Emotion-aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction

Wei Wei[1], Jiayi Liu[1], Xianling Mao[2], Guibing Guo[3], Feida Zhu[4], Pan Zhou[5], Yuchong Hu[6],

[1] Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology

[2] School of Computer Science and Technology, Beijing Institute of Technology

[3] Software College, Northeastern University

[4] School of Information Systems, Singapore Management University

[5] School of Electronic Information and Communications, Huazhong University of Science and Technology

[6] School of Computer Science and Technology, Huazhong University of Science and Technology

[1] {weiw, liujiayi7, panzhou, yuchonghu}@hust.edu.cn [2] maoxl@bit.edu.cn

[3] guogb@swc.neu.edu.cn [4] fdzhu@smu.edu.sg

ABSTRACT

The consistency of a response to a given post at ***semantic***-level and ***emotional***-level is essential for a dialogue system to deliver human-like interactions. However, this challenge is not well addressed in the literature, since most of the approaches neglect the emotional information conveyed by a post while generating responses. This article addresses this problem by proposing a *unified* end-to-end neural architecture, which is capable of simultaneously encoding the *semantics* and the *emotions* in a post for generating more intelligent responses with appropriately expressed emotions. Extensive experiments on real-world data demonstrate that the proposed method outperforms the state-of-the-art methods in terms of both content coherence and emotion appropriateness.

KEYWORDS

Dialogue generation; emotional conversation; emotional chatbot

ACM Reference Format:

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, Yuchong Hu. 2019. Emotion-aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357937

1 INTRODUCTION

Dialogue systems in practice are typically built for various purposes like emotional interaction, customer service or information acquisition, which can be roughly categorized into three classes, *i.e.,* chitchat chatbots, task-oriented chatbots and domain-specific chatbots. For example, a task-specific chatbot can serve as a customer consultant, while a chitchat chatbot is commonly designed for convincingly simulating how a human would respond as a conversational partner. In fact, most recent work on response generation in chitchat domain can be summarized as follows, *i.e.,* retrieval-based, matching-based, or statistical machine learning based approaches [13, 30, 42, 43]. Recently, with the increasing popularity of deep learning, many research efforts have been dedicated to employing an encode-decoder architecture *i.e.,* Sequence-to-sequence (*Seq2seq*) models [5, 36], to map a post to the corresponding response with little hand-crafted features or domain-specific knowledge for the conversation generation problem [33, 41]. Subsequently, several variants of *Seq2seq* models are also proposed to address different issues [23, 29, 31, 44].

Despite the great progress made in neural dialogue generation, a general fact is that few work has been reported to automatically incorporate emotional factors for dialogue systems. In fact, several empirical studies have proven that chatbots with the ability of emotional communication with humans are essential for enhancing user satisfaction [4, 21, 28]. To this end, it is highly valuable and desirable to develop an emotion-aware chatbot that is capable of perceiving/expressing emotions and emotionally interacting with the interlocutors. In literature, Zhou *et al.* [46] successfully build an emotional chat machine (ECM) that is capable of generating emotional responses according to a pre-defined emotion category, and several similar efforts are also made by [11, 25], such as [48] proposed by Zhou *et al.* that utilizes emojis to control the emotional response generation process within conditional variational auto-encoder (CAVE) framework.

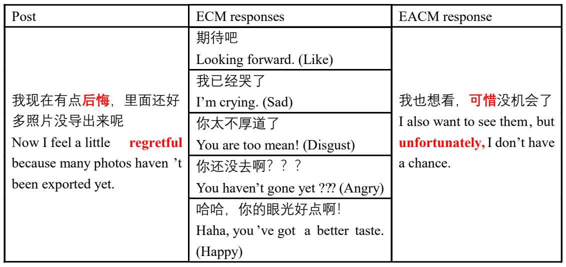
Nevertheless, these mentioned approaches cannot work well owing to the following facts: (i) These approaches solely generate the emotional responses based on a pre-specified label (or emoji) as shown in Figure 1, which is unrealistic in practice as the well-designed dialogue systems need to wait for a manually selected emotion category for response generation; (ii) The generation process apparently divided into two parts would significantly reduce the smoothness and quality of generating responses; and (iii) As shown in Figure 1, some emotionally-inappropriate responses (even conflicts) might apparently affect the interlocutor's satisfaction. Thereby, here we argue that fully exploiting the emotional information of the given post to supervise the learning process is definitely beneficial for automatically generating responses with the optimal

CIKM '19, November 3–7, 2019, Beijing, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6976-3/19/11…$15.00

https://doi.org/10.1145/3357384.3357937



Post	ECM responses	EACM response
我现在有点**后悔**，里面还好多照片没导出来呢 Now I feel a little **regretful** because many photos haven ’t been exported yet.	期待吧 Looking forward. (Like)	我也想看，**可惜**没机会了 I also want to see them, but **unfortunately,** I don’t have a chance.
	我已经哭了 I’m crying. (Sad)	
	你太不厚道了 You are too mean! (Disgust)	
	你还没去啊？？？ You haven’t gone yet ??? (Angry)	
	哈哈，你的眼光好点啊! Haha, you ’ve got a better taste. (Happy)	

Figure 1: Sampled dialogue generated from ECM and EACM.

emotion (*rf.* “EACM response” generated by our method shown in Figure 1).

Previous methods greatly contribute to emotion-aware conversation generation problem, however, they are insufficient and several issues emerge when trying to fully-address this problem. ***First***, it is not easy to model human emotion from a given sentence due to semantic sparsity. Psychological studies [26, 27] demonstrate that human emotion is quite complex and one sentence may contain multi-types of emotions with different intensities. For example, a hotel guest might write a comment like “The environment is not bad, however the location is too remote.” As such, solely using the post’s emotion label is insufficient and we need to appropriately extract the emotional information of the input post for representation. ***Second***, it is difficult for a model to decide the optimal response emotion for generation, and it is also not reasonable to directly map the post’s emotion label to the response’s emotion label, as the emotion selection process is determined not only by the post’s emotion but also by its semantic meaning. ***Third***, it is also problematic to design a unified model that can generate plausible emotional sentence without sacrificing grammatical fluency and semantic coherence [46]. Hence, the response generation problem faces a significant challenge: that is, how to effectively leverage the emotion and semantic of a given post to automatically learn the emotion interaction mode for emotion-aware response generation within a unified model.

In this paper, we propose a novel emotion-aware chat machine (*i.e.,* EACM for short), which is capable of perceiving other interlocutor’s feeling (*i.e.,* post’s emotion) and generating plausible response with the optimal emotion category (*i.e.,* response’s emotion). Specifically, EACM is based on a unified *Seq2seq* architecture with a *self-attention* enhanced emotion selector and an emotion-biased response generator, to simultaneously modeling the post’s emotional and semantic information for automatically generating appropriate response. Experiments on the public datasets demonstrate the effectiveness of our proposed method, in terms of two different types of evaluation metrics, *i.e., automatic metric* and *human evaluation*, which are used to measure the diversity of words in the generated sentences (*automatic metric*, indirectly reflecting the diversity of expressed emotions, *e.g., distinct-n*), and whether the generated responses’ emotions are appropriate according to human annotations (*human evaluation*). The main contributions of this research are summarized as follows:

(1) It advances the current emotion conversation generation problem from a new perspective, namely emotion-aware response generation, by taking account of the emotion interactions between interlocutors.
(2) It also proposes an innovative generation model (*i.e.,* EACM) that is capable of extracting post’s sentimental and semantic information for generating intelligible responses with appropriate emotions.
(3) It conducts extensive experiments on a real-word dataset, which demonstrates the proposed approach outperforms the state-of-the-art methods at both *semantic*-level and *emotional*-level.

2 RELATED WORKS

The current conversation generation approaches are mostly based on the basic Sequence-to-sequence (*Seq2seq*) framework, which has been proven [5, 36] that is able to effectively address the sequence-mapping issues. Following the success of *Seq2seq* model, methods based on such framework have been applied to various domains, such as machine translation [2] and image caption generation [1].

Indeed, there exist some attempts on improving the performance of such encoder-decoder architecture for machine translation problem. Bahdanau *et al.* [2] utilize Bi-directional Long Short-Term Memory (Bi-LSTM) network with attention mechanism for long-sentence generation, which is able to automatically search for relevant parts in the context. Luong *et al.* [20] thoroughly evaluate the effectiveness of different types of attention mechanisms, *i.e.,* global/local attention with different alignment functions. Furthermore, self-attention mechanism proposed by [18, 38] is proved effective for machine translation, which can yield large gains in terms of BLEU [24] as compared to the state-of-the-art methods. To cope with the increasing complexity in the decoding stage (caused by large-scale vocabularies), Jean *et al.* [12] consider to use sampled softmax methods and thus achieve encouraging results. These works have improved the generation performance of the *Seq2seq* model and speeded up decoding process, which build a solid foundation for the future studies based on this architecture.

There also exist many efforts dedicated to research on how to apply Seq2seq model for conversation systems [33, 35, 41], by regarding the conversation generation as a *monolingual translation* task, and later on several variants are proposed for a wide variety of domain-specific issues, such as hierarchical recurrent model [31, 32], topic-aware model [44]. Besides, several persona-based [15] models and identity-coherent models [29] are proposed to endow the chatbots with personality for addressing the context-consistency problem. There have been numerous attempts to generate more diverse and informative responses, such as Maximum Mutual Information (MMI) based model [14] and enhanced beam-search based model [16, 39]. Several approaches are also proposed for specific tasks, such as Zhou *et al.* [47] take account of static graph attention to incorporate commonsense knowledge for chatbots. Zhang *et al.* [45] propose different solutions for two classical conversation scenarios, *i.e.,* chit-chat and domain-specific conversation.

In recent years, many researches propose that emotion factors are of great significance in terms of successfully building human-like conversation generation models. Ghosh *et al.* [8] propose *affect language model* to generate texture conditioned on the given affect categories with controllable affect strength. Hu *et al.* [10] present a combined model of the Variational Auto-Encoder (VAE) and holistic attribute discriminators to generate sentences with certain types of sentiment and tense. However, these models are mainly built for emotional text generation task. Several proposals study the conversation generation problem with emotional factors, which are most related to our proposed conversation generation problem. Zhou *et al.* [46] develop an Emotional Chat Machine (ECM) model using three different mechanisms (*i.e., emotion embedding*, *internal memory* and *external memory*) to generate responses according to the designated emotion category. Similarly, Zhou *et al.* [48] propose a reinforcement learning approach within conditional variational auto-encoder framework to generate responses conditioned on the given emojis. In [11], Huang *et al.* propose three different models that are capable of injecting different emotion factors for response generation. Peng *et al.* [25] utilize Latent Dirichlet allocation (LDA) models to extract topic information for emotional conversation generation. However, all of such models are based on a designated emotion category to generation emotional responses, which need human beings to decide an optimal response emotion category for generation. Besides, the emotion information of the given post is not explicitly modeled, and thus the generated responses are not good enough. As opposed, our proposed model is capable of effectively leverage the emotion and semantics of a given post to automatically learn the emotion interaction mode for emotion-aware response generation.

3 PROPOSED MODEL

3.1 Preliminary: Sequence-to-Sequence Attention Model

In the literature, sequence-to-sequence (*Seq2seq*) model is widely adopted for dialogue generation [2, 5, 36]. In order to promote the quality of the generated sentences, the Seq2seq-attention model [2] is proposed for dynamically attending on the key information of the input post during decoding. In this paper, our approach is mainly based on *Seq2seq-attention* models for response generation and therefore we will first illustrate this basic model in principle.

The Seq2seq-attention model is typically a deep RNN-based architecture with an encoder and a decoder. The encoder takes the given post sequence $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$ (T is the length of the post) as inputs, and maps them into hidden representations $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)$. The decoder then decodes them to generate a possibly variably-sized word sequence, *i.e.,* $\boldsymbol{y} = \{y_1, y_2, \cdots, y_{T'}\}$, where $T^{'}$ is the length of the output sequence, and it may differ from T.

In more detail, the context representation $\boldsymbol{c}_t$ of the post sequence $\boldsymbol{x}$ is computed by parameterizing the encoder hidden vector $\boldsymbol{h}$ with different attention scores [2], that is,

$$\boldsymbol{c}_t = \sum_{j=1}^{T} \boldsymbol{\alpha}(s_{t-1}, \boldsymbol{h}_j) \cdot \boldsymbol{h}_j, \tag{1}$$

where $\boldsymbol{\alpha}(.,.)$ is a coefficient estimated by each encoder token's relevance to the predicting y_t. The decoder iteratively updates its state $\boldsymbol{s}_t$ using previously the generated word $\boldsymbol{y}_{t-1}$, namely,

$$\boldsymbol{s}_t = f(\boldsymbol{s}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{c}_t), \qquad t = 1, 2, \cdots, T^{'}, \tag{2}$$

where f is a non-linear transformation of RNN cells (*e.g.,* LSTM [9] or GRU [5]).

Then, the probability of generating the t-th token y_t conditioned on the input sequence $\boldsymbol{x}$ and the previous predicted word sequence $y_{1:t-1}$ is computed by

$$\Pr(y_t | y_{1:t-1}, \boldsymbol{x}) = g(\boldsymbol{y}_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t), \tag{3}$$

where $g(.)$ is a function (*e.g.,* softmax) to produce valid probability distribution for sampling the next word of the output sequence.

3.2 Problem Definition and Overview

In this paper, our task is to perceive the emotion involved in the input post and incorporate it into the generation process for *automatically* producing both *semantically reasonable* and *emotionally appropriate* response. Hence, our conversation generation problem is defined as, given an input post $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$ with its emotion $\boldsymbol{e_p}$, for a dialogue system, generate a corresponding response sequence $\boldsymbol{y} = \{y_1, y_2, \cdots, y_{T'}\}$ with proper emotion $\boldsymbol{e_r}$.

To address the problem, we propose an emotion-aware chat machine (EACM), which primarily consists of two subcomponents, the emotion selector and the response generator. More concretely, the emotion selector is in charge of the emotion selecting process, yielding an emotion distribution over $\mathcal{E}_c$ for the to-be-generated response based on the input post and its emotion:

$$\boldsymbol{e_r^*} \leftarrow \arg\max_{\boldsymbol{e_r} \in \mathcal{E}_c} \Pr(\boldsymbol{e_r} | \boldsymbol{x}, \boldsymbol{e_p}), \tag{4}$$

where $\mathcal{E}_c$ is the vector space of the emotions. Subsequently, the generator generates a corresponding response to the given input based on the obtained emotion $\boldsymbol{e_r^*}$ and the input post $\boldsymbol{x}$, namely,

$$\boldsymbol{y_t^*} \leftarrow \arg\max \Pr(\boldsymbol{y_t} | \boldsymbol{y}_{1:t-1}, \boldsymbol{x}, \boldsymbol{e_r^*}). \tag{5}$$

Remark. Actually, previous approaches usually assume that the emotion of the generated response is derived from a unique emotion category[1] (denoted by the one-hot vectors in $\mathcal{E}_c$). However, human emotions are intuitively more delicate, and thus we argue that the emotion of each response may not be limited in a single emotion category. To this end, we assume that the emotion probability distribution is over the entire vector space $\mathcal{E}_c$.

3.3 Self-attention Enhanced Emotion Selector

3.3.1 Self-attention Based Encoding Network. Our information encoding network consists of two parts: an emotion encoder and a semantics encoder. We leverage these two encoders by explicitly extracting emotional information and semantic information, and then feed them into a fusion prediction network. Specifically, the emotion encoder is implemented using a *GRU* network to extract information from post sequence $\boldsymbol{x} = (x_1, x_2, \cdots, x_T)$, and map them into hidden representations $\boldsymbol{h_e} = (h_e^1, h_e^2, \cdots, h_e^T)$ using the following formula:

$$\boldsymbol{h}_e^t = \mathrm{GRU}(\boldsymbol{h}_e^{t-1}, x_t), \tag{6}$$

where $\boldsymbol{h}_e^t$ denotes the hidden state of post's emotion information at time step t.

[1] Here we follow the work [46], where the emotion categories are {Angry, Disgust, Happy, Like, Sad, Other}.

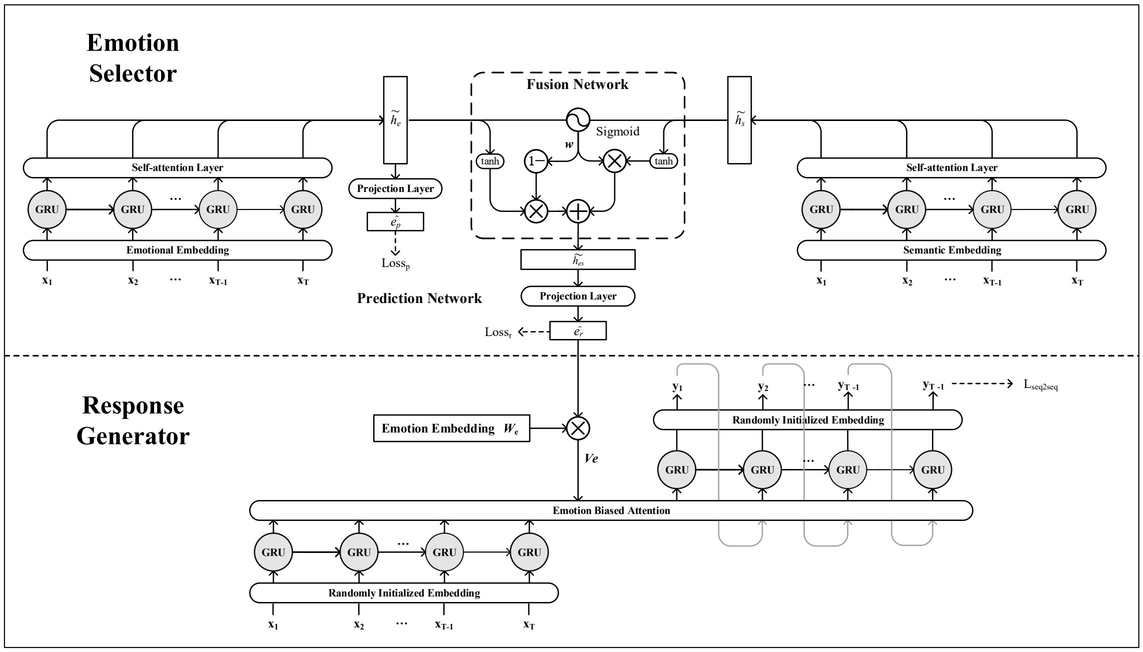


Figure 2: Overview of EACM.

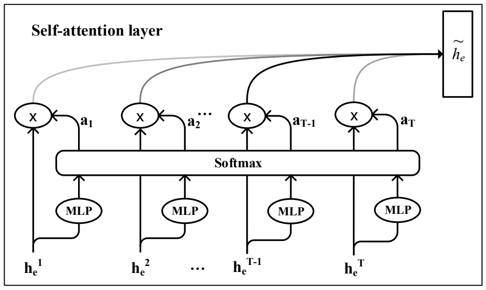


Figure 3: Details of self-attention layer for emotion encoder.

To enhance the representation power of the hidden states, we utilize *self-attention* mechanism [18] to enable the encoders to attend to emotion-rich words in the post, and then obtain the emotional hidden state $\widetilde{\boldsymbol{h_e}}$ by calculating:

$$\widetilde{\boldsymbol{h_e}} = \sum_{i=1}^{T} a_i \boldsymbol{h}_e^i, \tag{7}$$

where a_i is the weight of hidden state $\boldsymbol{h_e^i}$, which is calculated by feeding $\boldsymbol{h_e^i}$ into a multi-layer perceptron with a softmax layer to ensure that all the weights sum up to 1:

$$a_i = softmax(\boldsymbol{V_a} tanh(\boldsymbol{W_a}(\boldsymbol{h}_e^i)^\top)). \tag{8}$$

The network focuses on emotional information by imposing a cross entropy loss on the top of the emotion hidden state $\widetilde{\boldsymbol{h_e}}$, that is, passing the emotion hidden state through a linear layer and a sigmoid layer to project it into an emotion distribution over $\mathcal{E}_c$, and then calculating the cross entropy loss as follows,

$$\hat{\boldsymbol{e_p}} = \sigma(\boldsymbol{W_e}\widetilde{\boldsymbol{h_e}} + b), \tag{9}$$

$$\mathcal{L}_p = -\boldsymbol{e_p}\log(\hat{\boldsymbol{e_p}}), \tag{10}$$

where $\boldsymbol{e_p}$ is a multi-hot representation of post's emotion vector since it may contain various emotions and $\mathcal{L}_p$ is the loss function.

However, simply mapping $\hat{\boldsymbol{e_p}}$ to the response emotion category $\boldsymbol{e_r}$ is insufficient to model the emotion interaction process between partners, as we cannot choose emotion category only based on post's emotion category under some circumstances. In fact, some posts expressing negative feelings like sad are inappropriate to be replied with the same emotion, such as "It's a pity you can't come with us" or "I'm so sad that you broke my heart". Therefore, we not only consider the post's emotion information, but also take into account its semantic meaning by combing another GRU network (*i.e.,* semantics encoder) to encode post's semantic information for generation. Similarly, we get a weighted summation of the hidden states represented as $\widetilde{\boldsymbol{h_s}}$.

3.3.2 *Emotional and Semantic Word Embeddings.*

In order to force the emotion selector to focus on different aspects of auxiliary information of the given post, we apply the emotion embedding for the emotional encoder and the semantic embedding for the semantic encoder, respectively. In particular, we make use of sentiment specific word embedding (*SSWE*) [37] and *word2vec* embedding [22] in our model. More concretely, *SSWE* encodes sentiment information in the continuous representation of words, mapping words with the same sentiment to the neighbor word vectors, which is used in the emotion encoder to promote the ability of perceiving emotional information in post's utterance. Simultaneously, *word2vec* is used to extract semantic information from the post, and the two embeddings work interactively to guarantee the efficacy of the encoding network.

3.3.3 Fusion and Prediction Network. To construct the response emotion category e_r, we consider to use a fusion network to balance the contributions derived from different types of information, and employ a prediction network to select the response emotion categories based on such mixed information. Then we concatenate the obtained $\widetilde{\boldsymbol{h}_s}$ and $\widetilde{\boldsymbol{h}_e}$ and feed it into a sigmoid layer to yield a trade-off weight:

$$w = \sigma([\widetilde{\boldsymbol{h}_s}; \widetilde{\boldsymbol{h}_e}]), \tag{11}$$

$$\widetilde{\boldsymbol{h}'_e} = tanh(\widetilde{\boldsymbol{h}_e}), \tag{12}$$

$$\widetilde{\boldsymbol{h}'_s} = tanh(\widetilde{\boldsymbol{h}_s}). \tag{13}$$

The final representation is a weighted sum of the semantic hidden state and the emotional hidden state:

$$\widetilde{\boldsymbol{h}_{es}} = w \otimes \widetilde{\boldsymbol{h}'_s} + (1 - w) \otimes \widetilde{\boldsymbol{h}'_e}, \tag{14}$$

where $\otimes$ indicates element-wise multiplication. The final representation is fed into a prediction network to produce an emotion vector for generation, which is passed through MLP and then mapped into a probability distribution over the emotion categories:

$$\hat{\boldsymbol{e}_r} = \sigma(\boldsymbol{W_r} \widetilde{\boldsymbol{h}_{es}} + b), \tag{15}$$

$$\mathcal{L}_r = -\boldsymbol{e}_r \log(\hat{\boldsymbol{e}_r}), \tag{16}$$

where $\boldsymbol{e}_r$ is the multi-hot representation of the response emotion vector. $\hat{\boldsymbol{e}_r}$ is the final response emotion vector generated through the proposed emotion selector, which is then passed to the generator for emotional response generation. Intuitively, the emotion selector can adaptively determine the appropriate emotion in the emotion selection process for emotional response generation, by taking into account both the post's semantic information and emotional information.

3.4 Emotion-Biased Response Generator

To construct the generator, we consider to use an emotion-enhanced seq2seq model that is capable of balancing the emotional part with the semantic part and generate intelligible responses.

Thereby, we first generate the response emotion embedding $\boldsymbol{V}_e$ by multiplying $\hat{\boldsymbol{e}_r}$ with a randomly initialized matrix:

$$\boldsymbol{V}_e = \boldsymbol{W}_e \hat{\boldsymbol{e}_r}, \tag{17}$$

where $\boldsymbol{W}_e$ is the emotion embedding matrix, which is the latent emotional factors, *i.e.,* the high-level abstraction of emotion expressions by following Plutchik's assumptions [26]. As mentioned, a one-hot emotion embedding is inappropriate and thus here we do not use a softmax on $\hat{\boldsymbol{e}_r}$ to only pick an optimal emotion category for generation. As such, we call it as *soft-emotion injection* procedure, which is used to model the diversity of emotions.

By following the work [40], we use a new encoder to encode $\boldsymbol{x}$ for obtaining a sequence of hidden states $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)$ through a RNN network, and then generate the context vector $\boldsymbol{c}_t$ for decoding the current hidden state $\boldsymbol{s}_t$, via applying attention mechanism to re-assign an attended weight to each encoder hidden state $\boldsymbol{h}_i$.

Training #	**Posts**		219,162
	Responses	***No Emotion***	1,586,065
		Single Emotion	2,792,339
		Dual Emotion	53,545
Validation #	1,000		
Testing #	1,000		

Table 1: Details of ESTC dataset.

$$\begin{aligned} u_i^t &= \boldsymbol{v}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_i + \boldsymbol{W}_2 \boldsymbol{s}_t), & (18)\\ a_i^t &= \text{softmax}(u_i^t), & (19)\\ \boldsymbol{c}_t &= \sum_{i=1}^{T} a_i^t \boldsymbol{h}_i. & (20) \end{aligned}$$

At each time step t, the context vector encoded with attention mechanism enable our model to proactively search for salient information which is important for decoding over a long sentence. However, it neglects the emotion ($\boldsymbol{V}_e$) derived from the response during generation, and thus we propose an emotion-biased attention mechanism to rewritten Eq.(18),

$$u_i^t = \boldsymbol{v}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_i + \boldsymbol{W}_2 \boldsymbol{s}_t + \boldsymbol{W}_3 \boldsymbol{V}_e). \tag{21}$$

The context vector $\boldsymbol{c}_t$ is concatenated with $\boldsymbol{s}_t$ and forms a new hidden state $\boldsymbol{s}'_t$:

$$\boldsymbol{s}'_t = \boldsymbol{W}_4[\boldsymbol{s}_t; \boldsymbol{c}_t], \tag{22}$$

from which we make the prediction for each word; $\boldsymbol{s}_t$ is obtained by changing Eq. (2) into:

$$\boldsymbol{s}_t = \text{GRU}(\boldsymbol{s}'_{t-1}, [\boldsymbol{y}_{t-1}; \boldsymbol{V}_e]), \tag{23}$$

which fulfills the task of injecting emotion information while generating responses. To be consistent with previous conversation generation approaches, here we consider to use *cross entropy* to be the loss function, which is defined by

$$\begin{aligned} \mathcal{L}_{seq2seq}(\theta) &= -logP(\boldsymbol{y}|\boldsymbol{x}) & (24)\\ &= -\sum_{t=1}^{T'} logP(y_t \mid y_1, y_2, \cdots, y_{t-1}, \boldsymbol{c}_t, \boldsymbol{V}_e), & (25) \end{aligned}$$

where θ denotes the parameters.

3.5 Loss Function

The loss function of our model is a weighted summation of the semantic loss and sentiment loss:

$$\mathcal{L}_{EACM}(\theta) = \alpha \mathcal{L}_e + (1 - \alpha) \mathcal{L}_{seq2seq}, \tag{26}$$

where α is a balance factor, and $\mathcal{L}_e$ denotes the *emotional* loss, namely,

$$\mathcal{L}_e = \mathcal{L}_p + \mathcal{L}_r. \tag{27}$$

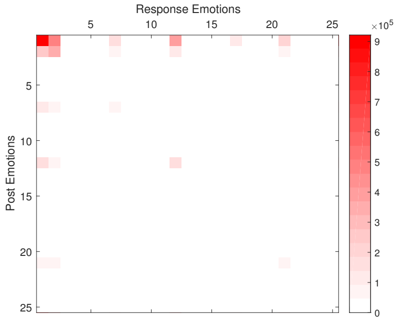


Figure 4: An illustration of emotion interaction pattens (EIPs). Numerical values on each axis represent different emotion categories. The darker each grid is, the more pairs (e_p,e_r) appear in such grid.

4 EXPERIMENTAL RESULTS

4.1 Dataset

We conduct our experiments on a public dataset, *i.e.,* Emotional Short-Text Conversation (ESTC) derived from STC [33], to evaluate our experimental results. In particular, we follow the work [46] to train an emotion classifier for assigning emotional labels to the sentences in the dataset.

ESTC, which contains over four million real-world conversations obtained from Chinese micro-blogging, *i.e.,* Weibo. The raw dataset contains $4,433,949$ post-comment pairs, from which $1,000$ pairs are extracted for validation and another $1,000$ pairs are used for testing. Details of this dataset is illustrated in Table 1.

Preprocessing. As the raw dataset (STC) does not have emotion labels, so we train an emotion classifier based on BERT [6] model over two different datasets, *i.e.,* NLPCC 2013[2] and NLPCC 2014[3] emotion classification datasets by following [46], which contain $29,417$ manually annotated data in total, and the best performance(accuracy) of 0.7257 is achieved at the $19,635$ step. Specifically, each sentence is marked with two labels, namely, a primary label and a secondary one. We preprocess the labels over the mentioned emotion categories, *i.e.,* (*like, disgust, sad, angry, happy, other*), note that here "*other*" indicates no any emotion information, and rare emotion categories like *fear* are removed. In particular, unlike [46] using solely one label for classification, we consider both of the emotion labels and thus regard it as a multi-label classification task. Under such circumstance, there are three cases appearing in the label sets, *i.e.,* no emotion labeled with (other, other), one emotion labeled with (emo1, other), as well as two emotions labeled with (emo1, emo2), respectively.

To evaluate the emotion perception ability of different approaches over the emotion categories, we build an emotion-rich dialogue set for a fair empirical comparison. Specifically, we randomly chose $1,000$ pairs whose primary emotion label is among the (*like, disgust, sad, angry or happy*) categories, with 200 pairs for each emotion, respectively. In addition, we also present an in-depth analysis the *emotion interaction pattern* (EIP) over conversations, in which each EIP is defined as a (e_p,e_r) pair for each conversation to reflect the transition pattern from the post's emotion to the response's emotion. Figure 4 shows a heatmap to depict the number of EIPs appearing in the dataset, and each row (*or* column) indicates the post's emotion $\boldsymbol{e_p}$ (*or* the response's emotion $\boldsymbol{e_r}$). From the figure we can observe that: (1) The darker each grid is, the more pairs ($\boldsymbol{e_p}$,$\boldsymbol{e_r}$) appearing in such grid; (2) The heat map is sparse since the EIPs of some post-response emotion pairs are overly rare to appear in our dataset.

Moreover, we also leverage "Weibo Sentiment Dataset" provided by Shujutang[4] to train the sentiment-specific word embeddings (SSWE), which consists of two million Weibo sentences with sentiment labels, and we remove some extra domain-specific punctuations like "@user" and "URLs".

4.2 Evaluation Metric

As reported in [19], BLEU might be improper to evaluate the conversation generation problem, as it correlates weakly with human judgements of the response quality, similar situations for METEOR [3] and ROUGE [17]. Besides, there still exists a challenge of automatically evaluating the generation model from emotion perspective. As a result, in this paper we adopt *distinct-1* and *distinct-2* by follow the work [14] to be as the metrics for evaluating the diversity of the generated responses, which measures the degree of diversity by computing the number of distinct uni-grams and bi-grams in the generated responses, and can indirectly reflect the degree of emotion diversity, as the generated sentence containing diverse emotions is more likely to have more abundant words in principle. In addition, we also carry out a manual evaluation for evaluate the performance of the generated responses at *emotional*-level and *semantic*-level separately with human intuition, and then the *response quality* is calculated by combining such two results at different levels for integrally assessing different models.

Automatic Evaluation. As mentioned, we consider to use *distinct-1* and *distinct-2* [14] to be as our automatic evaluation metric. *Distinct-n* is defined as the number of distinct n-grams in generated responses. The value is scaled by total number of generated tokens.

Human Evaluation. We randomly sampled 200 posts from the *test* set, and then aggregate the corresponding responses returned by each evaluated method, then three graduate students (whose research areas are not in text processing area) are invited for labeling. Each generated response is labeled from two different aspects, *i.e., emotion* and *semantics*. Specifically, from emotion perspective, each generated response is labeled with (score 0) if its emotion is apparently inappropriate (namely evident emotion collision,*e.g.,* angry-happy) to the given post, and (score 1) otherwise. From semantic perspective, we evaluate the generated results using the scoring metrics as follows. Note that if conflicts happen, the third annotator determines the final result.

- 1: If the generated sentence can be obversely considered as a appropriate response to the input post;
- 0: If the generated sentence is hard-to-perceive or has little relevance to the given post.

To conduct an integral assessment of the models at both *emotional*-level and *semantic*-level, we measure the *response quality* by

[2] http://tcci.ccf.org.cn/conference/2013/

[3] http://tcci.ccf.org.cn/conference/2014/

[4] http://www.datatang.com/data/45439

using the formula as follows,

$$Q_{response} = S_{sentiment} \wedge S_{semantics}, \tag{28}$$

where $Q_{response}$ reflects the *response quality* and $S_{sentiment}$, $S_{semantics}$ denote the sentiment score and semantic score, respectively, which is used to means the *response quality* of each case is equal to 1 if and only if both of its sentiment score and semantic score are scored as 1.

4.3 Baselines

We compare our model with the following baselines.

Seq2seq [36], the traditional *Seq2seq* model is adopted as one of our baselines.

ECM [46], as mentioned, *ECM* model is improper to directly be as the baseline since it cannot automatically select an appropriate emotion label to the respond. Thereby, we manually designate a most frequent response emotion to *ECM* for fairness comparison. Specifically, we train a post emotion classifier to automatically detect post's emotion, and then choose the corresponding response emotion category using the most frequent response's emotion to the detected post's emotion over EIPs.

Seq2seq-emb [11, 46], *Seq2seq* with emotion embedding (*Seq2seq-emb*) is also adopted in the same manner. This model encode the emotion category into an embedding vector, and then utilize it as an extra emotion input when decoding.

Intuitively, the generated responses from *ECM* and *Seq2seq-emb* can be viewed as the indication of the performance of simply incorporating the *EIP*s for modeling the emotional interactions among the conversation pairs.

4.4 Implementation Details

For all approaches, each encoder and decoder with 2-layers GRU cells containing 256 hidden units, and all of the parameters are not shared between such two different layers. The vocabulary size is set as 40,000, and the OOV (out-of-vocabulary) words are replaced with a special token *UNK*. The size of word embeddings is 200, which are randomly initialized. The emotion embedding is a 6 × 200-dimensional matrix (if used). The parameters of *imemory* and *ememory* in *ECM* are the same as the settings in [46]. We use stochastic gradient descent (SGD) with mini-batch for optimization when training, and the batch size and the learning rate are set as 128 and 0.5, respectively. The greedy search algorithm is adopted for each approach to generate responses. Additionally, for speeding up the training process, we leverage the well-trained *Seq2seq* model to initialize other methods.

The parameters for our proposed method are empirically set as follows: *SSWE* is trained by following the parameter settings in [37]. The length of hidden layer is set at 20, and we use AdaGrad [7] to update the trainable parameters and the learning rate is set as 0.1. The size of emotion embedding and word embedding are both set at 200. In particular, the *Word2vec* embedding is used based on *Tencent AI Lab Embedding*[5], which is pre-trained over 8 million high-quality Chinese words and phrases by using directional skip-gram method [34]. We use `jieba`[6] for word segmentation during the evaluation process.

4.5 Results and Discussion

In this section, we evaluate the effectiveness of generating emotional responses by our approach as comparison to the baseline methods.

Automatic Evaluation. From Table 2, we can observe that: (i) *ECM* performs worse than *Seq2seq*, the reason might be the emotion selection process is based on two-stage process, *i.e.,* post emotion detection process and response emotion selection process, which would significantly reduce the diversity and quality of emotion response generation due to the errors of emotion classification and the transition pattern modeling procedure. In particular, the emotion category *(other, other)* is more likely to be chosen than other emotion categories. We will present an in-depth analysis in Section 4.6. and (ii) Our proposed *EACM* consistently outperforms all of the baselines in terms of *distinct-1* and *distinct-2*. The results demonstrate that our emotion selection process is really effective in enhancing the ability of generating more diverse words.

Models	Distinct-1	Distinct-2
Seq2seq	0.0608	0.2104
Seq2seq-emb	0.0628	0.2370
ECM	0.0551	0.2022
EACM	**0.0745**	**0.2749**

Table 2: Automatic evaluation: distinct-1 and distinct-2.

Human Evaluation. From Table 3, we can observe that: *Seq2seq-emb* provides the worst performance as expected, this is because the generation process apparently interrupted would significantly reduce the accuracy and quality of generating responses and thus generate some hard-to-perceive sentences; *ECM* achieves a relatively better result, as it is good at modeling the emotion dynamics when decoding (*i.e., internal memory*) and assigning different generation probabilities to emotion/generic words for explicitly modeling emotion expressions (*i.e., external memory*); *Seq2seq-emb* results in a remarkable improvement over *Seq2seq-emb* in terms of *semantic score*, but it performs poorly when comparing *sentiment score*, which demonstrates the effectiveness of *emotion injection*, however the explicit two-stage procedure might reduce the smoothness of generated responses with low *semantic score.* Our proposed model *EACM* consistently outperforms all baseline methods, and the improvements are statistically significant on all metrics. For example, *EACM* outperforms *ECM* by 16.9%, 1.72% and 25.81% in terms of *semantic score*, *sentiment score* and *response quality*, respectively. The reason might due to the fact that *EACM* is capable of simultaneously encoding the semantics and the emotions in a post for generating appropriately expressed emotional response within a unified end-to-end neural architecture, which are benefit for alleviating the *hard emotion injection* problem (as compared to *soft emotion injection*).

[5] https://ai.tencent.com/ailab/nlp/embedding.html

[6] https://github.com/fxsjy/jieba

Models	$\bar{S}_{\text{semantics}}$	$\bar{S}_{\text{sentiment}}$	$\bar{Q}_{\text{response}}$
Seq2seq	0.390	0.815	0.360
Seq2seq-emb	0.280	0.795	0.250
ECM	0.355	0.870	0.310
EACM	**0.415**	**0.885**	**0.390**

Table 3: Human evaluation: averaged semantic score, sentiment score and response quality.

Method(%)	1-1	1-0	0-1	0-0
Seq2seq	36	45.5	3	15.5
Seq2seq-emb	25	54.5	3	17.5
ECM	31	**56**	4.5	**8.5**
EACM	**39**	49.5	**2.5**	9

Table 4: The percentage of the sentiment-semantic score given by human evaluation.

The percentage of the sentiment-semantics scores under the human evaluation is shown in Table 4. For *ECM*, the percentage of (0-0) degrades while the percentage of (1-0) increases as opposed to *Seq2seq*, which suggests that the effectivenss of EIP, *i.e.,* the most frequent response emotions have low probability to result in emotional conflicts, In addition, the percentage of (1-1) degrades while the percentage of (1-0) increases, which reflects that directly using emotion classifier to model emotion interaction process is insufficient. In comparison, *EACM* reduces the percentage of generating responses with wrong emotion and correct semantics (*i.e.,* 0-1) while increase the percentage of (1-1) correspondingly, which demonstrate that *EACM* is capable of successfully model the emotion interaction pattern among human conversations and meanwhile guarantee semantic correctness.

4.6 Case Study

In this section, we present an in-depth analysis of emotion-aware response generation results of our proposed approach. We select 3 samples (with the input posts and their corresponding responses) generated by different methods are shown in Figure 5, as can be seen:

Case 1. The results of *EACM* and *ECM* with (like, other) are similar and correct. However, responses given by *ECM* with other emotions are improper in semantics. Similar scenario for *Seq2seq-emb* with (Happy, other). However, the most frequent emotion to (Sad, other) is never be (Happy, other) over EIPs, and thus EIPs would fail to achieve the task.

Case 2. *EACM* can generate the relevant emotional response *i.e.,* (Other, other) for the post with (Other, other). However, all of *ECM* with different emotions seems improper for response (especially for *ECM* with (Happy,other)), which reflects directly using a designated emotion for generation might be a unreasonable way for modeling the emotion interaction pattern. In addition, *Seq2seq* cannot detect the emotion and thus generate a irrelevant response. As most of conversations belong to (other, other), and thus the diversity of such emotion category is quite complicated and hard for training.

Case 3. The emotion in the case is (Angry, other), however the responses provided by *ECM* with (Angry, other) is obviously incorrect in semantics, which demonstrate that simply using post's emotion is inappropriate for response generation. As compared, *EACM* generates the correct emotional response with a emotion-specific word (*i.e.,* "Holy crap"), which demonstrates *EACM* is good at controlling sentence emotion for generation.

5 CONCLUSION

In this paper, we propose an emotion-aware chat machine (EACM) to address the emotional response generation problem, which is composed of an emotion selector and a response generator. Specifically, a unified fusion-prediction network with self-attention mechanism is also developed to supervise the emotion selection process for generating a emotion-biased response. Extensive experiments conduct on a public dataset demonstrate the effectiveness of our proposed method as compared to baselines at both *semantic*-level and *emotional*-level, in terms of automatic evaluation and human evaluation metrics.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No.61602197, and in part by Equipment Pre-Research Fund for The 13th Five-year Plan under Grant No.41412050801.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*. 6077–6086.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop*. 65–72.

[4] Zoraida Callejas, David Griol, and Ramón López-Cózar. 2011. Predicting user mental states in spoken dialogue systems. *EJASP* (2011), 6.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* (2018).

[7] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR* 12, 7 (2011), 257–269.

[8] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. *arXiv* (2017).

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

[11] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic Dialogue Generation with Expressed Emotions. In *NAACL*. 49–54.

[12] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. *arXiv* (2014).

[13] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv* (2014).

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

[15] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*.

[16] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv* (2016).

Case	Posts	Seq2seq	Seq2seq-emb	ECM	EACM
1	**奇怪**，怎么没人说我这张照片把我拍得很帅呢 It is **strange** that no one thinks that I look handsome in this photo. (Sad, other)	你的照片是怎么拍的？ How did you take this photo? (Other, other)	我也想拍一张这样的照片 I also want to take a photo like this. (Like, other)	这张照片**拍的好，拍的很好** This photo **looks great ! Really great !** (Like, other)	这张照片**拍的好，拍的好** This photo **looks great ! Really great !** (Like, other)
			我也想知道，怎么回事？ I also want to know. What is happening? (Sad, other)	我也想知道，为什么我的眼里常含泪水，因为我爱你 I also want to know, why my eyes are always filled with tears. Because I love you. (Sad, other)	
			你的照片是谁拍的？ Who took this photo for you? (Disgust, other)	这是什么电影啊？！ Which movie is this? (Disgust, other)	
			我靠，这是什么情况？ Holy crap, what is happening? (Angry, other)	这是哪啊？怎么没拍到 Where is it? Why didn't you photograph it? (Angry, other)	
			哈哈，你也太**自恋**了吧 **Haha**, you are too **narcissistic**. (Happy, other)	哈哈，这照片拍的，太像了 Haha, these photos are too similar. (Happy, other)	
2	恩！我们会**整回来的**！ Sure, we are going to **fight it back!** (Other, other)	这是什么时候的事情啊？ When did it happen? (Other, other)	这是我的最爱的一次 It is my most favorable time. (Like, other)	不错不错，继续加油 Well done, fighting! (Like, other)	这是我们的地盘，我们**做主！** This is our site, we are the **master!** (Other, other)
			我也想知道这是什么情况 I also want to know what is happening? (Sad, other)	我也想知道，为什么要放弃治疗 I also want to know why you gave up your treatment. (Sad, other)	
			这是什么玩意啊？！ What is this ! (Disgust, other)	这是什么啊？真的假的？ What is this? Is this real? (Disgust, other)	
			这是什么情况？！？ What is happening?!? (Angry, other)	我想知道这是什么意思 I also want to know what does this mean. (Angry, other)	
			哈哈，这是我的梦想啊 Haha, this is my dream. (Happy, other)	哈哈，我也要去买一架 Haha, I'm going to buy one. (Happy, other)	
3	而且**居然**没有人管！至今仍风风光光逍遥法外！ And **unexpectedly** nobody is controlling him! He is still at large so far! (Angry, other)	这是什么情况？！？ What is happening?!? (Angry, other)	支持一下感觉挺不错的 It feels good to support it. (Like, other)	支持一下感觉挺不错的 It feels good to support it. (Like, other)	**我晕**，这是什么情况？ **Holy crap**, what's going on there? (Angry, other)
			可惜了，没有外援 It's a pity that there is no foreign aid. (Sad, other)	我也在上海，可惜没机会 I am also in Shanghai, but unfortunately I have no chance. (Sad, other)	
			你也不知道你是谁 You don't know who you are. (Disgust, other)	这是在说谁呢？！ Who is this talking about? ! (Disgust, other)	
			为什么不能给我们私信？ Why can't we give us a private message? (Angry, other)	晕，还没到货呢？！ What? Haven't been delivered yet? ! (Angry, other)	
			哈哈，这是在哪里啊？ Haha, where is this? (Happy, other)	哈哈，这个人很有才！ Haha, this person is very talented! (Happy, other)	

Figure 5: Case Study. 3 **Samples (with the given posts and the corresponding responses) generated by *Seq2seq, Seq2seq-emb, ECM* and *EACM*. Words that express appropriate emotion in responses are highlighted in red, along with their posts' corresponding emotion words.**

[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).

[18] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv* (2017).

[19] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

[21] Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *ISCA tutorial and research workshop*. 11–16.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv* (2013).

[23] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[25] Yehong Peng, Yizhen Fang, Zhiwen Xie, and Guangyou Zhou. 2019. Topic-enhanced emotional conversation generation with attention mechanism. *KBS* 163 (2019), 429–437.

[26] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[27] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.

[28] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *IJHCS* (2005), 231–245.

[29] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv* (2017).

[30] Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *EMNLP*. 583–593.

[31] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*. 3776–3784.

[32] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.

[33] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*.

[34] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *NAACL*. 175–180.

[35] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv* (2015).

[36] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.

[37] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*. 1555–1565.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 5998–6008.

[39] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. In *AAAI*.

[40] Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *NIPS*. 2773–2781.

[41] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

[42] Richard Wallace. 2003. The elements of AIML style. *Alice AI Foundation* (2003).

[43] Bruce Wilcox. 2011. Beyond Façade: Pattern matching for natural language applications. *GamaSutra. com* (2011).

[44] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation.. In *AAAI*. 3351–3357.

[45] Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Tailored Sequence to Sequence Models to Different Conversation Scenarios. In *ACL*. 1479–1488.

[46] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

[47] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention.. In *IJCAI*. 4623–4629.

[48] Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *arXiv* (2017).