EmotionFlow: Capture the Dialogue Level Emotion Transitions

IEEE ICASSP 2022

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, Longtao Huang

Institute of Information Engineering
Chinese Academy of Sciences
Alibaba Group

May 7, 2022
Motivation
What is ERC task

An example:

1) You liked it? You really liked it?
2) Oh, yeah!
3) Which part exactly?
4) The whole thing! Can we go?
5) What about the scene with the kangaroo?
6) I was surprised to see a kangaroo in a world war epic.
7) You fell asleep!
8) Don’t go, I’m sorry.

Figure: Example of ERC task.¹

¹https://github.com/declare-lab/awesome-emotion-recognition-in-conversations
What is the focus of this paper

Problem settings

1. single modal
   text only
2. real-time
   only use the past utterances to predict the emotion of current utterance
3. multi-party
   a conversation can contain more than 2 speakers
4. non-anonymous
   speakers’ habits can be learned

Focus

The spread effect of emotions at dialogue level.

Figure: The transition probability between emotions of current turn and next turn.
Model
How we design our model

**Semantic Context Modeling**

1. Roberta as encoder
2. QA-style input construction to capture speaker-specific features
   \[ X_t = [\langle s \rangle, s_{t-k}, u_{t-k}, s_{t-k+1}, \ldots, s_t, u_t, \langle /s \rangle, Q] \]
   Q = “How does \( s_t \) feel now?”
3. supervised signal training on CLS token
   *ensure that the outputs of this module are “probability” distributions.*

**Emotion Sequence Modeling**

A linear-chain CRF layer upon the outputs of the Semantic Context Modeling module.
full dialogue as input
each turn of dialogue is encoded separately then calculate the probability $x_i$
feed $x_1 \ldots x_n$ into CRF layer.

Figure: The overview of EmotionFlow.
Experiments
Datasets

Multimodal EmotionLines Dataset (MELD)

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversations</td>
<td>1038</td>
<td>114</td>
<td>280</td>
<td>1432</td>
</tr>
<tr>
<td>Utterances</td>
<td>9989</td>
<td>1109</td>
<td>2610</td>
<td>13708</td>
</tr>
<tr>
<td>Speakers</td>
<td>260</td>
<td>47</td>
<td>100</td>
<td>274</td>
</tr>
<tr>
<td>Speakers &gt;100</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Table: Statistics of MELD.

There are only 6 speakers that appeared more than 100 times in the dataset, which is good for the model to learn speakers’ features.
## Main Results

<table>
<thead>
<tr>
<th>Model</th>
<th>External Knowledge</th>
<th>Weighted F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DialogueGCN</td>
<td>✗</td>
<td>58.10</td>
</tr>
<tr>
<td>RGAT</td>
<td>✗</td>
<td>60.91</td>
</tr>
<tr>
<td>HiTrans</td>
<td>✗</td>
<td>61.94</td>
</tr>
<tr>
<td>DialogXL</td>
<td>✗</td>
<td>62.41</td>
</tr>
<tr>
<td>DAG-ERC</td>
<td>✗</td>
<td>63.65</td>
</tr>
<tr>
<td>TODKAT w/o KB</td>
<td>✗</td>
<td>63.97</td>
</tr>
<tr>
<td>EmotionFlow (Ours)</td>
<td>✗</td>
<td>65.05</td>
</tr>
<tr>
<td>KAIMTL</td>
<td>✓</td>
<td>58.97</td>
</tr>
<tr>
<td>KET</td>
<td>✓</td>
<td>58.18</td>
</tr>
<tr>
<td>COSMIC</td>
<td>✓</td>
<td>65.21</td>
</tr>
<tr>
<td>TODKAT</td>
<td>✓</td>
<td>65.47*</td>
</tr>
</tbody>
</table>

**Figure**: Performance comparisons on the MELD testset.
## Ablation Study

<table>
<thead>
<tr>
<th>Model</th>
<th>Weighted F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmotionFlow</td>
<td>65.05</td>
</tr>
<tr>
<td>EmotionFlow w/o CRF</td>
<td>63.70</td>
</tr>
<tr>
<td>EmotionFlow w/o QA</td>
<td>63.55</td>
</tr>
<tr>
<td>EmotionFlow w/o [CRF,QA]</td>
<td>62.35</td>
</tr>
</tbody>
</table>

**Figure:** Ablation study on MELD test set.
Conclusion and Future Work
Conclusion and Future Work

Conclusion

1. A novel model that can capture the spread effect of emotions via a CRF layer
2. QA-style input construction helps the model to learn speaker-specific features
3. A new state-of-the-art result on a widely used benchmark.

Future Work
Linear-chain CRF → probabilistic graphical model
Thanks for your attention!