JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 429-444 (2022)
DOI: 10.6688/JISE.202203_38(2).0009

TRGM: Generating Informative Responses for Open Domain Dialogue Systems

WANG GAO+, HONGTAO DENG+, XUN ZHU AND YUWEI WANG
School of Artificial Intelligence
Jianghan University
Wuhan, 430056 P.R. China
E-mail: {gaow; hongtaodeng; zhuxun; weberwang}@jhun.edu.cn

Sequence-to-sequence (seq2seq) neural network models are able to generate natural sounding conversational responses for open domain dialogue systems. However, these models tend to produce safe, universal responses (*e.g.*, I don't know) regardless of the input, which carry little information and can easily lead to the end of a conversation. In this paper, we propose a new Topic-driven Response Generation Model (TRGM). The proposed model leverages topic information to generate interesting and informative responses. Firstly, we design a topic generation model based on BERT to learn the topic information of the input. Then a response generation model utilizes a gate mechanism and a mixed probability model to integrate topic knowledge into a seq2seq model. We implement the two components using an end-to-end neural network and jointly train each component as a sub-task. Experimental results on a public dataset demonstrate that our method significantly outperforms state-of-the-art baselines on both automatic evaluation metrics and human judgment.

Keywords: response generation, open domain dialogue systems, topic model, bert, CRFTM

1. INTRODUCTION

With the rapid development of Artificial Intelligence (AI), Natural Language Processing (NLP) has been widely used in recommendation systems, topic evolution analysis, document summarization, social network analysis and so forth [1–4]. Dialogue systems, also known as conversation systems, virtual agents and chatbots, have gained increasing attention because of the promising potentials on applications such as intelligent customer service or virtual assistants [5]. Dialogue systems include goal-driven dialogue systems that are built for specific tasks in vertical domains, and open domain dialogue systems (also known as chatbots) whose purpose is to enable natural and human-like chit-chat with individuals on open domain issues [6]. Previous research on dialogue systems has focused on goal-driven dialogue systems. Recently, with the generation of a large amount of conversational data on the Internet and the great improvement of computing power, open domain dialog systems have attracted increasing attention in industry and academia.

Received April 17, 2020; revised August 2, 2020; accepted August 31, 2020.
Communicated by Berlin Chen.
+ Corresponding author.

Early open domain dialog systems were mainly based on retrieval methods. These methods first retrieve the conversational corpus, and then analyze the corpus by using a ranking model to find the most appropriate response. However, such systems are less efficient and can only produce responses that appear in the corpus. As a method of mapping one sequence to another, the sequence-to-sequence (seq2seq) framework has made remarkable progress in many NLP fields [7–9]. Essentially, the seq2seq framework is an encoder-decoder model. The encoder first converts the input sequence into a specific representation, and then the decoder converts it into an output sequence. A dialogue system can benefit directly from seq2seq because it requires a mapping between messages and responses. Therefore, learning a neural generative conversational model based on encoder-decoder from a large number of message-response pairs has become a common practice for building a chatbot. However, the neural generative conversational model tends to generate dull or repetitive, universal responses with little useful information, which often contain high frequency phrases such as "me too" or "I don't know". In addition, traditional seq2seq models only learn local information of the dialogue corpus, and generate responses based on the previous sentence of the dialogue, which is easy to generate inconsistent and context-independent responses. Although these responses are safe for replying to a lot of messages, they are boring and severely disrupt the user experience of chatbots.

In human-to-human conversation, people usually associate conversational information with topic-related concepts in their minds. Based on these topic-related concepts, they choose words for responses and organize sentences. For instance, in response to "I feel stressed", people tend to think it is a "stress" problem that can be alleviated by "listening to music" or "talking to friends". Based on the topic information, they would give more meaningful answers such as "you can listen to some music" rather than trivial answers such as "me too". The meaningful response allows other people to follow the topic and keep talking about how to release stress. "Stress", "music" and "friends" are thematic concepts related to the conversation, which indicate people's prior knowledge in the dialogue. When responding, people will choose information related to these concepts in their response, and even employ the concepts directly to build the basis of response sentences.

Inspired by this, in this paper, we propose a novel topic-driven model to deal with the above challenges. The main idea of our model comes from the answers to the following two questions: (1) How to combine topic modeling and deep learning techniques to learn more coherent topic representations for response generation? (2) How to incorporate topic knowledge into a seq2seq structure to generate informative responses for open domain dialog systems?

Specifically, we propose a Topic-driven Response Generation Model (TRGM), which has two components: a topic generation model and a response generation model. We employ neural networks to implement these two components, and jointly train them by using each component as a sub-task in a multi-task learning environment. The topic generation model utilizes the Bidirectional Encoder Representations from Transformers (BERT) model to learn topic vectors and topic words. BERT is a pre-trained language representation method that can be fine-tuned for downstream NLP tasks [10–12]. Furthermore, BERT considers not only individual words, but also the context. It enables the model to understand specific words in a sentence, which helps to learn a better topic rep-

resentation. The response generation model is built on a seq2seq framework that exploits topic-related information as prior knowledge for generating responses. In the encoding process, the model represents an input utterance as hidden vectors by an utterance encoder, and obtains topic vectors and topic words of the utterance from the topic generation model. Topic vectors and topic words are used to simulate topic-related concepts in people's minds. In decoder, TRGM not only leverages hidden states, but also uses the topic knowledge to generate each response word. In addition, our model utilizes a mixed probability model of a normal mode and a topic mode to increase the possibility that topic words appear in the response. The normal mode is a traditional seq2seq generation mode, while the topic mode selects words from the topic word set. We conduct extensive experiments on large scale open domain dialog data, and compare TRGM with different methods using both automatic evaluation and human comparison. Experimental results indicate the proposed model is capable of generating more informative responses, and significantly outperform baseline methods. The main contributions of this paper are summarized as follows:

1. We propose a novel Topic-driven Response Generation Model (TRGM) for open domain dialogue systems. TRGM employs a BERT-based topic generation model to learn topic-related information of the conversation. The topic generation model works similarly to an auto-encoder, in which words from utterances are taken as input and optimized for prediction.

2. TRGM utilizes a gate mechanism and a mixed probability model to naturally incorporate topic information into a seq2seq framework to increase the possibility that topic words appear in the response. To the best of our knowledge, this is the first work to integrate topic knowledge based on BERT into a seq2seq framework.

3. The performance of TRGM is evaluated on an open domain dialog corpus against state-of-the-art baseline models. Experimental results show that the proposed model can generate more informative responses, and outperforms other baselines in both automatic evaluation and human judgment.

2. RELATED WORK

The response generation of open domain dialog systems has gradually changed from a simple retrieval model based on similarity measurement to a seq2seq generation problem [13–17]. Although retrieval-based models do not generate grammatically incorrect responses, they can only produce output that has appeared before. Therefore, the performance of these models is limited by the size of the pre-constructed response repository [7]. On the other hand, generation-based models learn a language model and could produce highly coherent responses not seen in the training set. However, a common problem with these approaches is that they may produce very short or most generic output with insufficient information such as "me too".

There have been several recent studies to deal with this challenge. Shang *et al*. proposed a Neural Response Machine (NRM), which is a general encoder-decoder framework for response generation [18]. Nevertheless, NRM tends to produce universal or

trivial responses, usually involving high-frequency phrases like "thank you" or "I am fine". To solve the problem of universal responses produced by seq2seq methods, Li *et al*. proposed a neural network that uses Maximum Mutual Information (MMI) as the objective function, which measures the mutual dependence between input utterances and responses [13]. Experimental results validate that their MMI model is able to generate more appropriate and diverse responses. Xing *et al*. proposed a Hierarchical Recurrent Attention Network (HRAN) to model words and utterances in a unified framework [6]. Ghazvininejad *et al*. proposed a knowledge-based neural conversation model, which generates responses based not only on conversation history, but also on relevant factual content [19]. Qin *et al*. proposed a novel end-to-end neural dialogue model that jointly models on-demand machine reading and response generation, which can be regarded as an expansion of the knowledge-based neural conversation model [20]. However, the above approaches do not consider the topical relationship between input and responses, which helps generate interesting and informative responses.

There has been a lot of research on introducing topic information into open domain dialogue systems. Wu *et al*. proposed a topic aware convolutional neural tensor network, which not only utilizes a input vector and a response vector generated by Convolutional Neural Networks (CNN) to match between the input and the response, but also exploits two topic vectors to encode additional topic information [21]. After obtaining topic words by a pre-trained Latent Dirichlet Allocation (LDA) [22] model, the two topic vectors are linear combinations of these words of the input and the response, respectively. Similarly, Xing *et al*. designed a topic aware seq2seq model that uses topic words of messages obtained from LDA [23]. However, probabilistic topic models such as LDA are difficult to integrate directly into a seq2seq neural network. Our research has a similar motivation as [21, 23], but we do not leverage a pre-trained LDA model to learn topic words. In contrast, we adopt a topic generation model based on BERT to capture topic information, and integrate the topic generation model and the response generation model in a unified framework. To the best of our knowledge, this is the first attempt to incorporate BERT-based topic knowledge into a seq2seq network.

3. TOPIC-DRIVEN RESPONSE GENERATION MODEL

Fig. 1 illustrates the architecture of the proposed model. As shown in the figure, there are two components in TRGM: a topic generation model and a response generation model. The topic generation model learns topic information for the conversation, while the response generation model integrates the topic knowledge into a seq2seq framework to generate informative responses.

3.1 Topic Generation Model Component

In open domain dialogue systems, dialogues are mostly short texts, which are characterized by short length, sparse features and limited word co-occurrence information. Traditional topic models such as LDA, rely on mining word co-occurrence information in the document to perform topic modeling, which work poorly on short texts. In addition, topic models for short texts such as CRFTM and Twitter LDA [24, 25], are not fully applicable to neural dialogue models based on a seq2seq framework, and they cannot be

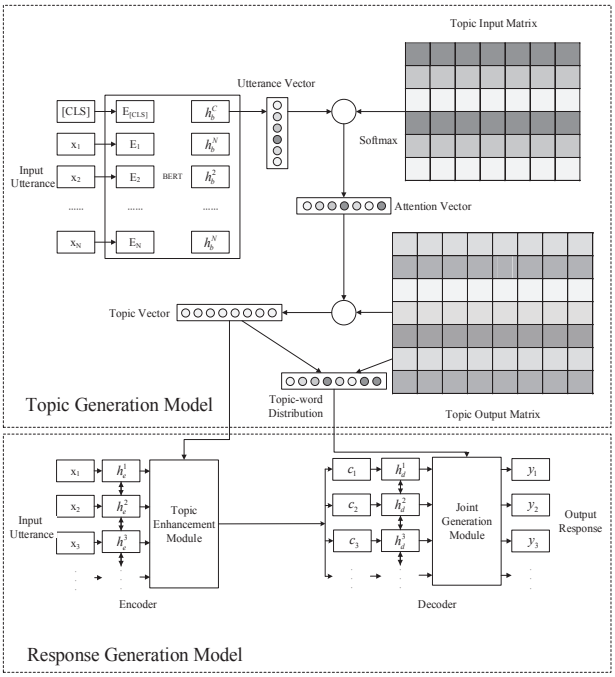


Fig. 1. The architecture of TRGM.

jointly trained, which may lead to error accumulation problems [26]. Therefore, in the topic generation model, BERT is used to encode input utterances. BERT is based on a Transformer structure [27], which has proven to be a better feature extraction network than CNN and RNN. After pre-training on a large-scale corpus, BERT can capture rich semantic information of words, which is helpful for training the proposed model.

Let $U = \{x_1, x_2, ..., x_i, ..., x_N\}$ be an input utterance, where N denotes the length of the utterance and x_i denotes the ith token in the utterance. All tokens are represented as token embeddings, segment embeddings and position embeddings. Token embeddings are embedding representations of words, and segment embeddings are employed to distinguish different utterances. Since the encoding of utterances is position independent, BERT leverages position embeddings to capture the sequential ordering information of utterances. Token set U is replaced by embedded tokens $E = \{E_1, E_2, ..., E_i, ..., E_N\}$ as input to the topic generation model during the embedding process. After that, BERT utilizes self-attention and multi-head attention to encode E into hidden states $H_b = \{h_b^1, h_b^2, ..., h_b^i, ..., h_b^N\}$. Attention functions can be thought of as mapping a series of key-value pairs and queries to the output [27]. In the self-attention mechanism, the weighted sum of values can be used to compute the output of queries. The weight of values can be computed by the dot product of keys and queries. Values, keys and queries are represented by matrices V, K and Q, and the out matrix is calculated as follows:

$$Attention(V, K, Q) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where d_k represents the dimension of keys and queries.

To increase the diversity, multi-head attention enables the BERT model to capture multiple relationships at different positions in an utterance. All heads are connected, and the multi-head attention is as follows:

$$\begin{aligned} MultiHead(Q,K,V) &= concat(head_1,...,head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{2}$$

where $W_i^O \in \mathbb{R}^{hd_{model} \times d_v}, W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^K \in \mathbb{R}^{d_{model} \times d_k}, W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ denote parameter matrices, and $concat(\cdot)$ represents a concatenation function. In the topic generation model, we set $h = 12, d_k = d_v = d_{model}/h = 64$. [CLS] is added to the input sequence as the first token, and its hidden state h_b^C is used as the input utterance vector V_i.

Topic-related information is stored in two lookup matrices: topic input matrix $\mathbb{I} \in \mathbb{R}^{k \times d_k}$ and topic output matrix $\mathbb{O} \in \mathbb{R}^{d_o \times k}$, where k denotes the number of topics and d_o represents the dimension of topic vectors. $\mathbb{I}$ and $\mathbb{O}$ are two parameter matrices that need to be learned. To align topic vectors and input utterance vector V_i, we calculate attention vector V_a by topic input matrix $\mathbb{I}$:

$$V_a = softmax(\mathbb{I}V_i), \tag{3}$$

where $V_a \in \mathbb{R}^k$. Inspired by the generative process of LDA, V_a represents a probability distribution over hidden topics for each utterance. Topic output matrix $\mathbb{O}$ and V_a can then be used to compute an utterance-topic representation:

$$V_o = \mathbb{O}V_a, \tag{4}$$

where $V_o \in \mathbb{R}^{d_o}$. V_o is essentially a weighted average of topic vectors , and the weight is determined by the attention vector V_a. Similar to topic models, our model considers an utterance as a mixture of probabilistic topics. Finally, TRGM links utterance-topic representation V_o to a full connection layer with a softmax output layer to predict each word in the utterance, and optimizes the model by using categorical cross-entropy loss.

With utterance-topic representation V_o and topic output matrix $\mathbb{O}$, the proposed model can also generate topic-word distributions like probabilistic topic models. Follow [28], we assume that each utterance belongs to only one hidden topic. This assumption may not apply to lengthy documents, but it is feasible for short text collections such as dialogs, which can alleviate the problem of sparseness. We first take the topic with the largest weight in V_o as the topic of the input utterance, and let $\mathbb{O}_i$ be the output vector of the ith hidden topic. Before calculating the softmax of the vocabulary, our model replace utterance-topic representation V_o with $\mathbb{O}_i$ to generate a multinomial distribution of words for the ith topic.

3.2 Response Generation Model Component

We utilize the topic generation model to obtain the topic vector and the topic-word probability distribution of the input utterance. In this section, TRGM introduces the topic information into a seq2seq dialog system, and uses the topic information to help the dialog system generate informative and more coherent responses.

Recently, the response generation model based on a seq2seq framework has become the mainstream method of open domain dialog systems. A standard seq2seq framework consists of an encoder and a decoder. The encoder is responsible for reading the input sequence and encoding it into an intermediate representation of the model, which is typically a vector called a context vector. The role of the decoder is to generate an output sequence based on the context vector. In this paper, the seq2seq structure of bidirectional RNN proposed by [29] is applied, which exploits Bidirectional Gated Recurrent Unit (BiGRU) to improve seq2seq models. It not only considers the information before the current word from front to back, but also considers the content after the current word from back to front, and generates a response based on both. To alleviate the problem of universal responses, TRGM introduces topic information into the seq2seq framework, and uses the topic knowledge to improve the response generation process of dialogue systems. The introduction of topic information enriches the content of responses and reduces the probability of generating a universal reply. In addition, the topic information also contains the context of the previous conversation, reducing the possibility of inconsistency.

Specifically, topic information can be applied through two modules: a topic enhancement module and a joint generation module. The topic enhancement module integrates topic vector V_o into the hidden state of the encoder, so that each hidden state can contain the topic information. Therefore, the semantic information contained in the hidden states of the encoder RNN is enriched. The joint generation module is based on a mixed probability model of two modes to predict words in the response, that is, a normal mode and a topic mode. The former is a traditional response generation probability mode, and the latter selects the generated words from topic words.

Topic enhancement module. Following [29], in the encoder, we employ a BiGRU to transform input utterance U into hidden state $H_e = \{h_e^1, h_e^2, ..., h_e^i, ..., h_e^N\}$ of equal length, and each word x_i is associated with the corresponding hidden state h_e^i. Formally, h_e^i is given by:

$$h_e^i = concat(\overleftarrow{h}_e^i, \overrightarrow{h}_e^i), \tag{5}$$

where $\overleftarrow{h}_e^i$ denotes the ith hidden state of a backward GRU, and $\overrightarrow{h}_e^i$ denotes the ith hidden state of a forward GRU. The backward GRU reads utterance U in reverse order (*i.e.*, from x_N to x_1), and compute $\overleftarrow{h}_e^i$ as:

$$\begin{aligned} z_i &= \sigma(W_z v_i + Q_z \overleftarrow{h}_e^{i-1}) \\ r_i &= \sigma(W_r v_i + Q_r \overleftarrow{h}_e^{i-1}) \\ s_i &= tanh(W_s v_i + Q_r(\overleftarrow{h}_e^{i-1} \circ r_i)) \\ \overleftarrow{h}_e^i &= (1 - z_i) \circ s_i + z_i \circ \overleftarrow{h}_e^{i-1}, \end{aligned} \tag{6}$$

where $\circ$ denotes element-wise product, v_i denotes the word embedding of x_i, $\sigma(\cdot)$ denotes a sigmoid function, r_i and z_i represent the reset and update activations respectively at time step i. W and Q are the parameters of TRGM. The forward GRU reads utterance U in order (*i.e.*, from x_1 to x_N), and $\overrightarrow{h}_e^i$ can be calculated in a similar way to the backward GRU.

We utilize a gate mechanism similar to Long Short-Term Memory (LSTM) [30] to allow the topic enhancement module to capture the impact of topic information on the response generation model:

$$\begin{aligned}
\widehat{i_i} &= \sigma(W_i V_o + Q_i h_e^i) \\
\widehat{f_i} &= \sigma(W_f V_o + Q_f h_e^i) \\
\widehat{o_i} &= \sigma(W_o V_o + Q_o h_e^i) \\
\widehat{m_i} &= tanh(W_m V_o + Q_m h_e^i) \\
m_i &= \widehat{f_i} \circ m_{i-1} + \widehat{i_i} \circ \widehat{m_i} \\
\widehat{h_e^i} &= \widehat{o_i} \circ tanh(m_i),
\end{aligned} \tag{7}$$

where m_i denotes memory cell. $\widehat{i_i}$, $\widehat{f_i}$ and $\widehat{o_i}$ are input, forget and output gates respectively. In the decoder, at time t, the context vector c_t can then be computed from the new hidden state $\widehat{h_e^i}$ by:

$$c_t = \sum_i^N a_{ti} \widehat{h_e^i} \tag{8}$$

where a_{ti} denotes an attention and can be computed by:

$$a_{ti} = \frac{exp(e_{ti})}{\sum_j^N exp(e_{tj})}; e_{ti} = \eta(h_d^t, \widehat{h_e^i}) \tag{9}$$

where h_d^t denotes the hidden state of the decoder at time t, η represent a multi-layer perceptron (MLP) with tanh as an activation function. In this way, we incorporate the topic information of the conversation into the encoder, and calculate context vector c that can affect the response generation process.

Joint generation module. Topic words are informative, and represent the context of the conversation. The joint generation module contains two modes that increase the likelihood of topic words appearing in responses. In the decoder, the generation probability of tth word y_t in the response sequence can be calculated by:

$$p(y_t|y_{t-1}, ..., ..., U) = p_N(y_t|y_{t-1}, h_d^t, c_t) + p_K(y_t|y_{t-1}, h_d^t, c_t, \phi) \tag{10}$$

where p_N denotes the normal mode, p_K denotes the topic mode, ϕ is the topic-word distribution of the topic of U. p_N and p_K can be defined by:

$$\begin{aligned}
p_N(y_t|y_{t-1}, h_d^t, c_t) &= \frac{1}{A} exp(\Theta_N(y_t)) \\
p_K(y_t|y_{t-1}, h_d^t, c_t, \phi) &= \frac{1}{A} exp(\Theta_K(y_t)) \\
A &= \sum_{w \in V} \Theta_N(w) + \sum_{w \in V} \Theta_K(w)
\end{aligned} \tag{11}$$

where Θ_N is the metric function of the normal mode, Θ_K is the metric function of the topic mode, V denotes vocabulary and A is a normalizer. The normal mode and the topic

mode are normalized by a softmax function. Therefore, they have a unified normalization term.

In normal mode, the metric function is defined as:

$$\Theta_N(y_t = w) = tanh(o_w^T(W_N^y y_{t-1} + W_N^h h_d^t + W_N^c c_t + b_N)), \tag{12}$$

where o_w denotes a one-hot vector of word w, W and b are the parameters of TRGM. The normal mode is similar to the generation process of seq2seq models, but the context vector c_t contains the topic information of the input utterance.

In topic mode, the metric function is defined as:

$$\Theta_K(y_t = w) = tanh(o_w^T(W_K^y y_{t-1} + W_K^h h_d^t + W_K^c c_t + W_K^\phi \phi_k + b_K)), \tag{13}$$

where ϕ_k denotes the probability distribution of word w in the topic of U. Eq. (13) means that in the topic mode, the probability of generating words in the response is more biased towards topic words. The more relevant the word is to the topic, the more likely it is to appear in the response.

Our model is able to generate a better first word associated with the topic in the response generation process. Because the first word is the starting point for the decoder generation model, it is very important and also plays a key role in whether the response is informative. If the first word is not chosen properly, it is difficult for the generation model to return an informative response. In traditional seq2seq models, since h_d^0 does not exist when $t = 0$, the generation of the first word in the response is completely determined by c_0 that only relys on $\{h_e^i\}_{i=1}^N$. In TRGM, the generation of the first word is not only determined by c_0, but also by ϕ that consists of topic-word distributions related to the input utterance.

4. EXPERIMENTS

In this section, we compare TRGM with baseline models by automatic evaluation and side-by-side human judgment.

4.1 Dataset

We evaluate the performance of response generation using a movie dialog corpus, which contains more than 220,000 conversations between movie characters. To reduce data sparsity, we performed the following preprocessing steps: (1) remove non-alphabetic characters; (2) convert uppercase letters to lowercase; (3) all dates and times are replaced by #dt#, all names are replaced by #person# and all numbers are replaced by #num#. Furthermore, to improve the coherence of topic vectors and topic words, we remove the stop words[1] of input utterances in the topic generation model. We choose 80% of the dataset as a training set, 10% as a testing set, and 10% as a validation set.

4.2 Baseline Methods

We compare TRGM with the following three response generation models and two variant of TRGM:

[1] Stop word list: http://www.nltk.org/

- **NRM-hyb** is a standard BiGRU seq2seq model with attention proposed by [18]. Using a hybrid approach, NRM-hyb can integrate local and global information into the process of response generation.
- **MMI** is a seq2seq neural network model for conversational response generation [13]. MMI uses maximum mutual information as the objective function to generate more interesting and diverse responses.
- **TopicAttention** is a topic aware seq2seq model, which utilizes topic information to produce responses by a joint attention mechanism [23].
- **TRGM-TE** is a variant of TRGM, which only leverages topic information by the topic enhancement module.
- **TRGM-JG** is a variant of TRGM, which only leverages topic information by the joint generation module.

4.3 Evaluation Measures

Accurately evaluating the effectiveness of a response generation model has always been an open problem, and this is not the focus of this paper. We choose three evaluation metrics in the existing work [5, 13, 23] and human evaluation.

Perplexity. Perplexity is a metric for evaluating probabilistic language models, which can also be used in open domain dialogue systems. The basic idea is that the utterances in the testing set are smooth. Therefore, the model with higher generation probability on the testing set has better effect. The perplexity is defined as follows:

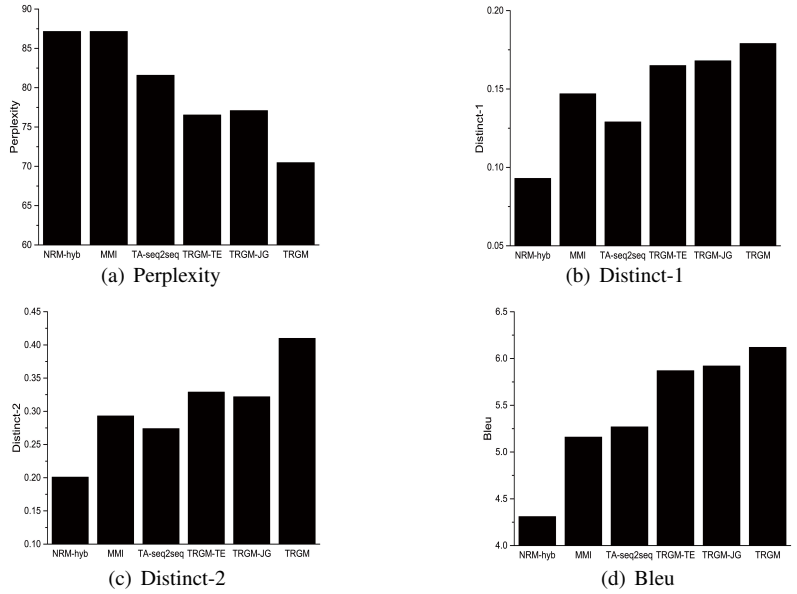
$$PPL = exp\{-\frac{1}{N}\sum_{i=1}^{N} log(p(y_i)\}, \tag{14}$$

where N denotes the length of the response. A lower perplexity score generally represents better generation performance.

Distinct-1 & Distinct-2. Distinct-1 and distinct-2 utilize the number of distinct unigrams and bigrams in the generated response to measure the diversity of responses. Following [13], the two metrics are calculated by dividing the numbers by total number of generated tokens. Distinct-1 and distinct-2 measure the amount of information and diversity of generated responses. A higher score indicates that a generated response is longer and contains more content.

Bleu. Bleu is a metric for automatic evaluation of machine translation. For open domain dialogue systems, bleu can measure the word overlap between a generated response and a real response. The more similar the response generated by the model is to the real response, the higher the bleu score, indicating better performance of the model.

Human evaluation. In addition to the above automatic metrics, we invited human annotators to evaluate the quality of different models. We first randomly shuffle the responses



(a) Perplexity (b) Distinct-1

(c) Distinct-2 (d) Bleu

Fig. 2. Results on automatic metrics of different models.

generated by different models. Five annotators, who are computational linguistics graduate students, judge the quality of the generated responses based on the following criteria: (1) Poor(+0): The generated response has obvious grammatical structure errors or garbled characters, and cannot be used as a response to the dialogue; (2) Medium(+1): The generated response can be used as a response to the dialogue, but the content of the response is not informative, and it may be a universal response such as "I don't know" or "Me too"; (3) Good(+2): The generated response is not only strongly related to the dialogue but also natural and smooth, and the content of the response is informative.

In the topic generation model component, we utilize $BERT_{BASE}$ [10] to encode input utterances. The hidden size h is set to 768, and the number of heads and layers are set to 12. Gelu activation is used in the model. We set the number of topics k to 200, and the dimension of topic vector d_o to 100. In the response generation model component, we exploit freely-available FastText[2] word embeddings as the input of the model. We set the dropout keep probability of attention and hidden layers to 0.8, and choose Adam as the model optimizer with a learning rate of 5e-5. For baseline models, we set the parameters according to their original papers.

4.4 Evaluation Results

Fig. 2 illustrates the experimental results of automatic metrics. It can be intuitively found from the results TRGM achieves the best performance. For perplexity, the proposed models TRGM-TE, TRGM-JG and TRGM are better than other baseline models, and the perplexity of TRGM is the lowest. The reason is that TRGM incorporates BERT-based topic information into a seq2seq model, which can increase the probability of topic words

[2]https://fasttext.cc/docs/en/english-vectors.html

appearing in the generated response, thereby reducing the perplexity score. Due to the use of a gate mechanism and a mixed probability model to introduce topic knowledge, TRGM-TE and TRGM-JG achieve a 10-11% improvement over NRM-hyb. Furthermore, TopicAttention utilizes the topic information obtained from a pre-trained LDA model to generate a reply by an attention mechanism. However, LDA-based topic extraction is not suitable for neural dialogue models, which may be the reason for its poor result.

For Distinct-1 and Distinct-2, TRGM achieves the best performance on both metrics. The experimental results further verify that the topic enhancement module and joint generation module are helpful to enrich the content of generated responses. Additionally, MMI simply relies on changing the objective function of the neural network without adding additional knowledge to the model. On the contrary, TRGM actively introduces the topic information into the response to avoid generating safe responses. TRGM-TE is better than TopicAttention, which indicates that topic extraction based on BERT can generate more informative and diversified responses.

For bleu, compared with NRM-hyb, the bleu scores of TRGM-TE and TRGM-JG increased by 34% and 36%, respectively. The reason may be that the diversity and complexity of input utterances cause the standard seq2seq model NRM-hyb to generate conservative universal responses [23]. Therefore, the bleu score of TopicAttention is higher than that of NRM-hyb and MMI. TRGM achieves the highest score. The experimental results show that the introduction of topic information enables the model to capture more semantic information of the utterance in the response generation process.

Table 1. Results on human evaluation of different models.

Model	Poor(+0)	Medium(+1)	Good(+2)	Average
NRM-hyb	35.15%	40.93%	23.92%	0.884
MMI	33.71%	35.04%	31.25%	0.9754
TopicAttention	35.18%	31.93%	35.18%	1.0229
TRGM-TE	31.41%	29.77%	38.82%	1.0741
TRGM-JG	29.20%	31.43%	39.37%	1.1017
TRGM	27.14%	30.76%	42.10%	1.1496

The results on human evaluation of six models are shown in Table 1. We randomly selected 4000 responses generated by each model for human evaluation. "Average" represents the arithmetic mean of scores. As shown in the table, TRGM is significantly better than all other baseline models. Compared with MMI, the propose models generate more results rated as "Good", which indicates topic information can help the seq2seq model to generate more informative responses. In addition, the average score of TRGM-TE is lower than TRGM-JG. This may be because without the joint generation module, the response generation model brings noise in the process of incorporating topic information, which leads to error accumulation during model training.

4.5 Case Study

Table 2 shows four example responses generated by TopicAttention, MMI, and the proposed model. All input utterances are randomly selected. We observe that MMI tends to produce reasonable responses to simple utterances. However, when the semantic complexity of utterances becomes higher, the responses transform into safe, universal forms,

such as "me too". Another observation is that the response produced by the proposed model carries rich topic information, which makes it more informative. Although TopicAttention also utilizes topic information to generate responses, it requires a large amount of external corpus to train an independent topic model. When only using conversation data, the response carries very little information, and may quickly lead to the end of a conversation. In addition, TRGM keeps the topic relationship between input and responses consistent, which may prevent the conversation from switching to new topics. This motivates us to find a solution to this problem in the future work.

Table 2. Responses generated by different models.

Utterance	TopicAttention	MMI	TRGM
i have lunch at home today.	we will have breakfast.	me too.	what for lunch?
i like doing business.	you can do this.	you are a good person.	i'm in money business and i always have ideas.
do you have music class today?	yes, you are an excellent student.	yes, we do.	no, we can go to the concert in the park.
you are laughing.	you look funny.	i am sorry.	i think she is very cute.

5. CONCLUSIONS

In this paper, we propose a novel model for generating informative responses for open domain dialogue systems, namely Topic-driven Response Generation Model (TRGM). TRGM first utilizes a topic generation model to capture topic information of input utterances, which helps to enrich the content of responses. Next, our model incorporates topic knowledge by using a response generation model to encourage topic words to appear in responses. We conduct extensive experiments on a public dialog corpus. The experimental results show the effectiveness of the proposed model compared with classic and state-of-the-art baselines. In the future, we would like to explore topic model structures that are more appropriate for open domain dialogue systems.

ACKNOWLEDGMENT

We are grateful to the anonymous reviewers, as well as to Fan Zhang, Gang Hu, Yuan Fang and Weiguang Han for helpful comments and suggestions on this work. This work was supported in part by Jianghan University Doctoral Research Startup Fund Project and Hubei Provincial Natural Science Foundation of China.

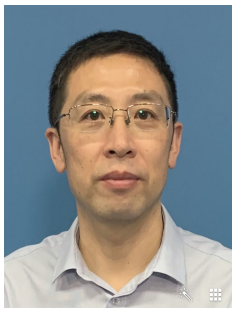
REFERENCES

1. Y. Chen and H. Su, "A distributed neural filter for finding depth-k skyline friends in social networks," *Journal of Information Science and Engineering*, Vol. 34, 2018, pp. 1097-1118.

2. W. Gao, M. Peng, H. Wang, Y. Zhang, W. Han, G. Hu, and Q. Xie, "Generation of topic evolution graphs from short text streams," *Neurocomputing*, Vol. 383, 2020, pp. 282-294.
3. C. Yang, J. Fan, and Y. Liu, "Multi-document summarization using probabilistic topic-based network models," *Journal of Information Science and Engineering*, Vol. 32, 2016, pp. 1613-1634.
4. H. Jin, C. Lin, H. Chen, and J. Liu, "Quickpoint: Efficiently identifying densest subgraphs in online social networks for event stream dissemination," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 32, 2020, pp. 332-346.
5. I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2016, pp. 3776-3783.
6. C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018, pp. 5610-5617.
7. L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, and J. Liu, "A hybrid retrieval-generation neural conversation model," in *Proceedings of ACM International Conference on Information and Knowledge Management*, 2019, pp. 1341-1350.
8. X. Peng, L. Song, D. Gildea, and G. Satta, "Sequence-to-sequence models for cache transition systems," in *Proceedings of Annual Meeting of Association for Computational Linguistics*, 2018, pp. 1842-1852.
9. L. Shen, P. Tai, C. Wu, and S. Lin, "Controlling sequence-to-sequence models – A demonstration on neural-based acrostic generator," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 43-48.
10. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171-4186.
11. W. Zhou, T. Ge, K. Xu, F. Wei, and M. Zhou, "Bert-based lexical substitution," in *Proceedings of Annual Meeting of Association for Computational Linguistics*, 2019, pp. 3368-3373.
12. H. Tseng, H. Chen, K. Chang, Y. Sung, and B. Chen, "An innovative bert-based readability model," in *Proceedings of the 2nd International Conference on Innovative Technologies and Learning*, 2019, pp. 301-308.
13. J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110-119.
14. Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots," *Computational Linguistics*, Vol. 45, 2019, pp. 163-197.
15. R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, 2016, pp. 55-64.

16. L. Yang, Q. Ai, J. Guo, and W. B. Croft, "aNMM: Ranking short answer texts with attention-based neural matching model," in *Proceedings of ACM International Conference on Information and Knowledge Management*, 2016, pp. 287-296.
17. B. Kim, K. Chung, J. Lee, J. Seo, and M. Koo, "A bi-lstm memory network for end-to-end goal-oriented dialog learning," *Computer Speech and Language*, Vol. 53, 2019, pp. 217-230.
18. L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proceedings of Annual Meeting of Association for Computational Linguistics*, 2015, pp. 1577-1586.
19. M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018, pp. 5110-5117.
20. L. Qin, M. Galley, C. Brockett, X. Liu, X. Gao, B. Dolan, Y. Choi, and J. Gao, "Conversing by reading: Contentful neural conversation with on-demand machine reading," in *Proceedings of Annual Meeting of Association for Computational Linguistics*, 2019, pp. 5427-5436.
21. Y. Wu, Z. Li, W. Wu, and M. Zhou, "Response selection with topic clues for retrieval-based chatbots," *Neurocomputing*, Vol. 316, 2018, pp. 251-261.
22. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
23. C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2017, pp. 3351-3357.
24. W. Gao, M. Peng, H. Wang, Y. Zhang, Q. Xie, and G. Tian, "Incorporating word embeddings into topic modeling of short text," *Knowledge and Information Systems*, Vol. 61, 2019, pp. 1123-1145.
25. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of European Conference on Advances in Information Retrieval*, 2011, pp. 338-349.
26. Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2015, pp. 2210-2216.
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.
28. C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, 2016, pp. 165-174.
29. K. Cho, B. V. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proceedings of Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 103-111.
30. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol. 9, 1997, pp. 1735-1780.

Wang Gao received the Ph.D. degree in Computer Software and Theory from Wuhan University, Wuhan, China, in 2019. He is currently a Lecturer at the School of Artificial Intelligence, Jianghan University. His main research interests include natural language processing and information retrieval.



Hongtao Deng is a Full Professor and Master's Supervisor with the School of Artificial Intelligence, Jianghan University. His current research interests include data mining, machine learning, information retrieval.

Xun Zhu is currently pursuing the Ph.D. degree at Wuhan University. She is currently an Assistant Professor at the School of Artificial Intelligence, Jianghan University. Her research interests include natural language processing and data mining.

Yuwei Wang received the M.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2015 and 2019. He is currently a Lecturer at the School of Artificial Intelligence, Jianghan University. His current research interests include computer vision and deep learning.