EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems

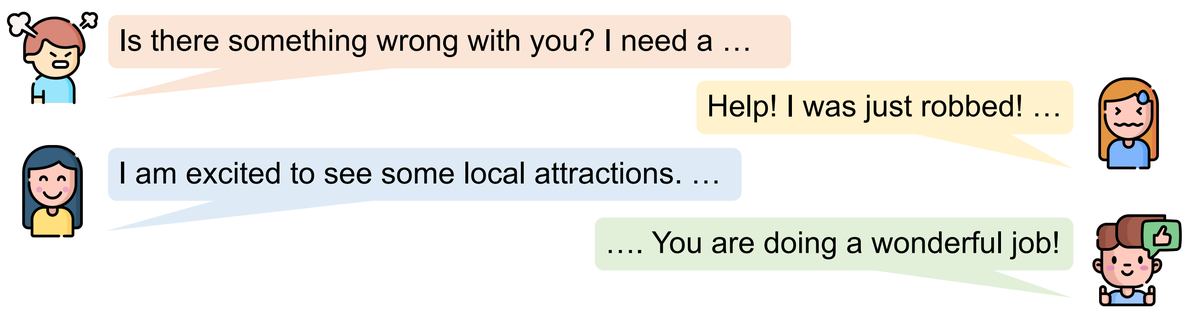
Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk and Milica Gašić

Heinrich Heine University Düsseldorf, Germany | {fengs, lubis, geishaus, linh, heckmi, niekerk, gasic}@hhu.de

Scan for EmoWOZ 👆

Introduction

Emotion makes a conversational AI human-like but is **largely overlooked in task-oriented dialogues**.

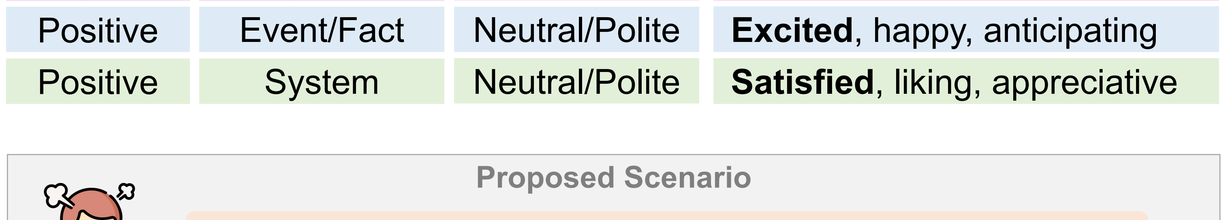


We present **EmoWOZ** to address emotions in task-oriented dialogues.

Existing Datasets	EmoWOZ
Limited Availability	Open-source
Small Size	Large-scale
Limited Linguistic Variation	Human to Machine + Human
Less Informative Labels	Tailored annotation scheme

Annotation Scheme

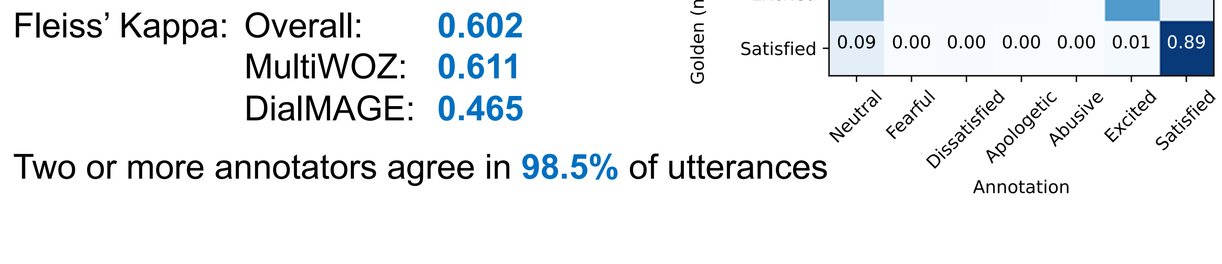
In task-oriented dialogues, emotion can be an indicator for **task performance**.



Our annotation scheme

- Inspired by **the Ortony, Clore and Collins (OCC) model** (Ortony et al., 1988): emotions as valenced reactions to various cognitive elicitors
- Considering **relevance** and **applicability** in task-oriented dialogues
- Implying **task performance**
- **7 emotion groups from 3 emotion aspects**

Valence	Elicitor	Conduct	Emotion Tokens
Neutral	-	-	**Neutral**
Negative	Event/Fact	Neutral/Polite	**Fearful**, sad, disappointed
Negative	System	Neutral/Polite	**Dissatisfied**, disliking
Negative	User	Neutral/Polite	**Apologetic**
Negative	System	Impolite	**Abusive**
Positive	Event/Fact	Neutral/Polite	**Excited**, happy, anticipating
Positive	System	Neutral/Polite	**Satisfied**, liking, appreciative



Dataset Construction

We annotate **user utterances** from two sources for **better emotion coverage**.

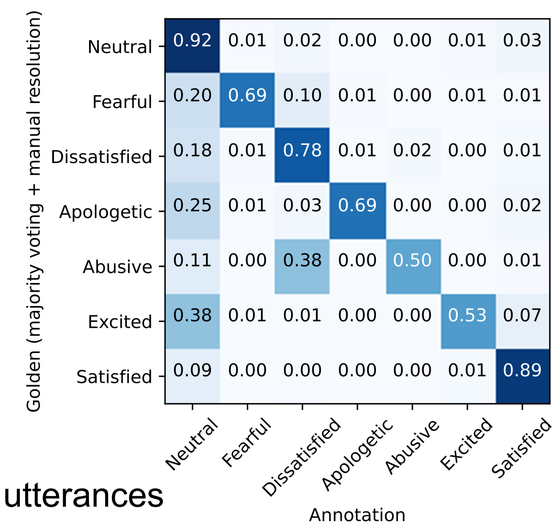
- **MultiWOZ (human-human):** 10k+ dialogues, 71k+ user utterances
- **DialMAGE (human-machine):** ~1k dialogues, 12k+ user utterances

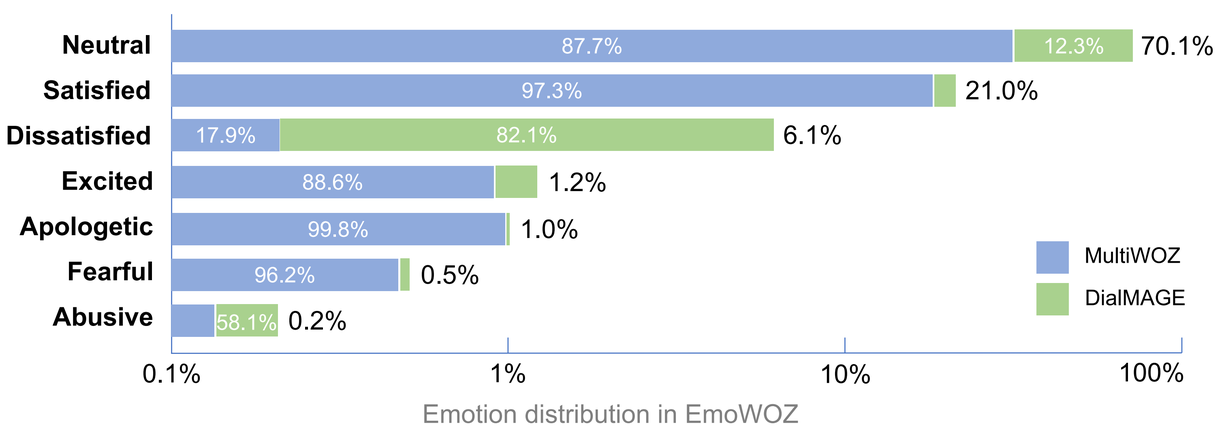
Amazon Mechanical Turk platform

- Three workers per utterance
- Qualification tests as tutorials
- Hidden tests
- Review of outliers
- Annotation limit per worker

Fleiss' Kappa: Overall: **0.602**
MultiWOZ: **0.611**
DialMAGE: **0.465**

Two or more annotators agree in **98.5%** of utterances





Emotion distribution in EmoWOZ

Experimental Results

Emotion recognition on EmoWOZ improved by dialogue context

	Macro F1 (w/o Neutral)	Weighted F1 (w/o Neutral)
BERT	50.1	73.5
ContextBERT	54.3	**79.7***
DialogueRNN (GloVe)	40.1	74.6
DialogueRNN (BERT)	52.1	75.5
COSMIC	**56.3**	77.1

Baseline model performance on EmoWOZ. * indicates statistical significant difference ($p < 0.05$).

Case Study #1

I need to arrive by 15:15 **[neutral]**

I have train TR4068 leaving at 5:35 and arriving at 5:52.

I want to confirm that I will arrive by 15:15? You stated, leaving at 5:32 and arriving at 5:52? **[dissatisfied]** **[To Classify]**

BERT❌ ContextBERT✅ DialogueRNN(GloVe)✅ DialogueRNN(BERT)✅ COSMIC✅

Case Study #2

I also need a taxi to go between the hotel and the restaurant. I'd like to leave the Gonville hotel by 9.15 **[neutral]**

When would you like to arrive by?

I just mentioned that I would like to leave by 9:15 please **[dissatisfied]** **[To Classify]**

BERT❌ ContextBERT❌ DialogueRNN(GloVe)❌ DialogueRNN(BERT)❌ COSMIC❌

Complementing datasets shown to be useful

	Test on MultiWOZ		Test on DialMAGE	
Training Set	Macro F1	Weighted F1	Macro F1	Weighted F1
MultiWOZ	**47.7**	**83.9**	33.6	14.5
DialMAGE	17.0	67.8	35.2	72.9
EmoWOZ	45.1	83.1	**50.0***	**73.5**

Average F1s without Neutral of ContextBERT. * indicates statistical significant difference ($p < 0.05$).

Dialogue State Tracking (DST) improved by multi-task learning with emotion recognition in conversations (ERC)

- TripPy DST (Heck et al., 2020)
- DST + ERC per training step

Training Task	JGA
DST	53.7
DST + ERC	**54.7***

Joint Goal Accuracy (JGA) on MultiWOZ. * indicates statistically significant difference ($p < 0.02$)

Conclusion

- Large-scale open-source dataset for emotion recognition in task-oriented dialogues
- New annotation scheme to support task-oriented behaviours
- In ERC experiments:
  - Dialogue context useful
  - Challenging to recognise implicit emotions
  - Features characteristic to task-oriented dialogues needed
- EmoWOZ useful to downstream task-oriented dialogue modelling tasks

Alexander von Humboldt Stiftung/Foundation
European Research Council Established by the European Commission
Google Cloud