Cluster-based Beam Search for Pointer-Generator Chatbot Grounded by Knowledge

Yik-Cheung Tam, Jiachen Ding\textsuperscript{2}, Cheng Niu, Jie Zhou
WeChat AI - Pattern Recognition Center, Tencent Inc
Shanghai Jiaotong University\textsuperscript{2}

\{wilsontam,niucheng,withtomzhou\}@tencent.com
ding222@sjtu.edu.cn\textsuperscript{2}
Agenda

• Introduction
• Proposed System
  • Model
  • Decoding Strategies
• Evaluation Results
• Conclusions
Introduction

• DSTC7 Track 2 challenge: Design a chatbot that is:
  • Grounded by unstructured external knowledge (facts)
  • Conversational-history aware
  • End-to-end trainable

• Diverse response generation
  • Relevant
  • Interesting
Issues

• How to integrate conversational history (H) and external knowledge (F)?

• How to generate diverse responses (R)?
Related Work

• Seq-to-Seq with copy mechanism
  • Chit-Chat: Gu 2016, See 2017
  • KB: Eric and Manning 2017

• Response diversity
  • MMI: Li 2016, Zhang 2018, Baheti 2018
  • Latent variables: Zhou 2017, Su 2018

• Beam search decoding
  • Diversity objective: Vijayakumar 2018

• Knowledge-grounded
  • Memory network: Ghazvininejad 2018, Madotto 2018

Many more papers..
Contributions

Model:
• Enable pointer generator to copy from history and facts

Decoding Strategies:
• Cluster-based beam search
• Safe response filtering using LM
Overall System Diagram
Pointer-Generator

• A Seq-to-Seq model with copy-mechanism [See et.al. 2017]
  • OOV can be copied to output

• At each decoding step, our model chooses to:
  1. Generate a token based on vocabulary V
  2. Copy a token from history
  3. Copy a token from facts
One decoding step: \( \text{Pr(} \text{word} \mid F, H) \)
Attention Mechanism

• BiLSTM history vectors: \( \{ h_1^H, h_2^H, ..., h_i^H, ..., h_L^H \} \)
• BiLSTM fact vectors: \( \{ h_1^F, h_2^F, ..., h_j^F, ..., h_T^F \} \)

Attention over history vectors at decoding step t

\[
e_{ti}^H = v_H^T \cdot \tanh(W_h^H \cdot h_i^H + W_r^H \cdot h_t^R + b^H)
\]

\[
\alpha_{ti}^H = \text{Softmax}(e_{ti}^H)
\]

\[
h_t^{H*} = \sum_{i=1}^{L} \alpha_{ti}^H h_i^H
\]

Attention over fact vectors at decoding step t

\[
e_{tj}^F = v_F^T \cdot \tanh(W_h^F \cdot h_j^F + W_r^F \cdot h_t^R + b^F)
\]

\[
\alpha_{tj}^F = \text{Softmax}(e_{tj}^F)
\]

\[
h_t^{F*} = \sum_{j=1}^{T} \alpha_{tj}^F h_j^F
\]
Mode prediction

- \{Generate, Copy from fact, Copy from history\}
- Concatenate all available features at decoder step t:
  - Attended fact vector
  - Attended history vector
  - Decoder hidden state
  - Last input word embedding
- Feed-Forward followed by Softmax over modes:

\[
Pr(mode = m|t, H, F) = \text{Softmax}(FF(h_t^{F*} \oplus h_t^{H*} \oplus h_t^R \oplus x_t))
\]
Output word distribution

• Linearly interpolate distributions from 3 modes:

\[ Pr(w|t, H, F) = \sum_{m=1}^{3} Pr(m|t, H, F) \cdot Pr_m(w|t, H, F) \]

• End-to-end trainable with cross-entropy
Decoding Issues

• Motivation: Safe responses are common in responses generated by Seq-to-Seq models
  • “This is the best thing I have ever seen”

• Observation from beam search:
  • Many responses are similar
    • “This is the best thing I have ever seen”
    • “This is the coolest thing I have ever seen”
  • Under a fixed beam budget, this is inefficient

Proposal: Cluster-based beam search
• Cluster similar partial hypotheses
• Prune per cluster
Beam Search with K-means

- Averaged word embedding to represent a candidate
- Apply K-means over extended candidates
- Prune candidates per cluster using Beam Size / K
- Remove repeated N-grams
- Filter out final meaningless candidates using LM

Algorithm 1: Beam search with K-means

\[
\text{Input: Beam size BS, Candidates } C \text{ initialized with start symbol}
\]
\[
\text{Output: Final response } rsp
\]
\[
\text{Data: Language model threshold } lm_{th}
\]

\[
\text{while Number of completed hypothesis does not reach BS or maximum decoding step is not reached do}
\]
\[
\text{for } i \text{ in BS do}
\]
\[
\text{tmpHyps=Top-N(Extend(C[i]), BS \times 2);}
\]
\[
\text{Remove hyps in tmpHyps with repeated N-grams or UNK;}
\]
\[
\text{Save tmpHyps to extended candidates;}
\]
\[
\text{end}
\]
\[
\text{Perform K-means over extended candidates;}
\]
\[
\text{for candidates in each cluster do}
\]
\[
\text{Sort candidates by partial log-prob scores;}
\]
\[
\text{Choose top BS/K candidates;}
\]
\[
\text{Put candidates with end symbol in } R;
\]
\[
\text{Put incomplete candidates in } C_{new};
\]
\[
\text{end}
\]
\[
C' \leftarrow C_{new}
\]
\[
\text{end}
\]
\[
\text{Sort } R \text{ according to log-prob scores;}
\]
\[
\text{for hyp in } R \text{ do}
\]
\[
\text{if } score_{lm}(hyp) < lm_{th} \text{ then}
\]
\[
\text{rsp } \leftarrow hyp;
\]
\[
\text{break;}
\]
\[
\text{end}
\]
\[
\text{end}
\]
Experimental Setup

- Glove word embedding
- Decode validation set to obtain N-best responses to train an LM for safe response filtering
- Single system

Table 1: Dataset statistics for DSTC7 Track 2.

<table>
<thead>
<tr>
<th></th>
<th>Train set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>1,408,951</td>
<td>4,542</td>
<td>13,108</td>
</tr>
</tbody>
</table>

Table 2: Hyper-parameter settings.

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>100,000</td>
</tr>
<tr>
<td>Word embedding size</td>
<td>300</td>
</tr>
<tr>
<td>LSTM hidden size</td>
<td>150</td>
</tr>
<tr>
<td>Batch size</td>
<td>128</td>
</tr>
<tr>
<td>Beam size (BS)</td>
<td>50</td>
</tr>
<tr>
<td>K-means clusters</td>
<td>10</td>
</tr>
<tr>
<td>N-gram repeat order</td>
<td>2</td>
</tr>
<tr>
<td>LM threshold ((l_{m_{th}}))</td>
<td>-35</td>
</tr>
<tr>
<td>Learning rate for Adam</td>
<td>0.0005</td>
</tr>
<tr>
<td>Maximum gradient norm</td>
<td>2</td>
</tr>
</tbody>
</table>
Official Evaluation Results

• Achieved best results on NIST-4, BLEU-4, Meteor
  • Cluster-based beam search help

Table 3: Automatic evaluation results. A total of 2208 test samples were evaluated. Best non-baseline results are marked in bold.

<table>
<thead>
<tr>
<th>Name</th>
<th>NIST-4</th>
<th>BLEU-4</th>
<th>Meteor</th>
<th>Entropy-4</th>
<th>Div-1</th>
<th>Div-2</th>
<th>Avg len</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (constant)</td>
<td>0.184</td>
<td>2.87%</td>
<td>7.48%</td>
<td>1.609</td>
<td>0.000</td>
<td>0.000</td>
<td>8</td>
</tr>
<tr>
<td>Baseline (random)</td>
<td>1.637</td>
<td>0.86%</td>
<td>5.91%</td>
<td>10.467</td>
<td>0.160</td>
<td>0.647</td>
<td>19.192</td>
</tr>
<tr>
<td>Baseline (seq2seq)</td>
<td>0.916</td>
<td>1.82%</td>
<td>6.96%</td>
<td>5.962</td>
<td>0.014</td>
<td>0.048</td>
<td>10.604</td>
</tr>
<tr>
<td>Team C/E</td>
<td>1.515</td>
<td>1.32%</td>
<td>6.43%</td>
<td>7.639</td>
<td>0.053</td>
<td>0.171</td>
<td>12.674</td>
</tr>
<tr>
<td>Team G</td>
<td>2.040</td>
<td>1.05%</td>
<td>7.48%</td>
<td><strong>10.057</strong></td>
<td>0.108</td>
<td><strong>0.449</strong></td>
<td>22.336</td>
</tr>
<tr>
<td>Our system w/ K-means</td>
<td><strong>2.523</strong></td>
<td>1.83%</td>
<td><strong>8.07%</strong></td>
<td>9.030</td>
<td>0.109</td>
<td>0.325</td>
<td>15.133</td>
</tr>
<tr>
<td>Our system w/o K-means</td>
<td>1.771</td>
<td><strong>1.94%</strong></td>
<td>7.64%</td>
<td>8.194</td>
<td>0.094</td>
<td>0.267</td>
<td>12.770</td>
</tr>
<tr>
<td>Human</td>
<td>2.650</td>
<td>3.13%</td>
<td>8.31%</td>
<td>10.445</td>
<td>0.167</td>
<td>0.670</td>
<td>18.757</td>
</tr>
</tbody>
</table>
Official Human Evaluation Results

- 5-level judgement, 3 judges
- Achieved best scores on “Interest and Informativeness” & Overall

<table>
<thead>
<tr>
<th>Model</th>
<th>Relevance</th>
<th>Mean Score</th>
<th>95% CI</th>
<th>Interest</th>
<th>Mean Score</th>
<th>95% CI</th>
<th>Overall</th>
<th>Mean Score</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (constant)</td>
<td></td>
<td>2.60</td>
<td>(2.560, 2.644)</td>
<td>2.32</td>
<td>(2.281, 2.364)</td>
<td>2.46</td>
<td>(2.424, 2.500)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (random)</td>
<td></td>
<td>2.32</td>
<td>(2.269, 2.371)</td>
<td>2.35</td>
<td>(2.303, 2.401)</td>
<td>2.34</td>
<td>(2.288, 2.384)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (seq2seq)</td>
<td></td>
<td>2.91</td>
<td>(2.858, 2.963)</td>
<td>2.68</td>
<td>(2.632, 2.730)</td>
<td>2.80</td>
<td>(2.748, 2.844)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team C/E (2nd place)</td>
<td></td>
<td><strong>3.05</strong></td>
<td><strong>(3.009, 3.093)</strong></td>
<td>2.77</td>
<td>(2.735, 2.812)</td>
<td>2.91</td>
<td>(2.875, 2.950)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team G (3rd place)</td>
<td></td>
<td>2.82</td>
<td>(2.771, 2.870)</td>
<td>2.57</td>
<td>(2.525, 2.619)</td>
<td>2.70</td>
<td>(2.650, 2.742)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our system (1st place)</td>
<td></td>
<td>2.99</td>
<td>(2.938, 3.042)</td>
<td><strong>2.87</strong></td>
<td><strong>(2.822, 2.922)</strong></td>
<td><strong>2.93</strong></td>
<td><strong>(2.882, 2.979)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>3.61</td>
<td>(3.554, 3.658)</td>
<td>3.49</td>
<td>(3.434, 3.539)</td>
<td>3.55</td>
<td>(3.497, 3.596)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case Study

- Response with K-means beam search is better

<table>
<thead>
<tr>
<th>Conversational history</th>
<th>til that in greek mythology zeus thought that the earth is overpopulated and trojan war was instigated to depopulate the earth.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance Interest</td>
<td>4.67, 4.67</td>
</tr>
<tr>
<td>w/ K-means</td>
<td>james cameron wrote a book about the origins of war on this.</td>
</tr>
<tr>
<td>w/o K-means</td>
<td>i’ve never heard of this, but i didn’t know it was aliens.</td>
</tr>
</tbody>
</table>
Bad case

- Low relevance

<table>
<thead>
<tr>
<th>Conversational history</th>
<th>till: Chinese philanthropist, run run shaw (1907-2014) gave $1 billion to fund education, aiding over 6000 construction projects. Because so many buildings (5000+) bear his name, yifu, it is often confused as a generic name.</th>
</tr>
</thead>
</table>
| Relevance Interest     | 1.33  
|                        | 3.00  |
| w/ K-means             | James Cameron is the best name ever. I’m pretty sure he voiced the Tianyi of Tesla.            |
| w/o K-means            | He’s a Chinese restaurant. It’s his name!                                                   |
Conclusions

• Proposed an end-to-end trainable system for submission
• Achieved the 1st place in the competition
• Response quality requires further improvement
  • Relevance
  • Interest
  • Response consistency
Thank you!

WeChat AI - Pattern Recognition Center, Tencent Inc