Neural Response Generation with Relevant Emotions for Short Text Conversation

Zhongxia Chen[1,2], Ruihua Song[3⋆], Xing Xie[2], Jian-Yun Nie[4], Xiting Wang[2], Fuzheng Zhang[5], and Enhong Chen[1]

[1] University of Science and Technology of China, Hefei, China
{czx87@mail.,cheneh@}ustc.edu.cn
[2] Microsoft Research Asia, Beijing, China
[3] Microsoft XiaoIce, Beijing, China
{rsong,xingx,xitwan}@microsoft.com
[4] University of Montreal, Montreal, Canada
[5] Meituan AI Lab, Beijing, China
nie@iro.umontreal.ca, zhangfuzheng@meituan.com

Abstract. Human conversations are often embedded with emotions. To simulate human conversations, the response generated by a chatbot not only has to be topically relevant to the post, but should also carry an appropriate emotion. In this paper, we conduct analysis based on social media data to investigate how emotions influence conversation generation. Based on observation, we propose methods to determine the appropriate emotions to be included in a response and to generate responses with the emotions. The encoder-decoder architecture is extended to incorporate emotions. We propose two implementations which train the two steps separately or jointly. An empirical study on a public dataset from STC at NTCIR-12 shows that our models outperform both a retrieval-based method and a generation model without emotion, indicating the importance of emotions in short text conversation generation and the effectiveness of our approach.

Keywords: Short Text Conversation · Emotion · Neural Response Generation · Attention Mechanism · Response Emotion Estimation.

1 Introduction

Conversation is emerging as a new mode of interaction between users and systems for important applications, such as chatbots [13]. Generating natural language conversations, or short text conversation (STC), is a challenging task in the artificial intelligence field. Many existing studies on STC target conversations on social media. The task is to generate a response (comment) that can reply to a user's previous post. The recent progress of neural networks [2,3,4] has demonstrated the great potential of constructing competitive generative models, which have been used in conversation generation [14,16,17].

⋆ Ruihua Song is the corresponding author.

	Text	Emotion	Rel		Text	Emotion	Rel
Post	刚刚看到一个骑摩托车的小伙子被撞飞 I just saw a young man on a motorbike being hit.	Neutrality		Post	今天又老了一岁 Today I become one year older again	Neutrality	
R1	哇，注意安全。 Wow. Be careful.	Surprise	Yes	R1	祝你生日快乐！ Happy birthday to you!	Happiness	Yes
R2	我没有摩托车，只有电动车，也很喜欢骑，嘿嘿嘿 I have no motorbike but only an electric motorcar, and I like to ride it (laugh)	Happiness	**No**	R2	唉... Sigh...	Sadness	Yes
				R3	时间过得真快！ Time flies!	Surprise	Yes

Fig. 1. Two example posts with corresponding response candidates. (Rel: relevance)

A good response should be fluent and related to the topic of the post. These are the evaluation criteria used in most existing studies. We observe, however, that another important aspect of human conversation - emotion - plays an important role in human conversation. Only several emotions are appropriate for responding to a given post. For example, for the left post "I just saw a young man on a motorbike being hit" in Figure 1, while *surprise* is an appropriate emotion, *happiness* is not suitable because the post is sharing bad news. The appropriate emotions are not only post-dependent but also diverse. For the right post "Today I become one year older again" in Figure 1, the following three comments express multiple emotions: *happiness*, *sadness* and *surprise* which are all suitable. Thus the emotion aspect should be incorporated in STC.

Recently there are existing studies [1,5,9,12,23,25] and tasks (e.g. NTCIR-14 CECG subtask) focusing on incorporating emotions into STC. However, they only focus on either emotion diversity or emotion appropriateness of generated responses. Most existing models are proposed to generate emotional responses of any given emotion. In real-world applications, such signals are usually lacking. The system should be able to select appropriate emotions to use in the response.

Our objective is not merely to generate comments that are topically relevant to a given post, but also emotionally suitable. To fully understand how the emotions are expressed in the conversation, we conduct an analysis on social media data. Based on the remarkable findings, we first propose a stepwise solution: given a user post, an RNN-based emotion relevance estimator determines the emotion preferences for responding to a post. After that, the encoder-decoder generator module generates comments relevant to the post with the determined emotion. We then rank the generated comments considering both emotion probability and generation quality. We further propose a joint emotion-aware neural response generation model where the two modules are trained together to enable knowledge transferring to each other in post context learning.

Experimental results show that both our stepwise model and the joint learning model outperform competing methods in generating responses and re-ranking retrieved comments. More importantly, our models produce more diverse responses with appropriate emotions.

Our main contributions in this paper are as follows:

- We propose methods to generate emotion-aware responses to mimic human conversations. Our models can determine the relevant emotions to reply to a user post and generate responses with appropriate emotions.
- Our experiments show superior performance with emotion-aware responses. This study also opens the door for designing STC systems with personality.

2 Related Work

Approaches to short text conversation can be classified into retrieval-based methods and generation-based methods.

Retrieval-based methods choose the suitable response from a large candidate dataset of short text responses. [10] integrates several semantic and syntactic features such as text similarities, topic words for matching and ranking candidate responses. Convolutional Neural Networks [8] and Long Short-Term Memory [19] are also introduced to extract sentence-level features. A limitation of retrieval-based approaches is that responses are limited to those seen in the repository.

Generation-based methods generate new responses. [17] constructs a sequence-to-sequence model with an RNN encoder-decoder structure. [14] proposes a responding machine based on the encoder-decoder model with an attention mechanism. This last approach is similar to ours, but without the emotion component. Although these models can generate relevant responses to the post context, they are deprived of other characteristics in human conversation such as emotion.

Recently, [23] proposes an emotional chatting machine which can react to the post with a required emotion, while [9] implements several strategies to embed emotion into sequence-to-sequence models. [25] incorporates reinforcement learning into emotional response generation based on a large dataset labeled by emojis. [5] designs an affect sampling method to force the neural network to generate emotionally relevant words. Although these studies show the possibility of generating a response capable of conveying an emotion, the approach is limited in that the emotion of the response should be determined manually by the user. In real-world applications, such signals are usually lacking. [12] tracks emotions in whole conversations and predicts the emotion for responding. [1] constructs affective loss functions to regularize the emotion of the response. However, these models cannot choose multiple relevant emotions and the predicted emotions are not explicit emotion categories. In this study, we aim to automatically determine the appropriate emotions to be expressed in a response and generate responses conveying these emotions. This is a significant extension of previous studies.

3 Analysis on Emotion in STC

To analyze whether and how an emotion plays a role in short-text conversation, we choose human conversations from the NTCIR-12 STC-1 collection, which is extracted from Weibo (a Twitter-like social media platform in China). We randomly sample 500 posts and 14,583 corresponding comments from the dataset.

Table 1. Emotion distribution in posts and comments in the repository.

Emotion	others	happiness	sadness	disgust	surprise	anger
Post	0.431	0.339	0.118	0.047	0.049	0.016
Comment	0.277	0.445	0.110	0.091	0.060	0.017

– **Emotion Definition** Following [6], the emotion is drilled down into six categories: *neutrality*, *happiness*, *sadness*, *disgust*, *surprise*, and *anger*.

Since the emotion of a short text sentence might be subjective, we hire three assessors to independently label each post and comment. They are asked to assign one of the six emotion classes to each sentence according to their first impression. The Fleiss' Kappa [7] of three assessors is 0.41, which equates to moderate agreement. This agreement level is expected because of the highly subjective nature of the judgments. In our analysis, we use the raw judgments of the three assessors and regard them as multiple labels.

Several facts about the use of emotions in STC are observed:

– **Human conversations are often tinged with emotions.** The distribution of emotions among posts and comments is shown in Table 1. We find that about 57% of posts and 72% of comments contain explicit emotions (other than the *others* category). This confirms our hypothesis that emotion plays an important role in conversations.
– **Multiple emotions can be expressed in human responses to the same post.** The reactions of users to a post are not homogeneous in emotion. Different users may feel differently and they may express different emotions in comments. To quantify this phenomenon, we count the distribution of comments with different emotions to each post and calculate its Shannon Entropy. The mean normalized entropy is 0.574, which roughly means that a post would receive comments with three different emotions if they have equal probabilities. This shows the large variation in user's reactions to the same post.
– **Different posts have different response emotion preferences.** A system comment could contain any of the possible emotions to be emotionally relevant. However, the possible emotions of comments facing a post may change

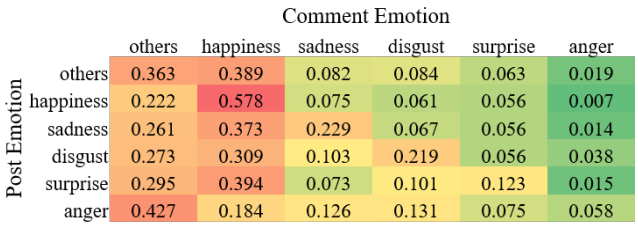

Fig. 2. Emotion transition heat map of sampled data. Each number represents the probabilities of responding to the post with this emotion when given the post emotion.

largely depending on the post. Figure 2 shows the emotion transition probabilities from post to comment (i.e. $P(comment\_emotion\ |post\_emotion)$). For posts with different emotions, the distributions of comment emotions are very different from each other. Meanwhile, for posts of the same category, the appropriate comment emotions may also change largely (see in Figure 1).

The above analyses show that the comments to a post should not only be topically relevant, but also emotionally relevant.

4 Emotion-Aware Response Generation

Given the input post $X = (x_1, ..., x_T)$, our goal is to generate a response $Y = (y_1, ..., y_t)$, and we want the emotion of Y ($\mathcal{E}_Y$) to be appropriate to the post X. In other words, we aim to maximize the generation probability of a response Y with the corresponding emotion $\mathcal{E}_Y$:

$$P(Y, \mathcal{E}_Y|X) = P(\mathcal{E}_Y|X) \times P(Y|\mathcal{E}_Y, X) \tag{1}$$

The whole response generation can be implemented in two different ways: (1) Two-step generation: We first determine the appropriate comment emotion(s) for an input post X, and then generate the comments corresponding to the emotion(s); (2) The two steps are trained jointly, with certain shared parameters. After generating responses with relevant emotions in the test stage, we rank all candidate responses with the overall scoring function based on generation probability to obtain final responses.

4.1 Two-Step Emotion-Aware Response Generation Model

Response Emotion Estimator A response emotion estimator is proposed to measure $P(\mathcal{E}_Y|X)$: how relevant an emotion $\mathcal{E}_Y$ is for responding to post X. Our implementation of this estimator is inspired by [18,24], which proposes an RNN-based text classification method with an attention mechanism. Figure 3(a) shows the architecture we implement. Similar to [18,24], we first create the hidden representations of the post X with an RNN encoder, followed by the use of an attention mechanism, a fully connected layer and finally a softmax to determine the probability of each emotion.

Following [2], we use a bidirectional recurrent neural network as the encoder. It consists of a forward RNN and a backward one. The overall hidden state h_j^e for word x_j in the post sequence is the concatenation of the forward and backward hidden states: $h_j^e = [\overrightarrow{h_j^e}^T; \overleftarrow{h_j^e}^T]^T$.

We then calculate the weighted hidden representation z:

$$z = \beta h^e \quad , \quad \beta_i = \frac{\exp(e_i)}{\sum_{k=1}^{T} \exp(e_k)} \quad , \quad e_i = v^T \tanh(U h_i^e + W_b \tanh(W_{s_b} \overleftarrow{h_1^e})) \tag{2}$$

Here $\overleftarrow{h_1^e}$ is the backward hidden state of the first word x_1, $W_{s_b}, W_b \in R^{n \times n}$, $U \in R^{n \times 2n}$ and $v \in R^n$ are weight matrices, n is the dimension of hidden states.

The representation z is then fed into a softmax fully connected layer:

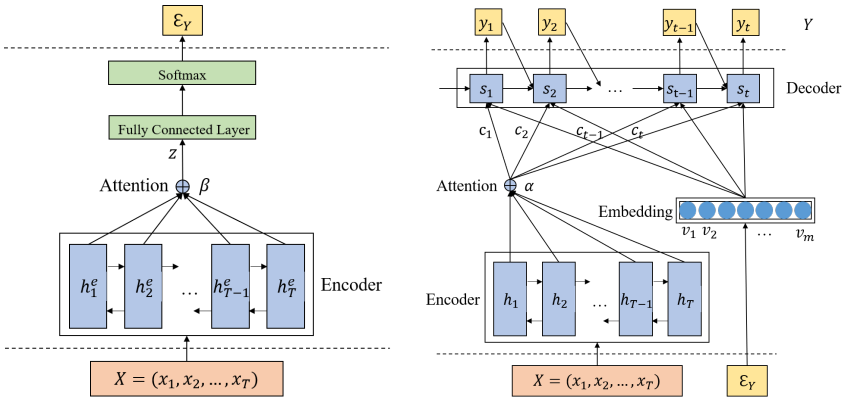

Fig. 3. Structure of (a) the response emotion relevance estimator and (b) the neural response generator. h^e and h are sets of hidden representations of the post X.

$$r = softmax(W_z z + b)$$

where $W_z \in R^{2n \times N_e}$, $b \in R^{N_e}$ and N_e is the number of emotions (6 in our case).

The probability that emotion $\mathcal{E}_Y$ is relevant to post X is $P(\mathcal{E}_Y|X) = r_{\mathcal{E}_Y}$. We use log-likelihood as the loss function of our emotion relevance estimator:

$$L(\theta_1) = \sum_{(X,\mathcal{E}_Y) \in S} \log P(\mathcal{E}_Y|X) \tag{3}$$

Emotion-Aware Response Generator We propose an emotion-aware response generator to obtain the response Y by maximizing $P(Y|\mathcal{E}_Y, X)$ with the given emotion $\mathcal{E}_Y$ for the given post X. The framework is shown in Figure 3(b).

- **Encoder** Similar to the encoder in our response emotion relevance estimator, we use another bidirectional recurrent neural network as the encoder. The overall hidden state for word x_j in the post sequence is: $h_j = [\overrightarrow{h_j}^T; \overleftarrow{h_j}^T]^T$.
- **Attention and Decoder** Similar to the traditional attention module [2], hidden states $h = (h_1, ..., h_T)$ are fed into the attention unit to obtain the context vector c_t at time t: $c_t = \sum_{j=1}^{T} \alpha_{tj} h_j$. Here the weight parameter α_{tj} is computed by

$$\alpha_{tj} = \frac{exp(r_{tj})}{\sum_{k=1}^{T} exp(r_{tk})} \quad , \quad r_{tj} = v_a^T \tanh(U_a h_j + W_a s_{t-1})$$

where s_{i-1} is the hidden state of the decoder at time $t-1$, the initial hidden state s_0 is computed by $s_0 = tanh(W_s \overleftarrow{h_1})$, $W_s, W_a \in R^{n \times n}$, $U_a \in R^{n \times 2n}$ and $v_a \in R^n$ are weight matrices, n is the dimension of RNN hidden states.

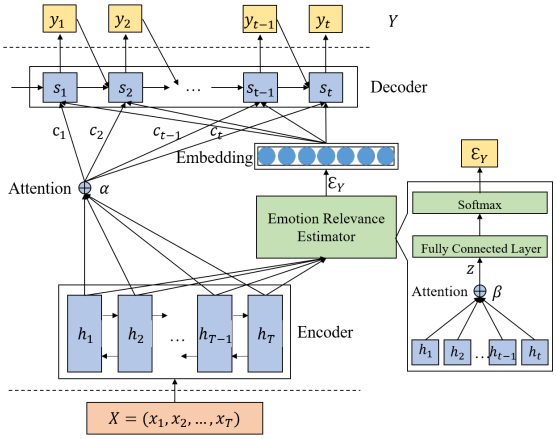

Fig. 4. Structure of jointly emotion-aware response generation model. h is the set of hidden representations for both emotion relevance estimation and generating responses.

We extend the standard decoder of the attention model with the emotion of output text $\mathcal{E}_Y$. Thus the probability of generating the t-th word y_t is:

$$p(y_t|y_{t-1}, .., y_1, \mathcal{E}_Y, X) = g(y_{t-1}, s_t, c_t, V_{\mathcal{E}_Y}) \tag{4}$$

where g is the softmax activation function, $V_{\mathcal{E}_Y}$ is the embedding of emotion $\mathcal{E}_Y$, $s_t = f(s_{t-1}, y_{t-1}, c_t, V_{\mathcal{E}_Y})$ is the hidden state at time t calculated by the RNN unit f.

– **Loss Function** We use the sum of log-likelihoods to train sequence decoding:

$$L(\theta_2) = \sum_{(X,Y)\in S} \log P(Y|\mathcal{E}_Y, X) = \sum_{(X,Y)\in S} \sum_{i=1}^{t} p(y_i|y_{i-1}, ..., y_1, \mathcal{E}_Y, X) \tag{5}$$

4.2 Joint Emotion-Aware Response Generation Model

It is intuitive that the same post would be represented in the same way at the hidden layer, so that both response emotion estimator and generator can share the same hidden representations.

Figure 4 shows the whole structure of the joint learning model. As the hidden representations are used for two tasks, we use a loss function that combines the two previous ones for the training:

$$L(\theta) = L(\theta_1) + \lambda_g L(\theta_2) = \sum_{(X,Y,\mathcal{E}_Y)\in S} \log P(\mathcal{E}_Y|X) + \lambda_g \log P(Y|\mathcal{E}_Y, X) \tag{6}$$

where λ_g is the weight of generation loss, which is set empirically.

Table 2. Statistics for the STC dataset

	Posts	Comments	Post-Comment Pairs
Training Repository	196,495	4,637,926	5,648,128
Validation Data	225	6,017	6,017
Test Data	100	22,856	26,096

4.3 Ranking Generated Results

To compare the fitness between responses generated with different possible emotion categories during testing, a scoring function for ranking is necessary. This function should fully consider two probabilities: the relevance of an emotion for the post X, and the generation of the comment sentence. As the length of comments may vary, the latter can change greatly. In order to better balance the influence of the sequence length, we replace the generation probability $P(Y|\mathcal{E}_Y, X)$ by the following average log-likelihood:

$$\hat{l}(Y|\mathcal{E}_Y, X) = \frac{1}{t}\sum_{i=1}^{t} \log p(y_i|y_{i-1}, ..., y_1, \mathcal{E}_Y, X)$$

Thus, the overall generation scoring function is as follows:

$$s(Y, \mathcal{E}_Y|X) = \lambda \log p(\mathcal{E}_Y|X) + (1-\lambda)\hat{l}(Y|\mathcal{E}_Y, X) \tag{7}$$

where λ is the weight parameter.

5 Experiments

5.1 Experiment Setup

Data The dataset comes from the NTCIR-12 STC task [15]. Table 2 gives some details from this dataset. The comments and their official evaluated scores for the test posts are collected by pooling the top ten results from all participants' submissions. There are three levels of judgment, L2, L1 and L0, which correspond to gain values of 3, 1, and 0. We also hire three assessors to label new generated comments in our experiments using the same criteria [15] and evaluation protocol as NTCIR. In the overall labeling period, the three assessors achieve 0.259 in terms of Fleiss' kappa [7]. This means they reach fair agreement. We choose the median value of three assessors' ratings as the final label.

For training the emotion-aware neural response generation model, the ground-truth emotion labels of the comments $\mathcal{E}_Y$ are necessary. We train the classifier to obtain emotion labels using Kim-CNN [11] on a large scale Weibo dataset which contains a total of 1,200,000 short texts with emoticons (e.g., smiley face). Emoticons have been used in a number of previous studies to determine the emotions of a text [20]. Similarly, we also map emoticons to emotions (e.g., smiley face to happiness). Then the labeled short text with emoticons removed are used for training. We test the emotion classifier on the labeled data created in the data analysis section. For the six categories, the overall accuracy is 0.503 which is acceptable for a 6-class classification task.

Training Details We choose the Jieba Chinese word breaker[6] to cut the short text sentences into word sequences. Since the distribution on words for posts and comments are different, we construct two vocabularies, each of which contains the 40,000 most frequent words for posts and comments. The max length of post and comment sentences is set to 40 words to reduce training costs.

We implement our model using Chainer[7]. Gated recurrent unit (GRU) are used for RNN encoder and decoder and the hidden size is set to 512. The word embedding length and emotion embedding length are 200 and 100, respectively. We use the Adadelta algorithm [21] as the training optimizer. We initialize model parameters by sampling from a uniform distribution between -0.1 and 0.1. λ_g and λ are empirically set to 1.0 and 0.5 by the validation dataset.

Measurements We use three official measures of the NTCIR-12 STC task [15]: Normalized Gain at Rank 1 (**nG@1**), Normalized Expected Reciprocal Rank at 10 (**nERR@10**) and **P+**. We further choose **Diversity** in [22] for measuring the generated result. For all metrics, high values represent good performance.

Baselines

- **Generation-based models** We compare our models with a generative model without involving emotion. This model is exactly the same as the local scheme of the Neural Responding Machine in [14] (denoted as **NRM_Loc**). We denote our two-step learning model as **ENRG_Split** and joint learning model as **ENRG_Joint**. To evaluate our model in detail, two contrasting implementations are proposed: 1) We use a uniform emotion distribution among all emotions. This uniform distribution is used to replace the emotion estimator in the joint learning model (denoted as **ENRG_Uniform**); 2) We use the emotion transition probabilities in Figure 2 to predict the response emotion according to the post emotion (denoted as **ENRG_Transition**). Here the post emotion is predicted by our emotion classifier trained before.
- **Retrieval-based models** The STC task in NTCIR-12 collects several retrieval-based results [15]. We therefore choose 1) **BUPT-C-R4** which is the best performer in the STC task; 2) **IR_base** which is our retrieval-based method that is submitted to the STC task and officially evaluated.

5.2 Evaluation on Generation Results

In this experiment, we compare the generation results of our four proposed ENRG models with the existing NRM_Loc model. For each ENRG model, we first calculate the probabilities of each candidate emotion being the responding emotion. Then we generate the top ten comments using beam search with a beam size = 30 for each responding emotion, and rank the results by the proposed

[6] https://github.com/fxsjy/jieba
[7] A flexible framework of neural networks for deep learning, http://chainer.org

Table 3. Evaluation result of generation methods. We conduct student t-tests between NRM_Loc and other methods and there is no significant difference. We also conduct t-tests between ENRG_Uniform and other methods. "⋆" means that p-value < 0.05.

Runs	Mean nG@1	Mean nERR@10	Mean P+	Diversity
NRM_Loc	0.3533	0.5166	0.5203	0.8503
ENRG_Uniform	0.3233	0.4786	0.4825	0.8488
ENRG_Transition	0.3667⋆	0.5277⋆	0.5345⋆	0.8535
ENRG_Split	0.3767	0.5410⋆	0.5351⋆	0.8356
ENRG_Joint	**0.3800⋆**	**0.5441⋆**	**0.5402⋆**	**0.8669**

scoring function. For NRM_Loc, we use the same beam search to generate the top ten comments without considering emotion. Results are shown in Table 3.

The table shows that our emotion-aware neural response generation models ENRG_Joint and ENRG_Split, as well as ENRG_Transition, outperform NRM_Loc on all three STC metrics. This result shows that emotion information does help in generating suitable comments when it is modeled reasonably. On the other hand, ENRG_Uniform performs worse than the model without any assumption about the comment emotion (NRM_Loc), which shows that an unreasonable assumption of the relevance of emotions would be of more harm than help.

ENRG_Joint not only leads to a significant improvement over ENRG_Uniform ($p < 0.05$ for nG@1, nERR@10 and P+), but also makes an improvement over ENRG_Transition on all metrics. This indicates that the appropriate comment emotion should be post-dependent.

Compared with ENRG_split, ENRG_Joint performs better in nG@1, nERR@10 and P+. This confirms the advantage of training two modules together. By sharing the same encoder parameters, the learning quality of the post context can be improved because it can benefit from both objectives. Among all generation models, ENRG_Joint can generate the most diverse responses, whereas ENRG_split has even lower diversity than NRM_Loc. This suggests that disconnecting emotion estimation and response generation may not be the best solution.

Case Study To understand how each method works, we provide some examples of the top-ranked responses in Figure 5. We can see that NRM_Loc only generates comments with popular positive emotions for the example post. Meanwhile, ENRG_Joint generates a *sad* comment "I can't go there, what a pity" which is also suitable for responding to the same post. This indicates that our model has a better ability to generate diversified comments of different appropriate emotion classes than a model that does not explicitly manage emotions.

5.3 Evaluation on Retrieval-Based Results

For each generative method, we re-rank the top ten comments given by the retrieval-based baseline method IR_base. All these re-ranking results can be evaluated by the test labeled data and compared with other retrieval-based re-

Post	马尔代夫的阿雅达Ayada度假村。想去的举个手呗 The resort of Ayada at Maldives. Raise your hands if you want to go there.	
	NRM_loc	ENRG_Prediction
C1	马尔代夫，一定去 Maldives, I must go there.	去马尔代夫，一定要去！ I want to go to Maldives, I must do it!
C2	我要去马尔代夫 I am going to Maldives.	去马尔代夫度假啊！ Go to Maldives on vacation!
C3	都想去马尔代夫 We all want to go to Maldives.	马尔代夫，好美 So beautiful Maldives is.
C4	这是我想去的地方 This is the place I want to visit.	**去不了啊，可惜** **I can't go there, what a pity.**

Fig. 5. Four comments generated by ENRG_Joint and NRM_Loc for the test post.

Table 4. Comparing the methods of applying different generation models in re-ranking our retrieval-based results with baseline methods. We conduct student t-tests between the IR_base and the other methods. "$\star$" means that $p < 0.05$.

Runs	Mean nG@1	Mean nERR@10	Mean P+
IR_base	0.3367	0.4592	0.4854
BUPT-C-R4	0.3567	0.4945	0.5082
NRM_Loc	0.3967	0.5169*	0.5470*
ENRG_Uniform	0.4133*	**0.5309***	**0.5624***
ENRG_Transition	0.4167*	0.5211*	0.5536*
ENRG_Split	**0.4200***	0.5201*	0.5563*
ENRG_Joint	**0.4200***	0.5240*	0.5565*

sults in NTCIR-12 STC-1. We compare our models with three baseline methods: IR_base, BUPT-C-R4 and NRM_Loc. Results are shown in Table 4.

The results indicate that any re-ranking method on top of the retrieved comments can improve performance, validating the hypothesis that an explicit consideration of emotions would help determine more appropriate comments.

Different from the previous results in comment generation, we no longer observe a clear superiority of ENRG_Joint and ENRG_Split over ENRG_Uniform and ENRG_Transition. We believe that the reason lies in the limited number of candidates which the re-ranking models have to work with. As the selection of the retrieved comments does not involve emotions explicitly, the comments may express very few emotions. Thus the effect of our emotion estimator is limited.

6 Conclusions

In this paper, we investigate how emotion influences responses to posts in human conversations. We propose an emotion-aware neural response generation solution for short text conversation. Our proposed approach performs the best in experiments of both generation and re-ranking scenarios on a public dataset. In the evaluation of generated results, our jointly learning model improves performance over the baseline generation-based method by 6.6% in nG@1, 5.3% in nERR@10 and 3.9% in P+. In re-ranking retrieval-based results, our method significantly beats the baselines and achieves 24.7% improvement in terms of nG@1.

References

1. Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L.: Affective neural response generation. In: European Conference on IR Research, ECIR. pp. 154–166 (2018)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
3. Cho, K., Merrienboer, B.V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014)
4. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014)
5. Colombo, P., Witon, W., Modi, A., Kennedy, J., Kapadia, M.: Affect-driven dialog generation. In: NAACL-HLT. pp. 3734–3743 (2019)
6. Ekman, P., Friesen, W.V., Ellsworth, P.: What emotion categories can observers judge from facial behavior? Emotion in the Human Face pp. 67–75 (1972)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
8. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: NIPS. pp. 2042–2050 (2014)
9. Huang, C., Zaiane, O., Trabelsi, A., Dziri, N.: Automatic dialogue generation with expressed emotions. In: NAACL HLT. pp. 49–54 (2018)
10. Ji, Z., Lu, Z., Li, H.: An information retrieval approach to short text conversation. CoRR **abs/1408.6988** (2014)
11. Kim, Y.: Convolutional neural networks for sentence classification. EMNLP (2014)
12. Lubis, N., Sakti, S., Yoshino, K., Nakamura, S.: Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. AAAI (2018)
13. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: EMNLP. pp. 583–593 (2011)
14. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: ACL. pp. 1577–1586 (2015)
15. Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., Miyao, Y.: Overview of the ntcir-12 short text conversation task. Proceedings of NTCIR-12 pp. 473–484 (2016)
16. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., Dolan, B.: A neural network approach to context-sensitive generation of conversational responses. In: NAACL HLT. pp. 196–205 (2015)
17. Vinyals, O., Le, Q.: A neural conversational model. arXiv:1506.05869 (2015)
18. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: EMNLP. pp. 606–615 (2016)
19. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: SIGIR. pp. 55–64 (2016)
20. Yuan, Z., Purver, M.: Predicting emotion labels for chinese microblog texts. In: Advances in Social Media Analysis, pp. 129–149 (2015)
21. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv:1212.5701 (2012)
22. Zhang, M., Hurley, N.: Avoiding monotony: Improving the diversity of recommendation lists. In: The 2nd ACM Conference on Recommender Systems (2008)
23. Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: Emotional conversation generation with internal and external memory. AAAI (2018)
24. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. ACL (2016)
25. Zhou, X., Wang, W.Y.: Mojitalk: Generating emotional responses at scale. In: ACL. pp. 1128–1137 (2018)