A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation

Jingyuan Li, Xiao Sun*
Anhui Province Key Laboratory of Affective Computing
and Advanced Intelligent Machine
Hefei University of Technology, Hefei, China
lijingyuan@mail.hfut.edu.cn, sunx@hfut.edu.cn

Abstract

Traditional neural language models tend to generate generic replies with poor logic and no emotion. In this paper, a syntactically constrained bidirectional-asynchronous approach for emotional conversation generation (E-SCBA) is proposed to address this issue. In our model, pre-generated emotion keywords and topic keywords are asynchronously introduced into the process of decoding. It is much different from most existing methods which generate replies from the first word to the last. Through experiments, the results indicate that our approach not only improves the diversity of replies, but gains a boost on both logic and emotion compared with baselines.

1 Introduction

In recent years, as artificial intelligence has developed rapidly, researchers are pursuing technologies with greater similarities to human intelligence. As a subjective factor, emotion performs an elemental difference between humans and machines. In other words, machines that could understand emotion would be more responsive to human needs. For example, in education, positive emotions improve students' learning efficiency (Kort et al., 2002). In healthcare, mood prediction can be used in mental health counseling to help anticipate and prevent suicide or depression (Taylor et al., 2017; Jaques et al., 2017). To make machine more intelligent, we must resolve the conundrum of emotional interactions.

There are tons of researches about conversation, an important channel for communication between humans. And lots of work has recently been carried out in open-domain conversation devoted to generating meaningful replies (Vinyals and Le, 2015; Li et al., 2016; Serban et al., 2016). Unfortunately, the factors considered in these methods only concerns topic, like (Xing et al., 2017), where they failed to take emotion into account. Unlike the former, the work in (Zhou et al., 2017) first addressed the emotional factor in large-scale conversation generation, and it showed that emotional replies obtain superior performances compared to the baselines that did not consider emotion. However, two defects still manifest themselves in the aforementioned models. **First**, all methods above only adopted a single factor (i.e., topic or emotion), because of which the bias of information can not comprehensively summarize the human conversations to achieve favorable results. **Second**, the way that generates replies from the first word to the last can lead to a decline in diversity, limited by the high-frequency generic words in the beginning (e.g., I and you), as argued in (Mou et al., 2016).

The deficiencies above inspire us to introduce a new approach called E-SCBA, studying both emotion and topic. Three main contributions are presented in this paper: (1) It conducts a study of compound information, which constitutes the syntactic constraint in the conversation generation. (2) Different from the work in (Mou et al., 2016), a bidirectional-asynchronous decoder with multi-stage strategy is proposed to utilize the syntactic constraint. It ensures the unobstructed communication between different information and allows a fine-grained control of the reply to address the problem of fluency and grammaticality as argued in (Ghosh et al., 2017; Zhou et al., 2017). (3) Our experiments show that E-SCBA work better on emotion, logic and diversity than the general seq2seq and other models that consider only a single factor during the generation.

2 Model

2.1 Overview

The whole process of emotional conversation generation consists of the following three steps:

*The corresponding author of this paper.

***Step I*:** Given a post, we first use two networks combined with category embeddings to respectively predict emotion keyword and topic keyword that should appear in the final reply (see Section 2.2).

***Step II*:** After the prediction, a newly designed decoder is used to introduce both keywords into the content[1], as shown in Figure 1. It first produces a sequence of hidden states based on the emotion keyword (Step I), and then uses an emotional attention mechanism to affect the generation of middle sequence, which is based on the topic keyword (Step II). The remaining two sides are ultimately generated by the combination of middle part and keywords (Step III). A detailed description is given in Section 2.3.

***Step III*:** Finally, a direction selector is used to arrange the generated reply in a logically correct order by selecting the better one from forward and backward forms of the reply generated in the last step (see Section 2.4).

In this work, we default that the replies contain at least one emotion keyword and one topic keyword, which are expected to appear in the dictionaries we used.

2.2 Keyword Predictor

The keywords to be selected are pre-stored in the prepared dictionaries. The adopted emotion dictionary was proposed by (Xu et al., 2008), which contains 27,466 emotion words divided into 7 categories: *Happy*, *Like*, *Surprise*, *Sad*, *Fear*, *Angry* and *Disgust*. The adopted topic dictionary was obtained by the LDA model (Blei et al., 2003), including 10 categories with 100 words for each category. And to avoid situations in which emotion and topic keywords are predicted to be the same word, all the overlapping words in these two dictionaries default to emotion keywords.

The prediction of emotion and topic keywords follows the similar path. We first derive topic category and emotion category from the post with two classifiers separately. To be more specific, the pretrained LDA model is used for the topic category inference. And the work in (Sun et al., 2018) is applied for emotion. The concrete model is an emotion transfer network. Given a specific external stimuli (e.g., a sentence), the network produce an emotional response, which is specifically an emotion category in this work. After this, combining the sum of hidden states $\tilde{h} = \sum_{i=1}^{T} h_i$ from encoder and the category embeddings $\boldsymbol{k} = \{k^{et}, k^{tp}\}$, keywords are predicted as follows:

$$p(w_{et}^k|\boldsymbol{x}, k^{et}) = softmax(\mathbf{W}_{et}^w[\tilde{h}; k^{et}]) \quad (1)$$

$$p(w_{tp}^k|\boldsymbol{x}, k^{tp}) = softmax(\mathbf{W}_{tp}^w[\tilde{h}; k^{tp}]) \quad (2)$$

where w_{et}^k and w_{tp}^k separately represent the emotion keyword and topic keyword that are expected to appear in the reply.

2.3 Bidirectional-Asynchronous Decoder

Due to the decoder architecture shown in Figure 1, we suppose the reply in this section is $\boldsymbol{y} = (\boldsymbol{y}^{ct}, w_{tp}^k, \boldsymbol{y}^{md}, w_{et}^k, \boldsymbol{y}^{ce})$[2] where $\boldsymbol{y}^{md}$ is the middle part between two keywords and $\boldsymbol{y}^{ct}$, $\boldsymbol{y}^{ce}$ represent the remaining sides connected to the topic keyword and emotion keyword. The generation of middle part $\boldsymbol{y}^{md} = (y_1^{md}, ..., y_K^{md})$ can be described as follows:

$$c_j^{et} = f_{att}^{et}(s_{j-1}^{tp}, \{s_i^{et}\}_{i=1}^{K'}) \quad (3)$$

$$p(\boldsymbol{y}^{md}|\boldsymbol{x}, \boldsymbol{w}^k) = \prod_{j=1}^{K} p(y_j^{md}|y_{j-1}^{md}, s_j^{tp}, c_j^{et}) \quad (4)$$

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

$$c_j^{et} = \sum_{i=1}^{K'} \alpha_{j,i}^{et} s_i^{et} \quad (5)$$

$$\alpha_{j,i}^{et} = \frac{\exp(e_{j,i}^{et})}{\sum_{t=1}^{K'} \exp(e_{j,t}^{et})} \quad (6)$$

$$e_{j,i}^{et} = (\mathbf{v}_\alpha^{md})^T tanh(\mathbf{W}_\alpha^{md} s_{j-1}^{tp} + \mathbf{U}_\alpha^{md} s_i^{et}) \quad (7)$$

where $e_{j,i}^{et}$ represents the impact scores of the emotion state s_i^{et} on the topic state s_{j-1}^{tp}.

After generating the middle part, we connect it with the keywords to form a new sequence. Two seq2seq models are used to encode the connected

[1] Syntactic constraint starts to work here, and can be intuitively interpreted as relative positions of emotion words and topic words, as well as different combinations between them.

[2] For the training data in opposite direction, we reversed the target replies to meet the requirement.

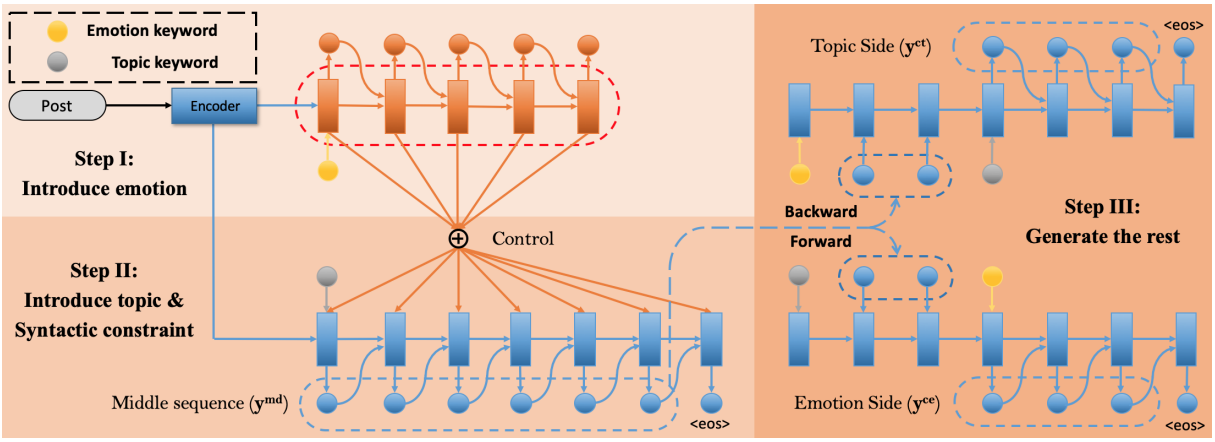

Figure 1: The process of generating replies in the test. The middle part of the reply is generated in Steps I and II, and the remaining two sides are generated in Step III. The RNN networks used in the decoder do not share the parameters with each other.

sequences and decode $\boldsymbol{y}^{ce} = (y_1^{ce}, ..., y_M^{ce})$ and $\boldsymbol{y}^{ct} = (y_1^{ct}, ..., y_N^{ct})$, as below:

$$p(\boldsymbol{y}^{ce}|\boldsymbol{w}^k, \boldsymbol{y}^{md}) = p(\boldsymbol{y}^{ce}|[w_{tp}^k, \boldsymbol{y}^{md,f}, w_{et}^k]) \quad (8)$$

$$p(\boldsymbol{y}^{ct}|\boldsymbol{w}^k, \boldsymbol{y}^{md}) = p(\boldsymbol{y}^{ct}|[w_{et}^k, \boldsymbol{y}^{md,b}, w_{tp}^k]) \quad (9)$$

where $\boldsymbol{y}^{md,f}$ and $\boldsymbol{y}^{md,b}$ are the forward and backward situations of the middle part, respectively.

2.4 Direction Selector

To make the samples meet the requirements of decoder, by default we place the topic keyword as the first keyword on the left and the emotion keyword on the right in training. However, in real situations, the topic keyword does not always appear before the emotion keyword, where we must determine correct direction by the machine.

By connecting the results in the preceding section, we get $\boldsymbol{y}^f = (\boldsymbol{y}^{ct,b}, w_{tp}^k, \boldsymbol{y}^{md,f}, w_{et}^k, \boldsymbol{y}^{ce,f})$ as the forward situation and $\boldsymbol{y}^b$ means the backward situation. GRU networks are used as encoders to process sequences in different situations, which do not share parameters. And the direction is predicted by:

$$p(d|\boldsymbol{y}^f, \boldsymbol{y}^b) = sigmoid(\mathbf{W}^d[\tilde{h}^{d,f}, \tilde{h}^{d,b}]) \quad (10)$$

$$\tilde{h}^{d,*} = \sum_{i=1}^{T'} \mathbf{GRU}(y_i^*) \quad (11)$$

where $* \in \{f, b\}$ means forward or backward. After the operation completes, one of the sequences $\boldsymbol{y}^f$ and $\boldsymbol{y}^b$ should conform to our expectations.

3 Experiment

3.1 Data

We evaluated and trained E-SCBA on the emotional conversation dataset NLPCC2017. There are a total of 1,119,201 Chinese post-reply pairs in the set. The dictionaries mentioned in Section 2.2 were used to mark the conversation. The cases whose replies contain both emotion keywords and topic keywords account for 42.6% (476,121) of the total[3], which are suitable data for the bidirectional-asynchronous decoder. We randomly sampled 8,000 for validation, 3,000 for testing and the rest for training. We also sampled another 60,000 pairs from the training set to train the LDA model[4] mentioned in Section 2.2. Besides, an error analysis is presented based on a Chinese movie subtitle dataset which is collected from the Internet.

3.2 Metrics

To evaluate our approach, we use the metrics as below:

Embedding-based Metrics: We measure the similarity computed by cosine distance between a candidate reply and the target reply using sentence-level embedding, following the work in (Liu et al., 2016; Serban et al., 2017).

[3]Please note that we did not use the original labels of the dataset, but the emotion categories of the keywords as labels to avoid unnecessary bias. For cases that contain multiple topic keywords or emotion keywords, we chose the keywords that appear less frequently to reduce imbalances.

[4]High frequency words and stop words, which have no benefit to the topics, were removed in advance.

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**
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**

Table 1: The results of human annotations (C = Consistency, L = Logic, E = Emotion).

Method	G–M	E–A	V–E	D–1	D–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 2: The results of automatic evaluation (G–E = Greedy Matching, E–A = Embedding Average, V–E = Vector Extrema).

Distinct Metrics: By computing the number of different unigrams (Distinct-1) and bigrams (Distinct-2), we measure information and diversity in the candidate replies, following the work in (Li et al., 2016; Xing et al., 2017).

Human Annotations: We asked four annotators to evaluate the replies[5] generated from our approach and baselines from *Consistency*, *Logic* and *Emotion*. *Consistency* measures fluency and grammaticality of the reply on a three-point scale: 0, 1, 2; *Logic* measures the degree to which the post and the reply logically match on a three-point scale[6] as above; *Emotion* judges whether the reply includes the right emotion. A score of 0 means the emotion is wrong or there is no emotion, and a score of 1 is the opposite.

3.3 Baselines

In the experiments, E-SCBA is compared with the following baselines:

S2S: the general seq2seq model with attention method (Bahdanau et al., 2014).

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

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

S2S-AW: the model uses an asynchronous method the same as (Mou et al., 2016).

The synchronous method in S2S-STW and S2S-SEW was mentioned in (Mou et al., 2015), acting as the contrast to the asynchronous models.

3.4 Results and Discussion

The results of automatic evaluation are shown in Table 2. Compared with the best model (S2S-AW) that considers only a single factor, E-SCBA makes significant improvement on the distinct metrics (+0.056 and +0.165), which verifies the effectiveness of taking both emotion and topic information into account to improve the diversity. Likewise, our approach also respectively achieves 0.042, 0.068 and 0.043 gains on G-M, E-A and V-E, benefiting from the compound information that captures the thrust of human conversation so that E-SCBA has a better ability to learn the goal distribution. Furthermore, the grades of the asynchronous models are higher than the synchronous models on both kinds of metrics, showing that the asynchronous method is a more suitable way for content-introducing conversation generation.

Table 1 depicts the human annotations (*t-test*: $p < 0.05$ for C and L, $p < 0.01$ for E). Overall, E-SCBA outperforms S2S-AW on all three metrics, where the compound information plays a positive role in the comprehensive promotion. However, in *Surprise* and *Angry*, the grades of *Consistency* and *Logic* are not satisfactory, since the data for them are much less than others (*Surprise* (1.2%)

[5] 700 conversations in total, 100 for each emotion category, were sampled randomly from the test set.

[6] If a reply is too short or turns up frequently, it would be annotated as either 0 or 1 (if the annotator thought the reply related to the post), like "Me too" and "I think so".

Post	Emotion	Chinese	English(translated)
受当局追捕，我们只能秘密活动。 Hunted by the authorities, we work in secret.	Disgust	有一种被嘲讽的感觉。	There is a **sense** of being **mocked**.
我一直在观察你们的制作过程。 I have been observing your production process.	Happy	我希望你继续努力	I **hope** you keep **working hard**
假如你那天阻止了他，事情 就不会这样了。 If you had stopped him that day, things would have been different.	Sad	不要回忆过去的日子了！	Don't **look back** on the **past**!
普通民众中也可以拥有某些 权利的概念。 The general public can also have some concepts of rights.	Like	可以作为他们生活的一部分	This can **be** a part of their **life**
我们现在也可以离开这里了。 We can also leave here now.	Sad	挺遗憾的。	What a **pity**.
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.	Disgust	我期待的是像你这样的流氓	I am **looking forward to** a **hooligan** like you

Table 3: Sampled conversations with a corresponding emotion from the Chinese movie subtitle data.

and *Angry* (0.7%)). Besides, the score of *Emotion* in *Surprise* has a big difference from others. We think the reason is that the characteristic of *Surprise* overlaps with other categories that have much more data, such as *Happy*, which interferes with the learning efficiency of the approach in *Surprise*. Meanwhile, it is harder for annotators to determine which one is the right emotion.

3.5 Case Study and Error Analysis

In this section, we sampled some typical cases from a Chinese movie subtitle dataset to do a further error analysis. The cases are shown in Table 3. The post of weibo and movie subtitle are applied in different scenes to obey different distributions. The weaker correlation between training sets and test sets can present a more reliable study.

The first three conversations are positive samples and others are negative samples that have content with flaws. For the reply in the antepenultimate line, its problem is the faint emotion. Since the emotion keyword in this sentence is a polysemic word, and it expresses a meaning with no emotion here. Under diverse circumstances, a polysemic word probably have different meanings, emotional or neutral. For example, the word "like" can be a generic word when it denotes *similar*, but it can also be an emotion word when it denotes *enjoy*. Same situation also occurs in Chinese. Besides, we notice that if the LDA model pick a meaningless topic keyword from the dictionary, our approach may have a difficulty in generating a diverse and long reply, as the reply in the penultimate line. The lack of information causes generic replies which are consisted of few words generated from the networks. The last line presents another limitation. The emotion keyword *hooligan* corresponds to the post and the topic keyword *looking forward to* is meaningful, but the combination of them, *looking forward to* a *hooligan*, does not conform to the normal logic. This situation is caused by the fact that two kinds of keywords are generated independently before decoding, and it may cause a mismatch. In the future, we will try to explore different network architectures to make keywords interact with each other during the generation.

4 Conclusion

In this paper, we proposed a novel conversation generation approach (E-SCBA) to make a more comprehensive optimization for the quality of reply, which introduces both emotion and topic knowledge into the generation. The newly designed decoder makes use of syntactic knowledge to constrain generation and ensures fluency and grammaticality of reply. Experiments show that our approach can generate replies that have rich diversity and feature both emotion and logic.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. 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)*, volume 1, pages 634–642.

Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. 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*.

B. Kort, R. Reilly, and R. W. Picard. 2002. 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*, pages 43–46.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119.

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

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Lili Mou, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2015. Backward and forward language modeling for constrained sentence generation. *arXiv preprint arXiv:1512.06612*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xiao Sun, Chen Zhang, Shuai Ding, and Changqin Quan. 2018. Detecting anomalous emotion through big data from social networks based on a deep learning method. *Multimedia Tools and Applications*, pages 1–22.

Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, volume 17, pages 3351–3357.

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

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.