Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Topic Aware Neural Response Generation

Chen Xing,[1,2]* **Wei Wu,**[4] **Yu Wu,**[3] **Jie Liu,**[1,2] †
Yalou Huang,[1,2] **Ming Zhou,**[4] **Wei-Ying Ma**[4]
[1]College of Computer and Control Engineering, Nankai University, Tianjin, China
[2]College of Software, Nankai University, Tianjin, China
[3]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[4]Microsoft Research, Beijing, China
{v-chxing, wuwei, v-wuyu, mingzhou, wyma}@microsoft.com {jliu,huangyl}@nankai.edu.cn

Abstract

We consider incorporating topic information into a sequence-to-sequence framework to generate informative and interesting responses for chatbots. To this end, we propose a topic aware sequence-to-sequence (TA-Seq2Seq) model. The model utilizes topics to simulate prior human knowledge that guides them to form informative and interesting responses in conversation, and leverages topic information in generation by a joint attention mechanism and a biased generation probability. The joint attention mechanism summarizes the hidden vectors of an input message as context vectors by message attention and synthesizes topic vectors by topic attention from the topic words of the message obtained from a pre-trained LDA model, with these vectors jointly affecting the generation of words in decoding. To increase the possibility of topic words appearing in responses, the model modifies the generation probability of topic words by adding an extra probability item to bias the overall distribution. Empirical studies on both automatic evaluation metrics and human annotations show that TA-Seq2Seq can generate more informative and interesting responses, significantly outperforming state-of-the-art response generation models.

Introduction

Human-computer conversation is a challenging task in AI and NLP. Existing conversation systems include task oriented dialog systems (Young et al. 2013) and non task oriented chatbots. Dialog systems aim to help people complete specific tasks such as ordering and tutoring, while chatbots are designed for realizing natural and human-like conversation with people regarding a wide range of issues in open domains (Perez-Marin 2011). Although previous research focused on dialog systems, recently, with the large amount of conversation data available on the Internet, chatbots are becoming a major focus of both academia and industry.

A common approach to building the conversation engine in a chatbot is learning a response generation model within a machine translation (MT) framework (Ritter, Cherry, and Dolan 2011; Sutskever, Vinyals, and Le 2014; Shang, Lu, and Li 2015; Sordoni et al. 2015a) from large scale social conversation data. Recently, neural network based methods have become mainstream because of their capability to capture semantic and syntactic relations between messages and responses in a scalable and end-to-end way. Sequence-to-sequence (Seq2Seq) with attention (Bahdanau, Cho, and Bengio 2014; Cho, Courville, and Bengio 2015) represents a state-of-the-art neural network model for response generation. To engage people in conversation, the response generation algorithm in a chatbot should generate responses that are not only natural and fluent, but also informative and interesting. MT models such as Seq2Seq with attention, however, tend to generate trivial responses like "me too", "I see", or "I don't know" (Li et al. 2015) due to the high frequency of these patterns in data. Although these responses are safe for replying to many messages, they are boring and carry little information. Such responses may quickly lead the conversation between human and machine to an end, severely hurting the user experience of a chatbot.

In this paper, we study the problem of response generation for chatbots. Particularly, we target the generation of informative and interesting responses that can help chatbots engage their users. Unlike Li et al. (Li et al. 2015) who try to passively avoid generating trivial responses by penalizing their generation probabilities, we consider solving the problem by actively bringing content into responses by topics. Given an input message, we predict possible topics that can be talked about in responses, and generate responses for the topics. The idea is inspired by our observation on conversations between humans. In human-human conversation, people often associate an input message with topically related concepts in their mind. Based on the concepts, they organize content and select words for their responses. For example, to reply to "my skin is so dry", people may think it is a "skin" problem and can be alleviated by "hydrating" and "moisturizing". Based on this knowledge, they may give more informative responses like "then hydrate and moisturize your skin" rather than trivial responses like "me too". The informative responses could let other people follow the topics and continue talking about skin care. "Skin", "hydrate", and "moisturize" are topical concepts related to the message. They represent people's prior knowledge in conversation. In responding, people will bring content that is relevant to the concepts to their responses and even directly

*The work was done when the first author was an intern in Microsoft Research Asia.

†He is the corresponding author.

use the concepts as building blocks to form their responses.

We consider simulating the way people respond to messages with topics, and propose a topic aware sequence-to-sequence (TA-Seq2Seq) model in order to leverage topic information as prior knowledge in response generation. TA-Seq2Seq is built on the sequence-to-sequence framework. In encoding, the model represents an input message as hidden vectors by a message encoder, and acquires embeddings of the topic words of the message from a pre-trained Twitter LDA model. The topic words are used as a simulation of topical concepts in people's minds, and obtained from a Twitter LDA model which is pre-trained using large scale social media data outside the conversation data. In decoding, each word is generated according to both the message and the topics through a joint attention mechanism. In joint attention, hidden vectors of the message are summarized as context vectors by message attention which follows the existing attention techniques, and embeddings of topic words are synthesized as topic vectors by topic attention. Different from existing attention, in topic attention, the weights of the topic words are calculated by taking the final state of the message as an extra input in order to strengthen the effect of the topic words relevant to the message. The joint attention lets the context vectors and the topic vectors jointly affect response generation, and makes words in responses not only relevant to the input message, but also relevant to the correlated topic information of the message. To model the behavior of people using topical concepts as "building blocks" of their responses, we modify the generation probability of a topic word by adding another probability item which biases the overall distribution and further increases the possibility of the topic word appearing in the response.

We conduct an empirical study on large scale data crawled from Baidu Tieba, and compare different methods with both automatic evaluation and human judgment. The results on both automatic evaluation metrics and human annotations show that TA-Seq2Seq can generate more informative, diverse, and topic relevant responses and significantly outperforms state-of-the-art methods for response generation.

The contributions of this paper include 1) a proposal for using topics as prior knowledge for response generation; 2) a proposal for a TA-Seq2Seq model that naturally incorporates topic information into the encoder-decoder structure; 3) empirical verification of the effectiveness of TA-Seq2Seq.

Background: sequence-to-sequence model and attention mechanism

Before introducing our model, let us first briefly review the Seq2Seq model and the attention mechanism.

Sequence-to-sequence model

In Seq2Seq, given a source sequence (message) $\mathbf{X} = (x_1, x_2, \ldots, x_T)$ and a target sequence (response) $\mathbf{Y} = (y_1, y_2, \ldots, y_{T'})$, the model maximizes the generation probability of $\mathbf{Y}$ conditioned on $\mathbf{X}$: $p(y_1, ..., y_{T'}|x_1, ..., x_T)$. Specifically, Seq2Seq is in an encoder-decoder structure. The encoder reads $\mathbf{X}$ word by word and represents it as a context vector $\mathbf{c}$ through a recurrent neural network (RNN), and then the decoder estimates the generation probability of $\mathbf{Y}$ with $\mathbf{c}$ as input. The objective function of Seq2Seq can be written as

$$p(y_1, ..., y_{T'}|x_1, ..., x_T) = p(y_1|\mathbf{c}) \prod_{t=2}^{T'} p(y_t|\mathbf{c}, y_1, ..., y_{t-1}).$$

The encoder RNN calculates the context vector $\mathbf{c}$ by

$$\mathbf{h}_t = f(x_t, \mathbf{h}_{t-1}); \mathbf{c} = \mathbf{h}_T,$$

where $\mathbf{h}_t$ is the hidden state at time t and f is a non-linear transformation which can be either a long-short term memory unit (LSTM) (Hochreiter and Schmidhuber 1997) or a gated recurrent unit (GRU) (Cho et al. 2014). In this work, we implement f using GRU which is parameterized as

$$\begin{aligned} \mathbf{z} &= \sigma(\mathbf{W}^z \mathbf{x}_t + \mathbf{U}^z \mathbf{h}_{t-1}) \\ \mathbf{r} &= \sigma(\mathbf{W}^r \mathbf{x}_t + \mathbf{U}^r \mathbf{h}_{t-1}) \\ \mathbf{s} &= tanh(\mathbf{W}^s \mathbf{x}_t + \mathbf{U}^s (\mathbf{h}_{t-1} \circ \mathbf{r})) \\ \mathbf{h}_t &= (1 - \mathbf{z}) \circ \mathbf{s} + \mathbf{z} \circ \mathbf{h}_{t-1} \end{aligned} \tag{1}$$

The decoder is a standard RNN language model except when conditioned on the context vector $\mathbf{c}$. The probability distribution $\mathbf{p}_t$ of candidate words at every time t is calculated as

$$\mathbf{s}_t = f(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}); \mathbf{p}_t = softmax(\mathbf{s}_t, y_{t-1})$$

where $\mathbf{s}_t$ is the hidden state of the decoder RNN at time t and y_{t-1} is the word at time $t-1$ in the response sequence.

Attention mechanism

The traditional Seq2Seq model assumes that every word is generated from the same context vector. In practice, however, different words in $\mathbf{Y}$ could be semantically related to different parts of $\mathbf{X}$. To tackle this issue, attention mechanism (Bahdanau, Cho, and Bengio 2014) is introduced into Seq2Seq. In Seq2Seq with attention, each y_i in $\mathbf{Y}$ corresponds to a context vector $\mathbf{c}_i$, and $\mathbf{c}_i$ is a weighted average of all hidden states $\{\mathbf{h}_t\}_{t=1}^T$ of the encoder. Formally, $\mathbf{c}_i$ is defined as

$$\mathbf{c}_i = \Sigma_{j=1}^T \alpha_{ij} \mathbf{h}_j, \tag{2}$$

where α_{ij} is given by

$$\alpha_{ij} = \frac{exp(e_{ij})}{\Sigma_{k=1}^T exp(e_{ik})}; e_{ij} = \eta(\mathbf{s}_{i-1}, \mathbf{h}_j) \tag{3}$$

η is usually implemented as a multi-layer perceptron (MLP) with tanh as an activation function.

Topic aware Seq2Seq model

Suppose that we have a data set $\mathcal{D} = \{(\mathbf{K}_i, \mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^N$ where $\mathbf{X}_i$ is a message, $\mathbf{Y}_i$ is a response, and $\mathbf{K}_i = (\mathbf{k}_{i,1}, \ldots, \mathbf{k}_{i,n})$ are the topic words of $\mathbf{X}_i$. Our goal is to learn a response generation model from $\mathcal{D}$, and thus given a new message $\mathbf{X}$ with topic words $\mathbf{K}$, the model can generate response candidates for $\mathbf{X}$.

To learn the model, we need to answer two questions: 1) how to obtain the topic words; 2) how to perform learning. In this section, we first describe our method on topic word acquisition, and then we give details of our model.

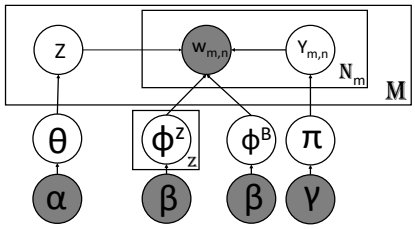

Figure 1: Graphic model of Twitter LDA

Topic word acquisition

We obtain topic words of a message from a Twitter LDA model (Zhao et al. 2011). Twitter LDA belongs to the family of probabilistic topic models (Blei, Ng, and Jordan 2003) and represents the state-of-the-art topic model for short texts (Zhao et al. 2011). The basic assumption of Twitter LDA is that each message corresponds to one topic, and each word in the message is either a background word or a topic word under the topic of the message. Figure 1 gives the graphical model of the Twitter LDA.

We estimate the parameters of the Twitter LDA using the collapsed Gibbs sampling algorithm (Zhao et al. 2011). After that, we use the model to assign a topic z to a message $\mathbf{X}$, pick the top n words ($n = 100$ in our experiments) with the highest probabilities under z, and remove universal words like "thank" and "you" to get the topic words $\mathbf{K}$ for $\mathbf{X}$.

In learning, we need a vector representation for each topic word. To this end, we first calculate a distribution for topic word w by Equation (4) where C_{wz} is the number of times that w is assigned to topic z in training. Then, we take the distributions as the vector representations of the topic words.

$$p(z|w) \propto \frac{C_{wz}}{\sum_{z'} C_{wz'}}. \tag{4}$$

In our experiments, we train a Twitter LDA model using large scale posts from Sina Weibo which is the largest microblogging service in China. The data provides topic knowledge apart from that in message-response pairs that we use to train the response generation model. The process is similar to how people learn to respond in conversation: they become aware of what can be talked about from Internet, especially from social media, and then use what they have learned as topics to form their responses in conversation.

Note that in addition to LDA, one can employ other techniques like tag recommendation (Wu et al. 2016) or keyword extraction (Wu et al. 2015) to generate topic words. One can also get topic words from other resources like wikipedia and other web documents. We leave the discussion of these extensions for future work.

Model

Figure 2 gives the structure of a topic aware sequence-to-sequence model (TA-Seq2Seq). TA-Seq2Seq is built on the sequence-to sequence framework, and leverages topic information using a joint attention mechanism and a biased generation probability.

Specifically, in encoding, a message encoder represents an input message $\mathbf{X}$ as a series of hidden vectors $\{\mathbf{h}_t\}_{t=1}^T$ by a bidirectional GRU-RNN from both ends[1]. GRU is defined in Equation (1). At the same time, a topic encoder obtains the embeddings of the topic words $\mathbf{K}$ of $\mathbf{X}$ by looking up an embedding table which is established according to Equation (4). With a little abuse of notations, we also use $(\mathbf{k}_1, \ldots, \mathbf{k}_n)$ to denote the the embeddings of words in $\mathbf{K}$. The meaning of $(\mathbf{k}_1, \ldots, \mathbf{k}_n)$ is clear in context.

In decoding, at step i, message vectors $\{\mathbf{h}_t\}_{t=1}^T$ are transformed to a context vector $\mathbf{c}_i$ by message attention given by Equation (2) and Equation (3), and embeddings of topic words $\{\mathbf{k}_j\}_{j=1}^n$ are linearly combined as a topic vector $\mathbf{o}_i$ by topic attention. The combination weight of $\mathbf{k}_j$ is given by

$$\alpha_{oj}^i = \frac{exp(\eta_o(\mathbf{s}_{i-1}, \mathbf{k}_j, \mathbf{h}_T))}{\Sigma_{j'=1}^n exp(\eta_o(\mathbf{s}_{i-1}, \mathbf{k}_{j'}, \mathbf{h}_T))}. \tag{5}$$

where $\mathbf{s}_{i-1}$ is the $i-1$-th hidden state in decoder, $\mathbf{h}_T$ is the final hidden state of the input message, and η_o is a multilayer perceptron. Compared to the traditional attention in Equation (2) and Equation (3), topic attention further leverages the final state of the message (i.e., $\mathbf{h}_T$) to weaken the effect of topic words that are irrelevant to the message in generation and highlight the importance of relevant topic words. As a result, the topic vectors $\{\mathbf{o}_i\}_{i=1}^{T'}$ are more correlated to the content of the input message and noise in topic words is controlled in generation. The message attention and the topic attention forms a joint attention mechanism which allows $\mathbf{c}_i$ and $\mathbf{o}_i$ to jointly affect the generation probability. The advantage of the joint attention is that it makes words in responses not only relevant to the message, but also relevant to the topics of the message.

We define the generation probability $p(y_i)$ as $p(y_i) = p_V(y_i) + p_K(y_i)$, where $p_V(y_i)$ and $p_K(y_i)$ are defined by

$$\begin{aligned}
p_V(y_i = w) &= \begin{cases} \frac{1}{Z} e^{\Psi_V(\mathbf{s}_i, \mathbf{y}_{i-1}, w)}, & w \in \mathbf{V} \cup \mathbf{K} \\ 0, & w \notin \mathbf{V} \cup \mathbf{K} \end{cases} \\
p_K(y_i = w) &= \begin{cases} \frac{1}{Z} e^{\Psi_K(\mathbf{s}_i, \mathbf{y}_{i-1}, \mathbf{c}_i, w)}, & w \in \mathbf{K} \\ 0, & w \notin \mathbf{K} \end{cases} \\
\mathbf{s}_i &= f(y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i, \mathbf{o}_i).
\end{aligned} \tag{6}$$

In Equation (6), $\mathbf{V}$ is a response vocabulary, and f is a GRU unit. $\Psi_V(\mathbf{s}_i, y_{i-1})$ and $\Psi_K(\mathbf{s}_i, y_{i-1}, c_i)$ are defined by

$$\begin{aligned}
\Psi_V(\mathbf{s}_i, y_{i-1}, w) &= \sigma(\mathbf{w}^T(\mathbf{W}_V^s \cdot \mathbf{s}_i + \mathbf{W}_V^y \cdot y_{i-1} + \mathbf{b}_V)), \\
\Psi_K(\mathbf{s}_i, y_{i-1}, \mathbf{c}_i, w) &= \sigma(\mathbf{w}^T(\mathbf{W}_K^s \cdot \mathbf{s}_i + \mathbf{W}_K^y \cdot y_{i-1} \\
&\qquad + \mathbf{W}_K^c \cdot \mathbf{c}_i + \mathbf{b}_K)).
\end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is tanh, $\mathbf{w}$ is a one-hot indicator vector of word w, and $\mathbf{W}_V^s$, $\mathbf{W}_K^s$, $\mathbf{W}_V^y$, $\mathbf{W}_K^y$, $\mathbf{b}_V$, and $\mathbf{b}_K$ are parameters. $Z = \Sigma_{v \in \mathbf{V}} e^{\Psi_V(\mathbf{s}_i, \mathbf{y}_{i-1}, v)} + \Sigma_{v' \in \mathbf{K}} e^{\Psi_K(\mathbf{s}_i, \mathbf{y}_{i-1}, \mathbf{c}_i, v')}$ is a normalizer.

Equation (6) means that the generation probability in TA-Seq2Seq is biased to topic words. For non topic words, the

[1] Hidden vectors from both directions are concatenated together.

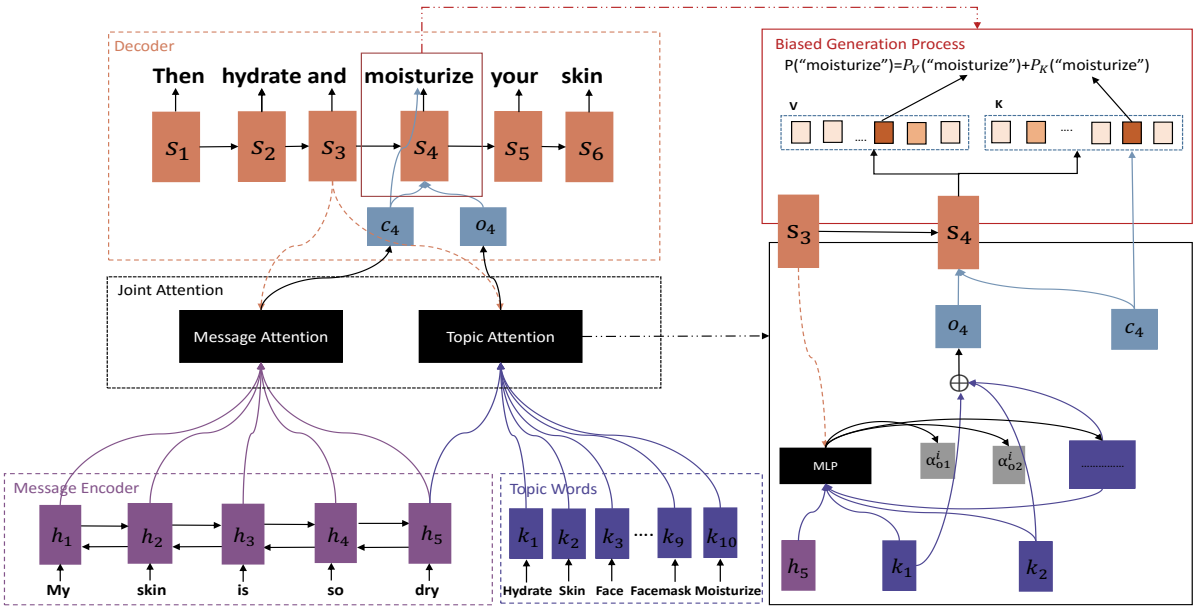

Figure 2: Structure of TA-Seq2Seq

probability (i.e., $p_V(y_i)$) is similar to that in sequence-to-sequence model but with the joint attention mechanism. For topic words, there is an extra probability item $p_K(y_i)$ that biases the overall distribution and further increases the possibility of the topic words appearing in responses. The extra probability is determined by the current hidden state of the decoder $\mathbf{s}_i$, the previous word in generation y_{i-1}, and the context vector $\mathbf{c}_i$. It means that given the generated parts and the input message, the more relevant a topic word is, the greater the likelihood that it will appear in the response.

An extra advantage of TA-Seq2Seq is that it makes a better first word in response generation. The first word matters a lot because it is the starting point of the language model of the decoder and plays a key role in making the whole response fluent. If the first word is wrongly chosen, then the sentence may never have a chance to go back to a proper response. In Seq2Seq with attention, the generation of the first word is totally determined by $\mathbf{c}_0$ which only depends on $\{\mathbf{h}_t\}_{t=1}^{T}$ since there is no $\mathbf{s}_{i-1}$ when $i = 0$. While in TA-Seq2Seq, the first word is generated not only by $\mathbf{c}_0$, but also by $\mathbf{o}_0$ which consists of topic information. Topic information can help calibrate the selection of the first word to make it more accurate.

We conduct topic learning and response generation in two separate steps rather than let them deeply couple like VHRED (Serban et al. 2016). This way we can leverage extra data from various sources (e.g., web and knowledge base) in response generation. For example, in this work, we estimate topic words from posts in Sina Weibo and provide extra topic information for message-response pairs.

We also encourage the appearance of topic words in responses in a very natural and flexible way by biasing the generation distribution. Through this method, our model allows appearance of multiple topic words rather than merely fixing a single key word in responses like what Mou et al. did in their work (Mou et al. 2016).

Experiments

We compare TA-Seq2Seq with state-of-the-art response generation models by both automatic evaluation and human judgment.

Experiment setup

We build a data set from Baidu Tieba which is the largest Chinese forum allowing users to post and comment on others' posts. We crawl 20 million post-comment pairs and used them to simulate message-response pairs in conversation. We removed pairs appearing more than 50 times to prevent them from dominating learning, and employ the Stanford Chinese word segmenter[2] to tokenize the remaining pairs. Pairs with a message or a response having more than 50 words were also removed. After this preprocessing, there are $15,209,588$ pairs left. From them, we randomly sample 5 million distinct message-response pairs[3] as training data, $10,000$ distinct pairs as validation data, and $1,000$ distinct messages with their responses as test data. Messages in the test pairs are used to generate responses, and responses in the test pairs are treated as ground truth to calculate the perplexity of generation models. There is no overlap among messages in training, validation, and testing. We keep the $30,000$ most frequent words in messages in the training data to construct a message vocabulary. The message vocabulary covers 98.8% of words appearing in messages. Similarly,

[2] http://nlp.stanford.edu/software/segmenter.shtml

[3] Any two pairs are different on messages or responses.

we construct a response vocabulary that contains the $30,000$ most frequent words in responses in the training data, covering 98.3% words in responses.

Models	+2	+1	0	Kappa
S2SA	32.3%	36.7%	31.0%	0.8116
S2SA-MMI	33.1%	34.8%	32.1%	0.7848
S2SA-TopicConcat	35.9%	29.3%	34.8%	0.6633
S2SA-TopicAttention	42.3%	27.6%	30.0%	0.8299
TA-Seq2Seq	44.7%	24.9%	30.4%	0.8417

Table 1: Human annotation results

We crawl 30 million posts from Sina Weibo to train a Twitter LDA model. We set the number of topics T as 200 and the hyperparameters of Twitter LDA as $\alpha = 1/T$, $\beta = 0.01$, $\gamma = 0.01$. For each topic, we select the top 100 words as topic words. To filter out universal words, we calculated word frequency using the 30 million posts, and remove the 2000 words with the highest frequency from the topic words. Words outside the topic words, the message vocabulary, and the response vocabulary are treated as "UNK".

Evaluation metrics

How to evaluate a response generation model is still an open problem but not the focus of the paper. Therefore, we follow the existing work and employ the following metrics:

Perplexity: following (Vinyals and Le 2015) and (Mikolov et al. 2010), we employ perplexity as an evaluation metric. Perplexity is defined by Equation (8). It measures how well the model predicts a response. A lower perplexity score indicates better generation performance. In this work, perplexity on validation (PPL-D in Table 2) is used to determine when to stop training. If the perplexity stops decreasing and the difference is smaller than 2.0 five times in validation, we think that the algorithm has reached its convergence and terminate training. We test the generation ability of different models by perplexity on the test data (PPL-T in Table 2).

$$PPL = exp\left\{-\frac{1}{N}\Sigma_{i=1}^{N}\log(p(\mathbf{Y_i}))\right\}. \quad (8)$$

Distinct-1 & distinct-2: we counted numbers of distinct unigrams and bigrams in the generated responses. We also follow (Li et al. 2015) and divide the numbers by total number of unigrams and bigrams. We denote the metrics (both the numbers and the ratios) as distinct-1 and distinct-2 respectively. The two metrics measure how informative and diverse the generated responses are. High numbers and high ratios mean that there is much content in the generated responses, and high numbers further indicate that the generated responses are long.

Human annotation: in addition to the automatic metrics above, we recruit human annotators to judge the quality of the generated responses of different models. Three labelers with rich Tieba experience are invited to do evaluation. Responses generated by different models (the top one response in beam search) are pooled and randomly shuffled for each labeler. Labelers refer to the test messages and judge the quality of the responses according to the following criteria:

+2: The response is not only relevant and natural, but also informative and interesting.

+1: The response can be used as a reply to the message, but is too universal like "Yes, I see" , "Me too" and "I don't know".

0: The response cannot be used as a reply to the message. It is either semantically irrelevant or disfluent (e.g., with grammatical errors or UNK). Agreements among labelers are calculated with Fleiss' kappa (Fleiss and Cohen 1973). Note that we do not choose BLEU (Papineni et al. 2002) as an evaluation metric, because Liu et al. (Liu et al. 2016) have proven that BLEU is not a proper metric for evaluating conversation models as there is weak correlation between BLEU and human judgment.

Models	PPL-D	PPL-T	distinct-1	distinct-2
S2SA	147.04	133.11	604/.091	1168/.207
S2SA-MMI	147.04	133.11	603/.151	1073/.378
S2SA-TopicConcat	150.45	132.12	898/.116	2197/.327
S2SA-TopicAttention	133.81	**119.55**	894/.106	2057/.277
TA-Seq2Seq	134.63	**122.82**	1355/.161	2970/.401

Table 2: Results on automatic metrics

Baselines

We considered the following baselines.

S2SA: the standard Seq2Seq model with attention.

S2SA-MMI: the best performing model in (Li et al. 2015).

S2SA-TopicConcat: to verify the effectiveness of the topic attention of TA-Seq2Seq, we replaced $\mathbf{o}_i$ given by the topic attention in $\mathbf{s}_i$ in Equation (6) by a simple topic vector. The simple topic vector is obtained by concatenating embeddings of topic words and transforming the concatenation to a vector that has the same dimension with the context vector by an MLP.

S2SA-TopicAttention: to verify the effectiveness of biased generation probability of TA-Seq2Seq, we kept the topic attention but removed the bias probability item which is specially designed for topic words from the generation probability in Equation (6).

Note that S2SA-TopicConcat and S2SA-TopicAttention are variants of our TA-Seq2Seq.

In all models, we set the dimensions of the hidden states of the encoder and the decoder as 1000, and the dimensions of word embeddings as 620. All models were initialized with isotropic Gaussian distributions $\mathcal{X} \sim \mathcal{N}(0, 0.01)$ and trained with an AdaDelta algorithm (Zeiler 2012) on a NVIDIA Tesla K40 GPU. The batch size is 128. We set the initial learning rate as 1.0 and reduced it by half if the perplexity on validation began to increase. We implemented the models with an open source deep learning tool Blocks[4], and shared the code of our model at https://github.com/LynetteXing1991.

[4]https://github.com/mila-udem/blocks

Message	TA-Seq2Seq	S2SA-MMI	S2SA
你也玩摄影 You also like photography	我不玩摄影，只是想拍个照片而已 I don't work on **photography**. I just want to take a few **pictures**.	额额额... Well well well....	额... Well...
我的左脸颊上有一些红血丝 There is some redness on my left cheek.	我也是敏感肌 My **skin** is also very **sensitive**.	我也是啊 Me too.	啊啊啊啊 Oh my god oh my god
大学生的话可以去实习吗 Can a college student apply for an internship	可以的，如果你着急的话可以先填个申请表 Yes, you can fill an **application form** first if you are in a hurry.	可以的可以的 Yes, you can.	可以的 Yes.
我的皮肤好干 My skin is so dry.	那就补水保湿吧 Then **hydrate** and **moisturize** your skin.	我也是啊 Me too.	啊啊啊啊 Oh my god oh my god.

Table 3: Case study

Evaluation Results

Table 1 shows the human annotation results. It is clear that topic aware models (S2SA-TopicConcat, S2SA-TopicAttention and TA-Seq2Seq) generate much more informative and interesting responses (responses labeled as "+2") and much less universal responses than the baseline models (S2SA and S2SA-MMI). Among them, TA-Seq2Seq achieves the best performance. Compared with S2SA-MMI, it increases 11.6% "+2" responses and reduces 9.9% "+1" responses. S2SA-TopicAttention performs better than S2SA-TopicConcat, meaning that the joint attention mechanism contributes more to response quality than the biased probability in generation. All models have a proportion of unsuitable responses (labeled as "0") around 30% but S2SA-TopicConcat and S2SA-MMI generate more bad responses. This is because without joint attention, noise in topics is brought to generation by the concatenation of topic word embeddings in S2SA-TopicConcat, and in S2SA-MMI, both good responses and bad responses are boosted in re-ranking. All models have high kappa scores, indicating that labelers reach high agreement regarding quality of responses. We also conduct a sign test between TA-Seq2Seq and the baseline models and results show that the improvement from our model is statistically significant (p-value < 0.01).

Table 2 gives the results of automatic metrics. TA-Seq2Seq and S2SA-TopicAttention achieve comparable perplexity on validation data and test data, and both of them are better than the baseline models. We conduct a t-test on PPL-T and the results show that the improvement is statistically significant (p-value < 0.01). On distinct-1 and distinct-2, all topic aware models perform better than the baseline models in terms of numbers of distinct n-grams (n=1,2). Among them, TA-Seq2Seq achieves the best performance in terms of both absolute numbers and ratios. The results further verify our claim that topic information is helpful for enriching the content of responses. Note that TopicConcat and TopicAttention are worse than S2SA-MMI on ratios of distinct n-grams. This is because responses from S2SA-MMI are generally shorter than those from TopicConcat and TopicAttention. The perplexities of S2SA and S2SA-MMI are the same because S2SA-MMI is an after-processing mechanism on the responses generated by S2SA. Thus we report the perplexity of S2SA to approximately represent the generation ability of S2SA-MMI.

Case study

Figure 3 compares TA-Seq2Seq with S2SA-MMI and S2SA using some examples. Topic words in the responses from TA-Seq2Seq are bolded. From the comparison, we can see that in TA-Seq2Seq, topic words not only help form the structure of responses, but also act as "building blocks" and lead to responses that carry rich information. For example, in Case 2, topic information provides prior knowledge to generation that redness on skin is usually caused by sensitivity of skin and helps form a targeted and informative response. On the other hand, although responses from S2SA-MMI and S2SA also echoed the message, they carry little information and easily lead the conversation to an end.

Related work

Based on the sequence-to-sequence framework, many generation models have been proposed to improve the quality of generated responses from different perspectives. For example, A. Sordoni et al. (Sordoni et al. 2015b) represent the utterances in previous turns as a context vector and incorporate the context vector into response generation. Li et al. (Li et al. 2016) try to build a personalized conversation engine by adding personal information as an extra input. Gu et al. (Gu et al. 2016) introduce copynet to simulate the repeating behavior of humans in conversation. Yao et al. (Yao, Zweig, and Peng 2015) add an extra RNN between the encoder and the decoder of the sequence-to-sequence model with attention to representing intentions. In this work, we consider incorporating topic information into the sequence-to-sequence model. Similar to Li et al. (Li et al. 2015), we also try to avoid safe responses in generation. The difference is that we solve the problem by actively bringing content into responses through topics and enriching information carried by the generated responses.

Conclusion

We propose a topic aware sequence-to-sequence (TA-Seq2Seq) model to incorporate topic information into response generation. The model leverages the topic information by a joint attention mechanism and a biased generation probability. Empirical studies on both automatic evaluation metrics and human annotations show that the model

can generate informative and diverse responses and significantly outperform state-of-the-art generation models.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 61105049 and No. U1633103), the Natural Science Foundation of Tianjin (No. 14JCQNJC00600), and the Science and Technology Planning Project of Tianjin (No. 13ZCZDGX01098).

References

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

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Cho, K.; van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* 103.

Cho, K.; Courville, A.; and Bengio, Y. 2015. Describing multimedia content using attention-based encoder-decoder networks. *Multimedia, IEEE Transactions on* 17(11):1875–1886.

Fleiss, J. L., and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*.

Gu, J.; Lu, Z.; Li, H.; and Li, V. O. 2016. Incorporating copying mechanism in sequence-to-sequence learning.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *INTERSPEECH*, volume 2, 3.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *CoRR* abs/1607.00970.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Perez-Marin, D. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*. IGI Global.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 583–593. Association for Computational Linguistics.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.; Gao, J.; and Dolan, B. 2015a. A neural network approach to context-sensitive generation of conversational responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, 196–205.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015b. A neural network approach to context-sensitive generation of conversational responses.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.

Wu, Y.; Wu, W.; Li, Z.; and Zhou, M. 2015. Mining query subtopics from questions in community question answering. In *AAAI*, 339–345.

Wu, Y.; Wu, W.; Li, Z.; and Zhou, M. 2016. Improving recommendation of tail tags for questions in community question answering. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yao, K.; Zweig, G.; and Peng, B. 2015. Attention with intention for a neural network conversation model. *Computer Science*.

Young, S.; Gasic, M.; Thomson, B.; and Williams, J. D. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer. 338–349.