TUA1 at the NTCIR-14 STC-3 Task

Yangyang Zhou, Zheng Liu, Xin Kang, Yunong Wu, and Fuji Ren

Tokushima University, Tokushima 770-8506, JP
{c501737062, c501837066, wuyunong}@tokushima-u.ac.jp
{kang-xin, ren}@is.tokushima-u.ac.jp

Abstract. In this paper, we describe the overview of our work in STC-3 Chinese Emotional Conversation Generation (CECG) subtask at NTCIR-14. We propose a Post & Emotion to Response (P&E2R) model to train emotions together with posts to obtain the responses. We then propose another model called Post to Response & Emotion to Response (P2R&E2R) model to separate the training of emotions from that of grammar and semantics on the basis of the prior model. We try to use these models to explore how to combine emotions with the generation model better. In the evaluation section, the average scores of our models are both over 0.8, which suggests that our proposed models have emotional output capabilities in Chinese.

Keywords: Emotion · Chinese · Seq2seq · LSTM · Beam search

Team Name. TUA1

Subtasks. Chinese Emotional Conversation Generation

1 Introduction

Emotional conversation generation is an interesting and challenging problem in natural language processing. STC-1 [8] is a retrieval-based task, and STC-2 includes both the retrieval and the generation task. This year, STC-3 [13] focuses on Chinese emotional conversation generation. The emotion categories are “anger”, “disgust”, “happiness”, “like”, “sadness”, “other”. The goal is to generate responses coherent with 5 other kinds of emotions except “other” based on the given posts.

In this subtask, we submit 2 runs of the P&E2R model and the P2R&E2R model. In the former model, we add the embedding of emotion categories in the encoding part of a simple seq2seq [9] model and concatenate them with the posts encoding results. Emotional labels are from the training set, which is pretrained from a classifier. As for the P&E2R model, we use 2 encoder-decoders and separate the emotion training from grammar and semantics training. The data used to train the labels classifier are used to generate responses from emotions in this model. We use beam search [10] in generation part, with beam size 5 in both models. Both of our submissions exceed the average score of 0.8, which suggests that our methods are effective for Chinese emotional conversation generation.

The rest of this paper is organized as follows. Section 2 briefly reviews related work of CECG. Section 3 describes the methods and details of building our models. We report our experiment and evaluation results in section 4. The conclusion is in section 5.

2 Related work

Recently, human-friendly expression in the research of human-computer interaction is gaining more and more attention. As an important field of human-computer interaction, dialogue generation also needs to integrate emotions. Ghosh, S [2] proposed the Affect-LM model based on deep learning to introduce emotion category information into the training process. Zhou H [14] proposed a memory network based emotional chatting machine, which introduced emotional factors into a Chinese dialogue generation system.

Our laboratory has been involved in the Chinese subtasks of STC-1 and STC-2 [12], and has obtained some achievements in terms of retrieval and generation. The data used including STC-3, are from Chinese Weibo. And our laboratory is also doing research [1] [7] on sentiment analyses for Chinese Weibo, which can help our responses generation be better.

3 Methods

We propose two different ways to incorporate emotion labels into the training of generating models. The P&E2R model directly concatenates emotion labels with posts, using only given training data. The P2R&E2R model trains the emotion labels separately from the posts, using additional data in the process of generating responses from emotions.

3.1 P&E2R model

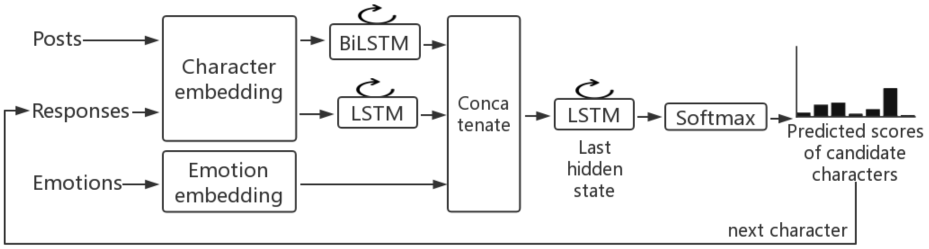

Fig. 1. P&E2R model

We concatenate the emotion category information with the simple seq2seq model as the P&E2R model. As can be seen from Fig.1, the posts and responses share the embedding weights at the beginning of the model. Because both posts and responses are in Chinese, this method ensures that the same character (or word) in post-response pairs has the same distribution when embedding, and the generated response can be more relevant to the post. For the posts contain all the information, we use Bi-LSTM [3] as the encoder for the post part. At the same time, the following characters in responses are masked, leaving only previous characters. We use LSTM [4] as the encoder for the response part. Emotion labels are embedded separately and concatenated with the encoding results from posts and responses.

Since the decoder is used to predict the current character, we use LSTM and put the last hidden state into the softmax classifier to predict the probability distributions. We use the cross-entropy loss function to reduce the distance between the results of the classifier and the probability distribution of ground truth. When choosing candidate responses, we use beam search, which is often better than greedy [11] search. After balancing time consumption and performance, we set the beam size to 5.

3.2 P2R&E2R model

Combining emotion labels with post-embedding results directly may affect the grammar and semantic training of responses by emotion classification. We consider how to concatenate emotion category information more legitimately with a simple seq2seq model rather than the way in the P&E2R model. Given the post and emotion label, when they are independent of each other, and according to Bayes' theorem, the conditional probability distribution of response is defined as:

$$\begin{aligned} P_{(Y|X,E)} &= \frac{P_{(X|Y)} \times P_{(E|Y)} \times P_{(Y)}}{P_{(X|E)}} \\ &= \frac{\frac{P_{(Y|X)} \times P_{(X)}}{P_{(Y)}} \times \frac{P_{(Y|E)} \times P_{(E)}}{P_{(Y)}} \times P_{(Y)}}{P_{(X|E)}} \\ &= \frac{P_{(Y|X)} \times P_{(Y|E)} \times P_{(E)}}{P_{(Y)} \times P_{(E|X)}} \end{aligned} \tag{1}$$

where X is the post, Y is the response, and E is the emotion category. We take the logarithmic of both sides of (1) as:

$$\log P_{(Y|X,E)} + \log P_{(Y)} = \log P_{(Y|X)} + \log P_{(Y|E)} + \log P_{(E)} - \log P_{(E|X)} \tag{2}$$

X and E are given in the task, that is to say, $\log P_{(E)}$ and $\log P_{(E|X)}$ are known quantities, which means:

$$\begin{aligned} \ell_{(\theta)} &= \log P_{(Y|X,E;\theta)} + \log P_{(Y;\theta)} \\ &\leq \max(\log P_{(Y|X;\theta)} + \log P_{(Y|E;\theta)}) + C \end{aligned} \tag{3}$$

where $L_{(\theta)}$ is a likelihood function of whether Y is semantically and emotionally related and whether it is a fluent grammatical response. C is known quantity

$\log P_{(E)} - \log P_{(E|X)}$. The maximum likelihood estimation θ^* is:

$$\begin{aligned}\theta^* &= \arg\max \ell_{(\theta)} \\ &= \arg\max(\log P_{(Y|X;\theta)} + \log P_{(Y|E;\theta)})\end{aligned} \tag{4}$$

As can be seen from (4), Post and emotion label are given independently in this task, so whether a response is emotionally relevant, semantically appropriate, and grammatically fluent can be expressed by a likelihood function of the probability distribution under post and emotion conditions respectively. Therefore, we propose the model in Fig.2.

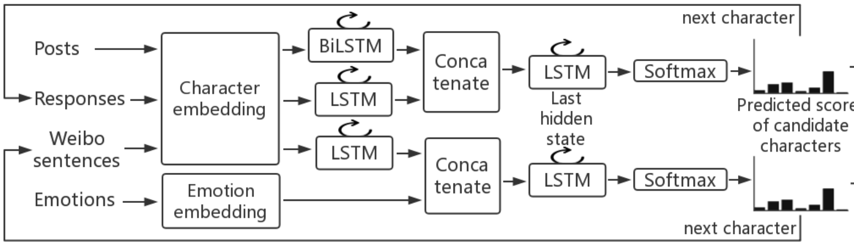

Fig. 2. P2R&E2R model

The emotion labels in the dataset are for reference only, and obtained by a simple classifier from Chinese Weibo. We choose to use this corpus to get the response probability distribution given emotion label. At the same time, we only use the post and response parts of the data provided by this task to do the grammar and semantics training.

The encoder and decoder in the post and response parts are the same as the P&E2R model. As for the emotion part, we predict the probability distribution of the current character by inputting masked Weibo sentence and the corresponding emotion label. In order to get the same range of distribution, we share the weights of embedding with the post and response part. Emotions are still embedded alone. Weibo sentences are like the responses. Due to the existence of the mask, we use LSTM as the encoder of Weibo sentence part, similar to the response part. Taking previous characters as input can make the response more like natural language. The decoder and the loss function are consistent with that of the P&E2R model. At last, the two probability distributions are added as a joint conditional probability distribution of response, and use the beam search to output candidate responses.

4 Experiments

4.1 Data analysis

The training set contains more than 1.7 million Weibo post-response pairs, and includes emotion labels of each post and response. All these sentences are tokenized. But after our statistics, we find that the length of the vocabulary is close to 32million, and the classification is quite difficult. We also try to remove the low-frequency words, but the number of words out of vocabulary occupy a high proportion, and still not good enough. A single Chinese character is meaningful, which is different from other languages such as English. Because the complexity of characters is lower than that of words, we rebuild the vocabulary with characters to improve the fluency of responses, instead of tokenized words.

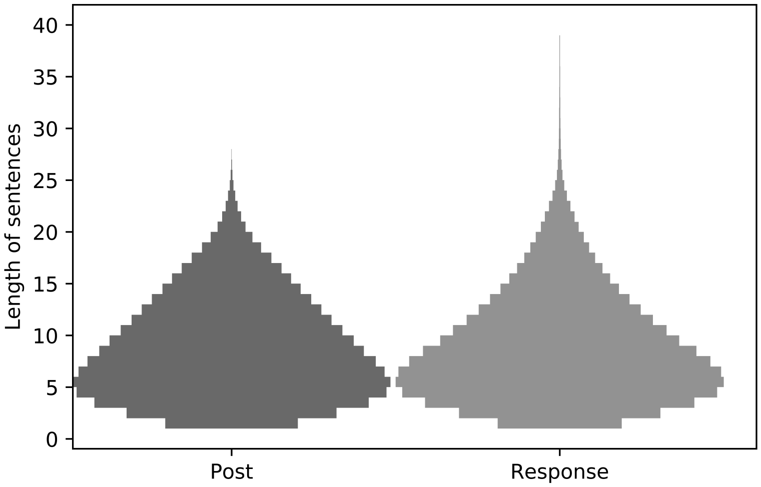

Fig. 3. Distribution of posts & responses length: the y-axis shows the length of the sentence (the number of characters in the sentence), and the x-axis shows the quantity of sentences in certain length (the wider the bar, the more sentences).

We also count the length of sequences in the training set. As shown in Fig.3, most sentences are between 1 and 15 in length. Considering the speed of training and generation, we limit the maximum length of response to 32. According to our statistics, about 0.32% of responses are longer than 32 characters, and truncating the excess should have little effect.

As for the extra dataset used for emotion part training, we choose the Chinese Weibo data from the NLPCC Emotion Classification Challenge [14], which is used for emotion classification. The dataset contains more than 40 thousand sentences and corresponding emotion labels.

4.2 Preprocessing

Before training, we first removed about 17 thousand post-response pairs that do not contain Chinese characters. Considering that a large number of repeated expressions can affect the learning of neural networks, we remove the extra duplicate words (including symbols) and remain 3 times at most.

We build vocabularies for posts and responses separately, where posts retain common characters with a frequency greater than 50 (approximately 3500 characters), responses keep common characters with a frequency more than 250 (approximately 2500 characters), and the rest are replaced by "out of vocabulary" symbol. We also add "start" and "end" symbols, and do padding and masking for each response.

4.3 Evaluation metrics

Since the existing automatic evaluation metrics for generation tasks such as BLEU [6] etc. are not suitable for dialogue generation, most of the dialogue generation results require manual evaluation, just like this task. The test set contains 200 posts, including 5 different emotion classes, with 40 posts each. Every post is required to generate a response for each emotion category, with a total of 1000 responses. When evaluating these responses, emotion consistency, coherence and fluency metrics are used. If it cannot be logically coherent or topic relevant or fluent in grammar, the response will get label 0. On the contrary, it will get label 1. Based on this, if it has the correct emotional expression, it will get label 2.

4.4 Results

	Like		Sad		Disgust		Anger		Happy		Total	
	1	2	1	2	1	2	1	2	1	2	1	2
Label 0	121	**109**	**84**	92	**82**	92	85	85	**71**	76	**443**	454
Label 1	11	24	31	40	105	82	110	107	36	25	293	278
Label 2	**68**	67	**85**	68	13	**26**	5	**8**	93	**99**	264	**268**
Average	0.735	**0.790**	**1.005**	0.880	0.655	**0.670**	0.600	**0.615**	1.110	**1.115**	**0.821**	0.814

Table 1. Evaluation results of our run submissions

Table 1 shows the evaluation results of 2 run submissions by our group, including the number of responses with different labels and the average scores. 1 stands for the P&E2R model, while 2 for the P2R&E2R model in the table. From the results of emotion classes, probably due to the quantitative imbalance of data, anger and disgust account for a small proportion in the training set, the performance of these 2 kinds is not as good as other classes, which is similar to the results of other groups. From the results of models, the response given by the P2R&E2R model is more subject-related, such as a keyword "movie(电影)" in the post. An example is given in table 2. We provide English translations of the texts using machine translation.

Post	Why do not **movie** theaters sell peripheral products [tears] 为什么**电影**院不卖周边呢 [眼泪]	
Responses	TUA1_1	TUA1_2
Like	I also like. 我也喜欢。	Because I like to watch **movie** 因为我喜欢看**电影**
Sad	[tears] I also want it [泪] 我也要	Because I also have not seen it [tears] 因为我也没看过 [泪流满面]
Disgust	What's going on? 这是什么情况啊?	Because you didn't watch the **movie** 因为你没有看**电影**啊
Anger	What do you mean? 什么意思?	What **movie**? 什么**电影**啊?
Happy	[laughing] [偷笑]	Ha ha, many people know, do not know what fun? 哈哈, 好多人都知道, 不知道有什么好玩的?

Table 2. Responses comparison of P&E2R and P2R&E2R model

Seq2seq model always tends to give generic and safe responses [5]. In our experiments, the emotions "like", "sad", "disgust" and "happy" always tend to generate sentences with emoji "[loving you]([爱你])", "[tears]([泪])", "[digging booger]([挖鼻屎])" and "[laughing]([偷笑])" respectively, while "anger" always tend to respond with questions such as "What happened?(怎么回事？)" and "What do you mean?(什么意思？)". It also implies that the emoji is strongly related to the expression of emotion, at least in Chinese Weibo.

In general, 2 submissions are very close in performance. Considering that the P2R&E2R model is trained separately, the probability distribution errors are cumulative, causing its response is inferior to that of the P&E2R model in grammar and semantics, but is slightly better than that of the P&E2R model in emotional expression.

5 Conclusions

This paper reports our work in STC-3 CECG subtask at NTCIR-14. After analyzing the training data, we use a character-based preprocessing method. We propose two different methods to generate responses with emotions. The P&E2R model is a method of concatenating emotion labels with posts, and the P2R&E2R model is a method by training posts and emotion labels respectively. Both methods utilize the beam search to improve the performance of responses. Both submissions get average scores above 0.8, proving that our proposed model is effective to some extent.

In the future, we will pay more attention to the diversity of response generation. For instance, adding a similarity penalty factor to intervene beam search may reduce the probability of generic and safe responses.

6 Acknowledgments

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

References

1. Gao, F., Sun, X., Wang, K., Ren, F.: Chinese micro-blog sentiment analysis based on semantic features and pad model. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). pp. 1–5. IEEE (2016)
2. Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S.: Affect-lm: A neural language model for customizable affective text generation. arXiv preprint arXiv:1704.06851 (2017)
3. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
7. Quan, C., Ren, F.: Visualizing emotions from chinese blogs by textual emotion analysis and recognition techniques. International Journal of Information Technology & Decision Making **15**(01), 215–234 (2016)
8. Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., Miyao, Y.: Overview of the ntcir-12 short text conversation task. In: NTCIR (2016)
9. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
10. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse beam search: Decoding diverse solutions from neural sequence models. arXiv preprint arXiv:1610.02424 (2016)
11. Wilt, C.M., Thayer, J.T., Ruml, W.: A comparison of greedy search algorithms. In: third annual symposium on combinatorial search (2010)
12. Wu, Y., Kang, X., Kita, K., Ren, F.: Tua1 at ntcir-13 short text conversation 2 task
13. Zhang, Y., Huang, M.: Overview of NTCIR-14 short text generation subtask: Emotion generation challenge. In: Proceedings of the 14th NTCIR Conference (2019)
14. 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 (2018)