This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

Emotional Conversation Generation Orientated Syntactically Constrained Bidirectional-Asynchronous Framework

Xiao Sun, *Member, IEEE,* Jingyuan Li, *Student Member, IEEE,* and Jianhua Tao, *Member, IEEE*

Abstract—The field of open-domain conversation generation using deep neural networks has attracted increasing attention from researchers for several years. However, traditional neural language models tend to generate safe, generic reply with poor logic and no emotion. In this paper, an emotional conversation generation orientated syntactically constrained bidirectional-asynchronous framework called E-SCBA is proposed to generate meaningful (logical and emotional) reply. In E-SCBA, pre-generated emotion keyword and topic keyword are asynchronously introduced into the reply during the generation, and the process of decoding is much different from the most existing methods that generates reply from the first word to the end. A newly designed bidirectional-asynchronous decoder with the multi-stage strategy is proposed to support this idea, which ensures the fluency and grammaticality of reply by making full use of syntactic constraint. Through the experiments, the results show that our framework not only improves the diversity of replies, but gains a boost on both logic and emotion compared with baselines as well.

Index Terms—Emotional conversation generation, bidirectional-asynchronous processing, syntactic constraint, compound information, affective computing.

1 Introduction

In recent years, as artificial intelligence has developed rapidly, researchers are no longer satisfied with simple prediction tasks, but pursuing technologies with greater similarity to human intelligence. As a subjective factor, emotion forms an elemental difference between humans and machines. Machines that could understand emotion would be more responsive to human needs. For instance, in education, positive emotions can improve the learning efficiency of students [1]. In the health field, mood prediction [2], [3] can be used in mental health counselling to help anticipate and prevent suicide and depression. Therefore, to make machine more intelligent, we must address the conundrum of emotional interactions.

Due to the increased social needs, the study of emotion has made great progress in computer vision [4], [5] and speech recognition [6], [7]. Text is also an important channel for sharing feelings with others, where many emotions are expressed in written or oral form. And as the most common form, emotion in conversation already had some traditional work [8], [9], showing its effectiveness in improving the quality of conversations. However, the rule-based methods used in these studies limit the room for further rise. Recently, neural language models with large-scale dataset has gained a great boost in open-domain conversation generation, which devoted to generating reply with rich content [10]. In [11], a special attention mechanism was used to introduce additional topic information into the generation and achieved significant improvement. Unfortunately, the factor considered in above method only related to topic, where they failed to take emotion factor into account. Unlike the former, the work in [12] first addressed the emotion factor into neural language models for conversation generation, and it showed that emotional replies obtain superior performances compared to the baselines that did not consider emotion. Despite these studies, there are still two defects. **First**, a conversation between humans always revolves around the special topic, but the content produces differences since it is affected by the emotional states (e.g., positive or negative) of the parties. Therefore, it is necessary to take both these two aspects into consideration in open-domain conversation generation. But to the best of our knowledge, the existing neural models did not have a study about this area. **Second**, the way in above methods that generates the reply from the first to the last word can lead to a decline in diversity, limited by the high frequency generic words in the beginning, as argued in [13].

Inspired by the above deficiencies, we propose an emotional conversation generation orientated syntactically constrained bidirectional-asynchronous framework called E-SCBA that has the following features:

- **Diversity**: E-SCBA introduces both emotion and topic information into the generation. With the support of the comprehensive information, our framework can generate emotional reply with richer content during a chat.
- **Feasibility**: Different from the work in [13], a newly-designed bidirectional-asynchronous decoder with a multi-stage generating strategy is proposed, which ensures the unobstructed communication between dif-

- *X. Sun and J. Li are with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China. E-mail: sunx@hfut.edu.cn; lijingyuan@mail.hfut.edu.cn.*
- *J. Tao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: jhtao@nlpr.ia.ac.cn.*
- *X. Sun and J. Li contributed equally to this study and share the first authorship.*

ferent information and allows a fine-grained control of the reply to address the problem of fluency and grammaticality as argued in [12], [14].
- **Scalability**: By using different information as constraints, our framework can be used as a customizable framework for controlling the content of generated conversations, meaning it is easily be scaled to other areas.

To summarize, there are three main contributions presented in this paper: (1) It conducts a study of compound information (i.e., both topic and emotion), which is used as the explicit syntactic constraint in the conversation generation. (2) It proposes a bidirectional-asynchronous decoder based the syntactic constraint, which allows fine-grained control of the generated conversation to endow the framework with feasibility. (3) Through the experiments, it reveals that E-SCBA achieves better scores on emotion, logic and diversity than does the general seq2seq and other models that consider only a single factor during the generation.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 describes the modules used in our framework and their principles in detail. Section 4 presents the results of experiments and provides a discussion. Section 5 summarizes the entire work and describes the possible future work.

2 RELATED WORK

2.1 Emotional Conversation Generation

A great deal of work has been conducted on conversation/dialogue generation. These studies can be categorized into open-domain [15], [16] and closed-domain [17]. Open-domain conversation generation is intended to generate fluency, grammaticality and meaningful replies to the given human utterances. Recently, a sequence prediction problem orientated method called sequence-to-sequence model [18], [19], [20] was proposed and rapidly used in machine translation and conversation generation, showing its superiority to the phrase-based method [21]. After that, a lot of optimization methods based on this model and RNN networks [22] emerged, including: replacing the maximum likelihood function [23], avoiding generic replies [24] and generating meaningful replies with higher diversity [11], [13].

However, a common flaw in the above work is that the emotion factor was not considered during the generation. It is impossible for a machine to be truly intelligent without emotional mechanisms, as argued in [25]. To address this problem, some studies had been explored this field. The work in [26] considers emotion-inherent responses and achieves superior performance compared to the baselines. In [9], the authors detected the emotion of user and generate a corresponding emotional reply, achieving a good performance. [27] introduced the syntactic information into the model to generate emotional replies, but the process was based on the small-scale data. The first attempt to use neural models for the emotional conversation generation is the work in [12]. Besides the emotion category embeddings in [14], it designed the other two memory mechanisms, integrating the emotion factor into the structure of networks successfully.

2.2 Sequence Generation Model

Sequence-to-Sequence (seq2seq) model is a kind of end-to-end machine optimized for the sequence issues. In recent years, it sparked the field of natural language processing to develop rapidly [28], [29], [30]. Given an input post $\boldsymbol{x} = (x_1, x_2, \cdots, x_T)$, an encoder is used to encode the post into a fixed-dimension vector, and a decoder is used to predict the target reply $\boldsymbol{y} = (y_1, y_2, \cdots, y_{T'})$ from the encoded vector. The advantage of this framework is that the sequence is not restricted to a fixed length, which makes it more flexible in practice. However, the vectors obtained by the encoder do not include sufficient information to generate a meaningful reply [20]. To address this problem, researchers have tried to use various methods to improve the quality of reply generated from the seq2seq model [31], [32].

One of the modified approaches is a content-introducing model proposed by [13]. This approach first generates a noun as keyword that reflects the gist of conversation. Then, the decoder works outward from the keyword rather than sequentially from the first to the last word. The keywords that can appear at arbitrary positions in this work introduce not only additional information but also the flexibility in structure. However, only a single information was adopted in this work, and it does not take emotion factor into account as well. In this paper, a complex syntactic constraint that includes both emotion and topic factors is used during the generation to obtain more diversified emotional replies.

3 MODEL

3.1 Overview

As shown in Figure. 1, given the post $\boldsymbol{x}$, the encoder is utilized to obtain the encoded vector as usual. After that, the process of emotional conversation generation consists of the following four steps:

***Step I*:** We first use the *Structure Detector* to predict whether the emotion keyword or topic keyword in our dictionaries should occur in the reply. If not, we use a general forward decoder to generate the reply. Otherwise, the model executes the next step.

***Step II*:** Based on the results of the first step, we generate the corresponding keywords (emotion keyword, topic keyword or both) that should appear in the reply with help of some prior knowledge (see Fig. 2).

***Step III*:** After all the prerequisite keywords have been predicted, the decoder is called. The process considers **two situations**. First, when only one keyword (emotion keyword or topic keyword) exists, a asynchronous decoder similar to [13] is used to generate the reply. Second, when two kinds of keywords exist, a newly designed bidirectional-asynchronous decoder is used to introduce both of them into the content, which first generate the middle part and then generate the rest two sides on the condition of it (see Fig. 4).

***Step IV*:** Finally, if the reply generated in the second step has two keywords, a direction selector is used to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

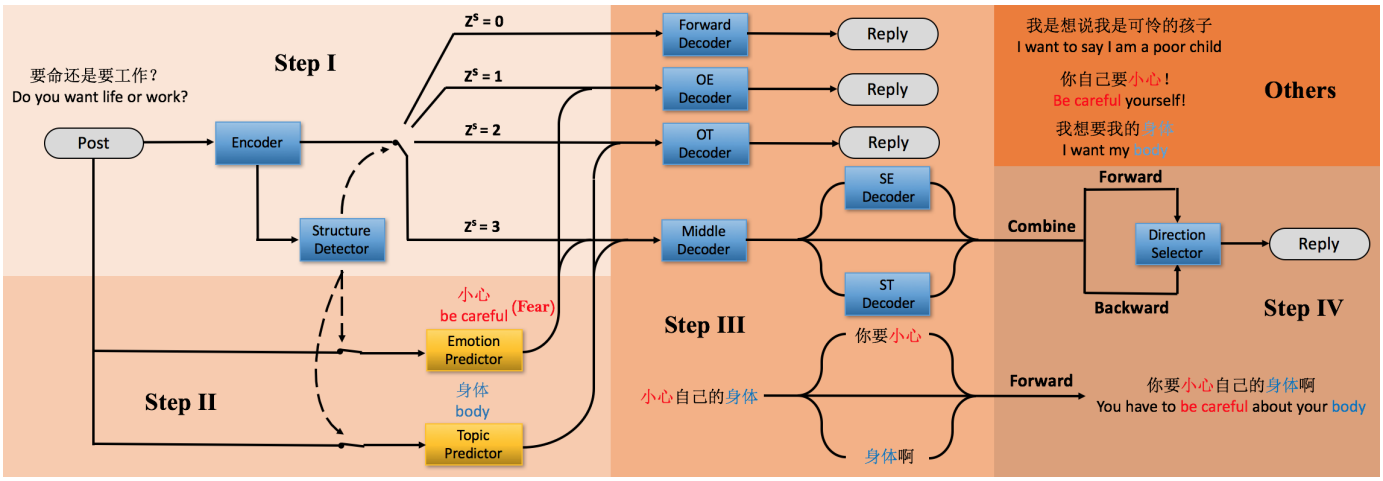


Fig. 1: An overview of E-SCBA. The process of decoding is divided into four cases based on the result of a *Structure Detector*. *OE* denotes generation based only on the emotion keyword constraint, and *OT* denotes generation based only on topic keywords. The keywords are obtained from pre-prepared dictionaries. The decoders in different situations do not share parameters.

arrange it in a correct logical order, where the reason will be described in Section 3.6 (see Fig. 5).

The example in Fig. 3 shows the process of generating reply by using the method we proposed in this paper. The topic machine and emotion machine are shown in Fig. 2, the decoder is shown in Fig. 4 and the direction selector is shown in Fig. 5. Note that we acquire the emotion and topic prior knowledge in different ways, which we elaborate on the detail in the following sections.

3.2 Post Encoding

The RNN used in the encoder is the gated recurrent units (GRUs) [33]. Given a sequence $\boldsymbol{x} = (x_1, x_2, ..., x_T)$, the recurrent hidden state will be updated by:

$$h_t = \mathbf{GRU}(h_{t-1}, x_t), \tag{1}$$

where x_t is the t-th word index and h_{t-1} is the hidden state at time $t-1$.

3.3 Structure Detector

This module aims at detecting whether the emotion keyword and topic keyword in our dictionaries should appear in the reply $\boldsymbol{y}$. As shown in Fig. 1, we define the following four specific cases:

- $z^s = 0$: No keyword: a general forward decoder is used to generate the reply;
- $z^s = 1$: Only an emotion keyword: an asynchronous decoder similar to [13] is used to generate the reply starting from the emotion keyword;
- $z^s = 2$: Only a topic keyword: an asynchronous decoder is the same as above but is used to generate the reply starting from the topic keyword;
- $z^s = 3$: Both emotion keyword and topic keyword: the bidirectional-asynchronous decoder we proposed is used to generate the reply.

Formally, given the post $\boldsymbol{x}$, we first obtain the hidden state sequence $\boldsymbol{h}$ from the encoder. Then the case number z^s is determined by a fully-connected layer as follows:

$$p(z^s = i|\boldsymbol{x}) = softmax(\mathbf{W}^s \cdot \tilde{h}) \tag{2}$$

where $i \in \{0, 1, 2, 3\}$, $\tilde{h} = \sum_{i=1}^{T} h_i$. For each post $\boldsymbol{x}$, this module is always called first. The multi-class classifier described by the above equation predicted the structure for the reply $\boldsymbol{y}$ so that we can do the following work.

3.4 Keyword Predictor

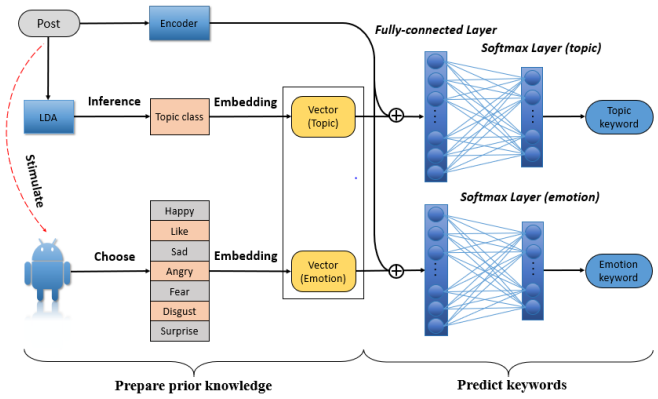


Fig. 2: The process of keywords generation.

The main role of the word predictor is to predict which keywords in our dictionaries should appear in the final reply. The adopted dictionaries are divided into an emotion dictionary and a topic dictionary. The emotion dictionary[1] we used is the work in [34], which contains 27,466 emotion keywords and includes seven categories: *Happy*, *Like*, *Sad*, *Angry*, *Fear*, *Disgust* and *Surprise*. The topic dictionary is

1. http://ir.dlut.edu.cn/EmotionOntologyDownload

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

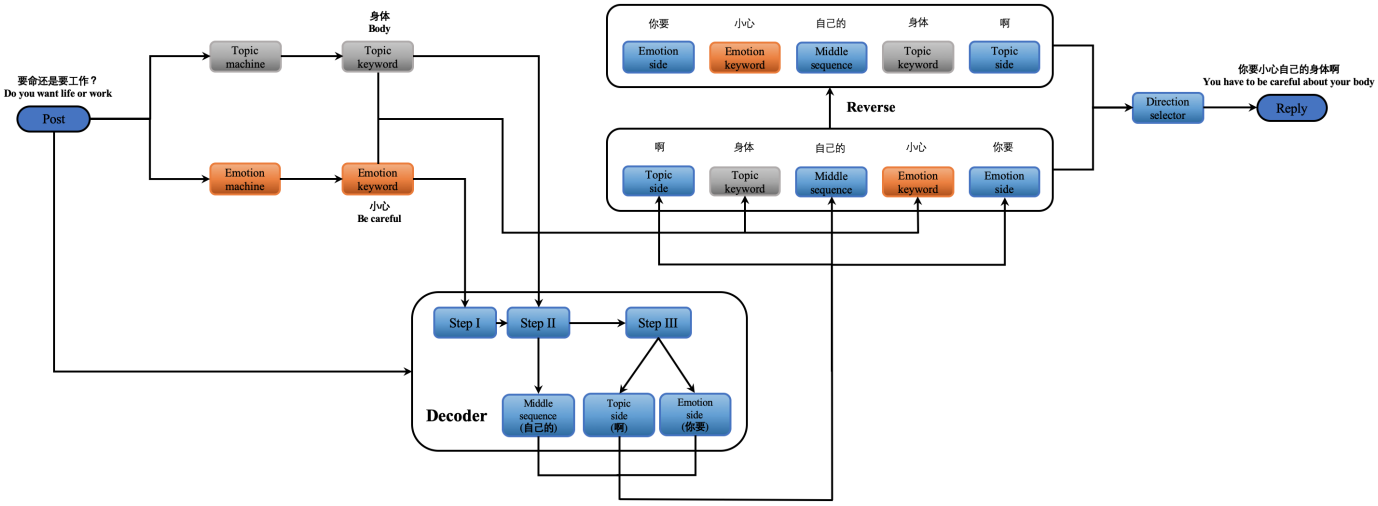


Fig. 3: A running example for the generation of reply.

obtained by a pre-trained Latent Dirichlet Allocation (LDA) model[2] from [35], including 10 categories and 100 keywords for each category[3].

As shown in Fig. 2, instead of using the hidden sequence of encoder to predict keywords directly, a sequence attention mechanism based on the prior knowledge is applied to complement the insufficient information in the encoder. Since the category (emotion or topic) is a high-level abstraction of knowledge representation [12], we intuitively used the category information in the above dictionaries as the prior knowledge mentioned here, which is derived from the following method in practice. For the emotion in the conversation, the work in [36], which gives a robot a personality defined by the **emotion transfer network** so that it can produce a corresponding emotional response to specific external stimuli, is used to generate a sign which marks it as the knowledge of emotion (in this work, we limited the responses to the seven types of emotions listed previously). For the topic in the conversation, post is processed by the pre-trained LDA model to infer the topic category.

To integrate the prior knowledge into the process, we usually compute the correlation between different kinds of knowledge embedding $\boldsymbol{k} = \{k^{et}, k^{tp}\}$ and each unit in the hidden sequence of encoder, representing it as a specific weight value. Formally, the details are described as follows:

$$c^{k,*} = \sum_{i=1}^{T} \alpha_i^{k,*} h_i, \tag{3}$$

$$\alpha_i^{k,*} = \frac{\exp(e_i^{k,*})}{\sum_{t=1}^{T} \exp(e_t^{k,*})}, \tag{4}$$

$$e_i^{k,*} = (\mathbf{v}_\alpha^{k,*})^T tanh(\mathbf{W}_\alpha^{k,*} k^* + \mathbf{U}_\alpha^{k,*} h_i), \tag{5}$$

where * $\in \{et, tp\}$ represents the aspect of topic or emotion. $\mathbf{v}_\alpha^{k,*}$, $\mathbf{W}_\alpha^{k,*}$ and $\mathbf{U}_\alpha^{k,*}$ are the trainable parameters. The information is concentrated in the weighted vector c_j^{et}, and the conditional probabilities of keywords are respectively calculated by:

$$p(w_{et}^k|\boldsymbol{x}, k^{et}) = softmax(\mathbf{W}_{et}^w c^{k,et}), \tag{6}$$

$$p(w_{tp}^k|\boldsymbol{x}, k^{tp}) = softmax(\mathbf{W}_{tp}^w c^{k,tp}), \tag{7}$$

where $c^{k,et}$ and $c^{k,tp}$ are the attention units computed by Equ. 3. Each of the above equations can be viewed as a multi-class classifier that produces a probability distribution over all the emotion keyword or the topic keyword.

3.5 Bidirectional-asynchronous Decoder

Since the final reply has both emotion keyword and topic keyword, a vital issue is how to achieve the conversation generation based on two keywords. For the situation having only one keyword, there is only one vacancy in the sentence, where the rest of the reply can be generated from both forward and backward at the same time. However, when multiple vacancies appear in the sentence, the method of synchronous generation is not feasible any more. Before getting the final result via argmax function, it is impossible to know the accurate position of the emotion keyword and the topic keyword in the reply, which means that the communication between different information is blocked if all the parts of the reply are generated synchronously. To address this problem, we propose a new bidirectional-asynchronous decoder that makes use of the syntactic constraint common to both the emotion and topic to generate the reply asynchronously from the keywords on both sides.

Suppose the reply is $\boldsymbol{y} = (\boldsymbol{y}^{ct}, w_{tp}^k, \boldsymbol{y}^{md}, w_{et}^k, \boldsymbol{y}^{ce})$ where $\boldsymbol{y}^{md}$ is the middle part of the reply between the two keywords and $\boldsymbol{y}^{ct}$, $\boldsymbol{y}^{ce}$ represent the parts connected to the middle part by the emotion keyword and topic keyword, respectively. As shown in Fig. 4, the generation of the middle sequence $\boldsymbol{y}^{md} = (y_1^{md}, y_2^{md}, ..., y_K^{md})$ is divided into an asynchronous strategy with two steps. Emotion keywords are first processed to form a sequence containing emotional information (Step I). And then we used an emotional attention mechanism, which is similar to [20]. It uses the hidden

2. http://gibbslda.sourceforge.net/

3. To avoid situations in which emotion keyword and topic keyword are predicted to be the same word, all the overlapping words in the two dictionaries default to emotion keywords.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

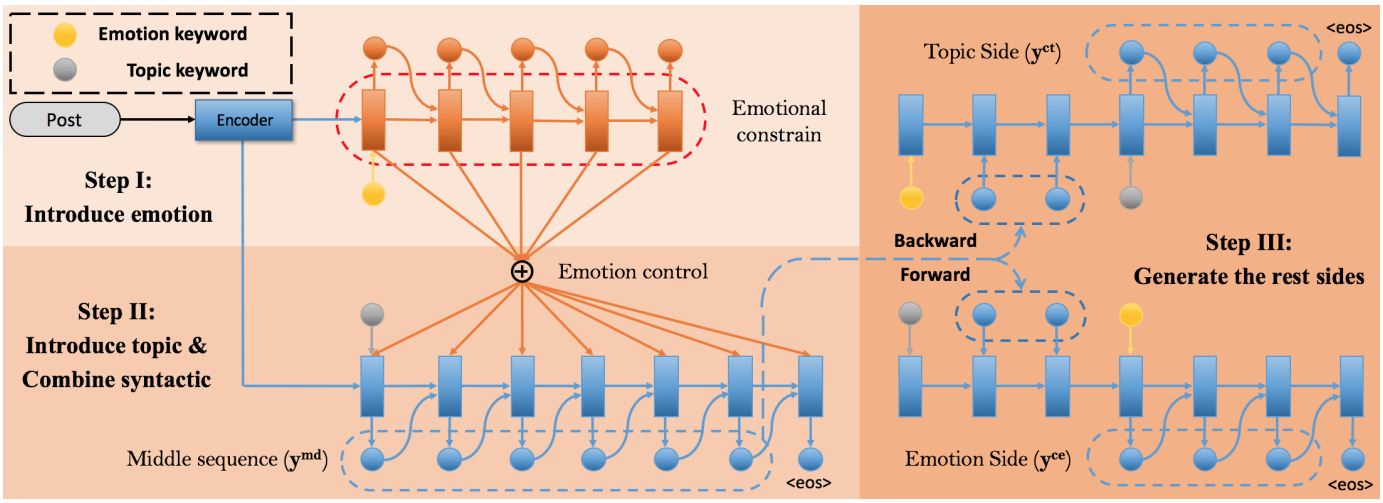


Fig. 4: The architecture of the bidirectional-asynchronous decoder. The middle part of the reply is generated in Steps I and II, and the remaining two sides are generated in Step III. The RNN networks used in the decoder do not share the parameters with each other.

sequence in Step I to control the generation in Step II. This process mimics the behavior of human speech, where the common topic as the objective factor is the gist of conversation and different emotions as the external subjective factor affect what we say. Formally, the definition is as below:

$$c_j^{m,et} = f_{att}^{et}(s_{j-1}^{tp}, \{s_i^{et}\}_{i=1}^{K'}), \tag{8}$$

$$p(\boldsymbol{y}^{md}|\boldsymbol{x}, \boldsymbol{w}^k) = \prod_{j=1}^{K} p(y_j^{md}|y_{<j}^{md}, < w_{et}^k, w_{tp}^k >) = \prod_{j=1}^{K} p(y_j^{md}|y_{j-1}^{md}, s_j^{tp}, c_j^{m,et}), \tag{9}$$

where $\boldsymbol{w}^k = < w_{et}^k, w_{tp}^k >$ represents the set of keywords, s_i^{et} and s_j^{tp} represent the decoding state of the step that introduce emotion and topic information, respectively. $c_j^{m,et}$ is the emotion attention unit at time j, computing by the emotion control function f_{att}^{et} as follows:

$$c_j^{m,et} = \sum_{i=1}^{K'} \alpha_{j,i}^{m,et} s_i^{et}, \tag{10}$$

$$\alpha_{j,i}^{m,et} = \frac{\exp(e_{j,i}^{m,et})}{\sum_{t=1}^{K'} \exp(e_{j,t}^{m,et})}, \tag{11}$$

$$e_{j,i}^{m,et} = (\mathbf{v}_\alpha^{md})^T tanh(\mathbf{W}_\alpha^{md} s_{j-1}^{tp} + \mathbf{U}_\alpha^{md} s_i^{et}), \tag{12}$$

where $\mathbf{v}_\alpha^{md}$, $\mathbf{W}_\alpha^{md}$ and $\mathbf{U}_\alpha^{md}$ are the trainable parameters in the function, $e_{j,i}^{m,et}$ represents the impact grades of the emotion state s_i^{et} on the topic state s_{j-1}^{tp}.

After generating the middle sequence, we connect it with all the keywords to form a new sequence. Since the starting point is different, two seq2seq that do not share parameters are used to encode the connected sequences, forward and backward, respectively. Then decode the sequences $\boldsymbol{y}^{ce} = (y_1^{ce}, y_2^{ce}, ..., y_M^{ce})$ and $\boldsymbol{y}^{ct} = (y_1^{ct}, y_2^{ct}, ..., y_N^{ct})$ connected with different keywords. The process can be defined as follows:

$$p(\boldsymbol{y}^{ce}|\boldsymbol{w}^k, \boldsymbol{y}^{md}) = p(\boldsymbol{y}^{ce}|g^f([w_{tp}^k, \boldsymbol{y}^{md,f}, w_{et}^k])) = \prod_{i=1}^{M} p^f(y_i^{ce}|y_{i-1}^{ce}, s_i^{ce}), \tag{13}$$

$$p(\boldsymbol{y}^{ct}|\boldsymbol{w}^k, \boldsymbol{y}^{md}) = p(\boldsymbol{y}^{ct}|g^b([w_{et}^k, \boldsymbol{y}^{md,b}, w_{tp}^k])) = \prod_{j=1}^{N} p^b(y_j^{ct}|y_{j-1}^{ct}, s_j^{ct}), \tag{14}$$

where $\boldsymbol{y}^{md,f}$ and $\boldsymbol{y}^{md,b}$ are the forward form and backward form of the middle part, respectively. g^f and g^b represent the GRU networks to encode the input sequences. Finally, we connect the results of the different parts together as a candidate reply for the next section.

3.6 Direction Selector

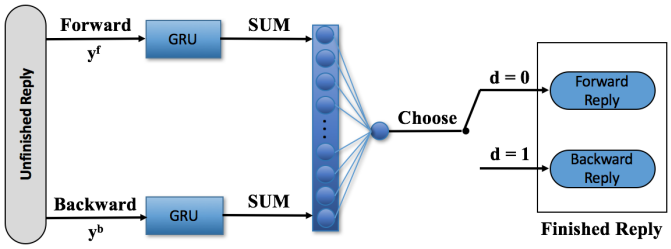


Fig. 5: The architecture of the direction selector, where d = 0 means the topic keyword appears on the left and the emotion keyword on the right (Forward), while d = 1 means the opposite (Backward).

To make the samples meet the requirements of our bidirectional-asynchronous decoder, by default we place the topic keyword as the first keyword on the left and the emotion keyword on the right in training. The uniform format facilitates the decoder to learn the syntax constraint

information between emotion and topic. However, in real situations, the topic keyword does not always appear before the emotion keyword, which means the machine must determine the correct direction from the forward and backward sequences.

As shown in Fig. 5, $\boldsymbol{y}^f$ and $\boldsymbol{y}^b$ are the forward and backward forms of the final result in the preceding section, respectively. One of them will be the reply we expect. The decision process is defined as follows:

$$p(z^d|\boldsymbol{y}^f, \boldsymbol{y}^b) = sigmoid(\mathbf{W}^d[\tilde{h}^{d,f}, \tilde{h}^{d,b}]), \tag{15}$$

$$\tilde{h}^{d,*} = \sum_{i=1}^{T'} \mathbf{GRU}(y_i^*). \tag{16}$$

The forward and backward sequence are first encoded by two GRU networks. Then, the sums of hidden states are connected to form the input to the decision network. Finally, a fully-connected layer with sigmoid function is used to select the direction of the reply. After all these operations complete, the finished reply $\boldsymbol{y}$ should conform to our expectations.

3.7 Loss Function

Since the framework is composed of multiple individual modules, the loss function consists of several parts and the data needs some extra processing. We describe these operations in their order of execution[4].

For *Structure Detector* and *Word Predictor*, the dictionaries mentioned in Section 3.4 are used to identify the keywords in each conversation in advance, including the word indexes and categories. Moreover, we divide the entire dataset U into four parts according to the results of mark, namely U^{nh}, U^{oe}, U^{ot}, and U^{bh} (as listed in Section 3.3). The loss functions of these two modules can be defined as follows:

$$\begin{aligned}
\xi_1 &= -\sum_{(\boldsymbol{x}, z^s)\in U} z^s \log p(z^s|\boldsymbol{x}), \\
\xi_2 &= -\sum_{(\boldsymbol{x}, k^{et}, w_{et}^k)\in(U^{oe}\cup U^{bh})} w_{et}^k \log p(w_{et}^k|\boldsymbol{x}, k^{et}) \\
&\quad -\sum_{(\boldsymbol{x}, k^{tp}, w_{tp}^k)\in(U^{ot}\cup U^{bh})} w_{tp}^k \log p(w_{tp}^k|\boldsymbol{x}, k^{tp}),
\end{aligned} \tag{17}$$

For the decoders, we split the reply $\boldsymbol{y}$ from the keywords and pad or trim each of them as the targets. The loss function for all the decoders is shown below:

$$\begin{aligned}
\xi_3 &= -\sum_{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}^k)\in U} \boldsymbol{y} \log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^k) \\
&= -\sum_{(\boldsymbol{x}, \boldsymbol{y})\in U^{nh}} \boldsymbol{y} \log p^{nh}(\boldsymbol{y}|\boldsymbol{x}) \\
&\quad -\sum_{(\boldsymbol{x}, \boldsymbol{y}, w_{et}^k)\in U^{oe}} \boldsymbol{y} \log p^{oe}(\boldsymbol{y}|\boldsymbol{x}, w_{et}^k) \\
&\quad -\sum_{(\boldsymbol{x}, \boldsymbol{y}, w_{tp}^k)\in U^{ot}} \boldsymbol{y} \log p^{ot}(\boldsymbol{y}|\boldsymbol{x}, w_{tp}^k) \\
&\quad -\sum_{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}^k)\in U^{bh}} \boldsymbol{y} \log p^{bh}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^k),
\end{aligned} \tag{18}$$

4. In our scenario, the rules of the labels are as follows: x = post, y = reply, w = keywords, k = prior knowledge and z^* = others.

The loss function for the decoders has four terms as shown above based on the result in *Structure Detector*. Each term represents the cross-entropy function defined on a sub-dataset.

For *Direction Selector*, we mark the correct directions in the dataset according to the relative positions of the emotion keyword and topic keyword. The loss is given by:

$$\begin{aligned}
\xi_4 = -\sum_{(\boldsymbol{x}, \boldsymbol{y}, z^d)\in U^{bh}} & z^d \log p(z^d|\boldsymbol{y}^f, \boldsymbol{y}^b) \\
- & \ (1 - z^d) \log(1 - p(z^d|\boldsymbol{y}^f, \boldsymbol{y}^b)),
\end{aligned} \tag{19}$$

where $z^{d,*} \in \{0, 1\}$ is a real number. All the labels in ξ_* are one-hot representations of the gold distribution over the dataset.

4 Experiment

4.1 Data Description

We evaluated and trained our framework on the emotional conversation dataset NLPCC2017[5] which was collected from Weibo. Besides, another Chinese movie subtitle dataset crawled from the Internet[6] was used as additional data for the case study to test the performance of framework on cross-domain dialogue and develop a detailed error analysis. Note that the movie subtitle dataset is noisy and only used for case study in our work.

As shown in Table 1, there are a total of 1,119,201 post-reply pairs in the dataset. The conversations requiring both types of keywords accounted for 42.6% of the total, which are available data for our bidirectional-asynchronous decoder, where we used 8,000 for validation, 3,000 for testing and the rest for training. For conversation generation in Chinese, the data with small noise is seldom. Therefore, instead of removing the infrequent categories (*Surprise* (1.2%) and *Angry* (0.7%)), we used the over-sampling method to construct additional training data for them: we replaced the emotion keywords in part of the training data with words that belong to the infrequent categories, as synonyms. These generated conversations were then used as supplemental training data. And for the prediction of topic keywords, we chose 60,000 pairs from our training data to train the LDA. The high frequency words and stop words, which have no bearing on the topics, were removed in advance. To consider the emotions, we controlled the proportion of each emotion category to ensure that the gaps between different categories in the training data were not too large.

4.2 Implementation Details

The encoders and decoders in our framework use GRU networks [19] with a 2-layer structure and 512 hidden units in each layer. The word embedding (with a size of 128) is pre-trained using Word2Vec on the Chinese Wiki Corpus[7]. The size of the dictionary is 32,000: it contains both generic words and keywords. The optimization method used is the Adam algorithm, where the learning rate starts at 0.005, the

5. http://tcci.ccf.org.cn/conference/2017/
6. http://www.zimuku.cn/
7. https://dumps.wikimedia.org/zhwiki/latest/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

Type	Number	Percentage
All	1,119,201	100.0%
B–H	476,121	42.6%
O–E	157,043	14.0%
O–T	410,767	36.7%
N–H	75,270	6.7%

TABLE 1: The results of the keywords annotations. B–H, O–E, O–T, and N–H denote the conversations whose replies have both an emotion keyword and a topic keyword, only an emotion keyword, only a topic keyword, and no keyword, respectively.

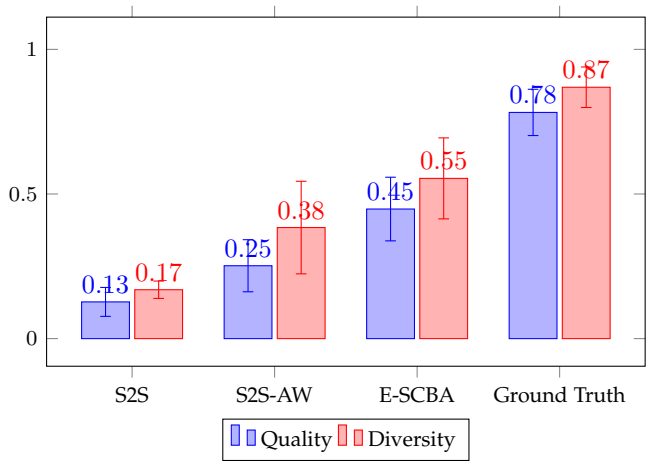


Fig. 6: The grades of quality and diversity in artificial evaluation. Error bars are plotted against the values of standard deviation.

decay rate is 0.99, and the other hyperparameters were set to their default values.

The dropout method was applied to avoid overfitting with a rate of 0.9 and early stopping on the validation set. To reduce the impact of *exposure bias* between training and inference, we adopted a sampling mechanism during training: the true tokens was used to train decoders at early stage for several epochs, and the sampled tokens from model replaced a part of true tokens as the work in [37] for further training. In particular, the input in the third step of the bidirectional-asynchronous decoder (see Fig. 4) applied the same mechanism. The whole process took about four weeks on two GTX1080 GPU machines. The implementation of our framework is based on the **TensorFlow** deep learning framework[8].

4.3 Artificial Evaluation

In this section, we asked the annotators to evaluate the results of our framework and baselines. In total, we used 700 conversations, 100 for each emotion category, which were sampled randomly from the test set. The baselines included the general seq2seq model and the asynchronous model with a single keyword (the same as [13]).

4.3.1 Detailed evaluation

In this experiment, we evaluated the replies from annotators scores, where the annotators were asked to score a reply based on the following metrics as proposed in [21]:

Consistency: Measure fluency and grammaticality of the reply on a three-point scale: 0, 1, 2.

Logic: Measure the degree to which the post and the reply logically match on a three-point scale: 0, 1, 2. Note that a too short or too high frequency reply would be annotated as either 0 or 1 (if the annotator thought the reply related to the post), like "Me too", or "I think so".

Emotion: Measure whether the target reply includes the right emotion. A score of 0 means the emotion in the reply is wrong or there is no emotion, and a score of 1 means the reply includes the right one.

Table 2 (*2-tailed t-test*: $p < 0.05$ for Consistency and Logic, $p < 0.01$ for Emotion) compares our bidirectional-asynchronous framework E-SCBA with the baselines. As we can see, E-SCBA outperforms the other models on all three metrics. For the baseline S2S-AW, the scores of emotion is completely inferior to E-SCBA (+0.245) that takes emotion factor into account, showing that emotion in the replies to be easier to perceive by introducing the emotion keywords. Additionally, the best performance on *Consistency* (+0.027) and *Logic* (+0.060) shows that the fluency and grammar of reply generated from E-SCBA is guaranteed. S2S-AW yields a higher score on these two metrics than the general seq2seq, but E-SCBA is more skillful in regulating information attributed to the effect of syntactic constraint.

However, even if E-SCBA produces higher quality replies than do the other baselines in most cases, the performance in *Surprise* or *Angry* categories is not satisfactory. Our framework achieves worse scores in *Consistency* (-0.043 and -0.095) and *Logic* (-0.014 and -0.038) classified as these two emotion categories. One reason is that these categories have much less training data than do the others. Although the consistency and logic of replies from all the models falls without sufficient data, it is more difficult to avoid overfitting using models with more complex structures. The emotions scores also have a large fractional span across different categories (largest: $\pm$0.370). We think the reason for this problem is that some emotion categories with lower scores have a high degree of similarity to others with higher scores, and a sentence may contain multiple emotions, such as *Surprise* and *Happy*, *Angry* and *Disgust*. For those categories that have much less data, E-SCBA may be unable to extract the features well. The result is that these emotions are difficult to express in the replies, which leads to lower scores. There is another noteworthy place about scores of emotion. Our model obtains lower scores in Consistency and Logic but a higher score in Emotion for the replies based on "surprise" and "angry". This is because each reply generated by E-SCBA has a keyword based on the corresponding emotion. When annotators notice emotional keywords, the keywords impress them deeply and motivate them to give higher scores of emotion even if the consistency and logic of the reply is insufficient.

8. https://github.com/tensorflow/tensorflow/

Metric / Method	Overall			Happy			Like			Surprise		
	C	L	E	C	L	E	C	L	E	C	L	E
S2S	1.301	0.776	0.197	1.368	0.924	0.285	1.341	0.757	0.217	1.186	0.723	0.076
S2S-AW	1.348	1.063	0.231	1.437	1.097	0.237	1.418	1.125	0.276	**1.213**	**0.916**	0.105
E-SCBA	**1.375**	**1.123**	**0.476**	**1.476**	**1.286**	**0.615**	**1.437**	**1.173**	**0.545**	1.197	0.902	**0.245**
Ground Truth	1.793	1.615	0.656	1.867	1.728	0.781	1.910	1.530	0.582	1.782	1.627	0.537

Metric / Method	Sad			Fear			Angry			Disgust		
	C	L	E	C	L	E	C	L	E	C	L	E
S2S	1.393	0.928	0.237	1.245	0.782	0.215	**1.205**	0.535	0.113	1.368	0.680	0.236
S2S-AW	1.423	1.196	0.293	1.260	1.105	0.272	1.198	**0.860**	0.182	1.488	1.145	0.253
E-SCBA	**1.497**	**1.268**	**0.525**	**1.268**	**1.124**	**0.453**	1.110	0.822	**0.347**	**1.637**	**1.289**	**0.603**
Ground Truth	1.808	1.547	0.593	1.725	1.593	0.657	1.655	1.638	0.653	1.803	1.643	0.787

TABLE 2: The grades of consistency, logic and emotion in artificial evaluation (C = Consistency, L = Logic, E = Emotion). E-SCBA is our framework, and *Ground Truth* is the real replies for reference.

4.3.2 Comprehensive evaluation

In the second experiment, we want the annotators can make an intuitive assessment of the quality and diversity. Therefore, some comprehensive indicators were adopted to evaluate the results, including:

Quality: Measure the quality of reply for the given post, **one score for each conversation**, where a score of 0 is bad and a score of 1 is good.

Diversity: Measure the diversity of replies generated from the different models on the test set using a continuous score range of (0, 1), **one score for each dataset that expresses a specific emotion**. The model with better diversity should get a higher score[9].

An average grade bar graph for the different models and the ground truth are shown in Fig. 6. E-SCBA obtains significant improvement (+0.17) on the diversity protocol in comparison with S2S-AW. Our framework utilizes both emotion keywords and topic keywords, meaning there are more options for keywords in the replies and the diversity of information increase exponentially. Besides, the bidirectional-asynchronous structure of decoder integrates these two kinds of information to further improve the quality of sentences (+0.20). Error bars show that the grades of some indicators are sway, where we think the subjective factor of human annotators is the cause of this disagreement.

4.4 Automatic Evaluation

In this section, we evaluated the generated reply on some objective indicators. Moreover, the diversity distribution of words in the reply is visualized in a heatmap. In addition to the baselines in the preceding section, we used two other baselines with a single keyword in the automatic experiment:

S2S-STW: the model uses a synchronous method that starts generating its reply solely and directly from the topic keyword.

9. We did not define what the "good" or "bad" reply is and the standard for diversity of replies, annotators had their own subjective discretion during the experiment.

Method	G–M	E–A	V–E	distinct-1	distinct-2
S2S	0.297	0.382	0.284	0.086	0.212
S2S-STW	0.328	0.433	0.327	0.135	0.343
S2S-SEW	0.322	0.421	0.319	0.146	0.364
S2S-AW	0.363	0.485	0.352	0.162	0.417
E-SCBA	**0.405**	**0.553**	**0.395**	**0.218**	**0.582**

TABLE 3: The scores of objective metrics. G–E = Greedy Matching, E–A = Embedding Average, V–E = Vector Extrema.

S2S-SEW: the model uses a synchronous method that starts generating its reply solely and directly from the emotion keyword.

The synchronous method in S2S-STW and S2S-SEW was mentioned in [38] to act as a contrast to the existing models with asynchronous structure.

4.4.1 Numerical analysis

The objective indicators we used in this experiment is defined as follows:

Embedding-based Metrics: measure the similarity computed by cosine distance between a candidate reply and the target reply using sentence-level embedding, following the work in [39], [40].

Distinct Metrics: measure the diversity and information in the candidate reply using the number of distinct unigrams and bigrams in the generated replies, called Distinct-1 and Distinct-2, respectively, following the work in [11], [23].

The result are listed in Table 3. Compared with S2S-STW and S2S-SEW, E-SCBA produces the highest scores on similarity (+0.077, +0.12, +0.068) and diversity (+0.072 and +0.218). One reason is that E-SCBA takes both the emotion and topic aspects of the conversations into account during the generation by introducing keywords, which is accurate and comprehensive. This focus causes our model to capture the thrust of the dialogue and makes the replies more in line with the true distribution. With the help of the bidirectional-asynchronous decoder, our framework takes advantage of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2923619, IEEE Transactions on Affective Computing

Post	Method		Chinese	English(translated)
浪费时间就是浪费生命。 Wasting time is a waste of life.	S2S		浪费时间就是浪费生命	Wasting time is a waste of life
	E-SCBA	Happy	**时间**要拿来**享受**生命	**Time** should be taken to **enjoy** life
		Like	我的**时间**很**值钱**啊！	My **time** is **worth** a lot of money!
		Sad	浪费**时间**真**可惜**	It is a **pity** to waste **time**
		Fear	浪费**时间**真**羞耻**	It is a **shame** to waste **time**
		Disgust	不要**浪费**我的**时间**！	Do not **waste** my **time**!
		Angry	**时间**的**脾气**怎么样?	How is the **temper** of **time**?
		Surprise	**原来时间**是这样的	The **original time** is like this
生活艰辛但确实很有趣， 每天沉浸其中是乐事。 Life is hard but it is truly interesting, and fun to be immersed in it every day.	S2S		生活是艰辛的	Life is tough
	E-SCBA	Happy	我沉浸在**生活**的**乐趣**中	I am immersed in the **fun** of **life**
		Like	没有**完美**的**生活**	There is no **perfect life**
		Sad	我的**生活**充满了**遗憾**	My **life** is full of **regret**
		Fear	**生活**让我感到**害怕**	**Life** makes me feel **scared**
		Disgust	我的**生活不及**你	My **life** is **inferior** to yours
		Angry	为**生活**提出**抗议**！	Make a **protest** for **life**!
		Surprise	**生活**真是**奇怪**啊	**Life** is really **strange**

TABLE 4: Sampled conversations with different emotions from the test data. The bold words with single underlines denote the topic keywords, and the bold words with wavy underlines denote the emotion keywords. For the same post, the topic keyword is fixed.

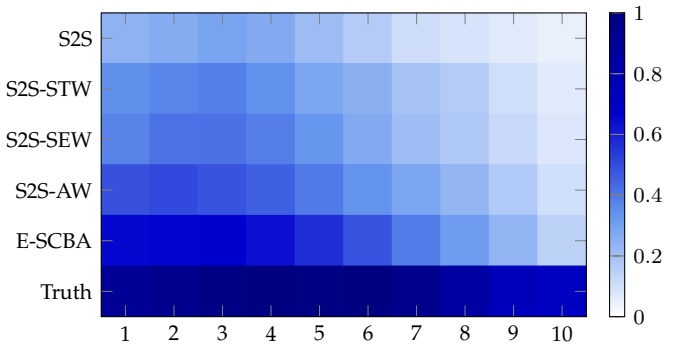


Fig. 7: The visualization of word distribution, where the positions with deeper color have a higher diversity.

the diverse mix of emotions and topics to obtain better performance in diversity of the replies.

Another interval with large improvement on diversity is that the seq2seq model starting from a pre-generated keyword (i.e., S2S-STW and S2S-SEW) outperforms the general seq2seq using a traditional forward decoder (+0.060 and +0.152), showing that it is better to start from a arbitrary position instead of from the first word to the end. Note that, this experiment does not include perplexity, since one of our goals is to diversify the generated replies, where perplexity is not a suitable metric for this work, as argued in [23].

4.4.2 Visual distribution

As shown in Fig. 7, we visualized the diversity distribution of words in different positions (1–10) of the generated reply. E-SCBA is committed to solve the problem of generic reply, which can be defined as a high frequency of certain replies to any posts, as well as a large number of identical words produced in the same place. In other words, a rich word diversity in the reply reflects a model that generates meaningful replies. The diversity of word positions in our scenario is defined as the number of different words in the same position. The results shown in the figure have been normalized.

Compared with other models, E-SCBA has much deeper color in the same positions, which means that the reply generated from our framework has richer content. The worst of all is the general seq2seq, whose words changes slightly in the first half. Besides the insufficient information from the post, the immutable sequential structure limits the potential of the model, resulting in the generic reply. In contrast, E-SCBA can insert the keywords and start generation from an arbitrary position, helping to reduce the probability of this situation. Overall, the color of E-SCBA fades more slowly and get a longer duration (1–7) than others, showing that our framework improves not only the quality of content but also the capacity of memory. For one thing, the bidirectional-asynchronous decoder allows the information to be transmitted from the middle to the sides during generation, which reduces the path of transmission and implicitly enhances the ability of the networks to memorize information. For another, the attention mechanism realizes the asynchronous introduction of topic keyword and emotion keyword, increasing the information capacity of the networks. In brief, machine studies that start generation from an arbitrary point instead of the beginning position is a good choice for achieving high diversity and persistence, because this may allow the machine to have more room to think about how to generate more meaningful replies.

4.5 Case Study and Error Analysis

4.5.1 Multiple Emotion Study

Even for the same post, there are different features for the generation with different emotions and may be multiple

Post	Emotion	Chinese	English(translated)
你能把她送到寄宿制学校是吧? You can send her to boarding school, right?	Happy	在家里更轻松啊！	It is more **relaxing at home**
和童话世界不同， 现实世界是没那么容易改变的。 Unlike the fairy tale world, the real world is not easy to change.	Like	你要**相信**我的**想法**	You must **believe** my **thoughts**
他就是一直话太多。 He always talks too much.	Disgust	他的**内心**很**糊涂**	His **heart** is very **confused**
如果回去了，我不能原谅我自己。 If I go back, I cannot forgive myself.	Sad	那就**回来道歉**！	Then **come back** and **apologize**!
再过一会我就找不到你了， 你必须马上脱离。 I can not find you in a few minutes, and you have to leave it right away.	Happy	**哈哈**，**发生**什么了?	**Haha**, what **happened**?
普通民众中也可以拥有某些权利的概念。 The general public can also has some of the concepts of rights.	Like	可以**作为**他们**生活**的一部分	This can **be** a part of their **life**
我不知道，当我没说。 I don't know, so I did not say anything.	Surprise	这很**神秘咯**	It's very **mysterious**.
那个人最期待的是像你这样的美女 送这个箱子。 The man is looking forward to having a beautiful girl like you to send this box to.	Disgust	我**期待**的是像你这样的**流氓**	I am **looking forward to** a **hooligan** like you

TABLE 5: Sampled conversations with a corresponding emotion from the Chinese movie subtitles data.

suitable replies. Therefore, we provide some examples with multiple emotion in Table 4. We still first determine the emotion of the post and then generate replies containing a specific emotion in the actual scene.

As we can see, the general seq2seq perfers to generate short and meaningless reply. In these examples, it extracted information from post correctly, but its ability to handle information is not flexible. The reply is more like a summary of post rather than a conversation. In the E-SCBA, the generated replies with different emotions hold great diversity and most of them catch the essence of the conversation. Concretely, the topic keyword as a gist closely connects the reply with the post. Furthermore, changing the emotion keywords used in the reply causes the decoder to generate replies with diverse content, which shows that emotion is one of the dominant factors in diversification. The keywords that represent strong emotion helped to better express the feelings in the reply. However, the reply generated from E-SCBA do not perform well in *Surprise*, since less available data puts it at a inferiority when competing with others in training. And for emotions that are less related to the post, sometimes E-SCBA is unable to produce emotion keywords that match the conversation, like the sencod case of *Angry* shown in the table.

4.5.2 Real-Situation Study

This section is based on a movie subtitles dataset. We sampled some representative examples to indicate potential problems in our framework and to do a further error analysis. Table 5 shows the replies generated from E-SCBA in a true situation, where a conversation has only one corresponding emotion.

The replies in the first four lines is the normal cases that represent the normal situation when E-SCBA performs well in the test. However, the replies in the last four lines, all have flaws of logic or emotion, where we divide them into four different types. **First**, for the reply in the count down to the fourth line, the classifier obviously does not analyze the emotion of the conversation correctly. This is a extreme case, and the wrong category negatively impacts the quality of the reply, especially for the emotion with opposite polarity. Although the purpose of this work is not to improve the accuracy of emotional classification, increasing data of the minority category to eliminate the imbalance in the future may be a measure to reduce the occurrence of such situation. **Second**, for the reply in the antepenultimate line, the emotion keyword has virtually no emotional tendency in this conversation. We think the main reason for this problem is the polysemy of words, meaning that a word can express different emotions or no emotion in different contexts. For instance, the word "like" is an emotion keyword when it denotes *enjoy*, but it is actually a generic word when it

means *similar*. Same situations always occur in Chinese. It is not appropriate to the expectation of conversation although the lack of emotion has only a slight effect on the logic. **Third**, for the reply in the penultimate line, its meaningless topic keyword (in Chinese) is dispensable in this case. This kind of short posts is suitable for the general seq2seq instead of E-SCBA if there is no mistake in the *Structure Detector*. Keywords generated by the LDA method can indeed cover most conversations, but the lack information of short sentences make the extraction much difficult. **Last but not least**, for the reply in the last line, the emotion keyword and topic keyword are incompatible, where the word *hooligan* and *looking forward to* is not a suitable combination. This situation is caused by the fact that emotion keyword and topic keyword are generated independently before decoding, and this independence may cause problems of suitability in the reply. Even though the emotion is appropriate and the topic keyword is meaningful, the reply does not match expected logic.

5 Conclusion and Future Work

In this paper, we proposed a syntactically constrained, bidirectional-asynchronous framework to introduce both emotion and topic knowledge into the process of generating emotional conversation. In contrast to the existing methods, which generate replies from the first word to the end, we started with the pre-generated keywords, which respectively represent the emotion and the gist of conversations, in the arbitrary positions. The newly designed decoder in E-SCBA makes use of syntactic relation to constrain the generation, which helps to ensure the fluency and grammaticality of reply. The results of both subjective and objective experiments show that our model can generate replies that feature both emotion, logic and higher diversity. Finally, the case study showed the advantages of E-SCBA and the error analysis revealed some potential issues during the generation.

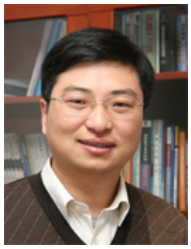
In the future, we will explore the scalability of this framework. In other words, the conversation generation framework we proposed can be more flexible than its use in this study. The knowledge can not only be emotion or topic but also verb tense, attitude (affirmative or negative) or other more complex concepts. Based on the given information, we can customize the framework to generate replies that meet the special requirements. Another place should be noticed is the size of topic dictionary. If the scale is small, the large topic can lead to loss of context information. But if the scale is large, the increased number of topic keywords will lead to a decrease in the accuracy of classifier. The balance between these two aspects also needs an in-depth analysis. Finally, we will introduce the reinforcement learning for a further research. In our work, The accuracy of the upstream model, such as emotion classifier, will greatly affect the quality of reply generated by the downstream model (decoder). A wrong answer of classification may result in poor quality of reply. We think the reinforcement learning is a good way to address this issue. Reward based on the quality of reply can directly affect the parameters update of the upstream model. This will stimulate the upstream model to be more inclined to generate better keyword to get more reward.

Acknowledgment

The authors would like to thank Zhou for the use of the NLPCC 2017 Shared Task Sample Data: Emotional Conversation Generation as training dataset. The work was supported by the State Key Program of the National Natural Science of China (61432004, 61461045). This work was partially supported by the China Postdoctoral Science Foundation funded project (2017T100447). This work was also supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Xiao Sun was born in 1980. He received his M.E. in 2004 from the Department of Computer Sciences and Engineering at Dalian University of Technology and received a double doctorate from Dalian University of Technology (2010) in China and the University of Tokushima (2009) in Japan. He is now working as an associate professor in the AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machines at Hefei University of Technology. His research interests include affective computing, natural language processing, machine learning and human-machine interaction.

Jingyuan Li was born in 1997. He is currently studying for the B.S. degree at the School of Computer and Information, Hefei University of Technology, China, where he will graduate in 2019. His research interests include natural language understanding, affective computing and dialogue generation.



Jianhua Tao received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include speech synthesis and recognition, human–computer interaction, and emotional information processing. He has authored over 60 papers in major journals and proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, Interspeech, ICME, ICPR, ICCV, and ICIP. In 2006, he was elected as the Vice Chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing, and an Executive Committee Member of the HUMAINE Association. He is the Editorial Board Member of the Journal on Multimodal User Interfaces and the International Journal of Synthetic Emotions, and the Steering Committee Member of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

REFERENCES

[1] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion," in *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, 2002, pp. 43–46.

[2] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, 2017.

[3] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, Texas*, 2017.

[4] Y. Liu, X. Hou, J. Chen, C. Yang, G. Su, and W. Dou, "Facial expression recognition and generation using sparse autoencoder," in *International Conference on Smart Computing*, 2014, pp. 125–130.

[5] A. Mollahosseini, B. Hassani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1509–1516.

[6] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[7] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, 2014.

[8] B. Martinovski and D. Traum, "Breakdown in human-machine interaction: the error is the clue," in *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, 2003, pp. 11–16.

[9] M. Skowron, "Affect listeners: Acquisition of affective states by means of conversational systems," in *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer, 2010, pp. 169–181.

[10] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models." in *AAAI*, 2016, pp. 3776–3784.

[11] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation." in *AAAI*, vol. 17, 2017, pp. 3351–3357.

[12] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," 2017.

[13] L. Mou, Y. Song, R. Yan, G. Li, L. Zhang, and Z. Jin, "Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3349–3358.

[14] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "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)*, vol. 1, 2017, pp. 634–642.

[15] S. Han, J. Bang, S. Ryu, and G. G. Lee, "Exploiting knowledge base to generate responses for natural language dialog listening agents." in *SIGDIAL Conference*, 2015, pp. 129–133.

[16] Z. Yu, Z. Xu, A. W. Black, and A. I. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems." in *SIGDIAL Conference*, 2016, pp. 404–412.

[17] T. Wen, D. Vandyke, N. Mrkšíc, M. Gašíc, L. Rojas-Barahona, P. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, vol. 1, 2017, pp. 438–449.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[19] V. M. B. G. C. B. D. B. F. S. H. B. Y. Cho, Kyunghyun, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1577–1586.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[23] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.

[24] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 994–1003.

[25] R. W. Picard, "Affective computing," *The MIT Press, Cambridge (MA)*, vol. 167, p. 170, 1997.

[26] S. L. Lutfi, F. Fernández-Martínez, J. M. Lucas-Cuesta, L. LóPez-LebóN, and J. M. Montero, "A satisfaction-based model for affect recognition from conversational features in spoken dialog systems," *Speech Communication*, vol. 55, no. 7-8, pp. 825–840, 2013.

[27] F. Keshtkar and D. Inkpen, "A pattern-based model for generating text to express emotion," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 11–21.

[28] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 280–290.

[29] K. Takabuchi, N. Iwahashi, and T. Kunishima, "A language acquisition method based on statistical machine translation for application to robots," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016 Joint IEEE International Conference on*. IEEE, 2016, pp. 300–301.

[30] J. Hou, S. Zhang, L. Dai, and H. Jiang, "Feedforward sequential memory networks based encoder-decoder model for machine translation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 622–625.

[31] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2157–2169.

[32] Q. Qian, M. Huang, and X. Zhu, "Assigning personality/identity to a chatting machine for coherent conversation generation," *arXiv preprint arXiv:1706.02861*, 2017.

[33] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Eprint Arxiv*, 2014.

[34] L. Xu, H. Lin, Y. Pan, H. Ren, and J. Chen, "Constructing the affective lexicon ontology," *Journal of the China Society for Scientific & Technical Information*, 2008.

[35] X.-H. Phan and C.-T. Nguyen, "Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda)," *Tech. rep.*, 2007.

[36] X. Sun, C. Zhang, S. Ding, and C. Quan, "Detecting anomalous emotion through big data from social networks based on a deep learning method," *Multimedia Tools and Applications*, pp. 1–22, 2018.

[37] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[38] L. Mou, R. Yan, G. Li, L. Zhang, and Z. Jin, "Backward and forward language modeling for constrained sentence generation," *arXiv preprint arXiv:1512.06612*, 2015.

[39] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2122–2132.

[40] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.