TKUIM at NTCIR-14 STC-3 CECG task

Shih-chieh Wei, Chi-bin Cheng, Guan-zhong-yi Cao, Yi-jing Chiang, Yi-jing Chiang, Chin-yi Wu, Shih-hsiang Lin, Kun-li Tsai

Tamkang University, Taiwan, R.O.C.
Abstract

For the **Chinese Emotional Conversation Generation** (CECG) sub-task of **Short Text Conversation** (STC-3) task, we (TKUIM) built a system consisting of two parts, the response generation subsystem and the emotion classification subsystem.

For the **response generation subsystem**, we trained five generative models using different parts of the training dataset. These models will output response candidates based on a Seq2Seq deep learning architecture with the attention mechanism.

For the **emotion classification subsystem**, we trained an emotion classifier with probability output for each emotion class. According to the desired response emotion class, a corresponding emotion classifier is used to select the most probable response from the previous response candidates. An emotion accept threshold and a default response library are set up for each response emotion class. When the selected response does not pass the emotion accept threshold, a default response from the library for that emotion class is output to replace the poorly generated response.

In this mission, we submitted only one valid result, which got an average total score of **0.726** within a maximum scale of 2.
The subtask aims to generate conversation text of a desired emotion class. In order to produce the conversation text required for the task, we built a system composed of response generation and emotion classification.
Methodology for Emotional Generative Model
Emotional Generative Model

The system architecture used in this task

Output response

Default emotional response

Score > Threshold

True

Response with score

Choose response with max score

Candidate response set with emotion score

Emotion classifier

Candidate response set

Response generation subsystem

Model-1

Model-2

Model-3

Model-4

Model-5

Response_label

Post
Emotional Generative Model

Each of the five Seq2Seq models is trained using 1/5 of the training data set
Each Seq2Seq model is composed of an encoder and a decoder (Luong et. al., 2015)
An emotion classifier for each class of emotion is trained respectively. Based on the desired response label, the corresponding emotion classifier is used to give a score for each candidate response.

Emotional Generative Model

Response with score

Choose response with max score

Candidate response set with emotion score

Emotion classifier

Candidate response set
The Emotion Classification Subsystem can produce the probability of the input sentence belonging to a certain emotion class. (Kim, 2014)
Emotional Generative Model

Based on the desired response label, the response with the maximum score must pass the **emotion accept threshold** for that class to be qualified for final output. Otherwise, a **default emotional response** for that class is used for final output.

![Diagram showing the process of emotional response generation.](Image)
The emotion accept threshold $T_{\text{emotion}}$ is used in the emotion classification subsystem to determine whether we accept the generated response based on the probability of the response belonging to the desired emotion class.
Experiment
The dataset is composed of 600,000 post/response pairs

[[[post, post_label],
  [response, response_label]],

[[post, post_label],
  [response, response_label]],
...

[[['现在 刷 朋友 圈 最大 的 快乐 就是 看 代购 们 各种 直播 。。。。。。', '5'],
  ['卧 槽 我 也 是’, ‘4’]],
...]}
Test and Evaluation Metric

IF Coherence and Fluency
  IF Emotion Consistency
    LABEL 2
  ELSE
    LABEL 1
ELSE
  LABEL 0

\[
\text{AvgOverallScore} = \frac{1}{N_t} \sum_{i=0}^{2} i \cdot \text{num}_i
\]

\(N_t\) is the total number of test questions
\(i\) is the label score
\(\text{num}_i\) is the number of test questions which has a label of \(i\)

\(\text{AvgOverallScore}\) has a range between 0 and 2, the higher the better.
### Experiment

#### Partial results submitted

<table>
<thead>
<tr>
<th>question</th>
<th>qEmo</th>
<th>Ans1 (like)</th>
<th>Ans2 (sad)</th>
<th>Ans3 (disgust)</th>
<th>Ans4 (anger)</th>
<th>Ans5 (happy)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>睡得不想想起，眼都肿了(打哈欠)</td>
<td>3(D)</td>
<td>我爱你</td>
<td>悲摧啊</td>
<td>是噢，我今天中午睡到3点半醒(哼)</td>
<td>什么情况？</td>
<td>哈哈，真的假的？</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>死了，还有好多没背，背了就忘。。怎么搞的</td>
<td>3(D)</td>
<td>我也觉得很黄！</td>
<td>唉，没法活了，不容易啊</td>
<td>哎···同感同感～～有我们文学系的多东西背吗(泪)</td>
<td>什么情况？</td>
<td>哈哈，真的假的？</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>心情坏得衣服都不想换脸也不要了就这样睡吧(哈欠)</td>
<td>3(D)</td>
<td>哈哈，你是我的错了</td>
<td>你不觉得我很委屈啊</td>
<td>哼]我也想睡 [怒]</td>
<td>什么情况？</td>
<td>不拘小节，我很欣赏</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>真是奇怪，这种时候居然还吃下去。。</td>
<td>3(D)</td>
<td>我也是哎，我也是</td>
<td>我还没吃呢，我还没吃呢</td>
<td>晕，啥时候应该吃不下去？</td>
<td>什么情况？</td>
<td>哈哈，真的假的？</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>讨厌被质问的感觉...</td>
<td>3(D)</td>
<td>哈哈，我也是我很无奈的</td>
<td>全部人都是这样想的</td>
<td>什么情况？</td>
<td>哈哈，真的假的？</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overall results and emotion-specific scores of our system

<table>
<thead>
<tr>
<th>Team Name</th>
<th>‘Like’ Score</th>
<th>‘Sad’ Score</th>
<th>‘Disgust’ Score</th>
<th>‘Anger’ Score</th>
<th>‘Happy’ Score</th>
<th>Overall Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TKUIM</td>
<td>0.82</td>
<td>0.65</td>
<td>0.63</td>
<td>0.63</td>
<td>0.875</td>
<td>0.726</td>
</tr>
</tbody>
</table>
Conclusion
Aiming to achieve Explainable Artificial Intelligence (XAI), our system is split into a response generation subsystem and an emotion classification subsystem. In this way, the results of each subsystem can be clearly observed for independent model design and tuning.

In the response generation subsystem, we produce 5 sentences from each of the 5 generative models to ensure the fluency of the response sentences and their coherence to the post sentence. The emotion classification subsystem ensures that the response of choice will meet the desired emotion class.

In the CECG subtask of the NTCIR-14 STC3 task, the average overall score of our TKUIM team is 0.726.

Our contribution is design and implementation of a dialogue system that can produce fluent response sentences of the desired emotion class.

Our emotion classification subsystem can be easily expanded by other emotion tags, like casualness, respect, or other specific commercial use cases.