Generating Stylistically Consistent Dialog Responses with Transfer Learning

Reina Akama\textsuperscript{1}
Kazuaki Inada\textsuperscript{1}, Naoya Inoue\textsuperscript{1}, Sosuke Kobayashi\textsuperscript{2}, Kentaro Inui\textsuperscript{1,3}

\textsuperscript{1} Graduate School of Information Sciences, Tohoku University
\textsuperscript{2} Preferred Networks, Inc.
\textsuperscript{3} RIKEN Center for Advanced Intelligence Project
Prior approaches

- Feed large-scale conversational data into sequence-to-sequence (seq2seq) model [Vinyals+‘15, Li+’16, etc.]
- **Problem:** generates **NOT** **stylistically consistent responses**
Our goal

- Generate dialog responses with **particular style consistently** while maintaining goodness of contents and fluency
Proposed method

- Apply **transfer learning** to seq2seq response generation model
- Propose **two-staged** training framework

1. **Dialog training**

2. **Style training**

<table>
<thead>
<tr>
<th>( V_d )</th>
<th>( \hat{V}_d )</th>
<th>( V_d )</th>
<th>( \hat{V}_d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>utterance pairs</td>
<td>vocabulary</td>
<td>Trans</td>
<td>Trans+alt</td>
</tr>
</tbody>
</table>

**Diagram Description:**
- The diagram illustrates the proposed method for generating stylistically consistent dialog responses using transfer learning.
- It consists of two main stages: **Dialog training** and **Style training**.
- In **Dialog training**, the model is trained with a sequence-to-sequence (seq2seq) architecture, where the input is the dialog corpus and the output is the generated text.
- The **Style training** stage involves a second seq2seq model that alternates between two vocabularies, \( V_d \) and \( \hat{V}_d \), to achieve stylistic consistency.
- The **Style corpus** is used for training the style model, which outputs either trans or trans+alt styles.
1st stage: Dialog training

- Focus on learning **what-to-say** and **fluent** responses
- Exploit cheap, large-scale but ** stylistically inconsistent** dialog corpus

**Dialog training**

- seq2seq
- $V_d$ $\tilde{V}_d$
- utterance pairs
- vocabulary
- Dialog corpus

**Style training**

- seq2seq
- $V_d$ $\tilde{V}_d$
- Trans
- Trans+alt
- Large dialog corpus **without** stylistic consistency
**2\textsuperscript{nd} stage: Style training**

- Focus on learning \textit{stylistically consistent} responses
- Fine-tune seq2seq model trained in dialog training
- Exploit expensive, small-scale but \textit{stylistically consistent} corpus

---

**Diagram:**

1. \textit{Dialog training}:
   - All the model parameters
     - Word embeddings, weight matrices in RNNs
   - \( V_d \)
   - \( \hat{V}_d \)
   - Utterance pairs
   - Vocabulary

2. \textit{Style training}:
   - seq2seq
   - \( V_d \)
   - \( \hat{V}_d \)
   - \( W_s \)
   - Style corpus

- Style training with transfer

---

**Text:**

**Small conversational corpus with stylistic consistency**
Seq2seq vocabulary in style training

**Issue:** how to create the seq2seq vocab. for style training?

1. **Trans**
   - Simply use the same vocabulary $V_d$ as Dialog training
   - $V_d$: the top $N_d$ most frequent words in dialog corpus

2. **Trans+alt**
   - Alter the vocabulary before style training
   - Replace infrequent words $\hat{V}_d$ from dialog corpus with frequent words $W_s$ in style corpus
Dataset

- Japanese single-turn dialog (i.e., utterance pairs)

- **Dialog corpus**
  - 3.7M utterance pairs extracted from tweet-reply chains

- **Style corpus**
  - Feminine and polite style
  - 12K utterance pairs extracted from TV subtitles
    - Only 0.3% of dialog corpus
  - TV program where *Tetsuko Kuroyanagi* chats with some guests

  - Japanese TV personality
  - Elderly woman
  - Speaks feminine and polite words
# Experiments setup

## Models

<table>
<thead>
<tr>
<th>features</th>
<th>models</th>
<th>Baselines</th>
<th></th>
<th>Proposals</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Base</td>
<td>Mix</td>
<td>Trans</td>
<td>Trans+alt</td>
</tr>
<tr>
<td>Dialog corpus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Style corpus</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vocabulary alternation</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Transfer learning</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

## Settings

- **Seq2seq**: 2 layer LSTMs with 2048 units
- **Optimization**: Adam (mini-batch size 64)
- **Word embeddings**: 1024 dim
- **Vocab. size**: 25K

## Vocabulary alternation (Trans+alt)

- **Vocab. size**: 1K words from style corpus ($W_s$)

---

Generating Stylistically Consistent Dialog Responses with Transfer Learning
Evaluation setup

Human evaluation
- 5 crowd workers judged each responses generated by each model
- Given input, generated response, style description, and 2 questions:
  Q1. Whether the response is grammatically and semantically appropriate
  Q2. Whether the style of response matches the given description

Evaluation metrics
- The percentage of responses judged as
  1. appropriate responses
  2. stylistically consistent

Generated: Oh, I’m surprised.
Style: feminine, polite
### Results

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Mix</th>
<th>Trans</th>
<th>Trans+alt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Appropriate responses (%)</strong></td>
<td>49.0</td>
<td>50.0</td>
<td>53.0</td>
<td>60.0<strong>B</strong></td>
</tr>
<tr>
<td><strong>Stylistically consistent (%)</strong></td>
<td>19.5</td>
<td>22.5</td>
<td>30.5<strong>B</strong></td>
<td>34.0<strong>B</strong></td>
</tr>
</tbody>
</table>

*B, *M: Statistical significance against Base, Mix. Sign test, p<0.05,0.01 for *, **.

- **Transfer learning** frameworks are successful
- **Increase stylistic consistency of** generated responses
- **Maintain the appropriateness of** responses
- **Vocabulary alternation** helps to make the more stylistically consistent responses
Conclusion

Summary

• Presented novel end-to-end framework to build **stylistically consistent** dialog response generation system
• Apply transfer learning to seq2seq response generation model
• Human evaluation demonstrated that proposed method produces stylistically consistent responses while maintaining appropriateness of responses

Future work

• Improve style training so that it can learn only the style of responses
• Explore effective way of creating style corpus
  • e.g., automatically collecting polite utterance from a large Twitter corpus
Appendixes
References


Generated responses

- Generated responses by out best model Trans+alt
- Feminine and polite style

<table>
<thead>
<tr>
<th>input</th>
<th>generated responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>お住まいはどちらでしょうか？ Where do you live?</td>
<td>京都です I live in Kyoto.</td>
</tr>
<tr>
<td>ドル安ですね The dollar is weak, isn’t it?</td>
<td>そうですね Yes, it is.</td>
</tr>
<tr>
<td>月が綺麗 The moon is beautiful.</td>
<td>すごい Marvelous!</td>
</tr>
</tbody>
</table>
Human evaluation

- Each model generated 200 responses for evaluation
  - input: 50 utterances
  - generated 4 responses for each utterance
- Each response is judged by 5 crowd workers
  - Via Yahoo! Crowd Sourcing (https://crowdsourcing.yahoo.co.jp/)
  - Given input, generated response, \textit{style description}
    and 2 yes-no questions
  - Final answer is determined by majority vote

\* only a description with several example
\* without the specific name of individual character