Emotional Chatting Machine

Bofei Zhang and **Shaoling Chen** and **Yuxuan He** and **Yu Cao**
Center for Data Science, New York University
{bz1030, sc6995, yh2857, yc3651}@nyu.edu

Abstract

A human-like chatbot with the ability to generate different responses given different emotions. However, current language studies focus only on coherence of the sentence. In this paper, we solve this problem by (1) employing Emotional Chatting Machine framework to build an emotional chatbot on English corpus, (2) improving ECM performance by applying global attention mechanism and improve the internal and external memory module. Experiments suggest that modified ECM produces more diverse and emotional response than traditional sequence-to-sequence method.

1 Introduction

Currently, natural language models for dialogue generation that simulates human conversation has been widely studied. However, expression and perception of emotion, which plays a central part of human communication receives little attention. Indeed, existing psychological studies of Prendinger and Ishizuka (2005b) have shown that addressing emotion in dialogue system can enhance user satisfaction. Motivated by these psychological studies, we want to build a emotional chatbot that can generate different responses given different emotion categories.

We base our research on Zhou et al. (2017)'s emotional chatting machine (ECM) framework. The ECM framework is an encoder-decoder model that incorporates emotion information in decoding process by (1) static emotion embedding, (2)an internal memory module, and (3) an external memory module.The specific model proposed in Zhou et al. (2017)'s paper was trained on Chinese corpus and renders noticeable performance. However, directly applying ECM on English corpus shows a poor performance. It cannot generate coherent or emotional sentences.

This problem might be caused by the linguistic differences between English and Chinese. First, word segmentation for Chinese encodes more information than that of English since Chinese text does not have natural delimiter (Xue et al., 2005). To enhance the context information of English token, we adopted global attention mechanism (Luong et al., 2015). Second, Morphology structure of English words leads to higher complexity in training compare to Chinese words (Xue et al., 2005). To solve this issue, we choose to use vocabulary from BNC Consortium (2007), which include various morphology words annotated by experts. Third, using static emotional embedding alone sacrifice grammar structure, and internal memory alone only has weak effect on emotion factor in experiments. To solve this issue, we combine both internal memory state and static emotion category embedding during decoding.

The modified ECM achieves best score in grammar and emotion aspects under human evaluation. To summarize our contribution:(1) It is the first paper implementing the ECM framework on English corpus. (2) It utilized three novel approaches to improve the original ECM model on attention mechanism, modified internal and external memory module to achieve best result.

2 Related Work

Back to early 2000s, studies have examined the effect of emphatic agent with subtle expressivity on the affective state of users. In the work of Prendinger and Ishizuka (2005a), it showed that an empathic character can soothe the nerve of user and potentially have positive effect on perception of the difficulty of the task. Partala and Surakka (2004) proposed that receiving encourage from dialogue system can improve user problem solving skills.

These psychological finding are limited to rule-based system and small amount data, which are not suitable for large scale dialogue generation. Recently, Zhou et al. (2017) proposed a model called emotional chatting machine (ECM) to generate response in conversation both appropriate in grammar and in emotions. The model is built on top of the traditional seq2seq framework Sutskever et al. (2014) using Gated Recurrent Unit in encoder and decoder(Chung et al., 2014). with a local attention mechanism from Bahdanau et al. (2014). To address the emotion factor, Zhou et al. (2017) introduces three changes comparing to seq2seq model. Firstly, it does a static emotion category embedding to model emotion category. Secondly, it added internal memory to capture the implicit emotion changes inside a sentence. Thirdly, it added a external memory module to explicitly choose emotion or generic words during decoding. It renders notable performance on Chinese corpus.

To allow seq2seq to generate better dialogue, current research focus on attention mechanism. Bahdanau et al. (2014) developed a local attention mechanism, where the decoder make a prediction based on the encoder output of current time step. On the other hand, Global attention mechanism proves to have better performance than local attention in various tasks such as machine translation (Luong et al., 2015).

As pointed out by Li et al. (2016), optimizing the cross entropy error loss will give maximum likelihood estimator of input sentence. However, it will favor generic response such as "I don't know", which will compromise the effect of emotion. To fix this, Li et al. (2016) implemented diversity-promoting algorithm in beam search. Changing objective function is another option to address this problem. Maximum Mutual Information(MMI) objective will add penalty to trade off the diverse response and generic response. As previous experiment shown, MMI objective for seq2seq will produce more informative response, yielding higher BLEU score than seq2seq with cross entropy error (Li et al., 2015).

3 Data

Sentiment Classifier We trained a bi-directional LSTM model (Christos Baziotis et al., 2018) to tag the data for training the main model. We used dataset of SemEval-2018 Task 1 (Mohammad et al., 2018) to train the sentiment analysis model, which consists of 7801 tweets and annotated with 4 emotion categories: anger, fear, joy, sadness. We also finetuned the pretrained BERT model (Devlin et al., 2018), which underperforms the bi-LSTM. This classifier was also used for ECM emotion evaluation.

Emotional Chatting Machine We use data incorporated from four datasets: BNC Corpus BNC Consortium (2007), Cornell Movie Corpus Danescu-Niculescu-Mizil and Lee (2011), Daily Dialogue Li et al. (2017) to train our model. Our dataset consists 233070 pairs of posts and responses for training 1000 pairs for validation, and 1000 pairs for testing. There are 15632 words in vocabulary, and 2526 words in external memory.

4 Methodology

Baseline Model The baseline model is a seq2seq model which is based on encoder-decoder framework (Sutskever et al., 2014). In this paper, we implemented encoder and decoder by bidirectional gated recurrent units (GRU) and unidirectional GRU respectively (Chung et al., 2014). The computation in our baseline follows exactly equations (1-4) of Zhou et al. (2017).

Modified Emotional Chatting Machine

Internal Memory with Static Emotion Embedding In our model, each emotion category is represented by two types of a real-valued dense vector, static emotional category embedding and internal memory from Zhou et al. (2017), which capture the emotion dynamics during decoding.

The detailed procedure of internal memory is illustrated in Figure 1. At each step t, ECM computes a read gate g_t^r and write gate g_t^w by following equation (6-7) from Zhou et al. (2017). The read and write gate controls read from and write into the internal memory by equation (8-9) of Zhou et al. (2017). Since g_t^w is in between 0 and 1, the internal emotion state will be erased certain amount at each step. At the end of sentence, this vector decays to 0. Unlike equation (10) of Zhou et al. (2017) which passed previous target word $e(y_{t-1})$, the previous state of the decoder h_{t-1}, the context vector of our model is c_t computed by global attention and the output from internal memory $M_{r,t}^I$ to the GRU, we feed static emotional category embedding v_e to GRU to update its state h_t:

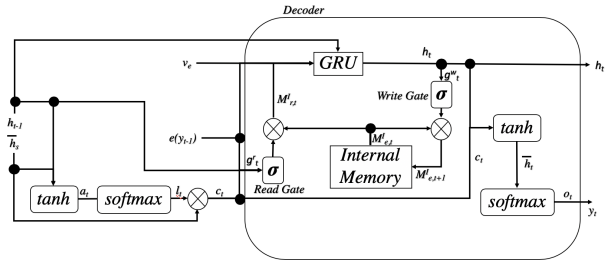

Figure 1: Data flow of the decoder with an internal memory

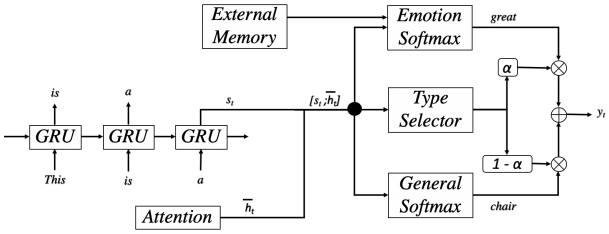

Figure 2: Data flow of the decoder with an external memory

$$h_t = GRU(h_{t-1}, [e_{y_{t-1}}; v_e; M^I_{r,t}; c_t]) \quad (1)$$

Rather than directly feeding into external memory, h_t will be feed into global attention to generate attentional output $\overline{h}_t$ and to update context vector c_t.

Global Attention The global attention generates location-based score by *concat* way from Luong et al. (2015). Given all hidden inputs from encoder $\overline{h}_s$ and hidden output from last time step h_{t-1}. the source side context vector c_t is computed as follows:

$$score(h_t, \overline{h}_s) = v_a^T tanh(W_a[h_t; \overline{h}_s]) \quad (2)$$

$$a_t = softmax(score(h_t, \overline{h}_s)) \quad (3)$$

$$c_t = a_t \times \overline{h}_s \quad (4)$$

The weighted sum of all encoder outputs gives context vector c_t. Then, based on Luong et al. (2015), we can combine the source-side context vector c_t and GRU output h_t as follows:

$$\overline{h}_t = tanh(W_c[c_t; h_t]) \quad (5)$$

We employ the attentional hidden input $\overline{h}_t$ to external memory. The original hidden output h_t, the resulting context vector c_t along with internal memory state $M^I_{e,t+1}$ from equation (9) of Zhou et al. (2017) are delivered to the next decoder.

External Memory Internal memory alone does not provide clear and observable correlation between emotion state and word selection. For examples, compared to *he* and *do*, *fantastic* and *stupid* express strong emotions. We add external memory module(EMem) to allow the model to choose emotion words and generic words by explicitly generated probability. We build external memory word list which contains word with corresponding emotion using the data from The Sentiment and Emotion Lexicons, a manually created sentiment word list by the experts of the National Research Council of Canada (Mohammad, 2016). During each decoding step, external memory explicitly supervises the probability of selection between generic word and emotion word. This process is illustrated in Figure 2. As discussed above, the attention output $\overline{h}_t$ is used to compute to final generic and emotion token distribution P_g and P_e respectively as follows:

$$\alpha_t = sigmoid(v_u^T \overline{h}_t) \quad (6)$$

$$P_g(y_t = w_g) = softmax(W_g^o \overline{h}_t) \quad (7)$$

$$P_e(y_t = w_e) = softmax(W_e^o \overline{h}_t) \quad (8)$$

$\alpha_t \in [0,1]$ is a scalar and considered a probability to choose word from emotion distribution P_e or generic distribution P_g. To combine these two distributions, Zhou et al. (2017) utilized functions involving complicated permutation and vector concatenation, which we believe are computationally inefficient and unstable in model performance. Therefore, we employ a novel approach to combine two distribution as follows:

$$y_t \sim o_t = P(y_t) = (1 - \alpha_t)P_g + \alpha_t P_e \quad (9)$$

Using the sum of these two distributions as the predicted distribution o_t.

Loss Function We employ the same loss function from equation (15) of Zhou et al. (2017).

5 Experiments and Results Analysis

We implemented proposed models in PyTorch. After hyper-parameter tuning, we chose the model with 4-layer GRU in encoder and decoder, and beam search with beam size 20 to retrieve responses. The diversity penalty we used is 1 as recommended by Li et al. (2016). We compare the results of our modified ECM, the modified ECM without external memory module against baseline model.

Automatic Metric Liu et al. (2016) argued that BLEU score is not suitable for dialogue generation. Therefore, we chose perplexity as our automatic metricsto evaluate sentence correctness at content level. We also used the emotion accuracy to evaluate the response's emotion using predicted emotion category of a generated response by the emotion classifier.

Model	Perplexity	Emotion Accuracy
Seq2Seq	12.28	24.84%
Modified ECM	12.03	25.27%
w/o EMem	11.32	23.71%

Table 1: Objective analysis

From Table 1, all models converge to comparable level, which indicates very good grammar correctness. The modified ECM model has the best performance on the emotion accuracy.

Human Evaluation We designed an online survey[1] to perform human evaluation to better understand the performance of different models. Given a pair of post and responses, participants were asked to select the emotion category and score the content relatedness, grammar and emotion intensity(scale of 1-5) of the response. 120 pairs of post and responses randomly generated from training data and three models were evaluated by 15 participants separately.

In reference to training data, Table 2 shows that external memory module boosted the performance in emotion and content coherence. Grammar score improves from 3.88 to 4.11, emotion intensity improves from 2.67 to 2.74 and emotion accuracy improves from 37% to 40%.

Model	Content Relatedness	Grammar	Emotion Intensity	Emotion Accuracy
Seq2Seq	3	3.88	2.67	37%
w/o EMem	2.48	3.57	2.42	40%
Modified ECM	2.81	4.11	2.74	40.70%
Training Data	3.09	4.39	2.84	51.6%

Table 2: Human Evaluation Results

Case Study As presented in the first case of Table 3, the modified ECM could generate various sentences in terms of the input emotion. Moreover, removing external memory will lead the model to generate same sentence for all emotion categories.

Error Analysis In Automatic Metric Table 1, the poor performance of sentiment classifier accuracy is related to limited training data and intrinsic subjective of emotion. In Human Evaluation Metric Table 2 content relatedness for the modified ECM model decreased 0.19 from baseline model, which shows that we scarified content relatedness slightly for emotion, but overall the modified ECM model is able to control the weight of emotion and generate responses appropriately. In Case Study Table 3, the modified ECM sadness and fear responses perform slightly worse than responses from modified ECM joy and anger due to the limitation of external memory module and intrinsic ambiguity in sentence emotions. First, we have insufficient amount of data in EMem word list and EMem world list is imbalanced(42% - joy, 28% - anger). Second, the sadness and fear responses for input "how is it coming" in our example demonstrated the ambiguity of sentence emotion.

[1]https://bofei.shinyapps.io/ECM-Experiment/

Post	Model	Emotion	Response
Do you love me?	Seq2Seq	-	I'm very fond of you
	Modified ECM	Joy	christ I love you .
		Anger	course I love you
		Sadness	do you love me.
		Fear	what kind of stupid question was that
How is it coming ?	Seq2Seq	-	I m fine . Thank you .
	Modified ECM	Joy	well . Thank you so much .
		Anger	better . I m tired
		Sadness	good . It is tired
		Fear	good . It is tired

Table 3: Case study for responses generated by baseline and modified ECM (red words are in external memory)

6 Conclusion and Future Work

In this paper, we employ the ECM framework to build an emotional chatbot on English corpus, improve ECM in both coherence and emotions by applying global at-tention mechanism and improve the internal and external memory module. The results suggest that modified ECM produces more diverse and emotional response than traditional sequence-to-sequence method.

In our future work, we will explore emotion interactions with ECM to improve engagement. The model has the potential to take user information, like preferences and personality, and the emotion of the inputs into account to decide the most appropriate emotion category for the response. In addition, the emotion tag for dataset has greatly influenced the performance. We will also try to collect more samples to optimize the sentiment classifier for better emotion tag results.

Collaboration

Chen trains the sentiment classifier. Zhang and He build the framework of the model and implement them in codes. Yu tries out some datasets and experiments. All members contribute in terms of literature review and paper write-up.

Code

- Please see ECM implementation https://github.com/bofei5675/ECM_NLU
- Please see sentiment classifier implementation https://colab.research.google.com/drive/1isDqph1hUJZqqFOUq8aBFtw1nPj94OlW
- Please see experiments in various pre-trained embedding https://github.com/jenniferhe/EmotionalChattingMachine

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

BNC Consortium. 2007. The british national corpus.

Nikos Athanasiou Christos Baziotis, Athanasia Kolovou Alexandra Chronopoulou, Nikolaos Ellinas1Shrikanth Narayanan Georgios Paraskevopoulos, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset.

Chia-Wei Liu, Ryan Lowe1, Iulian V. Serban, Michael Noseworthy, Laurent Charlin1, and Joelle Pineau1. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122 – 2132.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation.

Saif M. Mohammad. 2016. The sentiment and emotion lexicons.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

T. Partala and V Surakka. 2004. "the effects of affective interventions in humancomputer interaction". 16(2):95 – 309.

H. Prendinger and M. Ishizuka. 2005a. "using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game". 62(2):231 – 245.

Helmut Prendinger and Mitsuru Ishizuka. 2005b. The empathic companion: A character-based interface that addresses users' affective states. 19(3):267–285.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese TreeBank: Phrase structure annotation of a large corpus. 11(2):207–238.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory.