USTC at NTCIR-12 STC Task

Junbei Zhang  Junfeng Hou  Shiliang Zhang  Lirong Dai

University of Science and Technology of China

Presented by Junbei Zhang

2016.06.09

zjunbei@mail.ustc.edu.cn
• Introduction
• System Architecture
• Experiment and analysis
• Conclusion and future work
Task

- Short Text Conversation (STC)

Given a new post, can an appropriate (i.e. human-like) comment be returned by searching a post-comment repository?

Previous Model

- Neural Responding Machine for Short-Text Conversation [Shang et al. 2015]
- A neural conversational model [Vinyals et al. 2015]
• Introduction
• **System Architecture**
• Experiment and analysis
• Conclusion and future work
System Architecture

• Learning-to-rank
• **Lexical Features:**
  ‣ Query-Response Similarity
  ‣ Query-Post Similarity
  ‣ Transition-p2c
• **Semantic Features:**
  ‣ EncDec-Forward model
  ‣ EncDec-Reverse model
  ‣ Joint-Train model
• **Ranking:** linear RankingSVM
Lexical Features

• **Query-Response Similarity**
  - Map query and response to their own **TF-IDF score vector**
  - Calculate **cosine similarity** between query and response

• **Query-Post Similarity**
  - TF-IDF score vector
  - The cosine similarity between query and post is regarded as response score.
Transition-p2c

• Model the transition probability between post words and response words

• Example:
  ▶ Query: Where shall we eat?
  ▶ Query-Response: We eat two apples.
  ▶ Query-Post: Where shall we go? -> We go to the bank.
  ▶ Transition-p2c: Restaurant.
Transition-p2c

• Train:

\[ \text{Response-TF} \times \text{ResponseSet-IDF} \]

\[ \text{Post-TF} \times \text{PostSet-IDF} \]

Add

Transition Matrix
Transition-p2c

• Test:

Query-TF \* PostSet-IDF

Response-TF \* ResponseSet-IDF

Element-wise

Score

Transition Matrix
Semantic Features

- **EncDec-Forward model:**
  - seq2seq model to estimate $P(\text{Response} \mid \text{Post})$
• **EncDec-Reverse model:**
  - **Many-to-many:** one post → several responses, one response ← several posts
  - **Reconstruction:** max P(Post | Response)

```
A → B → C → <eos> → W → X → Y → Z → <eos>
```

Response  Post

```
hahaha
```
• Joint-Train model:
  ▪ **Generation & Reconstruction:** combine $P(\text{Response} \mid \text{Post})$ and $P(\text{Post} \mid \text{Response})$ in one model
  ▪ Decoder1 is regarded as **both a decoder and an encoder** (Encoder2)
• Linear RankingSVM
• **Merge all the scores** and outputs a final score for each query and response pair
• Introduction
• System Architecture
• **Experiment and analysis**
• Conclusion and future work
Implementation

• Subset:
  › Using 4 Models to get top 10000 responses each from the whole repository, and remove the same pairs
  › Query-Response Similarity, Query-Post Similarity, Transition-p2c and EncDec-Reverse

• loss function (Joint-Train):
  › loss = 0.2 * post-comment loss (encoder1&decoder1) + 0.8 * comment-post loss(encoder2&decoder2)
Results

• 5 Runs:
  • **USTC-C-R1**: Query-Response Similarity + Query-Post Similarity + EncDec-Forward + **EncDec-Reverse** + **Transition-p2c**
  • **USTC-C-R2**: Query-Response Similarity + Query-Post Similarity + EncDec-Forward + **EncDec-Reverse** + **JointTrain**
  • **USTC-C-R3**: Query-Response Similarity + Query-Post Similarity + EncDec-Forward + **Transition-p2c**
  • **USTC-C-R4**: Query-Response Similarity + Query-Post Similarity + EncDec-Forward + **EncDec-Reverse**
  • **USTC-C-R5**: Query-Response Similarity + Query-Post Similarity + EncDec-Forward
Results

• Official Results:

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@1</th>
<th>P+</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>R5</td>
<td>0.2867</td>
<td>0.4509</td>
<td>0.4160</td>
</tr>
<tr>
<td>R4</td>
<td>0.2767</td>
<td>0.4479</td>
<td>0.4181</td>
</tr>
<tr>
<td>R1</td>
<td>0.2733</td>
<td>0.4499</td>
<td>0.4169</td>
</tr>
<tr>
<td>R2</td>
<td>0.2567</td>
<td>0.4310</td>
<td>0.4001</td>
</tr>
<tr>
<td>R3</td>
<td>0.2267</td>
<td>0.4094</td>
<td>0.3848</td>
</tr>
</tbody>
</table>
Results

- Offline Results (using 6017 labeled data):

Table 2: STC (Chinese) training set results

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@1</th>
<th>P+</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>R5</td>
<td>0.4741</td>
<td>0.6529</td>
<td>0.6327</td>
</tr>
<tr>
<td>R4</td>
<td>0.4785</td>
<td>0.6582</td>
<td>0.6395</td>
</tr>
<tr>
<td>R3</td>
<td>0.4726</td>
<td>0.6570</td>
<td>0.6347</td>
</tr>
<tr>
<td>R2</td>
<td>0.4889</td>
<td>0.6625</td>
<td>0.6446</td>
</tr>
<tr>
<td>R1</td>
<td>0.4859</td>
<td>0.6618</td>
<td>0.6449</td>
</tr>
</tbody>
</table>
Results

• Inconsistent:

Figure 5: Diagram of the retrieval-based automatic response system.

A Dataset for Research on Short-Text Conversations. [Wang et al. EMNLP2013]
## Results

- **Transition-p2c Case Study:**

<table>
<thead>
<tr>
<th>post words</th>
<th>comment words</th>
<th>transition score</th>
</tr>
</thead>
<tbody>
<tr>
<td>运费 (freight)</td>
<td>代购 (purchasing agents)</td>
<td>0.3207</td>
</tr>
<tr>
<td>中型 (medium)</td>
<td>谢谢 (thanks)</td>
<td>0.1302</td>
</tr>
<tr>
<td>警报 (alarm)</td>
<td>口水 (saliva)</td>
<td>0.1273</td>
</tr>
<tr>
<td>元宵节 (Lantern Festival)</td>
<td>快乐 (happy)</td>
<td>0.1260</td>
</tr>
<tr>
<td>萌到 (sprout)</td>
<td>可爱 (lovely)</td>
<td>0.1180</td>
</tr>
<tr>
<td>拜年 (pay a New Year call)</td>
<td>新年快乐 (happy new year)</td>
<td>0.1177</td>
</tr>
<tr>
<td>王老吉 (Wong Lo Kat)</td>
<td>加多宝 (JDB Beverage)</td>
<td>0.1077</td>
</tr>
<tr>
<td>本地 (native)</td>
<td>流量 (traffic)</td>
<td>0.1066</td>
</tr>
<tr>
<td>小家伙 (kiddy)</td>
<td>可爱 (lovely)</td>
<td>0.1042</td>
</tr>
<tr>
<td>张国荣 (Leslie Cheung)</td>
<td>哥哥 (brother)</td>
<td>0.1007</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• System Architecture
• Experiment and analysis
• Conclusion and future work
Conclusions

• Our model:
  • **Lexical** Features:
    ‣ Query-Response Similarity
    ‣ Query-Post Similarity
    ‣ Transition-p2c
  • **Semantic** Features:
    ‣ EncDec-Forward model
    ‣ EncDec-Reverse model
    ‣ Joint-Train model
Conclusions

• The results in training set and cases show the efficiency of our method.

• Inconsistent: different evaluation indexes & the subset selection problem.
Future Work

• Try deep encoder-decoder model
• Try word2vec instead of TF-IDF