Overview

Problem: Automating replies to chat messages in the user's voice.
Solution: Classify the incoming message, pick a generator trained for that class, create the response.
Results: Generated replies that were plausible ~30% of the time.

Data

Personal: SMS (4.8k) and Whatsapp messages (3.6k).
Public: Marsan Ma Twitter chats (700k+)
Train/dev/test split: 0.8/0.1/0.1

Features

SMS (left) and Whatsapp (right) messages visualized with TSNE.

<table>
<thead>
<tr>
<th>Raw input</th>
<th>Feature</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each message</td>
<td>BOW vector</td>
<td>Classify messages based on their words</td>
</tr>
<tr>
<td>Each word</td>
<td>Word index</td>
<td>To transform into sequences / embeddings; count word occurrences</td>
</tr>
</tbody>
</table>

Models

K-means + Markov Chain (K+MC)
1. Train clusters/chains
   - Training data
   - K-means
   - Cluster 0
   - Cluster 0's Markov Chain
   - . . .
   - Cluster K
   - Cluster K's Markov Chain
   - . . .
2. Predict reply
   - New msg "Meet at 7?"
   - Find nearest cluster, J
   - Cluster J's Markov Chain
   - Generated reply "K"

Alternative: Seq2Seq+Attention (s2s+A)
State-of-the-art, generates meaningful sentences

Alternative (Baseline): K-means (K)
Like K+MC, but pick any response from nearest cluster. Mostly irrelevant responses, but showed that messages could be clustered meaningfully (k=11 better than lower k's). Clusters had themes like "making plans" and "catching up".

Results

"Plausibility": Whether the reply (1) sounds like user; (2) is relevant to the input message. Human evaluated.

<table>
<thead>
<tr>
<th>Model</th>
<th>Example</th>
<th>Sounds like user</th>
<th>Input-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>x: ok cool. hopping in singles y: i'm on the right hand side h: ohh why is there chocolate? thanks!</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td>K+MC</td>
<td>x: how's the price? y: yay april then! h: 15-20%? I'm really it's etc://</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td>s2s+A (w/ personal data)</td>
<td>x: eat first or swim? y: yep h: 20 if if if if korean korean cost</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>s2s+A (w/ public data)</td>
<td>x: busy at work? y: been playing com games ohgod h: i love you</td>
<td>0.05</td>
<td>0.45</td>
</tr>
</tbody>
</table>

K+MC replies were more plausible than s2s+A's but sentence structures were worse.

Discussion

Generated replies were usually bad since Both goals conflict with each other: "Sounds like user" ~ 1 / "Relevance to input"
Hard to train: too little data. Models did worst:
(a) MC on clusters vs all data; (b) s2s on personal vs public data.
Hard to evaluate: loss/ BLEU/ perplexity does not account for relevance/ personality.

Future Work

Find data-cheap method to permute s2s+A generated replies to personalized voice. Tried RL (MDP) on word characters but this did poorly.