Recent Advances towards Dialogue Systems in Open Domain

Dr. Rui Yan
Wangxuan Institute of Computer Technology, Peking University
ruiyan@pku.edu.cn
www.ruiyan.me
Conversational AI

- Human-computer conversation has been attracting increasing attention.
- Conversational agent (ChatBot)
  - e.g., Xiaoice (Microsoft), Turing Robot
- Virtual personal assistant
  - e.g., Cortana (Microsoft), Siri (Apple), Now (Google)
- E-commerce customer service robot
  - e.g., Alime (Alibaba), Jimi (JingDong)
Taxonomy of Dialogue Systems

- **Domain**
  - Vertical domain (Task driven)
    - Complete domain-specific tasks (e.g., hotel booking, weather enquiries, etc)
  - Open domain (Non-task driven)
    - Naturally and meaningfully converse with humans on any open domain topics

- **Technique**
  - Templated-based
  - Retrieval-based
  - Generation-based
  - Ensemble-based
Retrieval-based Approaches
Matching with Better Representation

I. Representations Go Deep

➢ Deep Attention Matching Network (DAM)
  • Representing utterances and responses by stacking multiple attention modules

- Self-attention: let an utterance attend itself
- Cross-attention: let an utterance and a response attend each other

Modeling long-term dependency among words

[Zhou et al., ACL 2018]
Matching with Better Representation

I. Representations Go Deep

➢ Deep Attention Matching Network (DAM)
  ● Representing utterances and responses by stacking multiple attention modules

---

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th>Douban</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
<td>$R_{10}@1$</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
<td>0.726</td>
</tr>
<tr>
<td>DAM</td>
<td>0.938</td>
<td>0.767</td>
</tr>
</tbody>
</table>

- **Cross-attention**: let an utterance and a response attend each other

---

Zhou et al., Multi-turn Response Selection for Chatbots with Deep Attention Matching Network. ACL’18
Matching with Better Representation

II. Representations Go Wide

➢ Multi-Representation Fusion Network (MRFN)
  • Fusing multiple types of representations are helpful, but how to fuse matters.

- Word2Vec
- Char-based Embedding
- CNN
- RNN
- Self-attention
- Cross-attention

[Tao et al., WSDM 2019]
Matching with Better Representation

II. Representations Go Wide

- Multi-Representation Fusion Network (MRFN)
  - Fusing multiple types of representations are helpful, but how to fuse matters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th>Douban</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
<td>$R_{10}@1$</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
<td>0.726</td>
</tr>
<tr>
<td>DAM</td>
<td>0.938</td>
<td>0.767</td>
</tr>
<tr>
<td>MRFN(FES)</td>
<td>0.930</td>
<td>0.742</td>
</tr>
<tr>
<td>MRFN(FIS)</td>
<td>0.936</td>
<td>0.762</td>
</tr>
<tr>
<td>MRFN(FLS)</td>
<td>0.945</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>+0.7%</td>
<td>+1.9%</td>
</tr>
</tbody>
</table>

[Tao et al., WSDM 2019]
III. Representations from Pre-Training

➢ Pre-training neural networks on large scale data sets as representations significantly improves the existing models.

Table 1: Model comparison on Ubuntu Corpus V1.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R_{10}@1$</th>
<th>$R_{10}@2$</th>
<th>$R_{10}@5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiView</td>
<td>0.662</td>
<td>0.801</td>
<td>0.951</td>
</tr>
<tr>
<td>DL2R</td>
<td>0.626</td>
<td>0.783</td>
<td>0.944</td>
</tr>
<tr>
<td>AK-DE-biGRU</td>
<td>0.747</td>
<td>0.868</td>
<td>0.972</td>
</tr>
<tr>
<td>SMN_{dynamic}</td>
<td>0.726</td>
<td>0.847</td>
<td>0.961</td>
</tr>
<tr>
<td>DUA</td>
<td>0.752</td>
<td>0.868</td>
<td>0.962</td>
</tr>
<tr>
<td>DAM</td>
<td>0.767</td>
<td>0.874</td>
<td>0.969</td>
</tr>
<tr>
<td>IMN</td>
<td>0.777</td>
<td>0.888</td>
<td>0.974</td>
</tr>
<tr>
<td>ESIM</td>
<td>0.796</td>
<td>0.894</td>
<td>0.975</td>
</tr>
<tr>
<td>MRFN_{FLS}</td>
<td>0.786</td>
<td>0.886</td>
<td><strong>0.976</strong></td>
</tr>
<tr>
<td>BERT_{base}</td>
<td>0.817</td>
<td>0.904</td>
<td>0.977</td>
</tr>
<tr>
<td>BERT-DPT</td>
<td>0.851</td>
<td>0.924</td>
<td>0.984</td>
</tr>
<tr>
<td>BERT-VFT</td>
<td><strong>0.855</strong></td>
<td><strong>0.928</strong></td>
<td><strong>0.985</strong></td>
</tr>
<tr>
<td>BERT-VFT(DA)</td>
<td><strong>0.858</strong></td>
<td><strong>0.931</strong></td>
<td><strong>0.985</strong></td>
</tr>
</tbody>
</table>

[Whang et al., arXiv]
Matching with Better Interaction

- Interaction-over-interaction network
  - Representations-[Interaction]$^K$-Aggregation

[Tao et al., ACL 2019]
Matching with Better Interaction

- Interaction-over-interaction network
  - Representations-[Interaction]^K -Aggregation

[Tao et al., ACL 2019]

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th>Douban</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(R_2@1)</td>
<td>(R_{10}@1)</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
<td>0.726</td>
</tr>
<tr>
<td>DAM</td>
<td>0.938</td>
<td>0.767</td>
</tr>
<tr>
<td>MRFN(FLS)</td>
<td>0.945</td>
<td>0.786</td>
</tr>
<tr>
<td>IOI</td>
<td>0.947</td>
<td>0.796</td>
</tr>
</tbody>
</table>

Initial Representation:
- Self-attention
- Interaction Operation
- Add Operation
Matching with External Knowledge

Knowledge is incorporated into matching through Pseudo Relevance Feedback

Knowledge is incorporated into matching through an Extra Matching Channel

[Yang et al., SIGIR 2018]
Learning a Better Matching model

- Learning with Co-Teaching – Denoising with Your Peer

Key Ideas

- Teaching: two models judge quality of training examples mutually. The knowledge is transferred between the two models through learning protocols.
- Learning: two models learn from their peers via the transferred learning protocols.
- Co-evolving: through teaching and learning, the two models get improved together.
- Resemble: two peer students who learn from different but related materials inspire each other during learning through knowledge exchange.

[Feng et al., ACL 2019]
Generation-based Approaches
Response Diversity

- Modeling the 1-to-n mapping by considering the correlation of different valid responses.

[Xu et al., ACL 2019]
Response Diversity

- Controlling multiple attributes in response generation (customize responses by tailoring the set of attributes)

Figure 1: Architecture of goal tracking memory enhanced sequence-to-sequence model.
Retrieval-then-Generation


Zhu et al., ACL 2019

Retrieval-based Method

Training Set

Index

Candidate#1: Could you tell me how this thing is cooked?
Candidate#2: Tiramisu is my favorite dessert. It’s so delicious.

N-best Response Candidates

Retrieval

Encoder

Decoder

Generator

Message: I made strawberry shortcake.

Policy Gradient


Discriminator

MLP

Concatenation

Average

\( z^X \)

\( z^{o_1} \) \( z^{o_2} \) ...

17
Retrieval-then-Generation

A prototype-then-edit paradigm for response generation

Table 1: An example of context-aware prototypes editing. Underlined words mean they do not appear in the original context, while words with strikethrough mean they are not in the prototype context. Words in bold represent they are modified in the revised response.

<table>
<thead>
<tr>
<th>Context</th>
<th>My friends and I went to some vegan place for dessert yesterday.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype context</td>
<td>My friends and I had Tofu and vegetables at a vegan place nearby yesterday.</td>
</tr>
<tr>
<td>Prototype response</td>
<td>Raw green vegetables are very beneficial for your health.</td>
</tr>
<tr>
<td>Revised response</td>
<td>Desserts are very bad for your health.</td>
</tr>
</tbody>
</table>

[Wu et al., AAAI 2019]
Retrieval-Generation Ensemble

- An adversarial learning framework for enhancing a retrieval-generation ensemble model (mutual enhanced)

[Zhang et al., SIGIR 2019]
More Challenges to Dialogue System

- Larger context (especially in multi-turn dialogue)
  - How to encode long context information
  - Does the neural model understand the context?
- How to select/generate a meaningful and coherent reply?
- How to incorporate domain knowledge, world knowledge?
- How to design fast (approximate) algorithms for efficient training and inference.
  - Current dialog models are clumsy and require long training time and huge computational power.
Looking Forward

- Model Design -> Model Learning
- Single Modality -> Multiple Modalities
- Big Data -> Small Data
  - Big data: mixed intentions, styles, characters, ……
  - Small data: specific intention, style, character, ……
Single Modality to Multiple Modalities

Conversation + Scene + Sentiment
Caption and Question
Image Scene + Sentiment
Facial Expression

Conversation Turn 1:
“Hanging out on Saturday.”

Conversation Turn 2:
“Did you have a good time?”

Huber et al., Emotional Dialogue Generation using Image-Grounded Language Models. CHI’18

Chu et al., A Face-to-Face Neural Conversation Model. CVPR’18
Q&A

ruiyan@pku.edu.cn