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

Introduction

Conversational Humor

- 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 (Attardo et al., 2013):

One-liners

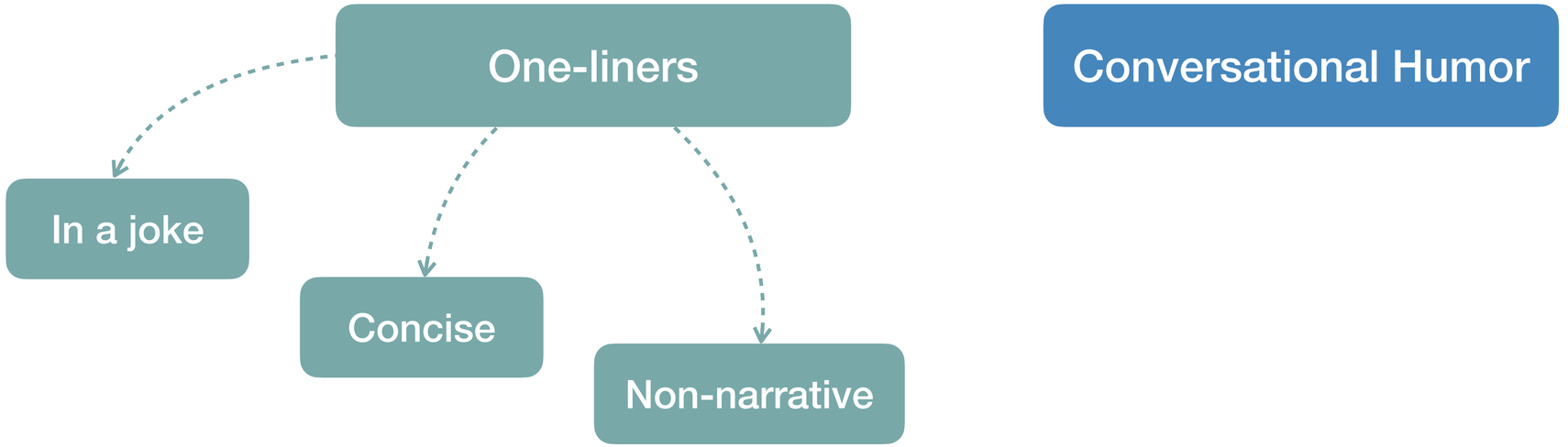
Conversational Humor

[1] Salvatore Attardo, Lucy Pickering, Fofo Lomotey, and Shigehito Menjo. 2013. Multimodality in conversational humor. Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association, 11(2):402416.

Introduction

Conversational Humor

- 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 (Attardo et al., 2013):

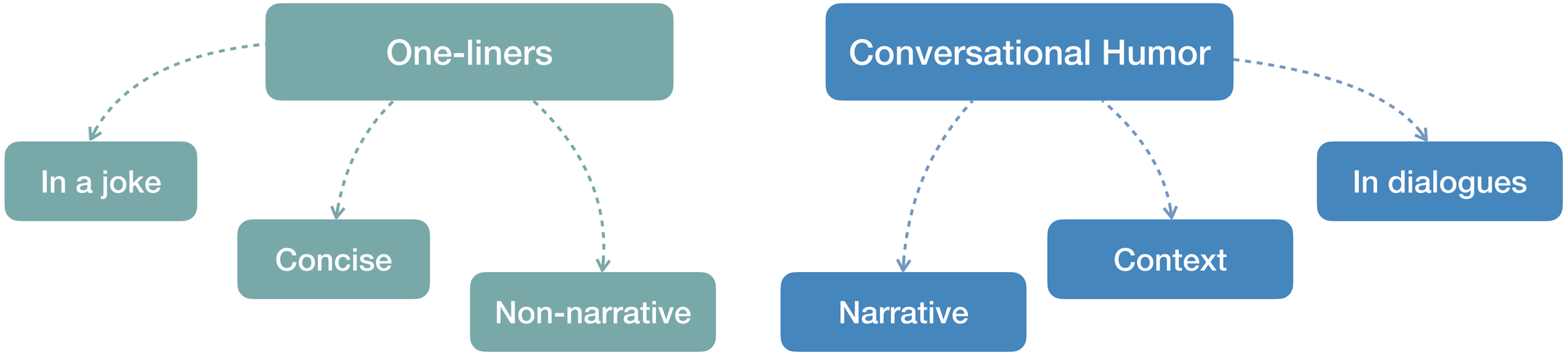


[1] Salvatore Attardo, Lucy Pickering, Fofo Lomotey, and Shigehito Menjo. 2013. Multimodality in conversational humor. Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association, 11(2):402416.

Introduction

Conversational Humor

- 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 (Attardo et al., 2013):



[1] Salvatore Attardo, Lucy Pickering, Fofo Lomotey, and Shigehito Menjo. 2013. Multimodality in conversational humor. Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association, 11(2):402416.

Introduction

Conversational Humor is Multimodal

Introduction

Conversational Humor is Multimodal

Introduction

Conversational Humor is Multimodal

Actors frequently employ humorous language and quirky tones, together with exaggerated expressions and actions to evoke humor.

Related Work

Unimodal

- Zhang and Liu (2014) collected humor dataset from Twitter
- Chen and Lee (2017) collected and annotated TED speech transcripts

[1] Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pages 889–898.

[2] Lei Chen and Chong Min Lee. 2017. Convolutional neural network for humor recognition. arXiv preprint arXiv:1702.02584.

Related Work

Unimodal

- Zhang and Liu (2014) collected humor dataset from Twitter
- Chen and Lee (2017) collected and annotated TED speech transcripts

In conversations, to enhance humorous expressions, people may combine the acoustic rhythm (quirky intonation, etc.) or the visual features (comical expressions or movements, etc.) interacting with the textual contents.

[1] Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pages 889–898.

[2] Lei Chen and Chong Min Lee. 2017. Convolutional neural network for humor recognition. arXiv preprint arXiv:1702.02584.

Related Work

Coarse-grained capture of the differentiation of multimodalities

- Chandrasekaran et al. (2016) analyzed humorous expressions in abstract scenes
- Boccignone et al. (2017) proposed a multimodal dataset to detect humor from images.
- Some research on multimodal conversational humor datasets has been built based on sitcoms, including *Friends* (Poria et al., 2018), *The Big Bang Theory* (Patro et al., 2021), *Seinfeld* (Bertero and Fung, 2016b) etc..

[1] Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4603–4612.

[2] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. Amhuse: a multimodal dataset for humour sensing. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 438–445.

[3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508

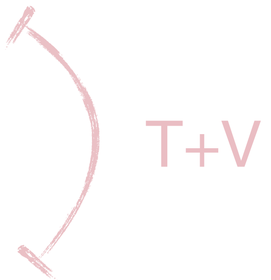
[4] Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 576–585.

[5] Dario Bertero and Pascale Fung. 2016b. Predicting humor response in dialogues from tv sitcoms. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5780–5784. IEEE.

Related Work

Coarse-grained capture of the differentiation of multimodalities

- Chandrasekaran et al. (2016) analyzed humorous expressions in abstract scenes
- Boccignone et al. (2017) proposed a multimodal dataset to detect humor from images.
- Some research on multimodal conversational humor datasets has been built based on sitcoms, including *Friends* (Poria et al., 2018), *The Big Bang Theory* (Patro et al., 2021), *Seinfeld* (Bertero and Fung, 2016b) etc..



[1] Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4603–4612.

[2] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. Amhuse: a multimodal dataset for humour sensing. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 438–445.

[3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508

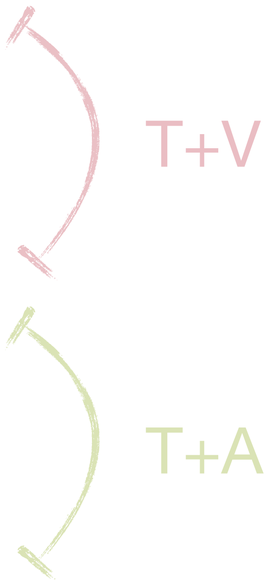
[4] Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 576–585.

[5] Dario Bertero and Pascale Fung. 2016b. Predicting humor response in dialogues from tv sitcoms. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5780–5784. IEEE.

Related Work

Coarse-grained capture of the differentiation of multimodalities

- Chandrasekaran et al. (2016) analyzed humorous expressions in abstract scenes
- Boccignone et al. (2017) proposed a multimodal dataset to detect humor from images.
- Some research on multimodal conversational humor datasets has been built based on sitcoms, including *Friends* (Poria et al., 2018), *The Big Bang Theory* (Patro et al., 2021), *Seinfeld* (Bertero and Fung, 2016b) etc..



[1] Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4603–4612.

[2] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. Amhuse: a multimodal dataset for humour sensing. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 438–445.

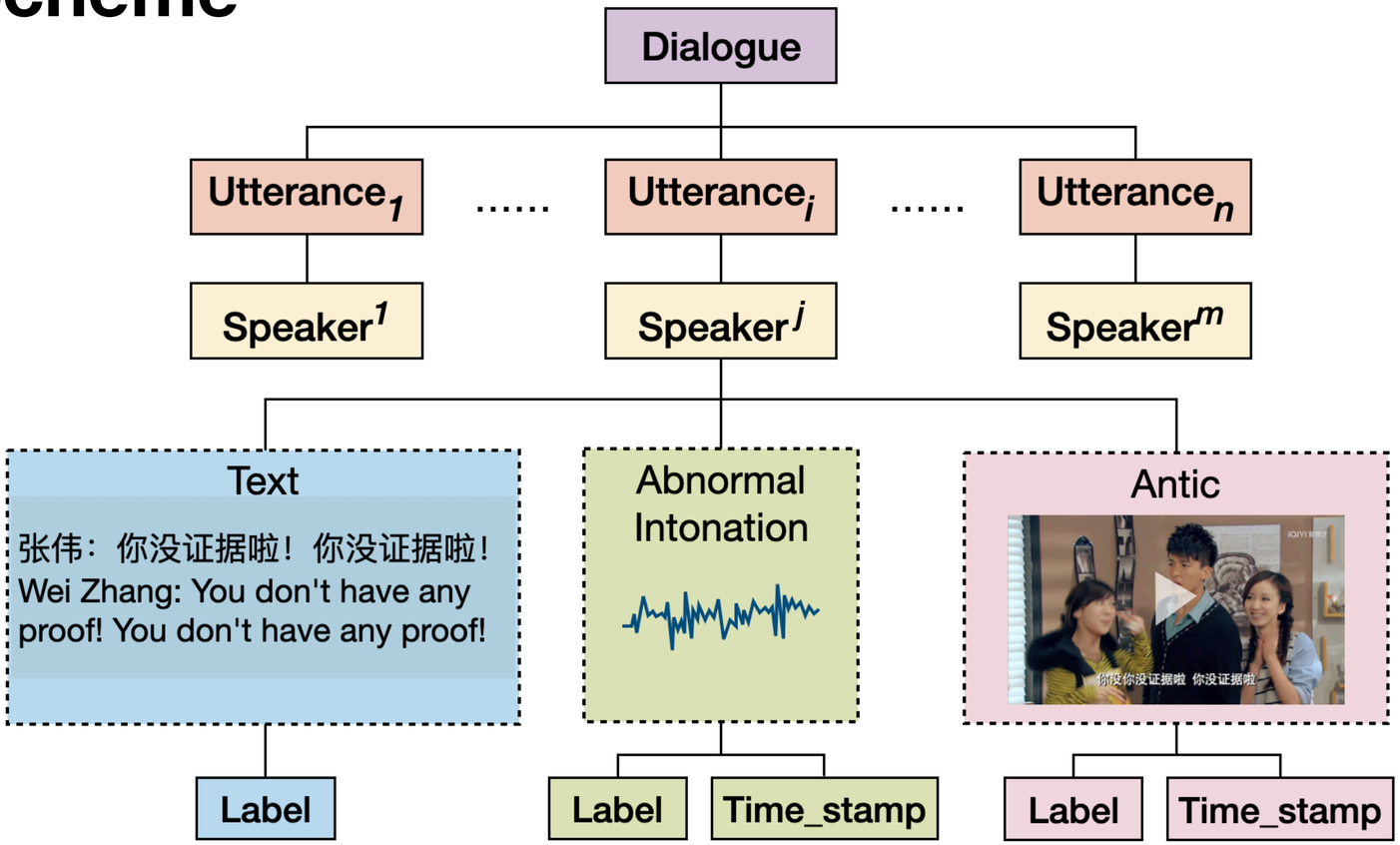
[3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508

[4] Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 576–585.

[5] Dario Bertero and Pascale Fung. 2016b. Predicting humor response in dialogues from tv sitcoms. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5780–5784. IEEE.

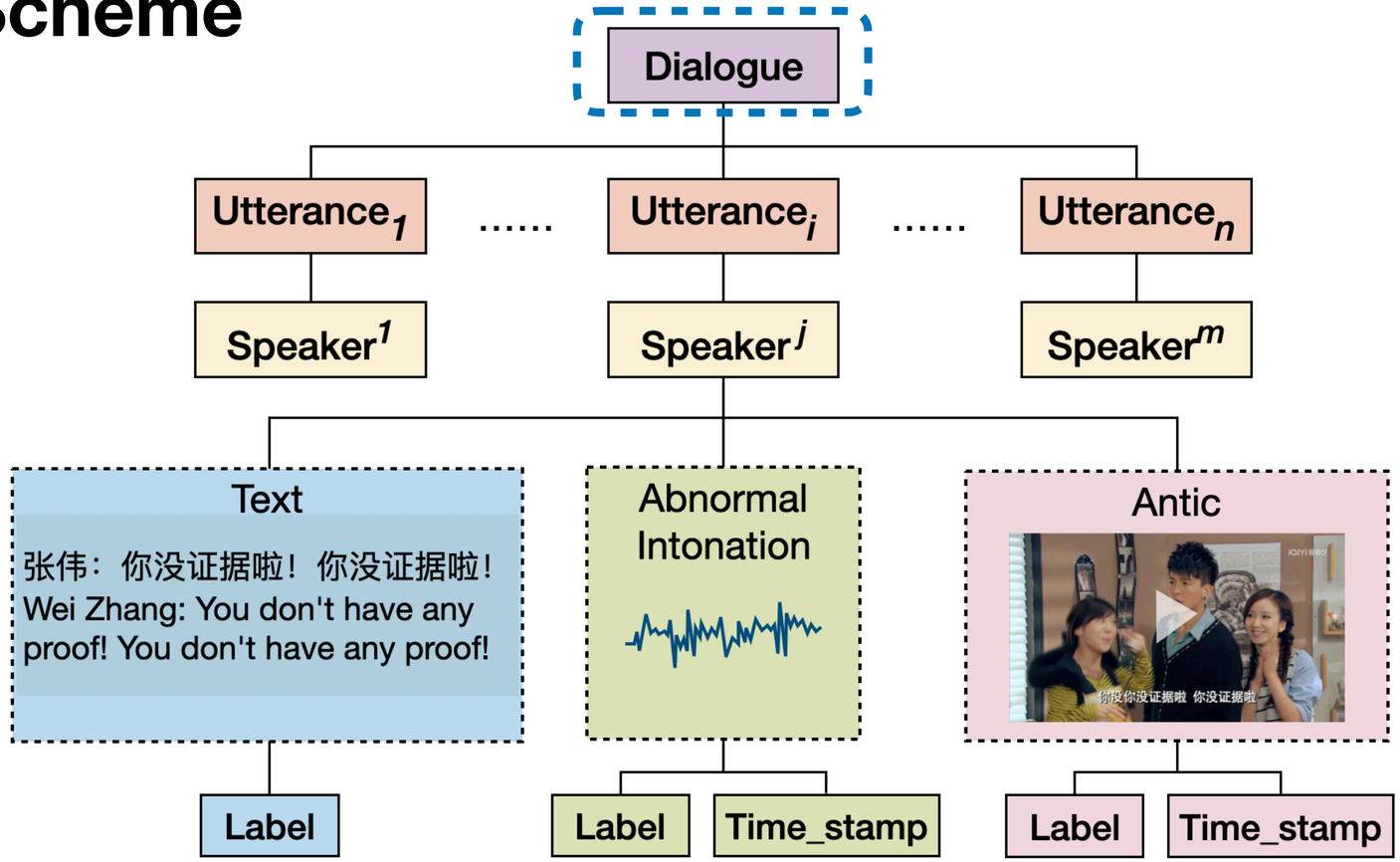
Corpus Construction

Annotation Scheme



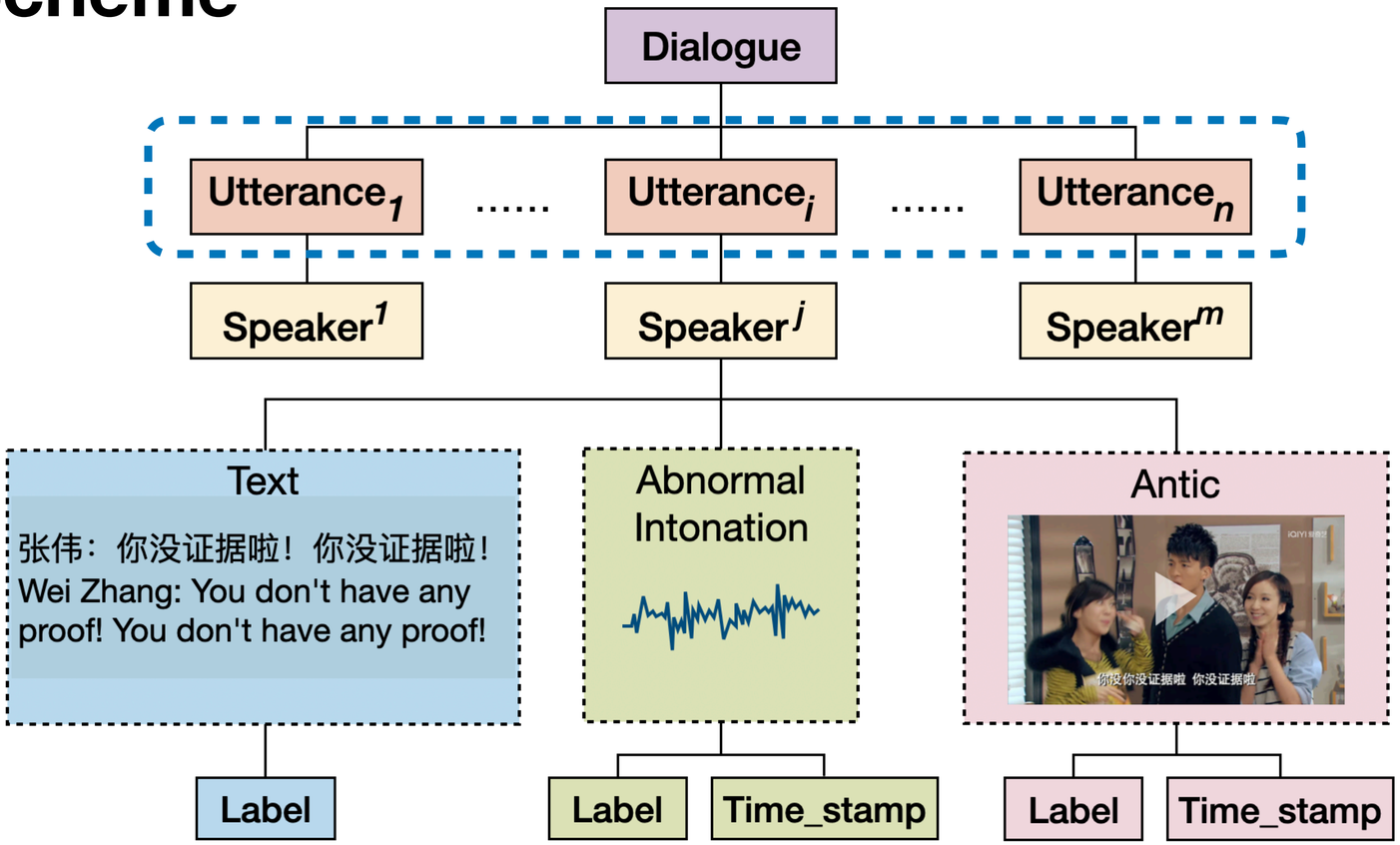
Corpus Construction

Annotation Scheme



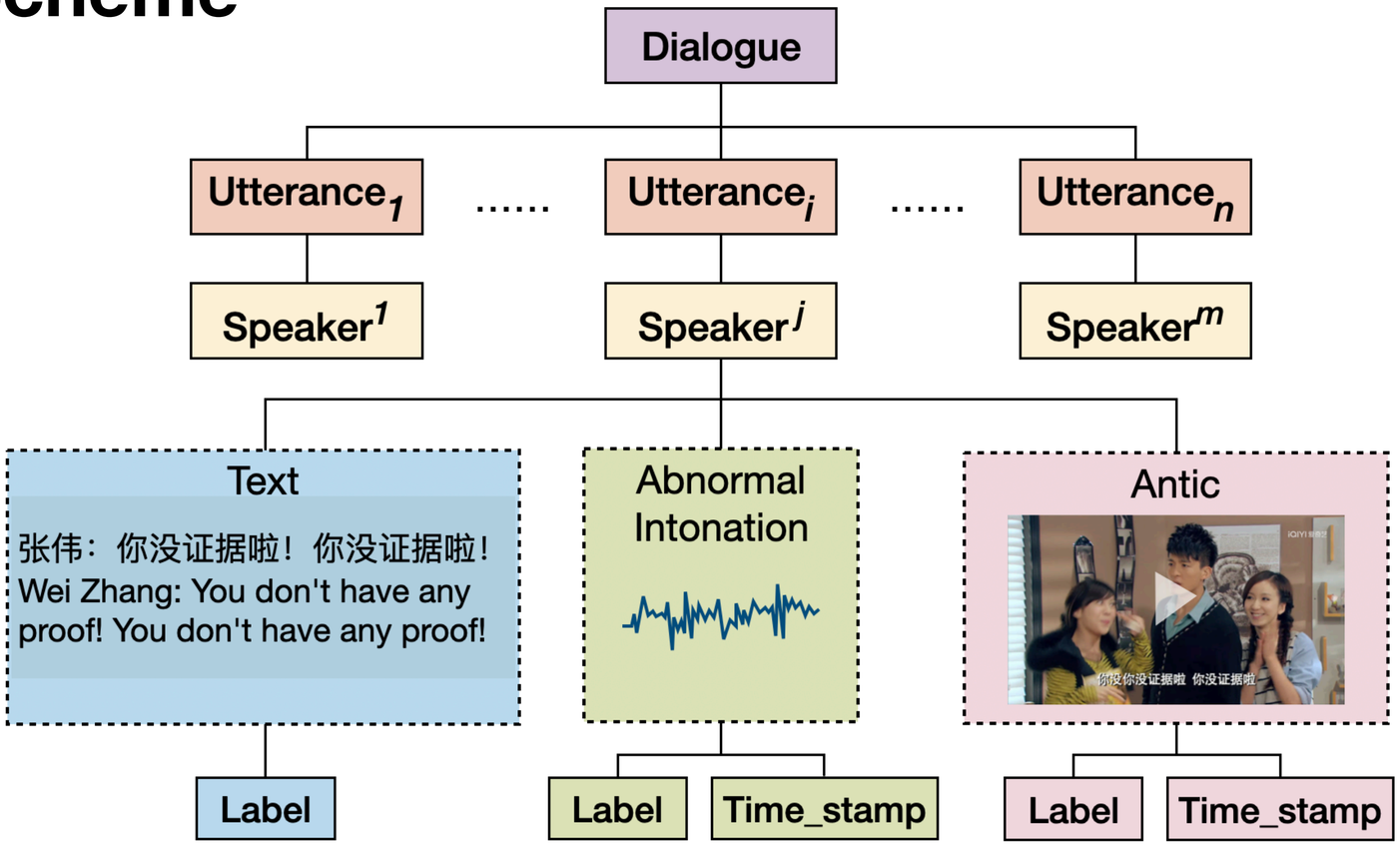
Corpus Construction

Annotation Scheme



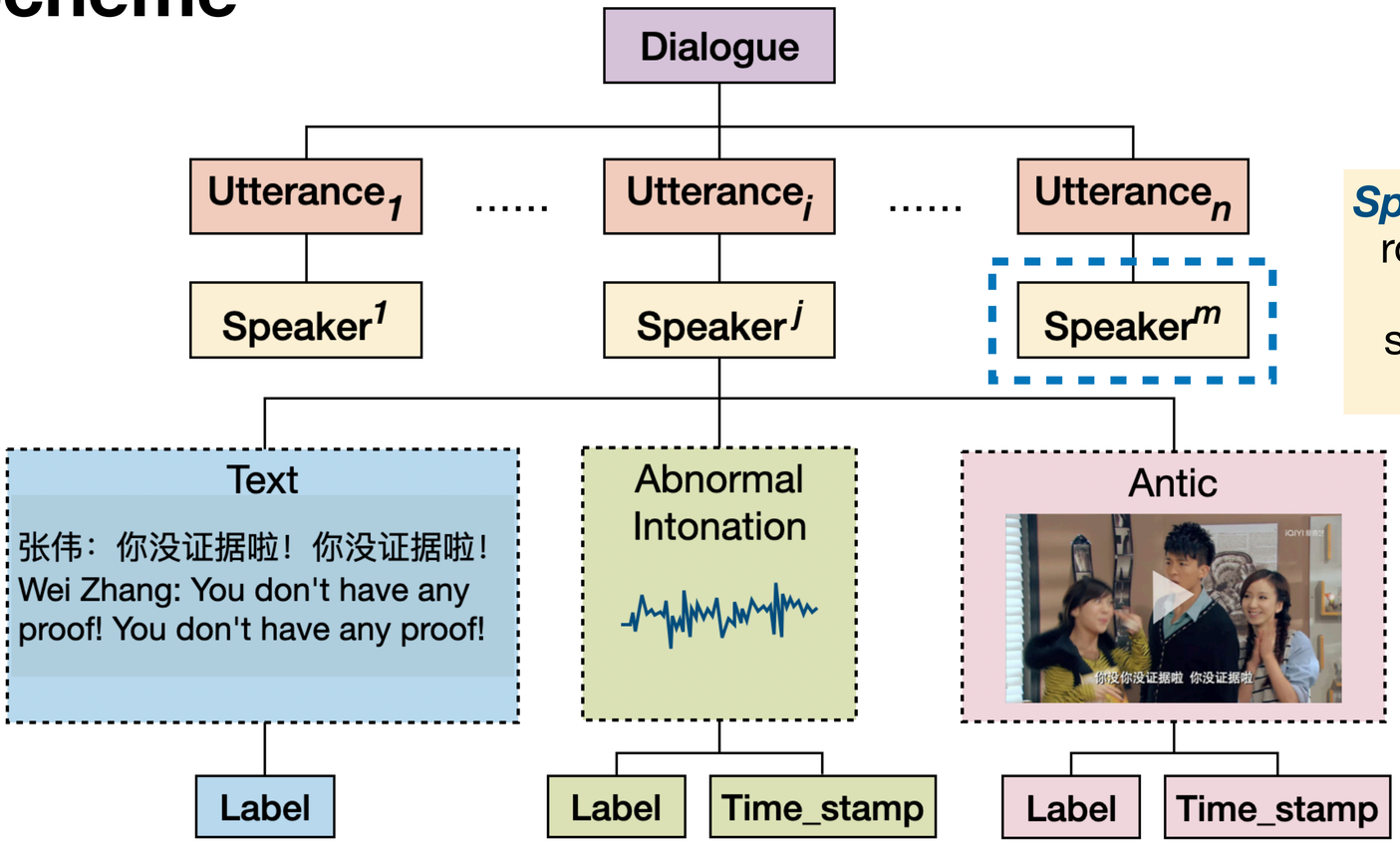
Corpus Construction

Annotation Scheme



Corpus Construction

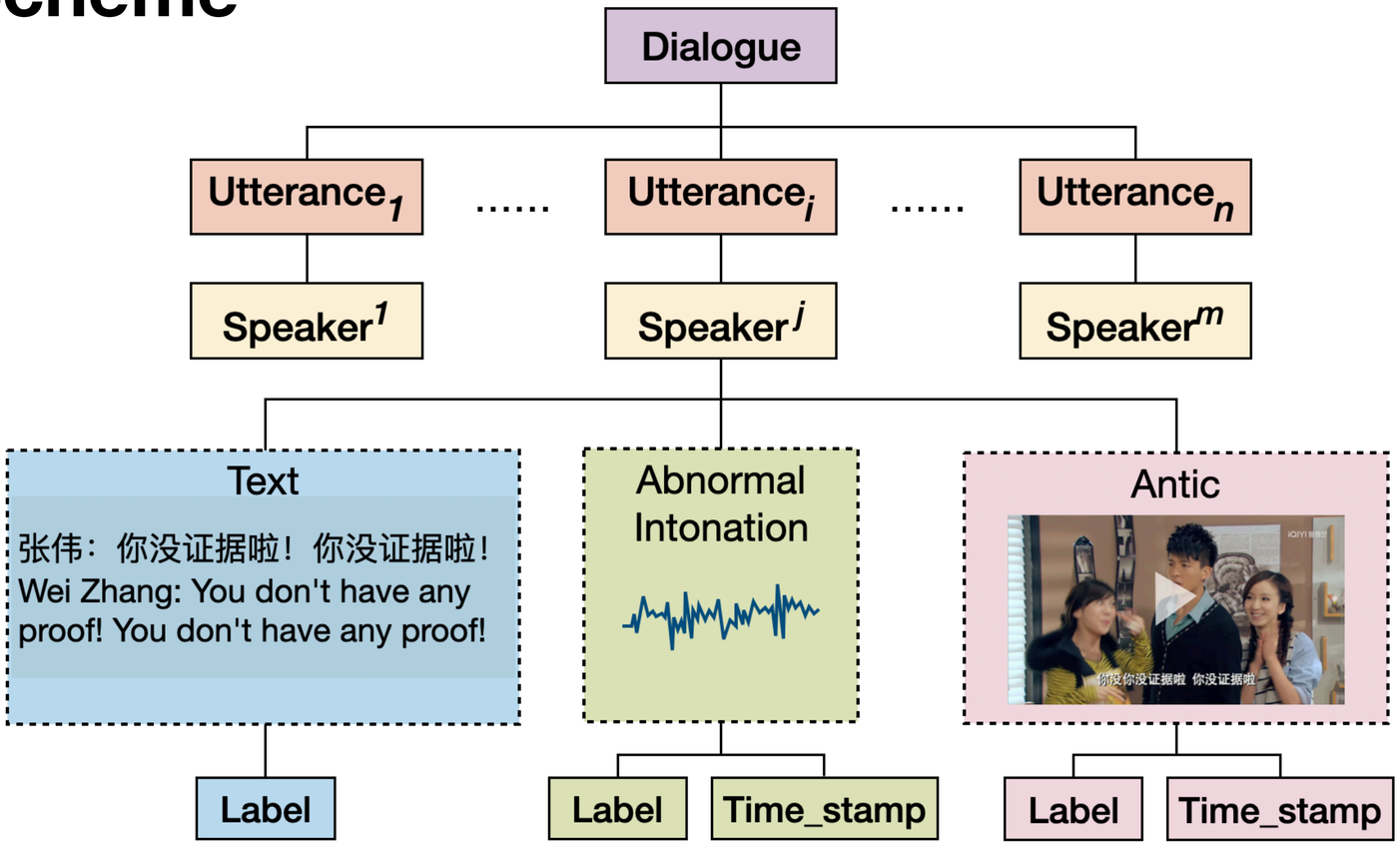
Annotation Scheme



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

Corpus Construction

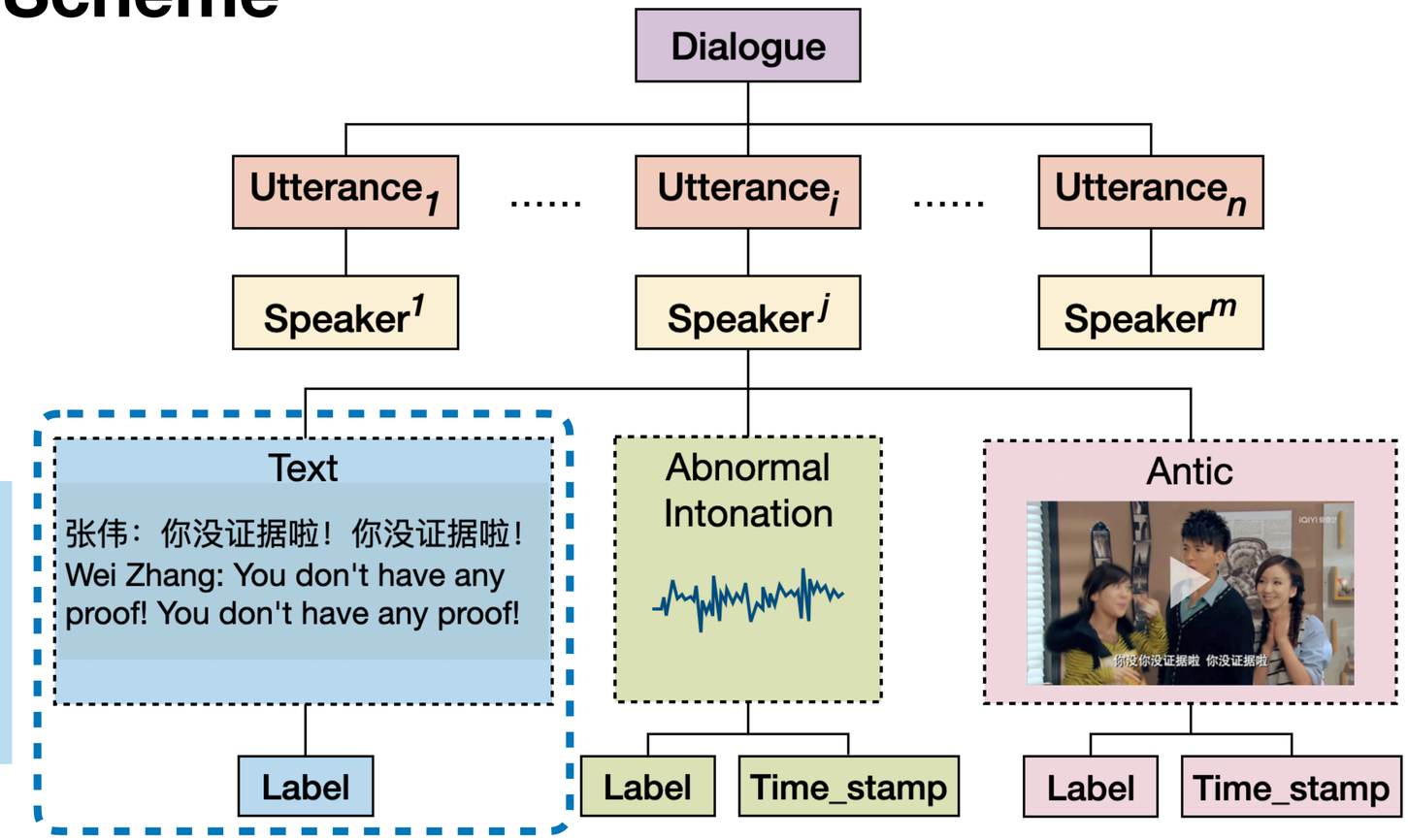
Annotation Scheme



Corpus Construction

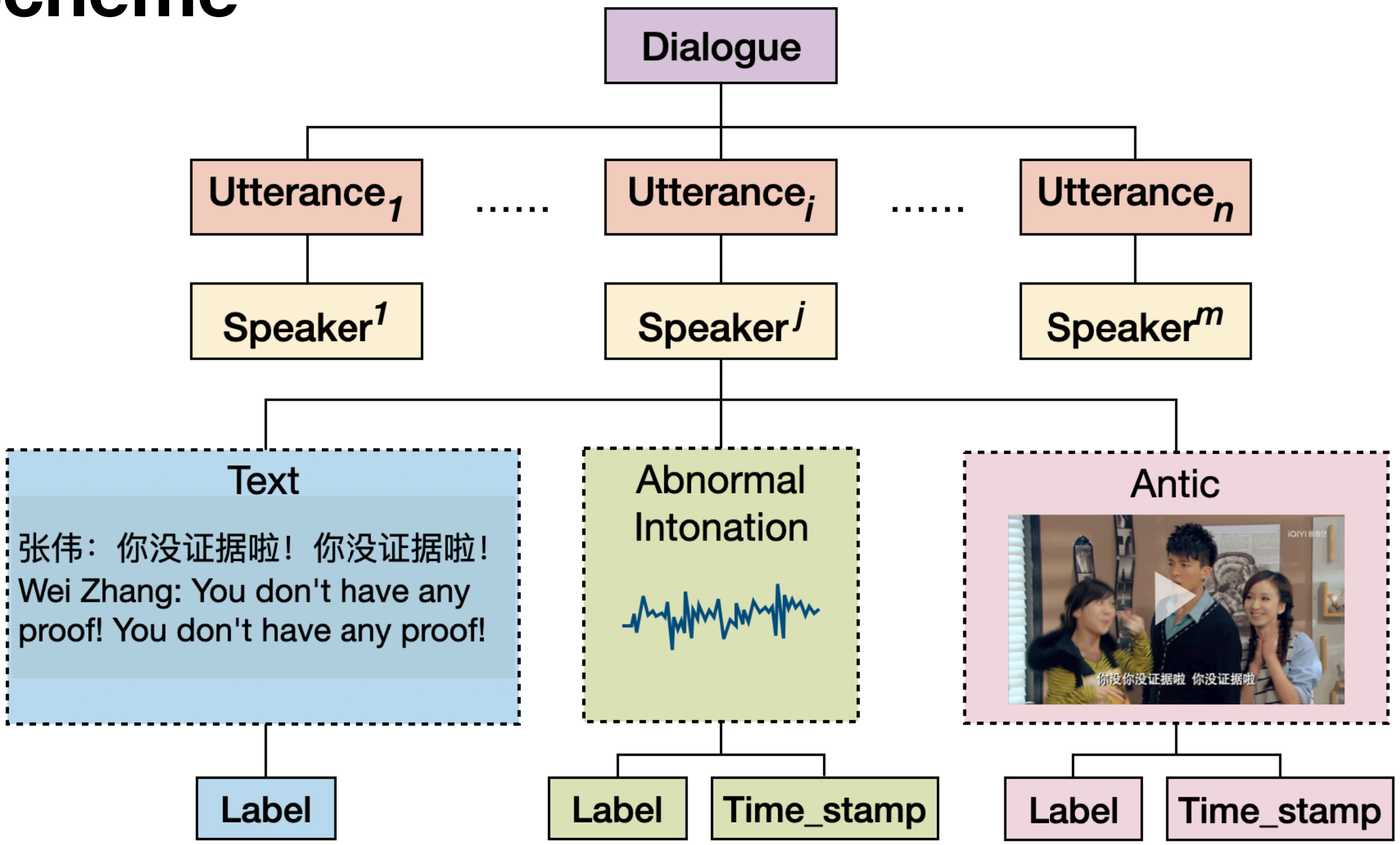
Annotation Scheme

Text represents the textual content of the utterance that indicates the specific manifestation of humor in the textual modality.



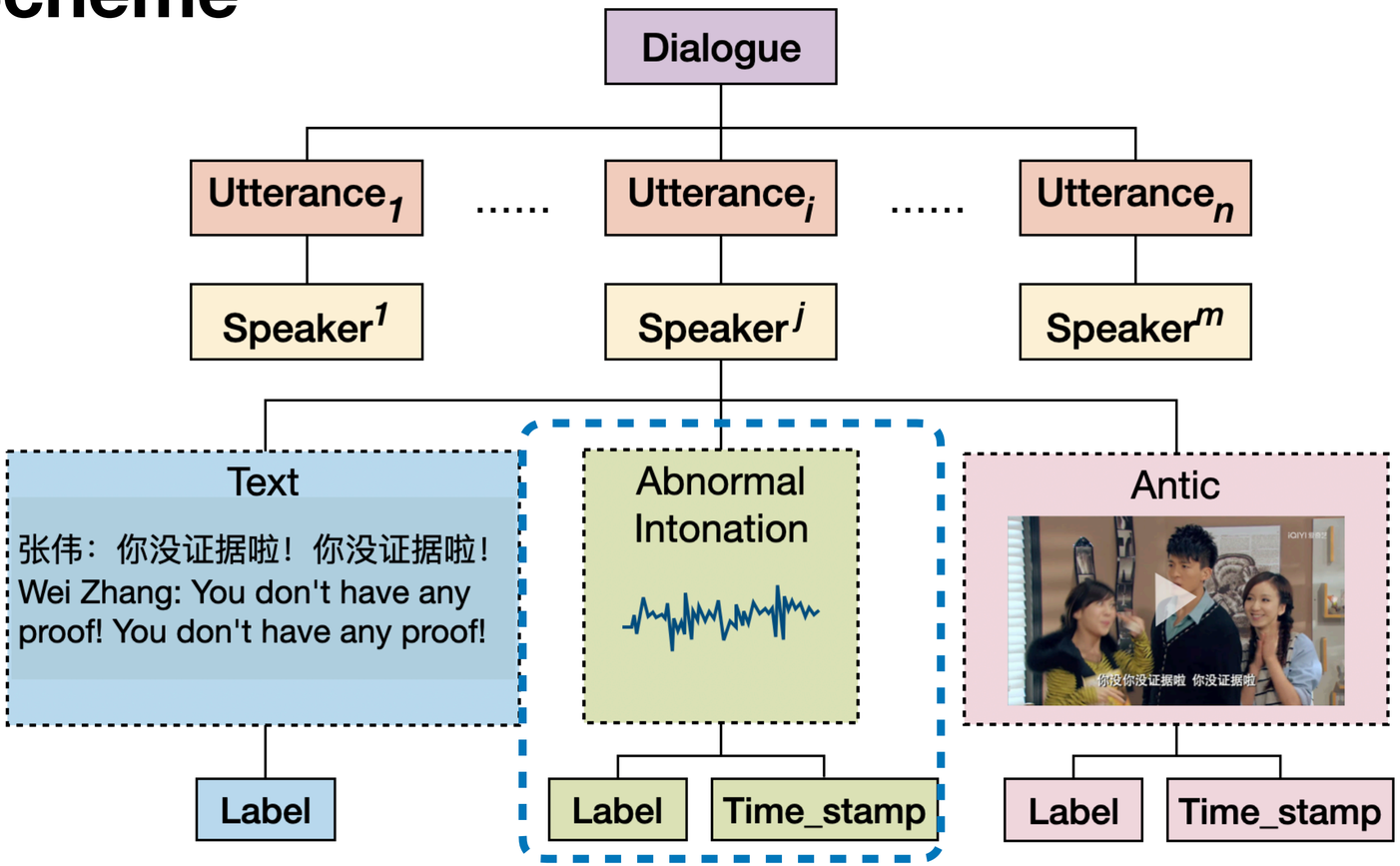
Corpus Construction

Annotation Scheme



Corpus Construction

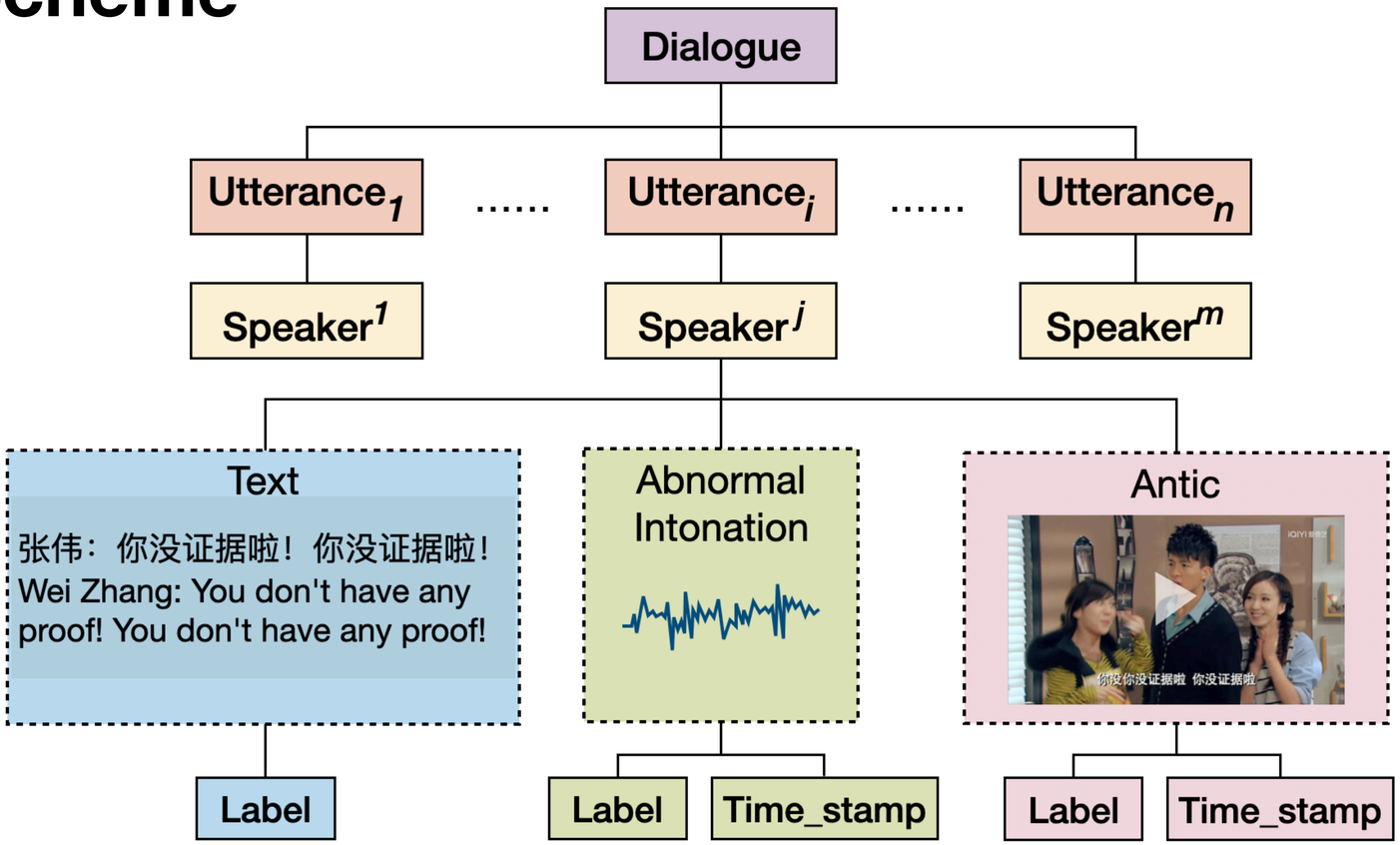
Annotation Scheme



Abnormal Intonation can enhance the speaker's emotional expression.

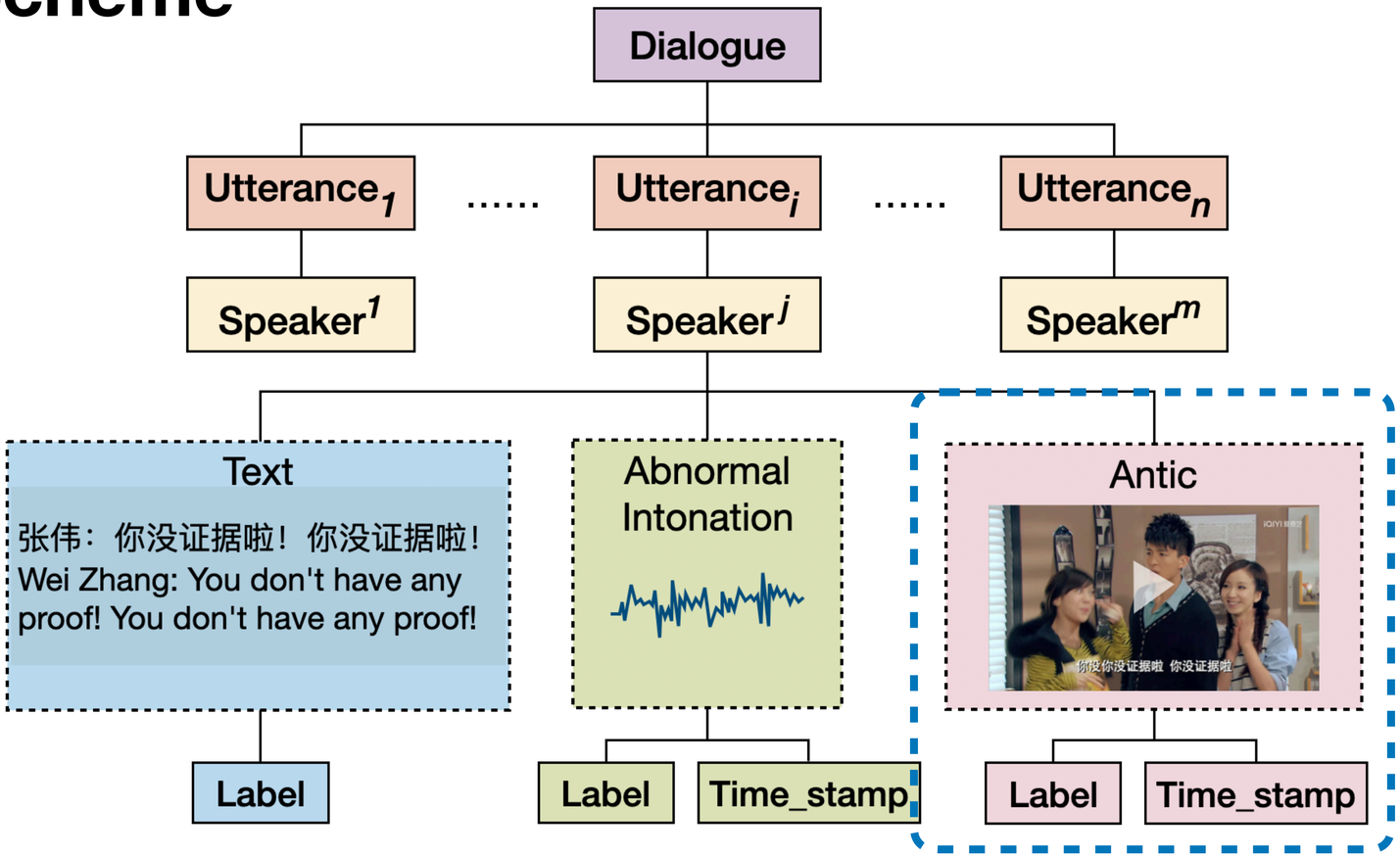
Corpus Construction

Annotation Scheme



Corpus Construction

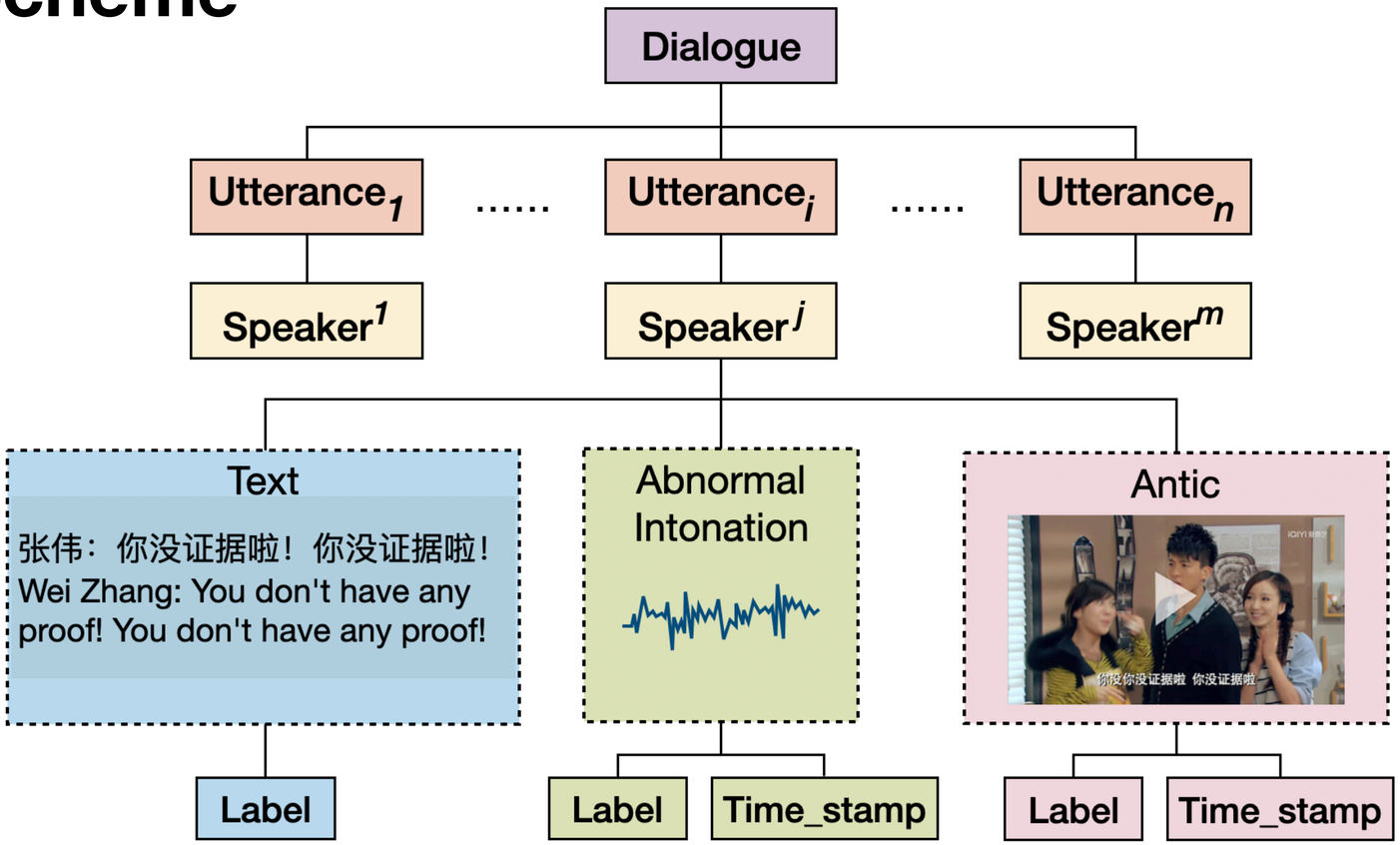
Annotation Scheme



Antic refers to the characters exhibit comical expressions or gestures during the conversation.

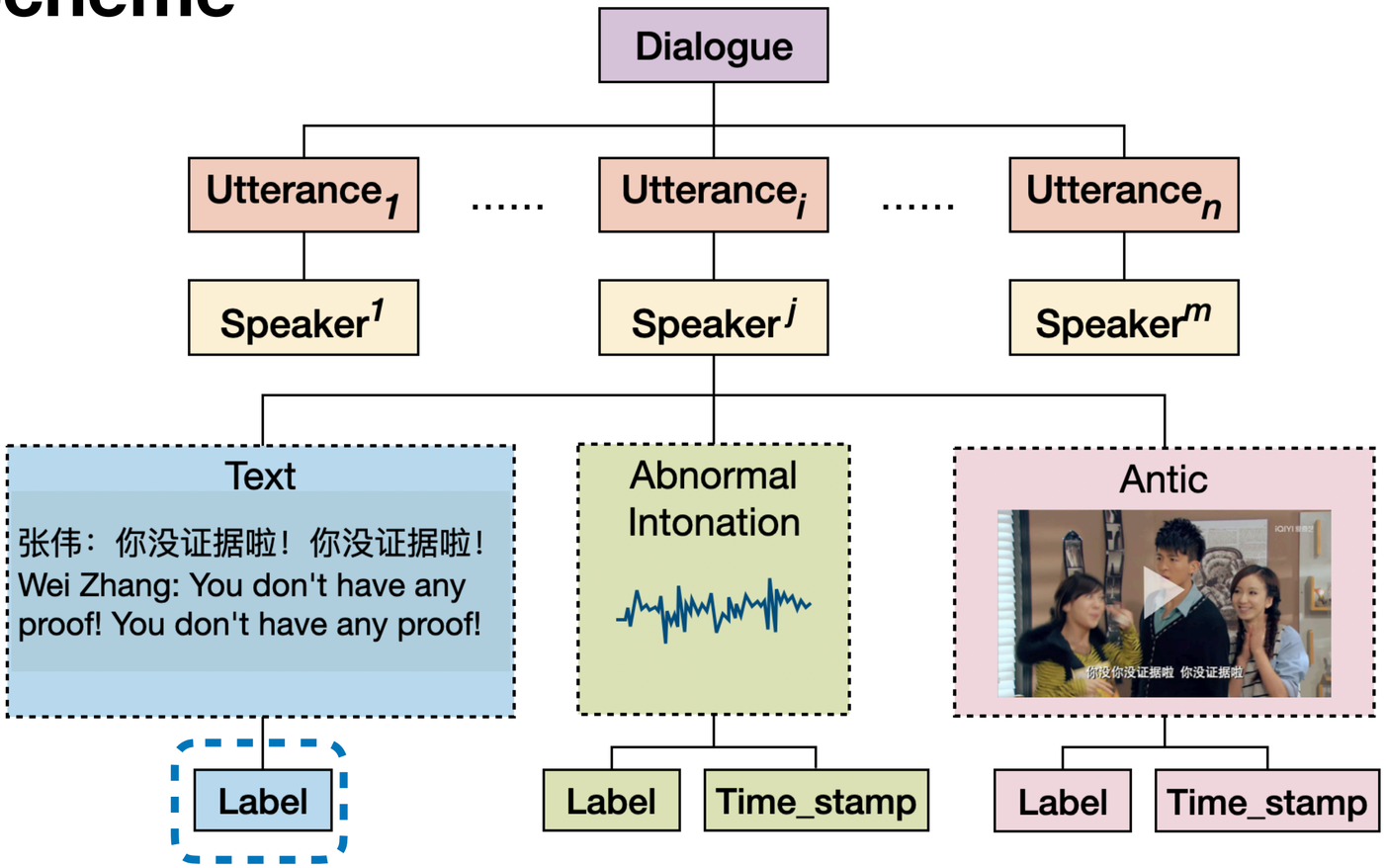
Corpus Construction

Annotation Scheme



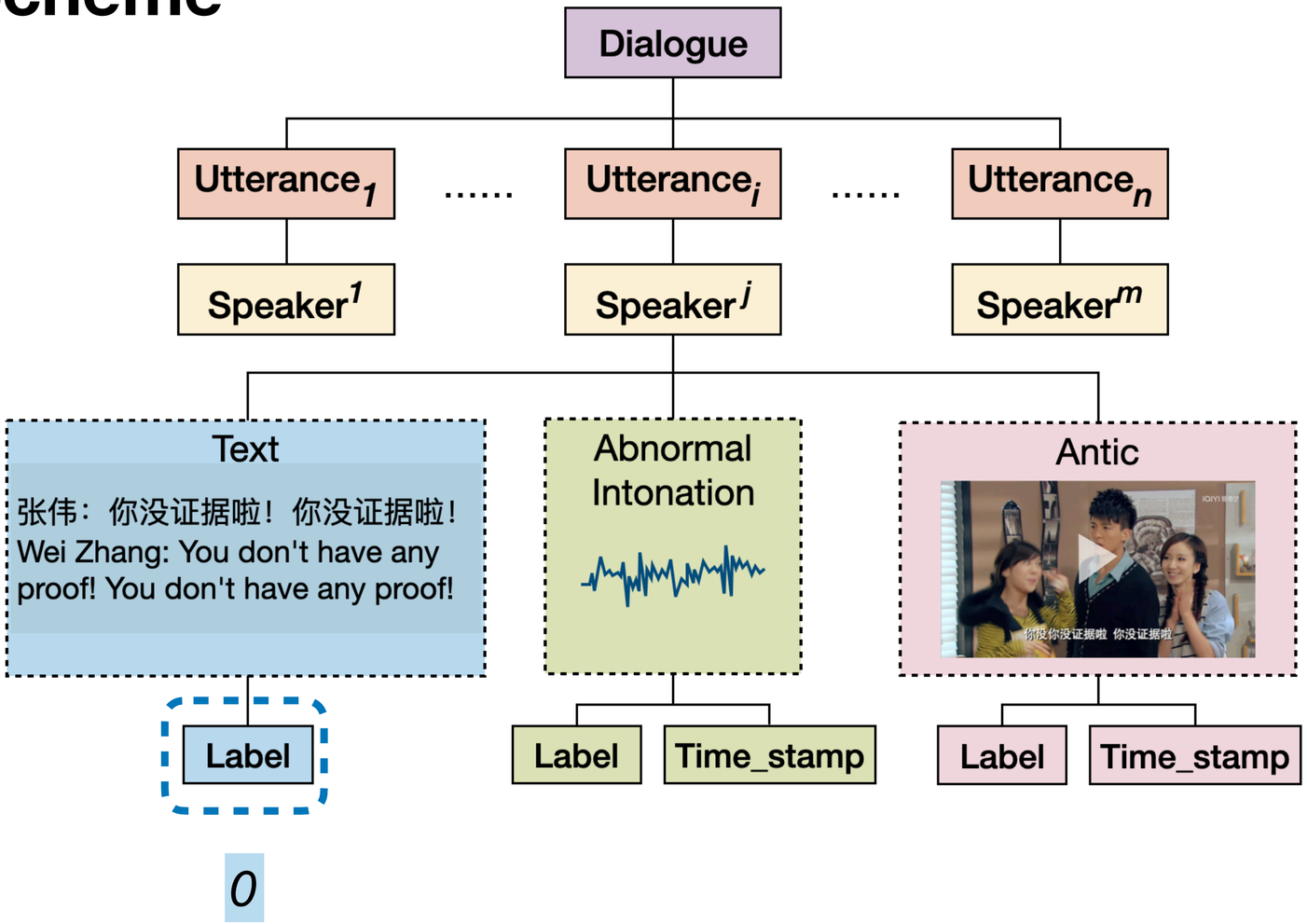
Corpus Construction

Annotation Scheme



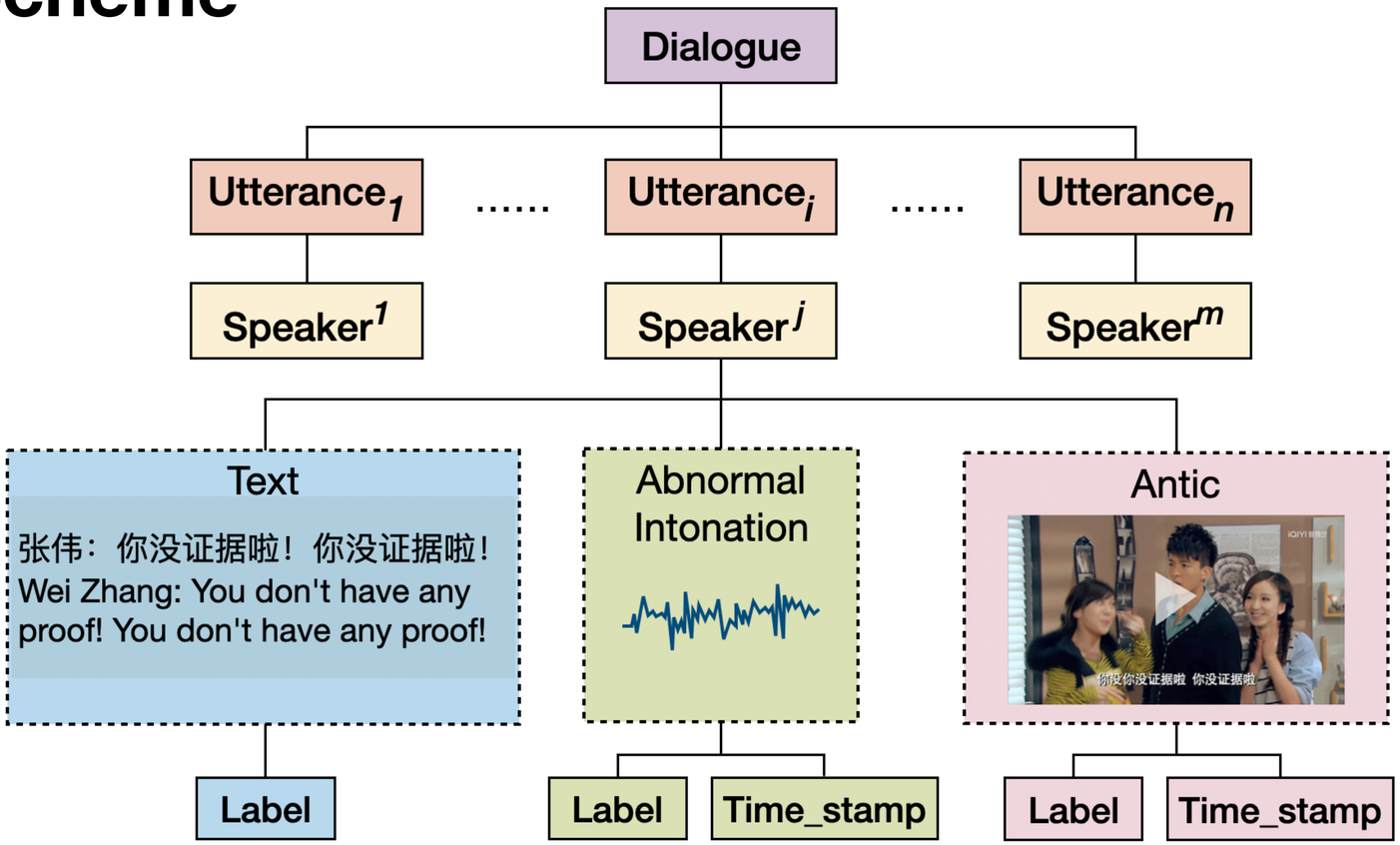
Corpus Construction

Annotation Scheme



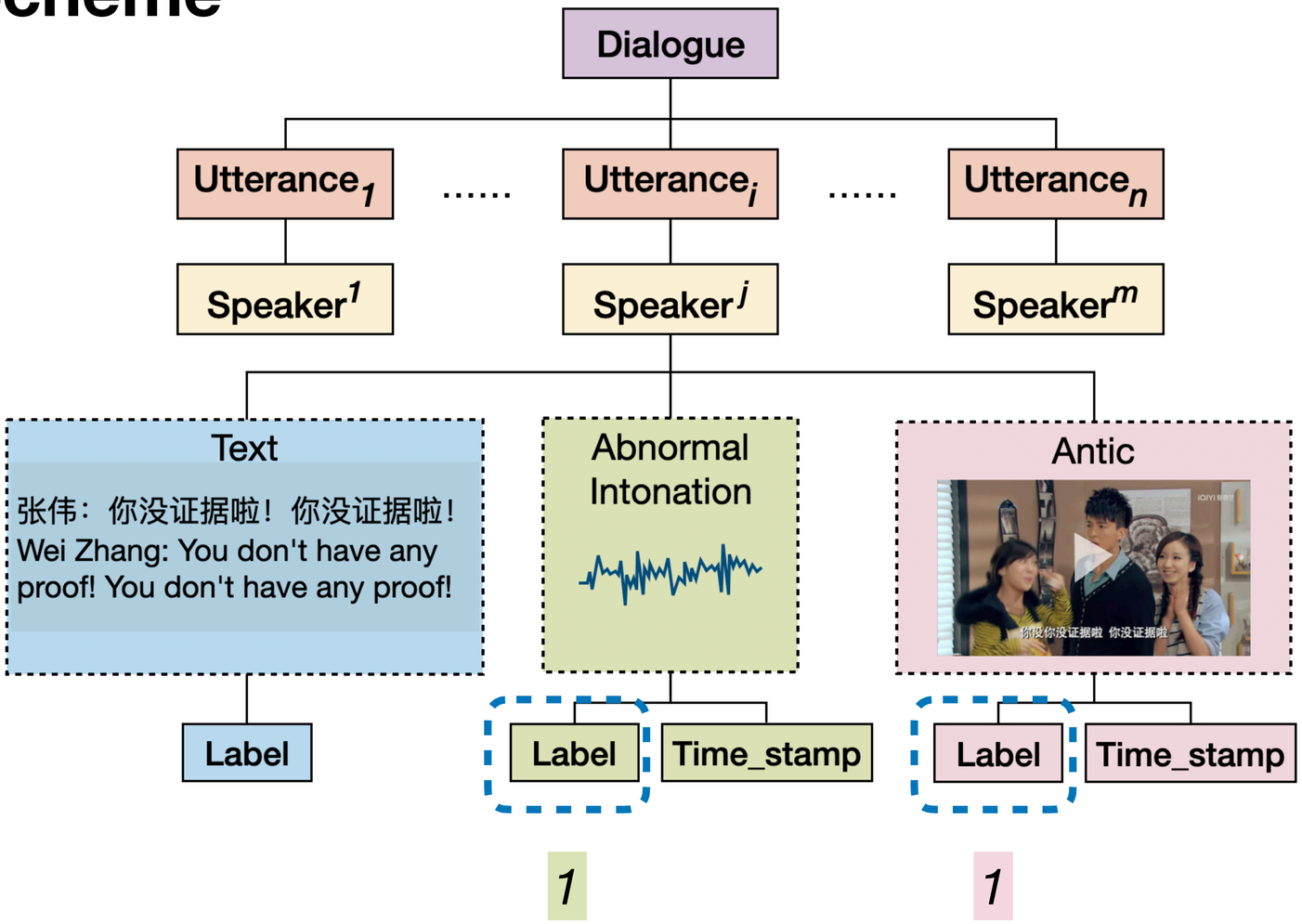
Corpus Construction

Annotation Scheme



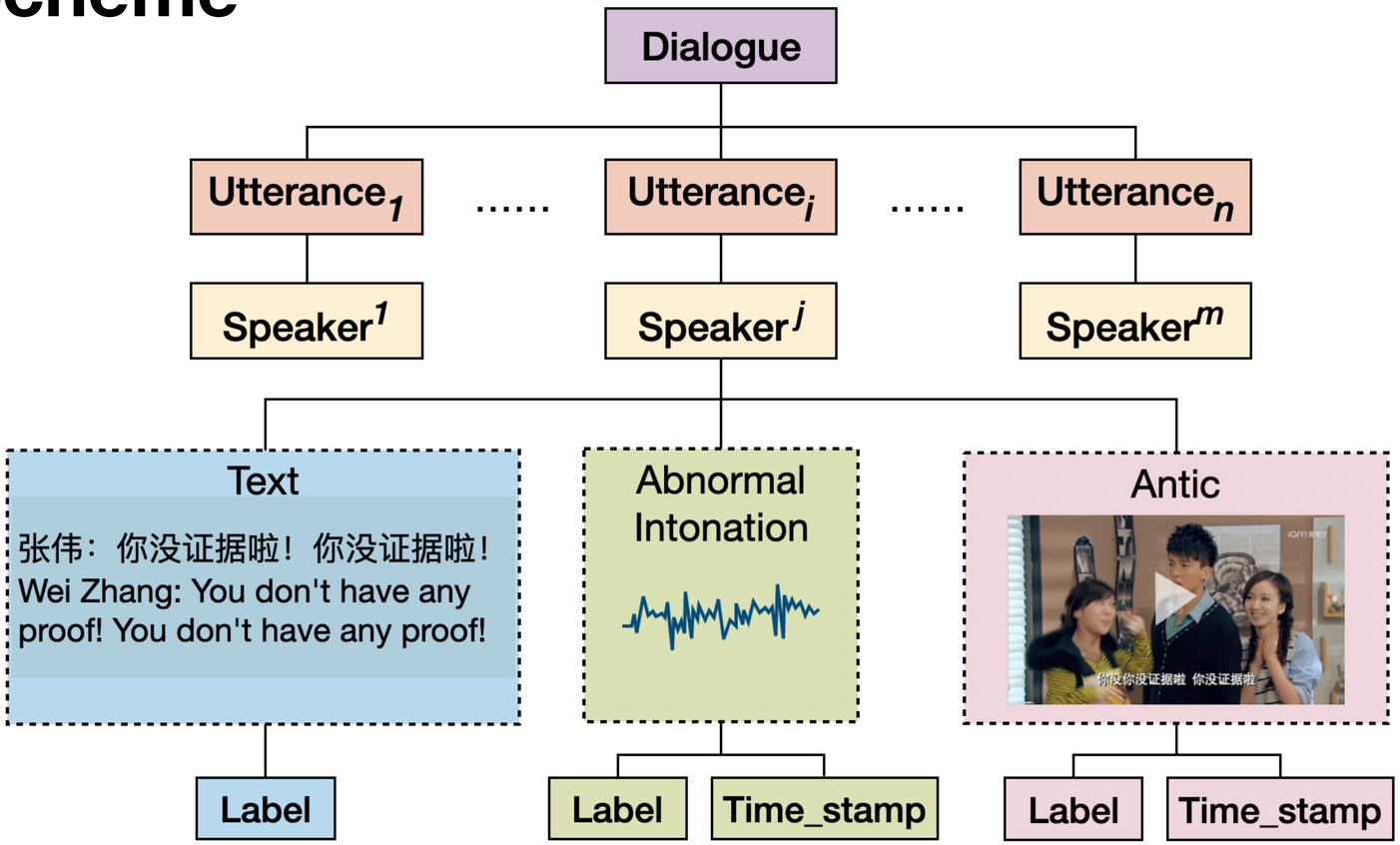
Corpus Construction

Annotation Scheme



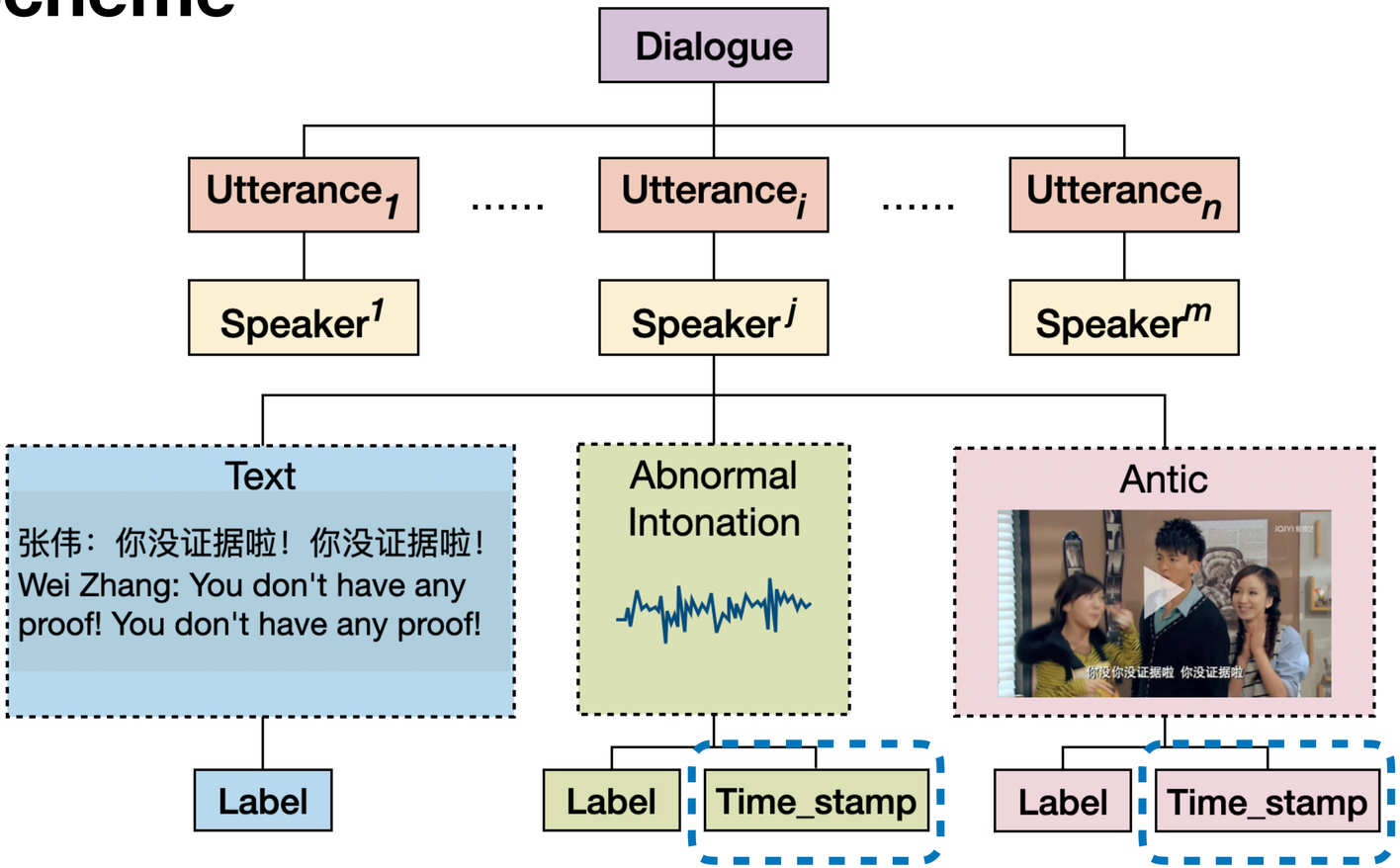
Corpus Construction

Annotation Scheme



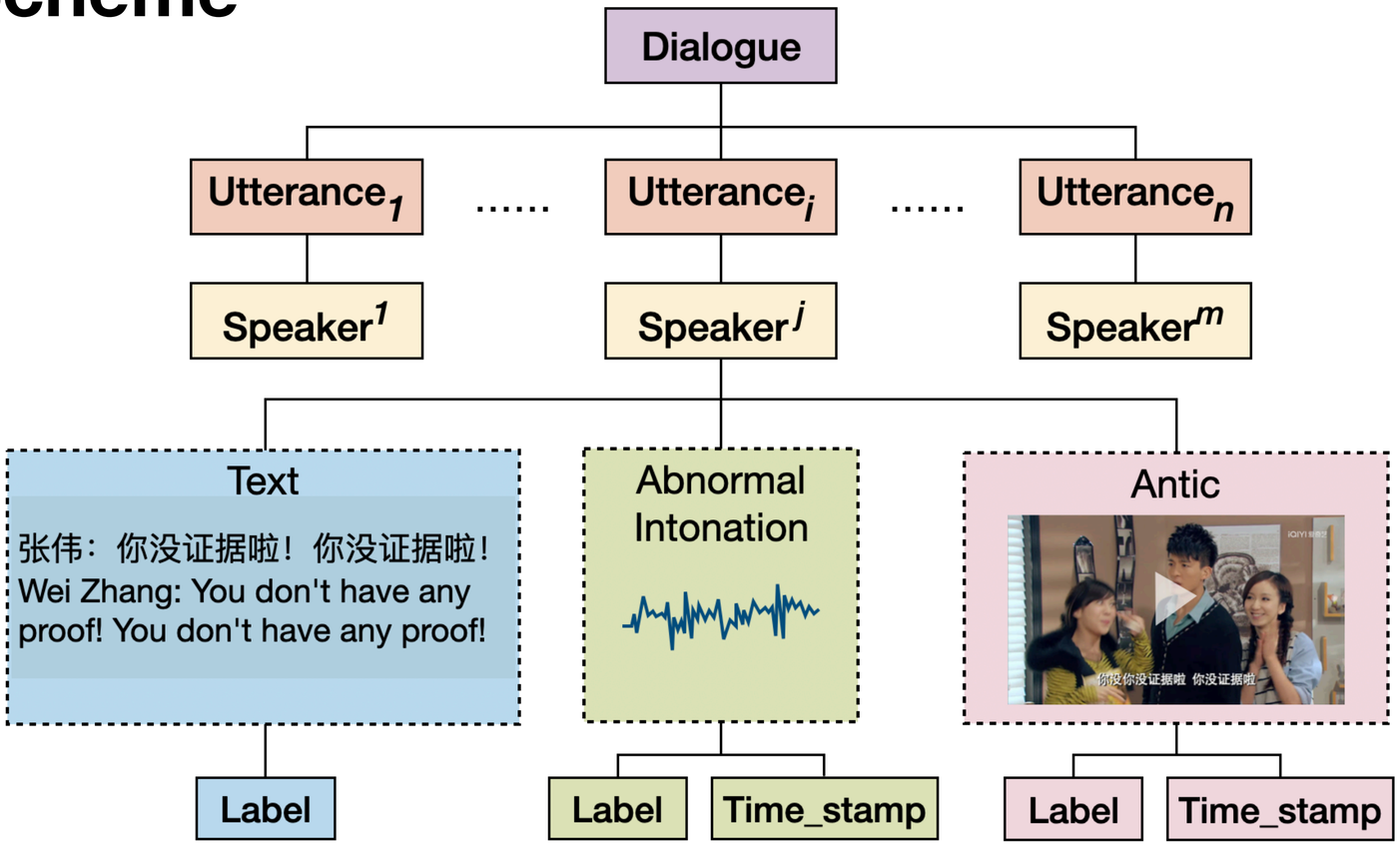
Corpus Construction

Annotation Scheme



Corpus Construction

Annotation Scheme



Corpus Construction

Annotation Process

Divide each episode into several dialogues based on different plots and scenes.

Record all the utterances in each dialogue and record the speaker and the content of each utterance.

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.

Corpus Construction

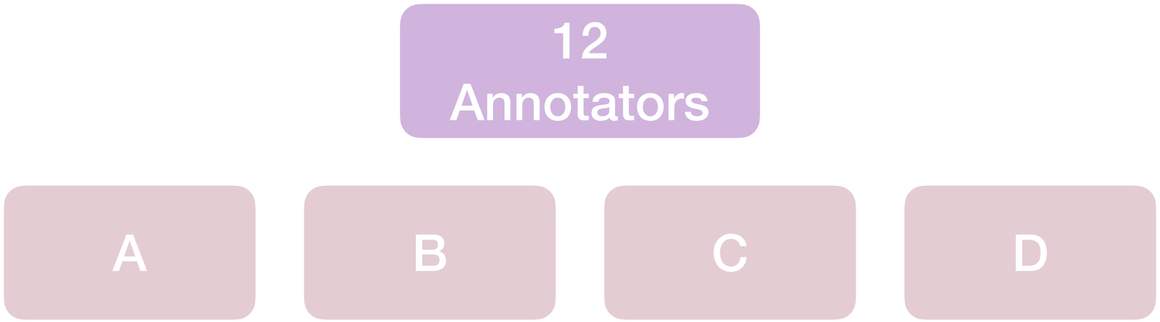
Annotation Process

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.

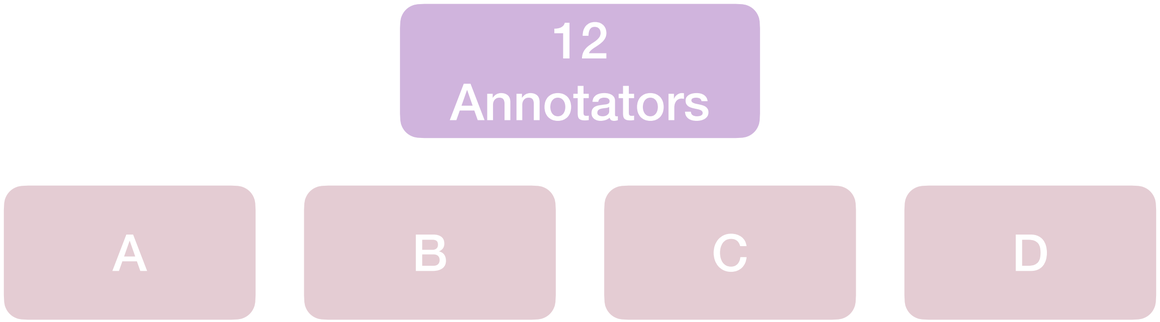
Corpus Construction

Annotation Process



Corpus Construction

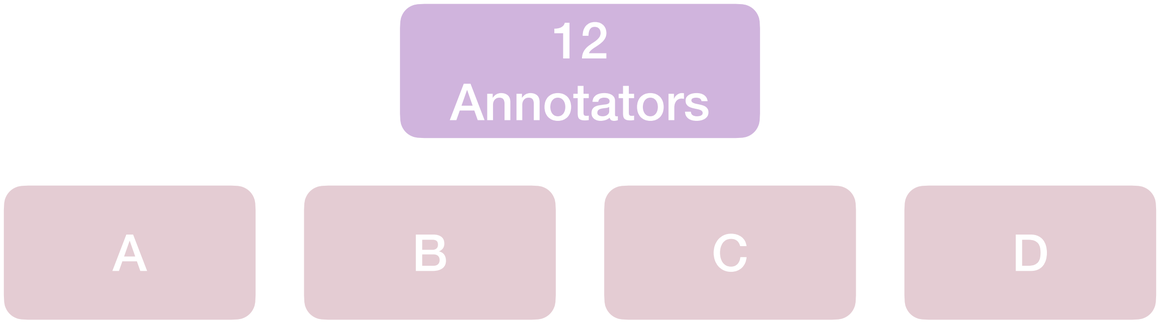
Annotation Process



If 3 annotators achieved the agreement, e.g., they all regarded the instance as humorous or non-humorous, the instance was labeled as 1 or 0, and the annotation was completed;

Corpus Construction

Annotation Process

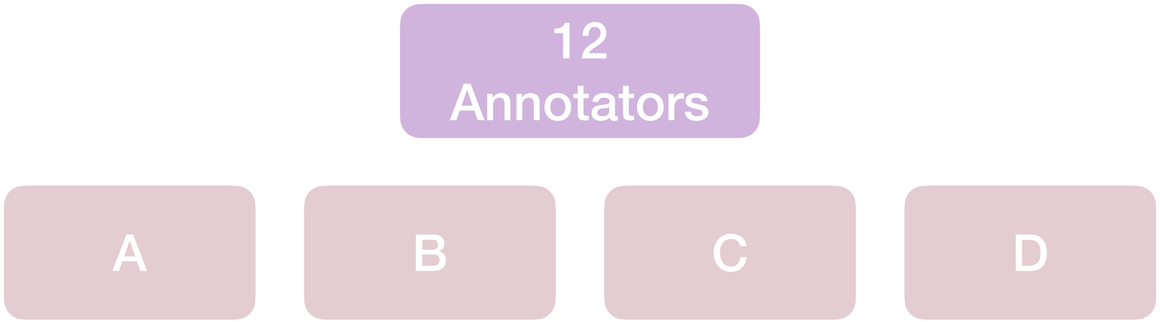


If 3 annotators achieved the agreement, e.g., they all regarded the instance as humorous or non-humorous, the instance was labeled as 1 or 0, and the annotation was completed;

If 2 annotators labeled the instance as humorous, the instance was labeled as 1 according to the majority rule;

Corpus Construction

Annotation Process



If 3 annotators achieved the agreement, e.g., they all regarded the instance as humorous or non-humorous, the instance was labeled as 1 or 0, and the annotation was completed;

If 2 annotators labeled the instance as humorous, the instance was labeled as 1 according to the majority rule;

If only 1 annotator considered the instance to be humorous, considering the contingency of humor, it required the other 9 annotators to annotate the instance and determine the final label by applying the majority rule.

Corpus Construction

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

Table 1: Statistics of MUCH corpus. Here, '#' denotes number, 'Avg.' denotes average, 'T' denotes text, 'V' denotes vision, and 'A' denotes acoustics.

Corpus Construction

Statistics

Dataset	Source	Annotation Process	Modalities	Language
UR-Funny (Hasan et al., 2019)	TED speech	Provide videos and their transcripts from the TED portal.	T, V, A	English
MuStARD (Castro et al., 2019)	Sitcom	Provide the utterance and the corresponding original fragment, while also proving contextual information.	T, V, A	English
TBBT (Kayatani et al., 2021)	Sitcom	Provide the utterance and the corresponding original fragement; Only the overall label.	T, V	English
MHD (Patro et al., 2021)	Sitcom	Annotation based on canned laughter.	T, V, A	English
M2H2 (Chauhan et al., 2021)	Sitcom	Provide the utterance and the corresponding original fragement; Only the overall label.	T, V, A	Hindi
MUCH (Ours)	Sitcom	Provide an overall label and labels for each of the three modalities for each utterance.	T, V, A	Chinese

Table 2: Comparison between MUCH and other multimodal humor datasets. It can be seen that compared with these datasets, MUCH has three respective annotations on the three modalities, not just the overall label.

Experiment

Approaches for Comparison

- **Unimodal Method:**
  - Textual modality method: BERT, RoBERTa
  - Visual modality method: ViT, OMNIVORE
  - Acoustic modality method: openSMILE
- **Multimodal Method:** CLIP

Experiment

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**

Conclusion

- We proposed a new multimodal conversational humor annotation scheme and manually annotated the MUCH corpus.
- The MUCH corpus was constructed based on a Chinese sitcom and includes three modalities: text, vision, and acoustics. It consists of 34,804 utterances in total, and 7,079 of them are humorous.
- We conducted several experiments based on some classical methods, including both unimodal and multimodal.

Thanks

Hongyu Guo
University of International Relations
2024/4/26