WUST at the NTCIR-14 STC-3 CECG Subtask

Zhanzhao Zhou, Maofu Liu, Zhenlian Zhang
School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
liumaofu@wust.edu.cn

Introduction

◆ Emotions in dialogue can significantly enhance the effect of human-computer interaction. An increasingly growing number of researchers pay attention to the generated responses that is not only appropriate in content but also adequate in emotion.

◆ In NTCIR-14, the STC-3 CECG Subtask is defined as follows: Given a Chinese post with a user-specified emotion category of the response to be generated, the goal is to generate a response Y that is coherent with the emotion category.

◆ We regard this task as an information retrieval problem. For the given post, we choose the best response by computing the text similarities between the given post and candidates.

System Architecture

1. Data Preprocessing

◆ Firstly, we remove the English and Cantonese dialogues in the original dataset. And we construct the inverted index table for the training dataset.

◆ Secondly, According to the emotion categories of responses, we divide the training dataset into six subsets and mark the corresponding emotion categories respectively, so the six training subsets are obtained.

◆ Finally, for each testing instance, we search for appropriate responses to construct the candidate dataset in the preceding corresponding training subset whose emotion class is the same as the testing instance emotion class by the inverted index table.

2. Matching

◆ TF-IDF (Term Frequency-Inverse Document Frequency) is an important indicator when extracting keywords. TF-IDF is used to assess the importance of a term to a document.

◆ This paper uses the Jieba, a word segmentation tool, to extract keywords by calculating TF-IDF weights.

Experiments

◆ The official evaluation results are showed as follows:

<table>
<thead>
<tr>
<th>Submissions</th>
<th>Emotions</th>
<th>Label0</th>
<th>Label1</th>
<th>Label2</th>
<th>Total</th>
<th>Overall</th>
<th>Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WUST</td>
<td></td>
<td>601</td>
<td>211</td>
<td>188</td>
<td>1000</td>
<td>587</td>
<td>0.587</td>
</tr>
<tr>
<td>Like</td>
<td></td>
<td>117</td>
<td>36</td>
<td>47</td>
<td>200</td>
<td>130</td>
<td>0.65</td>
</tr>
<tr>
<td>Sadness</td>
<td></td>
<td>124</td>
<td>31</td>
<td>45</td>
<td>200</td>
<td>121</td>
<td>0.605</td>
</tr>
<tr>
<td>Disgust</td>
<td></td>
<td>111</td>
<td>69</td>
<td>20</td>
<td>200</td>
<td>109</td>
<td>0.545</td>
</tr>
<tr>
<td>Anger</td>
<td></td>
<td>137</td>
<td>48</td>
<td>15</td>
<td>200</td>
<td>78</td>
<td>0.39</td>
</tr>
<tr>
<td>Happiness</td>
<td></td>
<td>112</td>
<td>27</td>
<td>61</td>
<td>200</td>
<td>149</td>
<td>0.745</td>
</tr>
</tbody>
</table>

◆ As the evaluation results show, the system can generate appropriate and reliable responses both in content and emotion and rank the fifth of all the participants. Retrieval based models should be competitive in this challenge.

◆ On the one hand, the system don’t capture semantic of the dialogues. On the other hand, if we can compare more ranking or re-ranking strategies, that would be better.

Conclusions

◆ In this paper, we take the information retrieval method to solve this task. Our proposed WUST system can generate appropriate and reliable responses both in content and emotion.

◆ In the future, we would like to propose a novel model using the generated-based approach. And the semantic relevance between post and response should be taken into consideration on proposing novel model. And the special mechanism for addressing emotional factor in dialogues should be proposed.