Modeling User Satisfaction Transitions in Dialogues from Overall Ratings

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, Toyomi Meguro
NTT Corporation
Objective

• Predict user satisfaction transitions during a dialogue
  – Useful for an analysis to improve dialogue systems
Related Work

• Plenty of work on predicting overall user satisfaction of a dialogue
  – PARADISE (Walker et al., Möller et al.)
  – N-gram based method (Hara et al.)

• Little work on predicting user satisfaction transitions during a dialogue
  – One exception: Engelbrecht et al., 2009
  – Models user satisfaction transitions using hidden Markov models (HMMs)
Engelbrecht et al., 2009

• Uses manually-labeled reference transitions

Problem: High cost for making reference data
Approach

- Train an HMM from dialogues with overall ratings
  - No utterance-level references are necessary
  - Requires only a single overall rating per dialogue
How to connect the states

• Connect with equal transition probabilities
  ⇒ Common dialogue acts (e.g. Greeting) are always predicted as the same rating

• Add a state (common state) trained from the data of all ratings to avoid evaluating rating-wide DAs

Too broad! may never come here in decoding

Need to sharpen the distribution
Solution: Concatenated Training

• We can sharpen the output distributions

We iterate this until all pairs converge against training data
Example

NOTE:
We ignore transition probabilities.
We assume the same initial probabilities.
EM algorithm updates probabilities based on the expected counts

- Rating 1 (bad)
  - Retrain
  - Dialogues (rating 1)
    - “Rude” 75 counts
    - “Greet” 25 counts

- Rating 0

```
Polite  Rude  Funny  Greet
75%     25%
```

75% of “Rude” output from Rating 1
25% of “Rude” output from Rating 0

```
Polite  Rude  Funny  Greet
82%     18%
```

75*0.25=18.75
25*0.5=12.5
⇒ 1.5:1

```
Polite  Rude  Funny  Greet
25% 25% 25% 25%
```

50% of “Greet” output from Rating 1
50% of “Greet” output from Rating 0

```
Polite  Rude  Funny  Greet
60% 40%
```

75*0.75=56.25
25*0.5=12.5
⇒ 4.5:1
Common utterances more likely to be output from common states

Rating-specific utterances more likely to be output from the states for each rating
We can predict the rating for dialogue acts more accurately.
Experiment

- Dialogue data (text chat) in two domains
  - **Animal Discussion** (human-system)
    - User and system discuss about animals
  - **Attentive Listening** (human-human)
    - Listener attentively listens to a speaker
  - Annotated with dialogue acts and US ratings (7-levels)

<table>
<thead>
<tr>
<th></th>
<th>Animal Discussion</th>
<th>Attentive Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td># dialogues with overall ratings</td>
<td>180</td>
<td>1260</td>
</tr>
<tr>
<td># dialogues with utterance-level ratings</td>
<td>90 (subset)</td>
<td>100 (subset)</td>
</tr>
<tr>
<td>Dialogue act types</td>
<td>29</td>
<td>40</td>
</tr>
</tbody>
</table>

- Comparison with a random baseline and an upper bound (Engelbrecht et al.) trained from reference transitions
Evaluation Criterion

- **Match rate per rating (MR/r)**
  - Equally treats easy and difficult-to-guess ratings
  - We want to predict rare but important ratings
  - Reliable than other metrics

(Higashinaka et al., 2010, IWSDS)

<table>
<thead>
<tr>
<th></th>
<th>U1</th>
<th>U2</th>
<th>U3</th>
<th>U4</th>
<th>U5</th>
<th>U6</th>
<th>U7</th>
<th>U8</th>
<th>U9</th>
<th>U10</th>
<th>U11</th>
<th>U12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hyp</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hyp’</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match rate</td>
<td>2/2</td>
<td>0/1</td>
<td>0/1</td>
<td>2/3</td>
<td>0/2</td>
<td>1/2</td>
<td>1/1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Avg</th>
<th>MR/r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.45</td>
</tr>
</tbody>
</table>

(Higashinaka et al., 2010, IWSDS)
## Results

### Animal Discussion domain

<table>
<thead>
<tr>
<th></th>
<th>random</th>
<th>simple</th>
<th>concat</th>
<th>Upper B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothness</td>
<td>0.143</td>
<td>0.137</td>
<td>0.177</td>
<td>0.217</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.143</td>
<td>0.156</td>
<td>0.164</td>
<td>0.231</td>
</tr>
<tr>
<td>Willingness</td>
<td>0.143</td>
<td>0.152</td>
<td>0.181</td>
<td>0.224</td>
</tr>
</tbody>
</table>

### Attentive Listening domain

<table>
<thead>
<tr>
<th></th>
<th>random</th>
<th>simple</th>
<th>concat</th>
<th>Upper B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothness</td>
<td>0.141</td>
<td>0.118</td>
<td>0.167</td>
<td>0.231</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.143</td>
<td>0.090</td>
<td>0.159</td>
<td>0.237</td>
</tr>
<tr>
<td>Good Listener</td>
<td>0.143</td>
<td>0.121</td>
<td>0.224</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Simple connection of rating-related states with equal transition probabilities

Common states with concatenated training improve prediction performance
Summary and Future work

• A novel approach for predicting user satisfaction transitions using HMMs
  – Uses only the dialogues with overall user satisfaction ratings
  – Reduces the cost for training prediction models
  – Can be a viable option for evaluation
    • Overall ratings can be obtained easily

• Future work
  – New emissions to improve prediction accuracy
  – Apply our HMMs to other tasks