Article

A Dynamic Emotional Session Generation Model Based on Seq2Seq and a Dictionary-Based Attention Mechanism

Qiangqiang Guo, Zhenfang Zhu , Qiang Lu , Dianyuan Zhang and Wenqing Wu

School of Information Science and Electrical Engineering, Shandong Jiao Tong University, Jinan 250357, China; gqqq777@163.com (Q.G.); lqsdjtu@163.com (Q.L.); csdzdy@163.com (D.Z.); winchywwq@163.com (W.W.)

* Correspondence: zhuzf@sdjtu.edu.cn; Tel.: +86-137-9310-0702

Received: 14 February 2020; Accepted: 8 March 2020; Published: 13 March 2020

Abstract: With the development of deep learning, the method of large-scale dialogue generation based on deep learning has received extensive attention. The current research has aimed to solve the problem of the quality of generated dialogue content, but has failed to fully consider the emotional factors of generated dialogue content. In order to solve the problem of emotional response in the open domain dialogue system, we proposed a dynamic emotional session generation model (DESG). On the basis of the Seq2Seq (sequence-to-sequence) framework, the model abbreviation incorporates a dictionary-based attention mechanism that encourages the substitution of words in response with synonyms in emotion dictionaries. Meanwhile, in order to improve the model, internal emotion regulator and emotion classifier mechanisms are introduced in order to build a large-scale emotion-session generation model. Experimental results show that our DESG model can not only produce an appropriate output sequence in terms of content (related grammar) for a given post and emotion category, but can also express the expected emotional response explicitly or implicitly.

Keywords: dialogue generation; emotional response; attention mechanism based on dictionary

1. Introduction

Designing human–machine dialogue systems (HDSs) and enabling computers to interact with humans through human languages is an important and challenging task in the field of artificial intelligence. In recent years, with the rapid growth of social data on the internet, data-driven open-domain dialogue systems have gradually become the focus of attention in the academic community, and HDSs have gradually changed from a service role to one of emotional partner [1]. With the development of deep learning, large-scale dialogue generation methods have been studied. These research efforts are aimed at solving the quality of generated dialogue contents, but failed to fully consider the emotional factors in it.

Emotional dialogue systems are a subdivision of HDSs that aim to enable dialogue systems through the ability to perceive and express emotions. As we all know, emotion is an important factor in allowing people to understand and express information. The Psychology Dictionary [2] holds that “Emotion is the attitude and experience of people reflecting whether the objective things meet their needs”. Chat dialogue systems with emotions can not only enhance user satisfaction [3], but also boost positive interactions [4] and prevent “misunderstandings” caused by emotional problems. In this paper, we propose a dynamic emotional session generation model (DESG) based on the Seq2Seq (sequence-to-sequence) model and a dictionary-based attention mechanism in response to the problem of emotional response in open-domain dialogue systems.

The main contributions of this study can be summarized as follow:

- It proposes to address the emotion factor in large-scale conversation generation.
- It proposes an end-to-end framework (called a DESG) to incorporate emotional influence into large-scale conversation generation via two novel mechanisms: an internal emotion regulator, and an emotion classifier.
- Experimental results with both automatic and human evaluations show that for a given post and an emotion category, our DESG can express the desired emotion explicitly (if possible) or implicitly (if necessary), while successfully generating meaningful responses with a coherent structure.

This paper is organized as follows. After introducing related works in Section 2, we elaborate on our proposed methods in Section 3, then perform an experimental evaluation in Section 4. Finally, we summarize our work and provide an outlook for future work in Section 5.

2. Related Work

Dialogue systems are an important research direction within artificial intelligence [5]. In recent years, data-driven open-domain chat dialogue systems [6] have gradually become a focus in the academic community as a response to the rapid growth of internet social data. Inspired by phrase-based statistical machine translation technology [7], Shang et al. [8] first adopted a sequence model framework in 2015 and brought about the short text dialogue generator NRM (Neural Responding Machine), which was based on the corpus of Sina Weibo. Research shows that NRM can generate grammatically correct and appropriate responses to more than 75% of input text. In 2016, Mei et al. [9] added a dynamic attention mechanism based on the RNN-based neural language model to achieve a single-round dialogue-generation model, which improved the model's effect. With the development of deep learning, large-scale dialogue generation methods have become increasingly studied. These research studies have aimed to improve the quality of generated dialogue content, but have failed to fully consider the emotional factors.

Emotional dialogue systems are a subdivision of HDSs that can perceive and express emotions. Inspired by the sequence model framework in deep learning, emotion-based dialogue generation systems have in recent years received greater attention. As an important part of human intelligence, EQ (emotional quotient) is defined as the ability to perceive, integrate, understand and regulate emotions [10]. We believe that a dialogue system with a certain "EQ" (emotion) is a much "smarter" dialogue system.

Recently, some attention has been given to generating responses with the specific properties like sentiments, tenses, or emotions. Hu et al. (2017) [11] proposed a text-generating model based on variational autoencoders (VAEs) in order to produce sentences that present a given sentiment or tense. Ghosh et al. (2017) [12] proposed an RNN-based language model to generate emotional sentences conditioned on their affect categories. Zhou and Wang (2018) [13] collected a large corpus of Twitter conversations including emojis (ideograms and smileys used in electronic messages) first, then used emojis to express emotions in the generated texts by trying several variants of conditional VAEs.

Li et al. (2016) [14] proposed the Seq2Seq (sequence-to-sequence) model, which proposes a general end-to-end sequence-learning method. In this method, the encoder maps the input sequence to a fixed-dimensional vector that is regarded as the semantic expression of the input sequence. The decoder generates the output sequence according to the semantic expression of the input sequence. Serban et al. (2016) [15] proposed the EmoEmb (emotion-embedding dialogue system) model, which represents each emotion category as a vector based on the Seq2Seq framework, inputting them into the decoder at each time step. Asghar et al. (2017) [16] used a VAD (Valence, Arousal, Dominance) dictionary in which the neural network was allowed to choose the emotions to be generated (by maximizing or minimizing emotional imbalances), meaning that their systems could not output sequences containing different emotions for the same input sequence, nor could they produce output sequences containing specific emotions.

Li et al. (2019) [17] proposed an adversarial empathetic dialogue system (EmpGAN) to evoke more emotion perceptivity during dialogue generation, with two mechanisms proposed to improve its performance. These two interactive discriminators utilize user feedback as an additional context for the interactions between dialogue context and generated response so as to optimize the long-term goal of empathetic conversation generation. Wu et al. (2019) [18] proposed a new dual-decoder architecture to generate different sentiment responses with respect to a single post. They connected two decoders to one encoder, simulating the manner in which different people reading the same post might have different thoughts and then generate different responses. Chen et al. (2019) [19] proposed methods to generate emotionally aware responses and mimic human conversations. These models were able to determine the relevant emotions by which to reply to a user post and generate responses appropriately. Zhou et al. (2018) [20] proposed the ECM (Emotional Chatting Machine) model, which is a large-scale generative sequence-to-sequence model that can respond to users emotionally. Based on the traditional sequence-to-sequence model, this method uses a static emotion vector expression, a dynamic emotional state memory network, and an external memory of emotional words, so that the ECM can output responses corresponding to emotions according to the specified emotion classification. However, the ECM must specify the emotion categories to be generated, which means that an external decision-maker is required. In order to solve the problem of needing an external decision-maker for the ECM model to work, Song et al. (2019) [21] proposed the EmoDS (emotional dialogue system) model, which increases the probability of emotional vocabulary at the correct time-step to seamlessly “insert” text. At the same time, by guiding the response generation process through the emotion classifier, it is ensured that the generated text contains specific emotions. Inspired by psychologists, we found that emotional responses in sentences are dynamic [22]. However, the EmoDS model does not fully consider the need for dynamic emotions in sentences.

Song et al. [21] found that there are at least two ways (explicit and implicit) to incorporate emotional information into a sentence. One is to describe the emotional state (such as “anger”, “disgust”, “satisfaction”, “happiness”, “sadness”, etc.) by explicitly using strong emotional words; the other is to increase the intensity of the emotional experience, not by using words from an emotional vocabulary, but by implicitly combining neutral vocabulary in different ways to express emotions.

As shown in Table 1, for each post, one emotional response is listed below. The emotional words associated with strong feelings are highlighted in a bold blue font.

Table 1. Two (explicit and implicit) varieties of emotional performance.

Post	Explicit	Implicit
I bought a beautiful dress yesterday!	Wearing a beautiful dress makes me happy!	Wow, you must feel walking on air!
The weather today is really nice.	I like sunny days.	I’m in a better mood today.

We propose a dynamic emotional session generation model (DESG) based on the Seq2Seq framework and a dictionary-based attention mechanism [23]. At the same time, in order to improve the model, two mechanisms are introduced: (1) An internal emotion regulator that is able to perform high-level abstract emotional performances and capture changes in implicit emotional states via embedded emotion categories. (2) An emotion classifier, which is used to guide the response generation process and ensure that the generated output sequence contains the appropriate emotion category. The DESG not only uses a variety of sorting algorithms to improve the diversity of output sequences, but also possess a relatively “accurate” emotion dictionary to express human experience and perception. The experimental results show that for a given input sequence and emotion category, our DESG model can not only generate a suitable output sequence in terms of content (relevant grammar), but also express the desired emotional response explicitly or implicitly.

3. Model

We assume a question $X = (x_1, x_2, \dots, x_M)$ and emotion category z , then divide the emotion category z into six emotions {anger, disgust, satisfaction, happiness, sadness, neutral} in order to generate a response $Y = (y_1, y_2, \dots, y_N)$ that contains a specific emotion category. We use a semi-supervised approach to build the required sentiment dictionary [21], which is constructed from a corpus of sentences with emotion categories and annotations. In this article, we refer $V = V_g \cup V_e$ to the sentiment dictionary with a general vocabulary V_g and an emotional vocabulary V_e . V_e can be further divided into several subsets, each of which stores words related to the emotion category z . For a given input sequence and emotion category, our DESG model can not only produce a suitable output sequence in terms of content (relevant grammar), but can also express the desired emotional response explicitly or implicitly in terms of emotion. A dynamic emotional session generation model (DESG) based on Seq2Seq and the dictionary-based attention mechanism are shown in Figure 1.

Figure 1. A dynamic emotional session generation (DESG) model based on Seq2Seq and a dictionary-based attention mechanism.

3.1. The Seq2Seq Model Based on a Dictionary-Based Attention Mechanism

The encoder part of the Seq2Seq framework uses the bi-directional LSTM (Bi-LSTM) [24] model. We assume that given an input sequence $X = (x_1, x_2, \dots, x_M)$, an output sequence $Y = (y_1, y_2, \dots, y_N)$. At a certain time node t , the encoder's hidden layer state h_t at the current time (t) is calculated by the hidden layer state h_{t-1} of the encoder at the time $t - 1$ and the current time input x_t , as shown in Equations (1)–(3):

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \quad (2)$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (3)$$

where $\vec{h}_t, \overleftarrow{h}_t$ are the hidden layer states of the encoder t of the forward LSTM and the backward LSTM, respectively; and h_t is the hidden layer state of the encoder at the current moment. $x_t \in R^d$ is the word embedding of x_t , and d is the dimension of word embedding.

In order to make the output sequence contain emotional information, we adopted a dictionary-based attention mechanism [25], which seamlessly “inserts” emotional words into the generated text at the correct time step. Specifically, first assume a given part of the input sequence and the generated output sequence, then find the emotional words in the sequence that are closely related to the words in the emotional dictionary. Calculate the weight of the dynamic influence on the emotional feature vector at the current moment, then seamlessly “insert” the desired emotional words into the output sequence. With this in mind, a unidirectional LSTM model based on a dictionary attention mechanism is adopted in the decoder part of the Seq2Seq framework. The dictionary-based attention mechanism is shown in Figure 2.

Figure 2. The Dictionary-based Attention Mechanism.

At a certain time node t , the decoder updates the output s_t according to the generated output sequence word vector y_{t-1} and the context vector c_t based on the dictionary-based attention mechanism, as shown in Equation (4):

$$s_t = \text{LSTM}([y_{t-1}; c_t]; s_{t-1}) \quad (4)$$

where s_t is the hidden layer state of the decoder, $s_0 = h_M$, and c_t is the embedding of c_t . The context vector, which is based on the dictionary-based attention mechanism, is the weighted sum of the word embedding in V_e^z of the given emotion category z , as shown in Equations (5)–(7):

$$c_t = \sum_k \alpha_{tk} * w_k^z \quad (5)$$

$$\alpha_{tk} = \frac{\exp(ctk)}{\sum_{l=1}^{T_z} \exp(ctl)} \quad (6)$$

$$ctk = \text{Sigmoid}(\alpha^T h_M + \beta^T s_{t-1} + \gamma^T w_k^z) \quad (7)$$

where w_k^z represents the k th word in V_e^z , T_z is the number of words in emotion category z , and α , β , and γ are trainable parameters. We use the attention mechanism model proposed by Luong et al. [26] to calculate the attention weight. For each emotional word in V_e^z , the attention mechanism weight at time node t is determined by three parts: the hidden layer state at the $t - 1$ moment of the decoder, the encoding h_M of the input sequence, and the k th word in the embedding.

Given the partial generated response and the input post, the more relevant an emotional word is, the more influence it will have on the emotion feature vector at the current time step. In this way, such lexicon-based attention leads to a higher probability of the emotional words being relevant to the current context.

To seamlessly “insert” emotional words into the output sequence, we estimate two probability distributions $P_e(y_t = w^e)$, $P_g(y_t = w^g)$, where z is the given emotion category, w^e is all the emotional words in the emotion dictionary V_z^e , and w^g is all the generic words in the general vocabulary list V_g , as shown in Equations (8)–(11):

$$P_e(y_t = w^e) = \text{Softmax}(W_e s_t) \quad (8)$$

$$P_g(y_t = w^g) = \text{Softmax}(W_g s_t) \quad (9)$$

$$\delta t = \text{Sigmoid}(\eta^T s_t) \quad (10)$$

$$y_t \sim P(y_t) = \begin{bmatrix} \delta t P_e(y_t = w^e) \\ (1 - \delta t) P_g(y_t = w^g) \end{bmatrix} \quad (11)$$

$\delta t \in (0, 1)$ controls the generation of emotional words as well as the weighting of common words, while W_e , W_g , and η are trainable parameters. The loss function of each sample is defined according to cross entropy loss, where f is a binary vector with all elements being zero, as shown in Equation (12):

$$LMCE = - \sum_{t=1}^N f_t \log(P(y_t)) \quad (12)$$

3.2. Internal Emotion Regulator

To generate emotional text sequences, we need to fully extract the static and dynamic emotional information features in the dialogue. The emotion categories in the text (e.g., happy, disgusted) are an overview and high-level abstraction of the emotional performance of the text. The most intuitive way to model emotions in a dialogue system is to take the emotion categories of the text sequence as additional inputs and represent each emotion category with a real-valued low-dimensional vector. For each emotion category z , we randomly initialize a vector of emotion category E_z , then learn the vectors of the emotion category through training.

The emotion category E_z , along with word embedding y_{t-1} , and the context vector c_t , are fed into the decoder to update the decoder’s state s_t , occur at a certain time node t , as shown in Equation (13):

$$s_t = \text{LSTM}([y_t - 1; c_t; E_z]; s_{t-1}) \quad (13)$$

The process by which the emotional categories are embedded is rather static [27], the emotion category embedding will not change during the generation process but which may sacrifice grammatical correctness of sentences as argued. Inspired by psychologists, we found that emotional responses in sentences are relatively short-lived and dynamic to changes.

In order to solve the problem of dynamic emotions occurring during the emotion category embedding process, we use an internal emotion regulator module to capture the dynamic emotion information during the decoding process that ensures the dynamic emotion problem is solved. Specifically, we simulate the expression process of emotions as follows: before the decoding process begins, each emotion category is assigned an internal emotional state; as each step moves forward, the emotional state will decline to a certain degree; once the decoding process is completed, the emotional state should decay to zero, which means the emotion has been fully expressed. The internal emotion regulator is shown in Figure 3.

Figure 3. Internal Emotion Regulator.

The read gate G_t^r is calculated at a certain time node t according to the generated output sequence word vectors y_{t-1} and s_{t-1} (the decoder hidden layer state at the moment $t-1$) and the context vector based on the dictionary-based attention mechanism, as shown in Equation (14). The write gate G_t^o is calculated based on the decoder's hidden state vector s_t , as shown in Equation (15).

$$G_t^r = \text{Sigmoid}(\omega_G^r [y_{t-1}; c_t; s_{t-1}]) \quad (14)$$

$$G_t^o = \text{Sigmoid}(\omega_G^o s_t) \quad (15)$$

The read and write gates are used to read from and write to memory respectively. Therefore, the emotional state will be erased by a certain amount G_t^o at each step. In the last step, the internal emotional state decays to zero. The process description is shown below:

$$M_{r,t}^I = G_t^r \otimes M_{z,t}^I \quad (16)$$

$$M_{r,t+1}^I = G_t^o \otimes M_{z,t}^I \quad (17)$$

where r/o means read/write, respectively; I means internal; and $\otimes$ is multiplication by an element. The decoder updates the s_t according to the generated output sequence word vectors y_{t-1} and s_{t-1} (the decoder's hidden layer state at moment $t-1$) and the dictionary-based attention mechanism context vector c_t , as well as the emotional state $M_{r,t}^I$, as shown in Equation (18):

$$s_t = \text{LSTM}([y_{t-1}; c_t; M_{r,t}^I]; s_{t-1}) \quad (18)$$

Define the loss function of each sample by the cross-entropy error, as shown in Equation (19):

$$L(\theta) = \|M_{z,M}^I\| \quad (19)$$

3.3. Emotion Classifier

The feelings can be put into words either by explicitly using strong emotional words associated with a specific category, or by implicitly combining neutral words sequentially in distinct ways. Therefore, we use sentence-level emotion classifiers to guide the generation process, which not only can increase the intensity of the emotional performance, but can also help identify emotional responses that do not contain any emotional words. The emotion classifier as an overall guide for sentiment expression is used only during training. The emotion classifier at the sentence level is shown in the upper right corner of Figure 1 and incorporates the details from Equation (20):

$$Q(\mu|Y) = \text{Softmax}(W * \frac{1}{N} \sum_{t=1}^N y_t) \quad (20)$$

where $W \in R^{K \times d}$ is the weight matrix, and K is the number of the emotion category. Due to the exponential relationship between search space and vocabulary quantity, and because the length of Y is unknown, it is impossible to enumerate all possible sequences. We sample several sequences based on the probability of the generation process so that it is not differentiable.

Furthermore, if we sample several sequences based on the probability of approximating the generation process, it is not differentiable. We use the concept of word embedding to approximate $Q(\mu|Y)$ [24]. Specifically, the word embedding we expect is a weighted sum of the embedding of all possible words at each time step. This is shown in Equation (21):

$$Ewe(t; X, z) = \sum_{yt \in V_g \cup V_e^z} P(yt) * (yt) \quad (21)$$

For each moment t , we enumerate all possible words in the union of V_g and V_e^z . The classification loss for each sample is shown in the Equations (22) and (23), where $P(E)$ is a vector representing the mood distribution required for a sample:

$$LCLA = -P(\mu) \log(Q(\mu|Y)) \quad (22)$$

$$Q(\mu|Y) = \text{Softmax}(W * \frac{1}{N} \sum_{i=1}^N Ewe(j; X, z)) \quad (23)$$

3.4. Loss Function

The training setup is divided into three goals: loss of generation, loss of an internal emotion regulator, and loss of an emotion classifier. The purpose of training $LMCE$ is to ensure that the generator can produce a meaningful response, while the purpose of training $L(\theta)$ is to ensure the grammatical correctness of sentences during the process of embedding emotion categories, and to solve the problem of dynamic emotions. The purpose of training $LCLA$ is to guide the generation process and ensure that specific emotions are properly expressed in the generated sequences. This is shown specifically in Equation (24), where the hyper parameter λ is of great importance to the emotion classifier.

$$L = LMCE + L(\theta) + \lambda LCLA \quad (24)$$

3.5. Sorting Algorithm

According to the research by Li et al. [28], although the cluster search algorithm can produce the N best output sequences, the contents of the output sequences are very similar. In view of this, we use a diverse sorting algorithm to increase diversity in the output sequence. Specifically, we first force the N candidate headwords in the output sequence to be different, then we determine these headwords. Finally, we generate the output sequence using a greedy algorithm. Experimental results show that our model can generate N best candidate sequences and increase the diversity of output sequences.

4. Results

In this section, we will describe the experimental datasets, experimental parameters and experimental comparison models, and analyze the DESG model.

4.1. Experimental Data and Preprocessing

We adopted the NLPCC dataset from the NLPCC2014 and NLPCC2017 emotion category tasks. The dataset had more than 3 million Weibo sentences collected with sentiment labels, including eight sentiment categories: anger (7.9%), aversion (11.9%), satisfaction (11.4%), happiness (19.1%), sadness (11.7%), fear (1.5%), surprise (3.3%), and neutrality (33.2%).

Since there are no large-scale ready-made emotional session data, we built a session dataset (STC) with emotion labels on the basis of the NLPCC dataset [8]. Specifically, we first removed the data noise and data containing the fear and surprise emotion categories (which provided only a small amount of data), then used a Bi-LSTM model to train the emotion classifier on the NLPCC dataset, using this classifier to perform emotion classification. Finally, we used the classifier to annotate the STC dataset in order to obtain a conversation dataset with emotion tags.

When constructing the STC dataset, we chose a Bi-LSTM model as the emotion classifier for the following reasons. According to the research of Zhou et al. [13], compared with other classifiers, the Bi-LSTM classifier has the best accuracy. The accuracy of the emotion classifier is shown in Table 2. To better train and test our model, we divided the STC dataset into training/validation/testing sets at a ratio of 9:0.5:0.5, respectively. Specific STC statistics are shown in Table 3.

Table 2. Accuracy of emotion classifiers.

Emotion Classifier (EC)	Accuracy
Lexicon-based	0.453
RNN	0.572
LSTM	0.597
Bi-LSTM	0.635

Table 3. Session dataset (STC) statistics.

Training Set						Validation Set	Test Set
Emotion Categories						Total Number	Total Number
Anger 194,685	Aversion 515,762	Satisfaction 324,516	Happiness 965,564	Sadness 474,863	Neutrality 1,164,349	3,639,739	202,207

4.2. Experimental Parameters

We adopted TensorFlow to implement the model in this paper. The encoder part of the model used a Bi-LSTM model and the decoder part applied a unidirectional LSTM model. The hidden state size of the LSTM in both the encoder and decoder were set to 256.

We used the Glove tool for word vector training. Although high-dimensional word vectors can more accurately reflect the semantic distribution of each word in a low-dimensional space and fully express semantic information, this will result in an increase in the parameters of the LSTM model, leading to higher risk of over-fitting. Given this, we set both the dimension of the word vector embedding and the emotion category embedding in the internal emotion regulator to 100. The general vocabularies mentioned in this article were constructed based on 30,000 common words. The emotional vocabularies of each emotion category were constructed using a semi-supervised method with a size of 200.

The stochastic gradient descent (SGD) method was employed in small batches [29], with the batch size and learning rate set to 64 and 0.5, respectively. The hyper-parameters λ in the emotion classifier are very important relative to the emotion classifier. Therefore, we set the experimental range of the λ value to $\{1 \times 10^{-1}, 1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$. A comparative analysis of the experimental results shows that the model works best with the hyper parameter λ being 1×10^{-2} .

4.3. Results and Analysis

4.3.1. Evaluation Index

The quality of the dialogue system needs to be evaluated with reasonable metrics. The following indicators were adopted in order to evaluate the language model:

- (1) BLEU the geometric average according to the accuracy of the n -word ($n = 1, 2, 3, 4$) and measures the similarity between the generated response and the real response [30]. BLEU in this paper refers to the default BLEU4.
- (2) A method based on word vector (embedding) that introduces three embedding-based measures: Greedy Matching, Embedding Average, and Vector Extrema [31]. These convert sequence mapping to a semantic vector expression, then the similarity between the generated response and the real response is calculated using a cosine similarity and other methods.
- (3) The level of expressed emotion in the output sequence is tested using emotion indicators Emotion-a and Emotion-w [21], where Emotion-a calculates the consistency of the prediction and real labels obtained by the Bi-LSTM emotion classifier, and emotion-w calculates the percentage of the corresponding emotional words in the output sequence.
- (4) In order to measure the diversity of the response calculating the ratio of single words to double (Distinct-1 and Distinct-2, respectively) in the generated responses [32].

4.3.2. Comparative Analysis of Experimental Results

In order to further verify the effectiveness of our proposed model, we compared it with related models at the forefront of research and selected four leading-edge emotional dialogue generating models that had achieved good results as baseline models. The baseline models included the following:

The Seq2Seq model [14]: This model proposes a general end-to-end sequence-learning method. In this method, the encoder maps the input sequence to a fixed-dimensional vector. The fixed-dimensional vector is regarded as the semantic expression of the input sequence. The decoder generates the output sequence according to the semantic expression of the input sequence.

The ECM (Emotional Chatting Machine) model [20]: Based on the traditional sequence-to-sequence model with static emotion vector expression, a dynamic emotional state memory network, and an external memory mechanism of emotional words, the ECM can output responses that correspond to the emotions defined by the designated emotion categories.

The EmoEmb (emotion embedding dialogue system) model [15]: This method represents each emotion category as a vector based on the Seq2Seq framework, and inputs them into the decoder at each time step.

The EmoDS (emotional dialogue system) model [21]: This model seamlessly “inserts” text by increasing the probability of emotional vocabularies at the correct time step, while guiding the response generation process through an emotion classifier to ensure that the generated text contains the specific emotions.

As shown in the comparison results of the baseline model experiments in Table 4, the DESG model outperformed other models in terms of the BLEU evaluation index, emotional performance, and response diversity; and was superior to the Seq2Seq, EmoEmb, and ECM models as an evaluation method based on word vectors, although it was not as good as the EmoDS model.

Table 4. Comparisons with baselines.

Baseline Models	BLEU	Word Vector Based Method			Emotional Performance		Responses Diversity	
	BLEU	Average	Greedy	Extreme	Emotion-a	Emotion-w	Distinct-1	Distinct-2
Seq2Seq	1.50	0.523	0.376	0.350	0.355	0.371	0.0038	0.012
EmoEmb	1.69	0.524	0.381	0.355	0.720	0.512	0.0054	0.0484
ECM	1.68	0.624	0.434	0.409	0.765	0.580	0.0090	0.0735
EmoDS	1.73	0.634	0.451	0.435	0.810	0.687	0.0113	0.0867
DESG	1.74	0.632	0.450	0.433	0.813	0.690	0.0121	0.0869

In terms of the BLEU evaluation index and the word-vector-based method, our model DESG improved significantly on the baseline model Seq2Seq, with improvements of 0.24, 0.11, 0.074, and 0.084 corresponding to BLEU, Average, Greedy, and Extreme, respectively. This indicates that the DESG model was able produce appropriate output sequences in terms of content (related grammar).

In respect to the emotional performance and response diversity indicators, the baseline Seq2Seq model performed quite poorly, as this model only generates short general responses and does not consider any emotional factors. In terms of emotional performance and response diversity, the DESG model improved significantly on to the Seq2Seq model, with improvements of 0.458, 0.319, 0.0083, and 0.0749 corresponding to Emotion-a, Emotion-w, Distinct-1, and Distinct-2, respectively. This indicates that our model was not only able to generate appropriate output sequences in terms of content (related grammar), but could also explicitly or implicitly express desired emotional responses.

To study the effects of different components, we performed an ablation experiment, removing one component from the DESG at a time.

DESG-MLE: DESG does not use an emotion classifier.

DESG-EV: DESG uses external emotional words, not internal emotional words.

DESG-BS: DESG uses a cluster search algorithm instead of our diverse sorting algorithm.

DESG-IM: DESG does not use the internal emotion regulator.

The comparison results of ablation experiments are shown in Table 5:

Table 5. Comparative results of ablation experiment.

Models	BLEU	Word Vector-Based Method			Emotional Performance		Responses Diversity	
	BLEU	Average	Greedy	Extreme	Emotion-a	Emotion-w	Distinct-1	Distinct-2
DESG-MLE	1.61	0.550	0.376	0.379	0.728	0.468	0.0054	0.0671
DESG-EV	1.64	0.569	0.398	0.387	0.749	0.472	0.0056	0.0657
DESG-BS	1.73	0.624	0.439	0.489	0.776	0.659	0.0050	0.0464
DESG-IM	1.72	0.615	0.430	0.480	0.761	0.649	0.0059	0.0670
DESG	1.74	0.633	0.450	0.434	0.813	0.690	0.0121	0.0869

As shown in the comparison results of the ablation experiments in Table 5, after removing the emotion classifier component, the DESG-MLE model had a significant decline in its emotional performance index. Because there was no emotion classifier present, the DESG-MLE model could only clearly express the target emotion in the generated response, and could not capture the emotion sequences that did not contain any emotional words.

For the DESG-EV model, the internal emotion dictionary was removed and the external emotion dictionary was used instead, leading to a significant decrease in its emotional performance index. This is because the external emotion dictionary and corpus share fewer words, causing the generation process to focus on generic vocabulary and more commonplace responses to be generated.

The DESG-BS model adopted the cluster search algorithm without utilizing the diversified sorting algorithm, leading to a significant decline in its response diversity and proving that the diversity sorting algorithm applied in our model could promote a diversity within generated responses.

The DESG-IM model without an internal emotion regulator showed a significant reduction in terms of the BLEU evaluation index, emotional performance, and diversity of responses, thus proving that an internal emotion regulator can not only guarantee the grammatical correctness of a generated response but also dynamically balance the weight between grammar and emotion.

In Table 6, a sample response to a single post is listed corresponding to each emotion category. The responses containing emotional words (highlighted in a blue font) were expressed explicitly, while those without were expressed implicitly.

Table 6. Case study for DESG.

Post	Model		Response
		Seq2Seq	What do you think of this photo?
Beach under the sunset is so beautiful.	DESG	anger aversion satisfaction happiness sadness	It is not beautiful at all. The photo is really ugly. I really want to go there. Both the sunset and beach are so beautiful! There is a sense of loneliness!
		Seq2Seq	The profession of a teacher is noble.
Teachers are the most glorious profession in the sun.	DESG	anger aversion satisfaction happiness sadness	I feel nothing at all. I hate the teachers. Of course! The teachers are so noble! I love all my teachers. It is sad that people don't respect teachers.
		Seq2Seq	I also want one.
The handmade wedding dress.	DESG	anger aversion satisfaction happiness sadness	The dress is not beautiful. It looks so bad. It is a creative design. I love this wedding dress so much! But I haven't got a chance to wear it.

The emotional interaction pattern (EIP) is defined as $P(e_r/e_p) = P(e_r, e_p)/P(e_p)$, where e_p is the emotion category of the problem, and e_r is the emotion category of the output sequence. The DESG model EIP test results are shown in Figure 4.

Figure 4. The emotional interaction pattern (EIP).

The experimental results shown in Figure 4 reveal that EIPs marked with darker shades appeared more frequently than those marked with lighter shades. From the above figure, we can observe some phenomena: (1) People tended to respond with the same emotional state as in the original post during communication (e.g., when a post expressed satisfaction, the response emotion was likely to be “satisfaction”). We see this is a common emotional interaction that indicate emotional empathy. (2) Emotions were diverse in the dialogue [33] (e.g., when an input sequence expressed happiness, the

response emotions had a smaller probability of “non-anger” emotion categories, revealing a diversity of emotions during the dialogue).

5. Conclusions

The emotional states contained in a text sequence can be expressed by explicitly using strong emotional words or by forming neutral words in different patterns. For this reason, this study proposed the dynamic emotional session generation (DESG) model to solve the emotional response problem in dialogue generation systems. The DESG incorporates a dictionary-based attention mechanism based on the Seq2Seq framework, while at the same time introducing internal emotion regulators and emotion classifier mechanisms in order to establish a large-scale emotion session generation model that can improve on existing models. To evaluate the language model, we adopted the BLEU evaluation index, a word vector-based method, and analyses of emotional performance and diversity of responses. As shown in the experimental results, our DESG model not only produced suitable output sequences in terms of content (relevant grammar) for a given post and emotion category, but also explicitly or implicitly expressed the desired emotional response.

Author Contributions: Writing—original draft, Q.G.; Writing—review & editing, Z.Z.; Data curation, Q.G.; Methodology, Q.G.; Software, W.W.; Funding acquisition, Q.G.; Project administration, Q.G.; Investigation, D.Z.; Validation, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported National Social Science Foundation (19BYY076); Science Foundation of Ministry of Education of China (14YJC860042), Shandong Provincial Social Science Planning Project (19BJCJ51, 18CXWJ01, 18BJYJ04).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Serban, I.V.; Lowe, R.; Charlin, L. Generative deep neural networks for dialogue: A short review. *arXiv* **2016**, arXiv:1611.06216.
2. Song, S. *Sun R. Ren P. Psychology Dictionary*; Guangxi People's Publishing House: Nanning, China, 1984.
3. Prendinger, H.; Mori, J.; Ishizuka, M. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *Int. J. Hum. Comput. Stud.* **2005**, *62*, 231–245. [[CrossRef](#)]
4. Prendinger, H.; Ishizuka, M. The empathic companion: A character-based interface that addresses users' affective states. *Appl. Artif. Intell.* **2005**, *19*, 267–285. [[CrossRef](#)]
5. Ritter, A.; Cherry, C.; Dolan, W.B. Data-driven response generation in social media. In Proceedings of the Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 27–31 July 2011; pp. 583–593.
6. Chen, C.; Zhu, Q.; Yan, R.; Liu, J. Survey on Deep Learning Based Open Domain Dialogue System. *Chin. J. Comput.* **2019**, *42*, 1439–1466.
7. Zens, R.; Och, F.J.; Ney, H. Phrase-based statistical machine translation. In Proceedings of the German Conference on Artificial Intelligence (KI), Aachen, Germany, 16–20 September 2002; Springer: Berlin/Heidelberg, Germany, 2002; pp. 18–32.
8. Shang, L.; Lu, Z.; Li, H. Neural responding machine for short-text conversation. In Proceedings of the International Joint Conference on Natural Language Processing (IJCAI), Beijing, China, 26–31 July 2015; pp. 1577–1586.
9. Mei, H.; Bansal, M.; Walter, M.R. Coherent dialogue with attention-based language models. In Proceedings of the National Conference on Artificial Intelligence (NAACL), San Diego, CA, USA, 12–17 June 2016; pp. 3252–3258.
10. Mayer, J.D.; Salovey, P. *What is Emotional Intelligence Emotional Development and Emotional Intelligence*; Basic Books: New York, NY, USA, 1997; pp. 30–31.
11. Zhiting, H.; Zichao, Y.; Xiaodan, L.; Ruslan, S.; Eric, P.X. Toward controlled generation of text. *arXiv* **2017**, arXiv:1703.00955.

12. Sayan, G.; Mathieu, C.; Eugene, L.; Louis-Philippe, M.; Stefan, S. Affect-lm: A neural language model for customizable affective text generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Beijing, China, 30 July 30–4 August 2017; pp. 634–642.
13. Xianda, Z.; William, Y.W. Mojitalk: Generating emotional responses at scale. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1128–1137.
14. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A diversity-promoting objective function for neural conversation models. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), San Diego, Ca, USA, 12–17 June 2016; pp. 110–119.
15. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
16. Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; Mou, L. Affective neural response generation. In Proceedings of the European Conference on Information Retrieval (ECIR), Grenoble, France, 25–29 March 2018; pp. 154–166.
17. Li, Q.; Chen, H.; Ren, Z.; Chen, Z.; Tu, Z.; Ma, J. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation. *arXiv* **2019**, arXiv:1911.08698.
18. Wu, X.; Wu, Y. A Simple Dual-decoder Model for Generating Response with Sentiment. *arXiv* **2019**, arXiv:1905.06597.
19. Chen, Z.; Song, R.; Xie, X.; Nie, J.Y.; Wang, X.; Zhang, F.; Chen, E. Neural Response Generation with Relevant Emotions for Short Text Conversation. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, China, 9–14 October 2019; Springer: Cham, Switzerland, 2019; pp. 117–129.
20. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.
21. Song, Z.; Zheng, X.; Liu, L.; Xu, M.; Huang, X.J. Generating Responses with a Specific Emotion in Dialog. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July 2019–2 August 2019; pp. 3685–3695.
22. Gross, J.J. The emerging field of emotion regulation: An integrative review. *Rev. Gen. Psychol.* **1998**, *2*, 71–299. [[CrossRef](#)]
23. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [[CrossRef](#)]
24. Lu, Q.; Zhu, Z.; Xu, F.; Guo, Q. Research on Bi-LSTM Chinese sentiment classification method based on grammar rules. *Data Anal. Knowl. Discov.* **2019**, *3*, 99–107.
25. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
26. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
27. Ghosh, S.; Chollet, M.; Laksana, E.; Morency, L.P.; Scherer, S. Affect-lm: A neural language model for customizable affective text generation. *arXiv* **2017**, arXiv:1704.06851.
28. Li, J.; Monroe, W.; Jurafsky, D. A simple, fast diverse decoding algorithm for neural generation. *arXiv* **2016**, arXiv:1611.08562.
29. Lai, T.L.; Robbins, H. Adaptive design and stochastic approximation. *Ann. Stat.* **1979**, *7*, 1196–1221. [[CrossRef](#)]
30. Abhaya, A.; Alon, L. Evaluation metrics for highcorrelation with human rankings of machine translation output. In Proceedings of the Third Workshop on Statistical Machine Translation (WMT), Columbus, OH, USA, 19 June 2008; pp. 115–118.

31. Liu, C.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 2122–2132.
32. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Phoenix, AZ, USA, 12–17 February 2016; pp. 3776–3784.
33. Li, N.; Liu, B.; Han, Z.; Liu, Y.S.; Fu, J. Emotion reinforced visual storytelling. In Proceedings of the International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 297–305.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).