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

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

<table>
<thead>
<tr>
<th>Limited Availability</th>
<th>Small Size</th>
<th>Limited Linguistic Variation</th>
<th>Less Informative Labels</th>
</tr>
</thead>
</table>

EmoWOZ

- Open-source
- Large-scale
- Human to Machine + Human
- 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

<table>
<thead>
<tr>
<th>Valence</th>
<th>Elicitor</th>
<th>Conduct</th>
<th>Emotion Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>-</td>
<td>Neutral</td>
<td>Neutral, Polite</td>
</tr>
<tr>
<td>Negative</td>
<td>Event/Fact</td>
<td>Neutral/Polite</td>
<td>Fearful, sad, disappointed</td>
</tr>
<tr>
<td>Negative</td>
<td>System</td>
<td>Neutral/Polite</td>
<td>Dissatisfied, disliking</td>
</tr>
<tr>
<td>Negative</td>
<td>User</td>
<td>Neutral/Polite</td>
<td>Apologetic</td>
</tr>
<tr>
<td>Negative</td>
<td>System</td>
<td>Impolite</td>
<td>Abusive</td>
</tr>
<tr>
<td>Positive</td>
<td>Event/Fact</td>
<td>Neutral/Polite</td>
<td>Excited, happy, anticipating</td>
</tr>
<tr>
<td>Positive</td>
<td>System</td>
<td>Neutral/Polite</td>
<td>Satisfied, liking, appreciative</td>
</tr>
</tbody>
</table>

Proposed Scenario

... (the user seems disappointed because the system made an error)

Aha, I must have made mistakes. I should apologise and correct myself.

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

Experimental Results

Emotion recognition on EmoWOZ improved by dialogue context

<table>
<thead>
<tr>
<th></th>
<th>Macro F1 (w/o Neutral)</th>
<th>Weighted F1 (w/o Neutral)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>50.1</td>
<td>73.5</td>
</tr>
<tr>
<td>ContextBERT</td>
<td>54.3</td>
<td>79.7*</td>
</tr>
<tr>
<td>DialogueRNN (GloVe)</td>
<td>40.1</td>
<td>74.6</td>
</tr>
<tr>
<td>DialogueRNN (BERT)</td>
<td>52.1</td>
<td>75.5</td>
</tr>
<tr>
<td>COSMIC</td>
<td>56.3</td>
<td>77.1</td>
</tr>
</tbody>
</table>

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? [disatisfied] [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 [disatisfied] [To Classify]

BERT ✓ ContextBERT ✔ DialogueRNN(GloVe) ✔ DialogueRNN(BERT) ✔ COSMIC ✔

Complementing datasets shown to be useful

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

<table>
<thead>
<tr>
<th>Training Task</th>
<th>JGA</th>
<th>DST</th>
<th>DST + ERC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiWOZ</td>
<td>47.7</td>
<td>83.9</td>
<td>33.6</td>
</tr>
<tr>
<td>DialMAGE</td>
<td>17.0</td>
<td>67.8</td>
<td>35.2</td>
</tr>
<tr>
<td>EmoWOZ</td>
<td>45.1</td>
<td>83.1</td>
<td>50.0*</td>
</tr>
</tbody>
</table>

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

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

Icons made by Freepik, The Noun Project, Flaticon, and Skeuomorph from www.flaticon.com