TKUIM AT NTCIR-14 STC-3

Shih-chieh Wei[1], Chi-bin Cheng[2], Guang-zhong-yi Cao, Yi-jing Chiang,
Chin-yi Wu, Shih-hsiang Lin, Kun-li Tsai
Tamkang University, Taiwan, R.O.C.
{sekewei[1], cbcheng[2]}@mail.tku.edu.tw

ABSTRACT

In this work, we will report how we (TKUIM) built a system for the sub-task CECG of STC-3. Our system mainly consists of two parts, the response generation subsystem and the emotion classification subsystem. For the response generation subsystem, we trained five generative models using different training parameters. 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.

Team Name. TKUIM
Subtasks. Chinese Emotional Conversation Generation (CECG)
Keywords: Natural Language Generation, Deep Learning, Sentiment Analysis

EMOTIONAL GENERATIVE MODEL

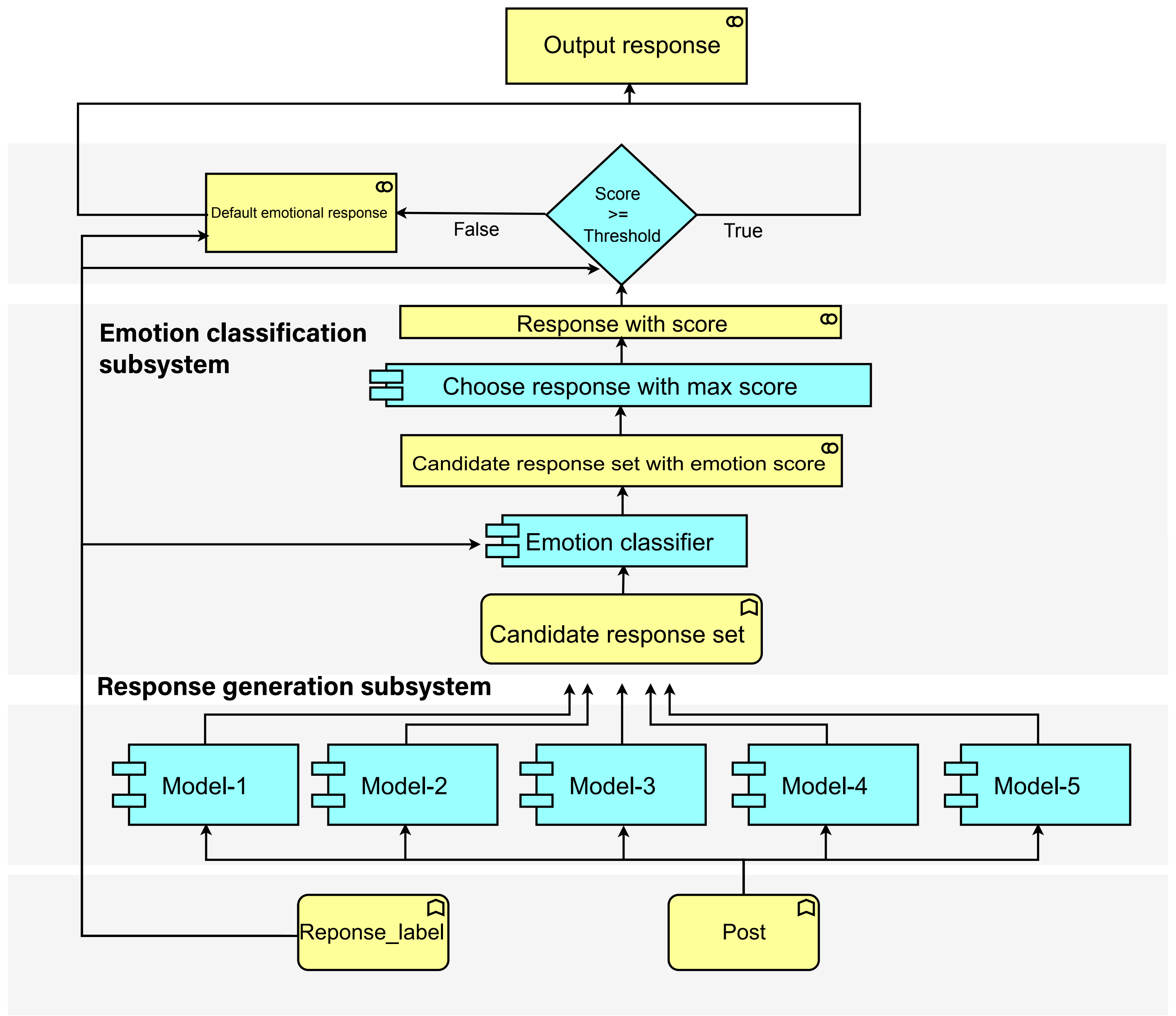
Output response
Default emotional response
False
Score
>=
Threshold
True
Emotion classification subsystem
Response with score
Choose response with max score
Candidate response set with emotion score
Emotion classifier
Candidate response set
Response generation subsystem
Model-1
Model-2
Model-3
Model-4
Model-5
Reponse_label
Post

Figure 1. The system architecture used in this task

The system architecture used in this task is shown in Figure 1.

The lower part of Figure 1 is our response generation subsystem, and the upper part is the emotion classification subsystem. For each post we will generate the response through the five Seq2Seq models, which constitute the candidate response set, all of which are responses to the same input. The candidate response set will then be input to the emotion classification subsystem, producing the expected probabilities for a desired emotion class (excluding emotion class 'Others'). Then we choose the response with the maximum probability. For this probability, we set an emotion accept threshold for each emotion class. The threshold determines whether the response can be the final output. If the response with the highest probability cannot pass the emotion accept threshold for the desired emotion class, it will be replaced with a random sentence from the library of default emotional responses for the desired emotion class.

EXPERIMENT

We divided the 600,000 official post-response pairs into five segments of training sets, and assigned them to five experimental groups for training five generative models using different parameters. The generated response sentences of the five models are collected as the candidate response set, which is input to the next subsystem.

PRE-PROCESSES

For the response generation subsystem, our focus is on the fluency of the response and its coherence to the post. Therefore, we trained our model without use of the emotion label, but only extracted the post-response pairs as the training data for the response generation subsystem. As for emotion classification subsystem, we mainly focus on the correlation between the conversation text and its emotion label. So for this part of the training, we do not distinguish between posts and responses. We take all of the posts, responses, and their emotion labels to train five classifiers for each emotion class, respectively. For a specific emotion classifier, the conversation text with a label of that emotion class is marked as positive. The conversation text with labels of other emotion classes is marked as negative. The emotion classifier outputs the probability of the input text belonging to that emotion class. In addition, we give the same number of positive and negative sentences to guarantee fairness in the training process.

The format of the data set is as follows:

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

A [[post, post_label], [response, response_label]] is called a post-response pair. The data set has a total of 600,000 post-response pairs. The 'post_label' denotes the emotion class of the post. The 'response_label' denotes the emotion class of the response. Both label values vary from 0 to 5, corresponding to the classes of 'Other', 'Like', 'Sadness', 'Disgust', 'Anger', and 'Happiness', respectively.

Take the first post-response pair of the data set as an example.

```
[['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']]
```

In this post-response pair, the post is ‘现在 刷 朋友 圈 最大 的 快乐 就是 看 代购们 各种 直播 。 。 。 。 。’, which has an emotion class of 5, or 'Happiness'. The response is ' 卧 槽 我 也 是 ', which has an emotion class of 4, or 'Anger'.

EMOTION ACCEPT THRESHOLD & DEFAULT EMOTIONAL RESPONSES

The emotion accept threshold T 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. The exact value of T for each emotion class depends on a certain quartile value of the output probability of an emotion classifier for the training set. For example, there are a total of 155,758 sentences whose emotion class is 1 or 'Like'. The emotion classifier for 'Like' will produce 155,758 probability values between 0 and 1. After these values are sorted in descending order, a certain quartile value is decided as the emotion accept threshold for the 'Like' emotion. In this way, the emotion accept thresholds for the five emotion classes can be different for tuning purposes. In addition, for each emotion class, a library of default emotional responses is manually selected from the official training data set. When each sentence in the candidate response set is assigned a probability by the desired emotion classifier, the sentence with the highest probability is selected as the response. However, if the response with the highest probability cannot pass the emotion accept threshold for the desired emotion class, it will be replaced with a sentence from the library of default emotional responses for the desired emotion class.

TEST AND EVALUATION METRIC

The submitted post-response pairs are evaluated by the following metrics.
Emotion Consistency: whether the emotion class of a generated response is the same as the pre-specified class.
Coherence: whether the response is appropriate in terms of both logical coherence and topic relevance in content.
Fluency: whether the response is fluent in grammar and acceptable as a natural language response.
The labeling procedure is shown by the following pseudocode:

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

All the submitted results were resorted to Baidu Data Crowdsourcing Service for manual evaluation, and each post-response pair is annotated by three curators. The average overall score for each run is computed by the following formula.

$$AvgOverallScore = \frac{1}{N_t}\sum_{i=0}^{2} i * num_i$$

where num_i is the number of pairs which has a label of i for each submission run, and N_t is the total number of pairs for each run. Therefore, the highest $AvgOverallScore$ for each run will be 2.

Since the data set has a relatively small proportion of sentences belonging to emotion class 4 or 'Anger', the response generative model is accepted at a lower rate for this class, resulting in more frequent use of the default sentence for the emotion class 'Anger'. Table 1 shows part of the result we submitted in this competition. The column names from left to right denote the post, the emotion class of the post, and five responses to the post. 'ans1' to 'ans5' are the responses corresponding to emotion classes 'Like', 'Sadness', 'Disgust', 'Anger', and 'Happiness' in order. The red rectangle in 'ans4' shows the responses of emotion class 'Anger'. The blue rectangle in 'ans5' shows the responses of emotion class 'Happiness'. It can be found that we use a lot of default sentences in these two emotion classes.

CONCLUSION

Table 1. Partial result submitted

	question	Ques Emo	ans1	ans2	ans3	ans4	ans5
70	睡 得 不 想起，眼 都 不 想 睁 肿 么 办 [打 哈 欠]	3	我 爱 你	悲 摧 啊	是 噢，我 今天 中午 睡 到 3 点 半醒 [哼]	什么 情况 ？ [怒]	哈哈 ， 真的 假 的 ？
71	死 了，还有 好多 没 背，背 了 就 忘 。 。 。 怎 么 搞 的	3	我 也 觉得 很 好 看 ！	唉 ， 没法 活 了 ， 不 容易 啊	哎 · · · 同感 同感 ~ ~ 有 我们 文学 系 的 多 东西 背 吗 [泪]	什么 情况 ？ [怒]	哈哈 ， 真的 假 的 ？
72	心情 坏 得 衣服 都 不 想 换 脸 也 不要 了 就 这样 睡 吧 [哈欠]	3	哈哈 ， 你 是 我 的 错 了	你 不 觉得 我 很 委屈 啊	[哼] 我 也 想 睡	什么 情况 ？ [怒]	不拘小节，我 很 欣赏
73	真是 奇怪，这种 时候 我 居然 还 吃 的 下 去 。 。 。	3	我 也 是 哎 ， 我 也 是	我 还 没 吃 呢 ， 我 还 没 吃 呢	晕，啥 时候 应 该 吃 不 下去 ？	什么 情况 ？ [怒]	哈哈 ， 真的 假 的 ？

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, the average total score of our TKUIM team is 0.726. For emotion class 'Like', the average score is 0.82. For emotion class 'Sadness', the average score is 0.65. For emotion class 'Disgust', the average score is 0.63. For emotion class 'Anger', the average score is 0.63. Finally, for emotion class 'Happiness', the average score is 0.875. In this work 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.

ACKNOWLEDGMENTS

This research was supported in part by the TKU research grant. We would like to thank the support from College of Business and Management, Tamkang University, Taiwan, R.O.C.