AI-NTPU Emotional Chatting Machine at the NTCIR-14 STC-3 CECG Task

Chih-Chien Wang. AI-NTPU [1], Min-Yuh Day. AI-NTPU [2], Wei-Jin Gao. AI-NTPU [3], Yen-Cheng Chiu. AI-NTPU [3], and Chun-Lian Wu. AI-NTPU [3]

[1]National Taipei University, ROC wangson@mail.ntpu.edu.tw

[2]Tamkang University, ROC myday@mail.tku.edu.tw

[3]National Taipei University, ROC {s710636303, s710736104, s710736401}@webmail.ntpu.edu.tw

Abstract. This paper gives an overview of our work for the NII Testbeds and Community for Information access Research (NTCIR)-14 Short Text Conversation (STC)-3 Chinese Emotional Conversation Generation (CECG) subtask. In NTCIR-14 STC3, emotion in the post-comment pairs was considered in both retrieval-based and generation-based approaches. In this subtask, we developed an emotion classification and two approaches, generation-based and retrieval-based approaches, to create responses to post. In STC3 CECG subtask repository, each post and comment were label with one emotions tag: Like, sadness, disgust, anger, and happiness. The posts and comment with emotion that not mentioned above were label as "other". We develop an emotion classification model to label the comment we created. The purpose of this subtask is to create comment that is coherence and fluency to the post. Besides, the created post should be emotion consistency. In the retrieval-based approach, we used Apache Solr to search an appropriate comment to the post. In generation-based approach, we use attention-based sequence to sequence (Seq2Seq) model to create new comment to each post. For emotion classification model, we used Multilayer Perceptron (MLP). In the paper, we provide our procedure in detail for creating new comment to post. We provide two submissions: retrieval-based approach with emotion and generation-based with emotion. However, due to the format issue, the evaluation results of submission for generation-based with emotion is not provided. For the purpose of self-improvement, we provide our self-evaluation results to both two submissions. Further improvement suggestions are also provided in the paper.

Team Name. AI_NTPU

Subtasks. Chinese Emotional Conversation Generation

Keywords: Short Text Conversation, Information Retrieval, Solr, Seq2Seq, Emotion Analysis, Long Short Term Memory (LSTM)

1 Introduction

With the increasing use of mobile devices, social media and instant messengers, it had been an important way that people communicate via short text. For the purpose of developing chatbots or voice assistants, it is important issue to develop an automatic mechanism to response users' dialogues (posts, queries, or questions). It would be fantastic if chatbot or voice assistants can provide appropriate, meaningful, and relevance responses (comments) to the dialogues during short text conversation.

NTCIR-12 included Short Text Conversation (STC-1) task which aims to provide responses (comments) to the post. STC1 reused a large repository of post-comment pairs by retrieval-based approach to respond to new posts. NTCIR-12 STC-1 proposed an approach that using large amount of short conversation from social media such as Twitter (Japanese) and Weibo (Chinese) to retrieve comments to new post. The generated retrieval-based responses (comments) was evaluate according to the relevance of the responses (http://ntcir12.noahlab.com.hk/stc.htm).

NTCIR-13 STC2 used both retrieval-based and generation-based approaches to respond to new posts. From NTCIR-13 STC-2 subtask, in addition to the retrieval-based approach, the generation-based approach was also considered. The criteria for assessing relevance of comments (responses) to post also considered both the fluency and grammatical correctness [1].

There has been a rising tendency in artificial intelligence (AI) research to create a robot which is capable of acting and talking to human at the human label. For the purpose of constructing a conversation system acting like a human being, it is essential for AI systems to have the ability to perceive, integrate, understand, and regulate emotions. Perceiving and understanding emotions are important to human behavior [2]. Thus, In NTCIR-14 STC3 Chinese Emotional Conversation Generation (CECG) subtask, emotion in the post-comment pairs was considered in both retrieval-based and generation-based approaches.

We participant the CECG subtask of NTCIR-14 STC-3, the goal of CECG subtask is to generate appropriate comments (responses) for given new post. CECG consider the generated comments (responses) should be appropriate not only in content but also in emotion. Therefore, emotion consistency is the challenge of the STC3 CECG subtask [3]. Here is the definition of subtask: To generate a comment for every specified emotion from given Chinese post. There are 5 specified emotions in emotion categories, which includes:

Like, sadness, disgust, anger, and happiness. Model constructing is based on retrieval-based method and generation-based method.

In this subtask, we developed an emotion classification model and two approaches of generation-based and retrieval-based to create responses to post. In STC3 CECG subtask repository, each post and comment were label with one emotions tag: Like, sadness, disgust, anger, and happiness.

This paper is organized as follows, we will discuss our approach of the retrieval-based method with emotion model and generation-based method with emotion model in Section 2 and Section 3. And then move on to Section 4 is about emotion classification model. In Section 5. We compare these approaches, evaluation results, and describe challenges we met during the period of this subtask, then we give the conclusions.

2 RETRIEVAL-BASED APPROACH

In the retrieval-based method, we used Apache Solr (http://lucene.apache.org/solr/) to retrieve the indexed post-comment pairs. We pre-process the post-comment pairs and indexed them in Apache Solr. We used the term of provided new post to search the indexed posts in Apache Solr and used the fetched search results as potential candidate comments. We accumulated the inverse term frequency for each candidate comments and computed the cosine similarity between the new post and candidate comments. We developed a relevance score which multiplied the accumulated inverse term frequency by the cosine similarity. We provide the candidate comments which match the assigned emotion category and with highest relevance score as generated comment. The system architecture and the detailed procedure for retrieval-based approach are as followings.

2.1 System Architecture

We used Solr to index the corpus. Before indexing it, we perform word segmentation, text analysis, and remove stop words. Then, we complete the Solr index building. When a new post provided, we search the Solr index, and obtain the fetched potential candidate comments. We calculated the accumulated inverse term frequency. We also computed the cosine similarity between the new post and the candidate comments. We multiplied accumulated inverse term frequency by cosine similarity as the relevance score. The candidate comment that match the assigned emotion and is with highest relevance score is treated as the generated comment. The system architecture for retrieval-base method is as Figure 1.

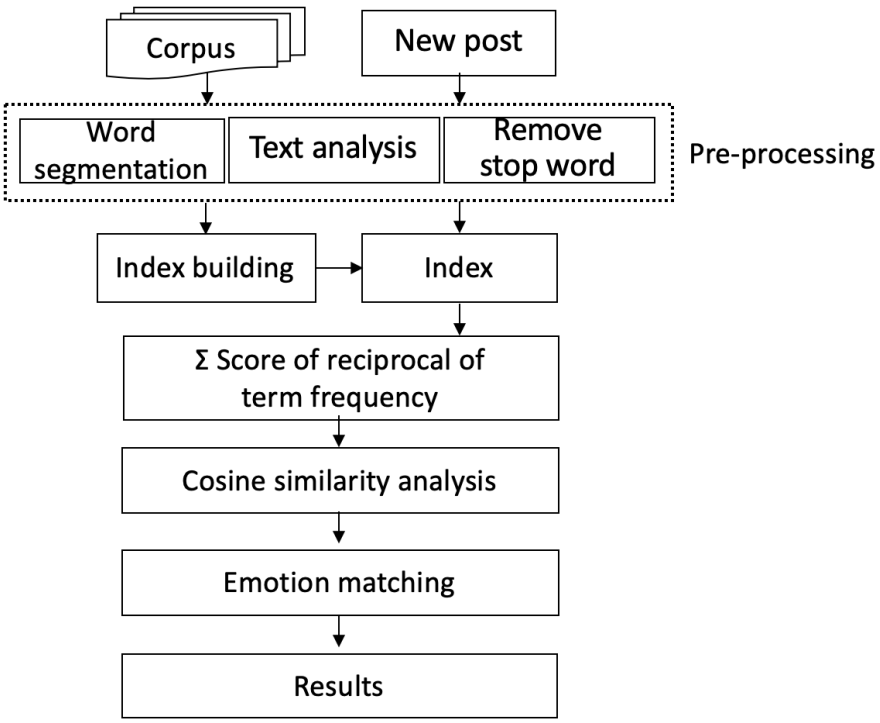

Figure 1. Retrieval-based System Architecture

2.2 Pre-processing

Firstly, we conducted a pre-processing procedure to analyze the given repository, including word segmentation, term frequency calculation, and remove some post-comment pairs. We calculate term frequency for each term. The term frequency represents the important of the term, the high frequency means that the term appear frequently. If a term is with high frequency, this term is not so important since that it appear in many post-comment pairs. The reciprocal of term frequency appear in the provided corpus would be used as an indicator for the importance of the term in the ranking stage I.

Then, we remove some types of comments that we believe they are not quite possible to be perfect candidates for any new post. Here are three types of comments that we remove from repository:

● Comments with high frequency.
We removed comments with high frequency because the comments are repeatedly used in many post-comment pairs. It might be a perfunctory reply or a greeting sentence. We hope our systems can provide meaningful comments rather than just perfunctory or greeting comments.
● Comments with special characters
We did not consider the special characters, although it is still meaningful in many cases. If we hope to develop the system into a voice assistant, these kinds of special characters cannot be used in voice.
● Comments with emotion label of other emotion: 0
As definition of CECG subtask, only label 1 to 5 in specified emotion categories were included. The post-comment pairs of "other emotion" were not included.

2.3 Indexing

Our system was built with Solr. Solr is open source search platform on Apache Lucene. In this stage, we import the given repository to system with defined schema. We use RESTful API (https://www.restapitutorial.com/) to define schema, features index, text analysis, and stop words filtering via configuration files. We used all terms (words) from the provided new post one by one to search the Solr. If the term appears in the post of post-comment pair, we fetched the "comment" (rather than post) as potential candidates for generated comments.

2.4 Ranking Stage I

After obtaining search results for each terms (words) appear in the new post, we assigned the reciprocal of term frequency as the "inverse frequency score" to each search results. We accumulated the "inverse frequency score" if the fetched comment appeared in search results of more than one term. Each comment would contain accumulated inverse frequency score. We ranked the comments based on the accumulated inverse frequency score. To improve computing efficiency, we keep only top 500 comments as potential candidates.

2.5 Ranking Stage II

Usually, the terms used in post may also appear in the comment. Once post and comment share the same term, the semantic meaning of post may similar with that of comment. Cosine similarity is a widely applied metric in information retrieval, which calculate the cosine of the angle between two texts' term vectors.
We calculated the Cosine similarity between the new post and the 500 candidate comments. The relevance score for each candidate comments was calculated using the formula of the Cosine similarity multiplied by the accumulated weight value of the reciprocal of term frequency.

2.6 Sorting Stage

We sort the candidate comments by the relevance scores, the higher the score, the higher the candidate is relevant to the new post. However, in STC3, we had to consider the match of emotion category for the generated comments. The generated comment which match the assigned emotion and with the highest relevance score was treated as the generated comment.

3. GENERATION-BASED APPROACH

We employed an attention-based sequence to sequence (Seq2Seq) network model [4.5] for the generation-based approach. We describe the generation-based approach in detail, as followings.

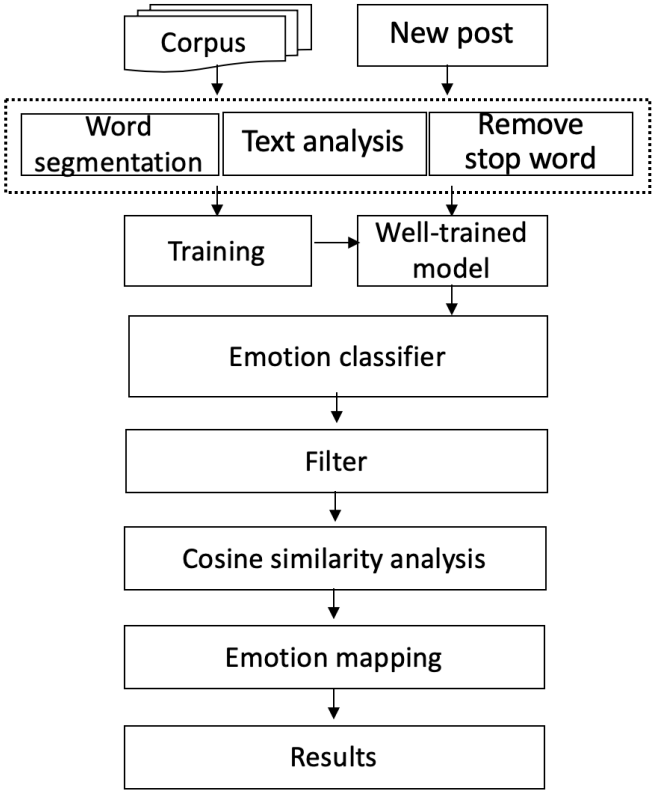

Figure 2. Generation-based System Architecture

3.1 System Architecture

Before training the model, we perform word segmentation, text analysis, and remove stop words. Then, we used an attention-based sequence to sequence (Seq2Seq) network model which take Long Short Term Memory (LSTM) as encoder and decoder to train the model using the provided corpus. When a new post provided, we used the Seq2Seq to generate candidate comment. We computed the cosine similarity between the new post and the generated candidate comments. We used the emotion classification model to tag the emotion of the generated comments. The candidate comment that match the assigned emotion and is with highest cosine similarity is treated as the generated comment. The system architecture for generation-base method is as Figure 2.

3.2 Pre-processing

The generation-based approach performed the similar pre-processing procedure as the retrieval-based approach. We performed word segmentation, term frequency calculation, and remove some post-comment pairs. We also removed some comments that we believe they are not quite possible to be perfect candidates for any new post, as we did in the retrieval-based approach.

3.3 Model Training

We feed the given repository which finished pre-processing to generation system and train the attention-based Seq2Seq model with following settings of hyperparameter:

- Batch size: 128
- Epoch: 220
- Dropout rate: 0.4

3.4 Comments Generating and Emotion Tagging

We put the given test datasets (new post) into well-trained generation-based method model and generate 1000 comments for each given new post. All of these comments were labeled with emotion tags via our emotion classification model, which was based on Multilayer perceptron (MLP) neural network.

3.5 Filtering and Truncating

We filtered and truncated out some comments that we believed they were mainly meaningless and could not be perfect candidate comments to new post. Followings were two filters and truncators for this purpose.

- Deduplicate consecutively repeated sentence segments.

For instance, "天天训训训训训训训训训" (Sleep sleep sleep sleep sleep sleep sleep sleep every day) will be truncated not to repeat more than three times as "天天训训训" (Sleep sleep sleep every day). (Note: Chinese character 训 in Cantonese is the same meaning of Chinese character 睡 in mandarin.

- Consecutively repeated special characters.

For instance, "。。。。。。。。。。。。。。。。。。。。" will be truncated not to repeat more than three times as "。。。".

3.6 Ranking and Sorting

We generated 1000 candidate comments and calculated the cosine similarity between the new post and the candidate comments. We sort the candidate comments by the cosine similarity, the higher the cosine similarity, the higher the candidate comment is relevant to the new post. However, in STC3, we had to consider the match of emotion category for the generated comments. The generated comments which match the assigned emotion and with the highest cosine similarity is treated as the comment to the new post.

4 Emotion Classification model

Literature had figured out that addressing emotion in conversational systems can enhance user satisfaction [6]. Recently, neural network models trained on large-scale data have obtained some achievements on research of conversation generation [7-9]. Same as STC-2, most of these language models aim to improve the content quality of conversation generation [10]. To address emotion in existing neural models for conversation generation, we build an emotion classification model to make generated comments of given new post appropriate in emotion by using matching via specified emotion categories.

Our system is based on Multilayer perceptron (MLP) neural network [11]. We employed it as our model of emotion classification model [12] to deal with multiclass problem such as specified emotion categories. Next, we describe the emotion classification model in detail. We use given repository to train our system with steps as Figure 3. The well-trained model could help to classify and label the generated comments specified emotion, as Figure 4.

4.1 System Architecture

We use given repository to train our system with steps as Figure 3. The well-trained model could help to classify and label the generated comments specified emotion, as Figure 4.

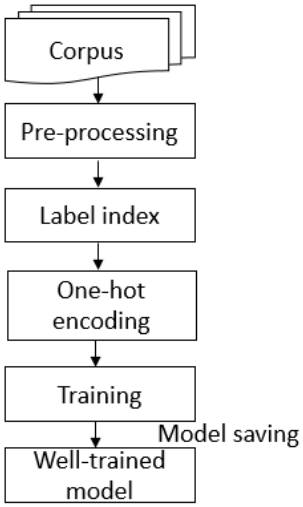

Figure 3. Flowchart for Training Emotion Classification model

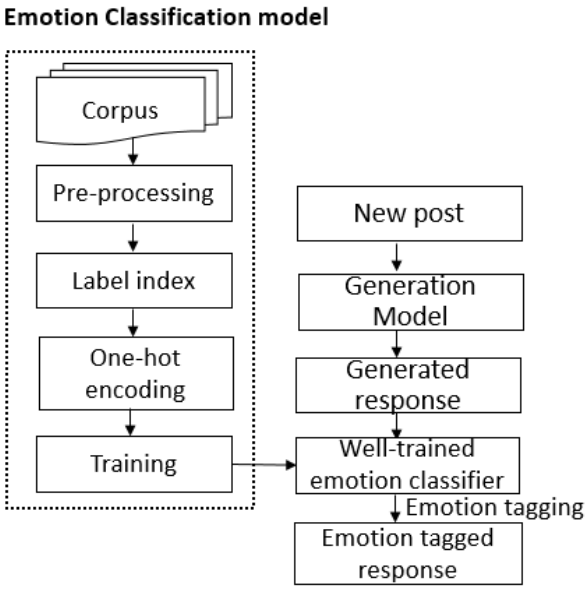

Figure 4. Emotion tagging flowchart

4.2 Model Training

We trained emotion classification model by the following steps: preprocessing, label indexing, one-hot encoding, and training. Regarding to the performance of this model, loss value is 0.675, and accuracy is 0.841, confusion matrix visualization with true and predicted labels is as Figure 5, and hyperparameter settings is as following:

- Batch size: 32
- Epoch: 50
- Dropout rate: 0.4

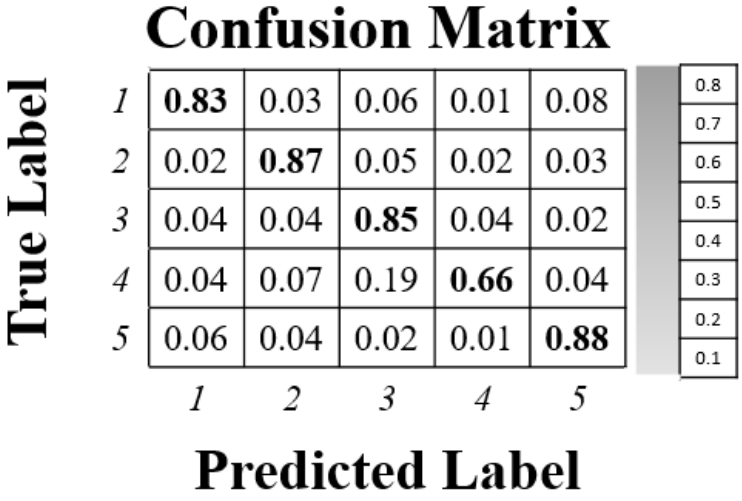

Figure 5. Confusion matrix with true and predicted labels

4.3 Emotion Categories

We obtained a well-trained emotion classification model with emotion categories. Five emotion categories were included with number (like:1, sad:2, disgust:3, angry:4, happiness:5). Table 1 provides examples of emotion classify results for post. Once the system generated comments for given new post, each comment will be classified and labeled with one specified emotion category.

Table 1. Example for Emotion Categories

Post	Emotion
运气不错,没有晚点太多,回到北京了 Fortunately, not too late, come back to Beijing.	1
晚上吃太多了 现在好难过躺也不是趴也不是[悲伤] Eating too much tonight, neither laying down nor prone is comfortable now.[Sad]	2
不能吃太多的东西了,发现自己很难控制食量 Cannot eat too much, it's so hard to control diets by myself.	3
这个月花太多钱了[抓狂] Spending too much money this month.[Mad]	4
每次去新世界都会看到很多俊男靓女啊[太开心] There're many handsome boys and beautiful girls while go to New World every time.[Thrilled]	5

5 EXPERIMENTAL RESULTS

The STC3 CECG provide 200 new posts for test purpose. There are five emotion categories. Each new post should have five generated comments, one for each five emotion categories. Thus, we generate 1000 comments for the 200 new posts.

In STC3 CECG, the submitted post-response pairs are evaluated by coherence and emotion consistency. If the response (comment) is appropriate in terms of both logically coherent and topic relevant content, the comment would be annotated as label 1 (coherence and fluency), the response (comment) would get 1 point. If the response (comment) is inappropriate in terms of either logically coherent or topic relevant, the comment would be annotated as label 0, and the response (comment) would not get any

point. For the coherence and fluency response (comment) is with emotion of the pre-specified emotion, the comment would be annotated as label 2 (emotion consistency), the response (comment) would get 2 point. Each comment is annotated as 0, 1, or 2 point, official labeling procedure is as Figure 6[3]. There are totally 200 new posts and 1000 generated comments; five comments for each 200 new post. Thus, the total score ranges from 0 to 2000.The official formula of overall score and average score computing is listed as Figure 7[3].

The STC3 CECG only allowed each team to submit one run for the retrieval-based approach and one run for generation-based approach. In addition, the STC3 CECG subtask organizer only provided overall evaluation results. No detailed report for the evaluation results was available. For the purpose of self-improvement, we performed similar denotation procedure for ourselves, which included denotation by our team members.

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

Figure 6. Official labeling procedure

$$\text{OverallScore} = \sum_{i=0}^{2} i * num_i$$

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

Figure 7. Official score computing formula

5.1 Results of retrieval-based approach

5.1.1 Evaluation results by organizers

We submitted one run for the retrieval-based approach. According to the evaluation results provided by STC3 CECG organizer, among the 1000 generated comments, 716 comments were annotated as label 0; 200 comments were annotated as label 1; and 84 comments were annotated as label 2. The overall score was 368. The average score was 0.368.

5.1.2 Evaluation results by ourselves

The generated comments were evaluated by three annotators. We use Kendall's coefficient of concordance (Kendall's W) [13] to assessing agreement among annotators. The results of evaluation, self-evaluation, and Kendall's W test of self-evaluation are listed in Table 2. The Kendall's W test of self-evaluation is 0.731, which revealed high consistency among annotation results of the three annotators.

According to the self-evaluation results provided by ourselves, among the 1000 comments, 560 comments were annotated as label 0, 208 comments were annotated as label 1, and 195 comments were annotated as label 2. The overall score was 598. The average score was 0.598.

The difference in scores between the evaluation results provided by organizer and by ourselves just reflected the fact the different standard label for annotation. The annotators of organizer and annotators of our team were different. Logically coherent, topic relevant, and emotion consistency are all subjective criteria for annotation. Our self-evaluation annotators may hold an easy standard for these three criteria. It is not an easy task to hold same standard label if the criteria is subjective.

Table 2. Evaluation Results

Approach	Evaluator	Label 0	Label 1	Label 2	Total	Score	Average Score	Kendall's W test
Retrieval-based	STC3 CECG organizers	716	200	84	1000	368	0.368	n/a
Retrieval-based	Team AI_NTPU	597	208	195	1000	598	0.598	0.731
Generation-based	STC3 CECG organizers	n/a	n/a	n/a	1000	n/a	n/a	n/a
Generation-based	Team AI_NTPU	873	85	42	1000	169	0.169	0.896

Note: STC3 CECG organizer did not provide feedback and evaluation due to format issue of our generated comments. Our system did not provide enough amount of the 1000 generated comments.
Self-evaluation results for retrieval-based approach with emotion are available at:
https://drive.google.com/open?id=1d4_hXN6M-GVaQjXpPAwS4ZHQ2PUHPEle4bd7QFr188s
Self-evaluation results for generation-based approach with emotion are available at:
https://drive.google.com/open?id=1pg1wbVXjB149AlNzB2pNZJ1LfBcTWbtJQvyGuzPzfKE

5.2 Results of generation-based approach

5.2.1 Evaluation results by organizers

We submitted one run for the generation-based method with emotion. However, STC3 CECG organizer did not provide feedback and evaluation due to format issue of our generated comments.

5.2.2 Evaluation results by ourselves

The generated comments were evaluated by three annotators. We use Kendall's coefficient of concordance (Kendall's W) to assessing agreement among annotators. The results of evaluation, self-evaluation, and Kendall's W test of self-evaluation are listed in Table 2. The Kendall's W test of self-evaluation is 0.896, which revealed high consistency among annotation results of the three annotators.
According to the self-evaluation results provided by ourselves, among the 1000 comments, 873 comments were annotated as label 0, 85 comments were annotated as label 1, and 42 comments were annotated as label 2. The overall score was 169. The average score was 0.169.

6 Conclusions

In this paper, we describe the procedure for complete the STC-3 CECG subtask of NTCIR 14. We adopted retrieval-based method and generation-based method to automatically generate responses in short text conversation. In the retrieval-based method, we adopted Solr to retrieve corpus. For generation-based method, we employ the attention-based Seq2Seq model to generate comments for new posts. We used MLP to develop emotion classification model to make sure all generated comments match the assigned emotion category.

We found that some of our generated comments satisfied the standard of content coherence but not satisfied the standard of emotion consistency. Thus, there is still some rooms for improvement of emotion consistency on both retrieval-based and generated-based results. In the future, we can try other neural network model in emotion classification model such as Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) to improve the accuracy and precision of emotion classification.
For the improvement of coherence and fluency of generated comments, we can try to employ new language model such as Bidirectional Encoder Representation from Transformers (BERT) or combine generation and retrieval method in the future study.

Besides, in CECG subtask, we found several posts with same contents in test dataset, as Table 3. Our system did not consider the existence of same posts. Instead, when pre-process the test dataset, our system mistakenly skipped the repeated new post. This is the reason why our generation-based approach only provided less than 1000 generated comments. STC3 CECG organizer did not provide feedback and evaluation results due to the reason that we did not provide enough amount of the 1000 generated comments.5.1 Results of retrieval-based approach.

Table 3. Repeat posts in test dataset

Number	Post	Emotion
81	这 两 个 星期 , 心情 很 压抑 … [悲伤] [悲伤]	2
82	这 两 个 星期 , 心情 很 压抑 … [悲伤] [悲伤]	2
131	好 , 去 西门 狠狠 的 吃 一 顿 , 嘿嘿嘿	1
132	好 , 去 西门 狠狠 的 吃 一 顿 , 嘿嘿嘿	1
176	今天 值班 下班 之后 要 去 做 红娘 嘻嘻 ~ ~ ~ ~ 希望 可以 成功 哈	5
177	今天 值班 下班 之后 要 去 做 红娘 嘻嘻 ~ ~ ~ ~ 希望 可以 成功 哈	5

7 ACKNOWLEDGEMENTS

Our thanks to NTCIR-14 task organizers for their efforts, especially thanks to AI lab of Prof. Minlie Huang for their hard work on organizing NTCIR-14 STC-3 CECG subtask. Authors would like to thank by Taiwan Ministry of Science & Technology research grant (MOST 107-2410-H-305 -036 -MY2) for partial financial sponsorship to this paper.

REFERENCES

1. Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., Miyao, Y., Arase, Y., &Nomoto,M.: Overview of the NTCIR-13 Short Text Conversation Task. In: Proceedings of NTCIR-14. pp. 1 (2017)
2. Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B.: Emotional chatting machine: Emotional conversation generation with internal and external memory. In Thirty-Second AAAI Conference on Artificial Intelligence. pp. 730-738 (2018, April)
3. Zhang, Y., & Huang, M. (2019). Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge. Proceedings of NTCIR-14, p to appear.
4. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).
5. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

6. Prendinger, H., Mori, J., & Ishizuka, M.: Recognizing, modeling, and responding to users' affective states. In International Conference on User Modeling. In: Springer, Berlin, Heidelberg. pp. 60-69 (2005, July)
7. Vinyals, O., & Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869. (2015)
8. Shang, L., Lu, Z., & Li, H. Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364. (2015)
9. Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. C., & Bengio, Y.: A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In: AAAI pp. 3295-3301 (2017, February)
10. Chen, Q., Zhu, X., Ling, Z., Wei, S., & Jiang, H.: Enhancing and combining sequential and tree LSTM for natural language inference. arXiv preprint arXiv:1609.06038. (2016)
11. Pal, S. K., & Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. In: IEEE Transactions on neural networks, 3(5), pp. 683-697 (1992).
12. Kopinski, T., Magand, S., Handmann, U., & Gepperth, A.: A pragmatic approach to multi-class classification. In: 2015 International Joint Conference on Neural Networks (IJCNN) pp. 1-8. IEEE. (2015, July)
13. Legendre, P (2005) Species Associations: The Kendall Coefficient of Concordance Revisited. Journal of Agricultural, Biological and Environmental Statistics, 10(2), 226–245.