content of the utterance that indicates the specific manifestation of humor in the textual modality. This attribute is labeled as *1* when textual humor occurs and labeled as *0* otherwise.

Abnormal Intonation can enhance the speaker's emotional expression. This attribute is labeled as *1* when an abnormal intonation occurs and labeled as 0 otherwise.

Antic refers to the characters exhibit comical expressions or gestures during the conversation. When comical expressions or gestures occur, Antic is labeled as *1*; otherwise, it is labeled as *0*. It allows for humor recognition in the visual modality through the annotation of comical actions and expressions.

3. Annotation Process

1. Divide each episode into several dialogues based on different plots and scenes.
2. Record all the utterances in each dialogue and record the speaker and the content of each utterance.
3. Judge whether each utterance embodies humor in textual, visual, and acoustic modalities following the proposed scheme.

When humor is expressed visually or acoustically, the time_stamp of the corresponding utterance is also provided.

4. Statistics

Filed		Value
# Dialogue		1,626
# Utterance		34,804
# Speaker		423
# Humorous utterance		7,079
Total duration in hour		62
Avg. duration of dialogue (minutes)		2.76
Avg. duration of utterance (seconds)		2.89
# Humor in unimodal	T	3,661
	V	647
	A	855
# Humor in multimodal	T+V	615
	T+A	514
	V+A	347
	T+V+A	440

The MUCH corpus consists of 4 seasons of iPartment, 80 episodes in total, with each episode approximately 45 minutes long. It consists of 34,804 utterances in total, and 7,079 of them are humorous. Of the humor utterances, 5,163 utterances generate humor by unimodal expressions, and 1,916 required two or more modalities to generate humor.

Dataset	Annotation Process	Modalities	Language
UR-Funny	Provide videos and their transcripts from the TED portal.	T, V, A	English
MuStARD	Provide the utterance and the corresponding original fragment, while also proving contextual information.	T, V, A	English
TBBT	Provide the utterance and the corresponding original fragement; Only the overall label.	T, V	English
MHD	Annotation based on canned laughter.	T, V, A	English
M2H2	Provide the utterance and the corresponding original fragment; Only the overall label.	T, V, A	Hindi
MUCH (Ours)	Provide an overall label and labels for each of the three modalities for each utterance.	T, V, A	Chinese

We also compare the MUCH with other current multimodal humor datasets. It can be seen that the corpus annotation schme we proposed can label humor for different modalities separately, not just the overall humor.

5. Experimental Results

Modality		Method	Acc.(%)	P(%)	R(%)	F1(%)
Unimodal	Text	BERT	65.18	60.72	61.44	61.08
		RoBERTa	**79.17**	**70.37**	**65.28**	**67.73**
	Vision	ViT	65.72	65.17	60.25	62.61
		OMNIVORE	60.13	60.37	59.12	59.47
	Acoustic	openSMILLE	56.41	55.93	61.01	58.36
Multimodal		CLIP	**82.96**	**74.86**	**77.05**	**75.94**

The experimental results of the MUCH corpus in both unimodal and multimodal methods demonstrated the applicability of the corpus annotation scheme we have constructed.

Method using multimodality outperformed methods that used unimodalities. Between text and nonverbal behaviors, text proved to be the most important modality. In most cases, multimodal methods are performs better than text alone for humor recognition.