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, Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System,

Wuhan 430065, China
liumaofu@wust.edu.cn

Abstract. As an important influencing factor of human-computer interaction experience, the research on the generation of emotional dialogue has aroused the widespread concern of researchers. We participated in the STC-3 (Short Text Conversation) CECG (Chinese Emotional Conversation Generation) subtask of NTCIR-14 and proposed a retrieval-based emotional dialogue system. The WUST system includes three modules, i.e. candidate generation, candidate matching, and candidate ranking. The system traverses the candidates and computes the text similarity one by one between the given post and candidate dialogue, and finally sorts the candidates by their scores calculated by a linear function. The highlight of the system is the ability to generate reliable responses in contrast to the generated-based system. 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.

Team Name: WUST.

Subtask: STC-3 Chinese Emotional Conversation Generation.

Keywords: Chinese Emotional Conversation Generation, Linear Function, Reliable.

1 Introduction

In human-computer interaction, a large quantity of researchers only focuses on the logical coherence and topic relevance of dialogue but pay little attention to the emotions contained in the dialogue. In fact, the understanding of the emotional expression is not only the significant cognitive behavior of human being but also the key to enhancing human-computer interaction [1]. Though plenty of models have been proposed for conversation generation from large-scale social data, it is still quite challenging to generate emotional responses. In order to study emotions in dialogue, we participated in the STC-3 CECG subtask of NTCIR-14. This task provides a large number of single-round conversations marked with emotion class labels. In this challenge, participants

are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. Nevertheless, in the way to address the emotional factor in dialogue generation, there are many difficulties to be solved. On one hand, the large-scale and high-quality emotional dialogue corpora are very rare or even difficult to obtain. And the emotional training dataset supplied by this task is achieved through an emotional classification model with an accuracy rate of 60%, which encourages participants to train their own classification models. On the other hand, the excess consideration of emotions in dialogue means losing a part of the logical coherence.

Some researchers have tried to address the emotional factor in dialogue and propose some models [2, 3], but the models are on the basis of the small-scale dataset or rule-based method. With the large-scale application of deep learning and the coming of the big data era, the generated-based dialogue systems can achieve good performance in terms of dialogue quality. The emotional chatting machine (ECM) model, proposed by Huang et al. [4], can generate responses appropriate not only in content but also in emotion, but ECM needs to specify an emotion class instead of deciding the most appropriate emotion category for the response. The Affect-LM model proposed by Ghosh et al. [5] can generate expressive text at varying degrees of emotional strength without affecting grammatical correctness. Asghar et al. [6] proposed a model which can produce more natural and emotionally rich responses without specifying the emotion class. In general, all the preceding models used a generation-based approach that generated short, universal, and even grammatical-free responses, which greatly damage the quality of the response.

This paper describes our WUST system built in the CECG task. We regard this task as an information retrieval problem [7]. WUST system can retrieve proper responses both in content and emotion. The original training dataset is too large, and it takes too much time to calculate the text similarity. Therefore, it is necessary to construct a smaller candidate dataset. Section 2 describes the system structure. Section 3 discusses the experimental results. And we summarize this paper in Section 4.

2 System Description

The retrieval-based dialogue system is shown in Figure 1. In the candidate generation module, all the training data need to be preprocessed to construct a candidate dataset. In the candidate matching module, the text similarity will be presented by the cosine similarity. In the candidate ranking module, a linear function is used for grading different text similarity methods and selecting the proper method. Finally, the candidate dialogues will be ranked by the final method.

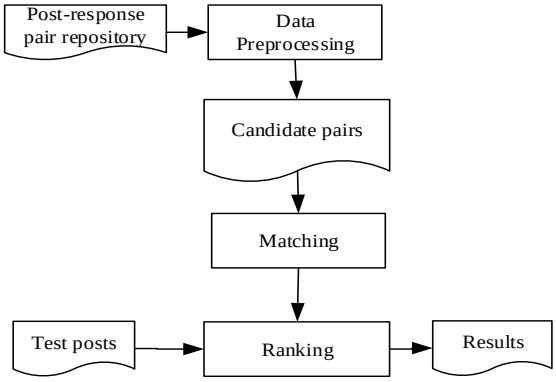

Fig. 1. System architecture.

2.1 Data Preprocessing

Training data is provided by STC-3 CECG subtask of NTCIR-14. This training dataset consists of 1.7 million post-response pairs crawled from Weibo. However, there are many messy conversations in the dataset, and we remove the English and Cantonese ones. Firstly, 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. Thirdly, 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.

The inverted index is a method of reverse mapping words to the documents, which plays an important role in information retrieval. In the inverted index table, each term records the frequency of occurrence of the word and a list of document indexes. This paper builds an inverted index table based on the training dataset, which can help quickly construct the candidate datasets, and thus save the running time.

2.2 Matching

In this module, calculating the text similarity is significant. TF-IDF (Term Frequency-Inverse Documentation Frequency) is an important indicator when extracting keywords. TF-IDF is used to assess the importance of a term to a document. The importance of the term is proportional to the number of the term occurrences in the document and inversely proportional to the term occurrences in the corpus. TF (Term Frequency) is the frequency of occurrence of a term in a document. The less the number of the term occurrences in the corpus is, the larger the IDF (Inverse Documentation Frequency) is, indicating that the term has greater discriminating power.

It is significant to convert a sentence into a sentence vector. This paper uses the Jieba[1] , a word segmentation tool, to implement this conversion. The tool can extract

[1] https://github.com/fxsjy/jiebademo

keywords by calculating TF-IDF weights. At the same time, the TF-IDF weights of the keywords can be used in the conversion of sentence vectors. Firstly, the extracted keywords of the two sentences are combined into a union set. The size of the union set is the size of the sentence vector. Secondly, if a sentence contains a keyword of the union set, its weight should be added into the sentence vector. Otherwise, zero should be added into it. In this way, the lengths of two sentence vectors are the same.

In this paper, the vector space model (VSM) [8] is used to represent the dialogue. The VSM expresses the query and each candidate dialogue into a vector in the same vector space and the text similarity is computed by the cosine similarity. This paper holds the assumption that if the posts are similar, then the response of the candidate post can serve as the response for the given post. First, the given post and the post of the candidate dialogue are converted to the TF-IDF vectors respectively, and the text similarity is computed between the two vectors by the cosine similarity,

$$sim_{Q2P}(q,p) = \frac{\vec{q}^T\vec{p}}{\|\vec{q}\|\|\vec{p}\|} \quad (1)$$

where $\vec{q}$ and $\vec{p}$ are the TF-IDF vectors of query q and post p respectively and $\|\vec{q}\|$ and $\|\vec{p}\|$ are the norm of the vectors of $\vec{q}$ and $\vec{p}$ respectively. In cases where posts of the candidate dialogues are the only consideration, many appropriate responses of candidate dialogues will be missed. So we also convert the response in the candidate dialog to a TF-IDF vector and take the same aforementioned works to reach the text similarity,

$$sim_{Q2R}(q,r) = \frac{\vec{q}^T\vec{r}}{\|\vec{q}\|\|\vec{r}\|} \quad (2)$$

where $\vec{q}$ and $\vec{r}$ are the TF-IDF vector of query q and response r respectively and $\|\vec{q}\|$ and $\|r\|$ are the norm of the vectors of $\vec{q}$ and $\vec{r}$ respectively.

2.3 Ranking

This paper uses a simple linear function to compute the scores of the candidate dialogues and rank them. This function is defined as:

$$rank(q,c) = \alpha * sim_{Q2P}(q,p) + (1-\alpha) * sim_{Q2R}(q,r) \quad (3)$$

where the range of α is from 0 to 1. We adjusted the parameter α for several times and found that when α is 0.85, our system has achieved the best performance.

3 Experiments

3.1 Dataset

Table 1 shows the number of training instances for the six emotion categories before and after training data preprocessing respectively. The task provided a number of 1,719,207 training instances and a number of 200 testing instances. This task requires the generation of five emotional categories of responses for each post in the testing

dataset.

Table 1. The original data and clean data

Type of Response	Original Data	Clean Data
Other	338038	325809
Like	283548	274004
Sadness	257400	244467
Disgust	317245	306556
Anger	192100	181238
Happiness	330876	307733
Total	1719207	1639807

3.2 Evaluation Metrics

How to evaluate the quality of dialogue generated by the dialogue system is a key point in the research, which is mainly divided into objective evaluation indicators and subjective evaluation indicators. The objective evaluation indicators are divided into the word overlap evaluation index and the word vector evaluation index. The objective evaluation index is mechanically rigid and cannot capture the semantic information well. The evaluation of subjective evaluation indicators can overcome this shortcoming.

This task selects subjective evaluation indicators. The rating of this evaluation index mainly considers three factors:

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 logically coherent and topic relevant content.

Fluency: whether the response is fluent in grammar and acceptable as a natural language response.

If the generated responses satisfy the fluency and coherence conditions, you can get 1 point, or you can not get scores. And if the generated responses also satisfy the emotion consistency condition, you can get another 1 point. There are 3 evaluators to decide whether the generated responses can reach the requirements, and only when at least two of the three evaluators reach an agreement can they give the corresponding score. There are two kinds of the formulas used for the grade the models, which are defined as:

$$OverallScore = \sum_{i=0}^{2} i * num_i \quad (4)$$

$$AverageScore = \frac{1}{N_t}\sum_{i=0}^{2} i * num_i \quad (5)$$

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.

3.3 Experimental Results

In the evaluation result [9], Our WUST system scores 587 marks in Overall Score. And for the happiness class, WUST system generates many appropriate responses and rank the top in contrast to other emotion classes. It is principally because the corresponding training dataset is enormous. Contrary to the emotion of happiness, the training dataset of disgust is the smallest dataset. Obviously, there is no doubt that the larger the training dataset, the better the model is. The Overall Scores of other emotion categories can prove our conclusion. From our evaluation results, we found that there are 601 instances in the testing dataset which is marked with zero scores. So there is still a huge room for improvement in this system. Table 2 shows the official evaluation results of our system.

Table 2. The official evaluation results

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

Our WUST system proposed in this paper uses the cosine similarity to calculate text similarity. Although it has ranked the fifth among all the participants, there are some problems in this method. In some cases, the given post and the candidate dialogue have no common words but similar words that have the same or similar meaning. And the candidate dialogues contain similar words will be missed in this way. This is a typical lexical gap phenomenon. Table 3 shows this phenomenon.

Table 3. An example of the lexical gap problem

Query q	手机被偷了。呵呵。 The mobile phone was stolen. Ha-ha.
Post	**Response**
无语我的 qq 被盗求拯救!讨厌怒 Oh my god, my qq was stolen! That is terrible!	孩子、愿阿门保佑你! May God bless you!
考试成绩不理想！ The exam results are not ideal!	想开点吧 Take it easy.

4 Conclusions

In this paper, we construct the WUST system to study the emotional factor in dialogue generation from the large-scale dataset. This WUST system can generate appropriate and reliable responses both in content and emotion. But the method of text similarity calculation is too simple to capture the semantic information. In addition, the aforementioned lexical gap phenomenon also has a bad effect on the content of the responses. In the future, we would like to propose a novel model using the generated-based approach. And how to capture the semantic information and alleviate the lexical gap phenomenon is our chief work. Only in this way, we can advance our study in addressing the emotional factor in dialogues.

References

1. Prendinger H, Mori J, Ishizuka M.: Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game[J]. International journal of human-computer studies, 2005, 62(2): 231-245.
2. Skowron M.: Affect listeners: Acquisition of affective states by means of conversational systems[M]. Development of Multimodal Interfaces: Active Listening and Synchrony. Springer, Berlin, Heidelberg. 2010: 169-181.
3. Polzin T S, Waibel A.: Emotion-sensitive human-computer interfaces[C]. ISCA tutorial and research workshop (ITRW) on speech and emotion. 2000: 201-206.
4. Zhou H, Huang M, Zhang T, et al.: Emotional chatting machine: Emotional conversation generation with internal and external memory[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
5. Ghosh S, Chollet M, Laksana E, et al.: Affect-lm: A neural language model for customizable affective text generation[C]. ACL, 2017: 634-642.
6. Asghar N, Poupart P, Hoey J, et al.: Affective neural response generation[C]. European Conference on Information Retrieval. Springer, Cham, 2018: 154-166.
7. Ji Z, Lu Z, Li H.: An Information Retrieval Approach to Short Text Conversation[J]. Computer Science, 2014.
8. Manning C, Raghavan P, Hinrich Schütze.: Introduction to Information Retrieval[M]. Cambridge University Press, 2008:83-89.
9. Yaoqin Zhang, Minlie Huang.: Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge. In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies, 2019.