# Topic Augmented Neural Response Generation with a Joint Attention Mechanism

**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

## Abstract

We consider incorporating topic information as prior knowledge into the sequence to sequence (Seq2Seq) network structure with attention mechanism for response generation in chatbots. To this end, we propose a topic augmented joint attention based Seq2Seq (TAJA-Seq2Seq) model. In TAJA-Seq2Seq, information from input posts and information from topics related to the posts are simultaneously embedded into vector spaces by a content encoder and a topic encoder respectively. The two kinds of information interact with each other and help calibrate weights of each other in the joint attention mechanism in TAJA2Seq2Seq, and jointly determine the generation of responses in decoding. The model simulates how people behave in conversation and can generate well-focused and informative responses with the help of topic information. Empirical study on large scale human judged generation results show that our model outperforms Seq2Seq with attention on both response quality and diversity.

## 1 Introduction

Chatbots, which are designed for natural and human-like conversation with people in open domains, have become hot in recent years. Many companies include Microsoft, Apple, Facebook and Google have released their chatbot products. The core of a chatbot is a response generation engine. In recent two years, with the success of long short-term memory recurrent neural network (LSTM-RNN) (Hochreiter and Schmidhuber, 1997) in capturing long-term information in sequences, a lot of effort on building a response generator has been paid to neural network models. A popular network structure for response generation is the encoder-decoder structure (Sutskever et al., 2014; Shang et al., 2015) or more commonly referred to as "sequence to sequence" (Seq2Seq) model. In Seq2Seq, an input post is encoded by the encoder as a context vector. Then, a language model like decoder decodes the semantic information in the vector and generates final responses. On top of Seq2Seq, attention mechanism (Bahdanau et al., 2014; Cho et al., 2015) which is first proposed for machine translation (MT), is then added to further improve generation quality. In Seq2Seq with attention, different words in decoding are generated from different context vectors. Each context vector is a linear combination of the hidden states of the encoder with weights reflecting importance of different parts of the input post. These weights are distinct for every corresponding word in the generated response. These structures have achieved great success on response generation and outperforms traditional retrieval based approaches (Shang et al., 2015) to a great deal.

In this paper, we consider incorporating topic information as prior knowledge into Seq2Seq with attention for response generation in chatbots. The idea is inspired by our observation on conversation between humans. In human-human conversation, people often associate an input post with topically related concepts and create their responses according to these concepts. For example, to respond to "I've watched Orphan" [1], people who know the movie may think it is a thriller, horrible and fright-

[1] Orphan is a movie `https://en.wikipedia.org/wiki/Orphan_(fi`

ening. Then based on this knowledge, they may give more targeted and well-focused responses like ”it is a thriller” or ”this movie is so frightening”, instead of plain responses like ”I haven’t watched it” or ”I don’t know”. “Thriller”, “horrible”, and “frightening” are concepts under the topic of the post. They represent people’s prior knowledge regarding to the input post, and help them form their responses. We would like to model this process of response generation with topics and use topics to enhance the performance of Seq2Seq with attention.

We propose a topic augmented joint attention based Seq2Seq (TAJA-Seq2Seq) model to leverage topic information in response generation. TAJA-Seq2Seq is equipped with two encoders, and each one has an attention module. The first encoder sequentially represents different parts of an input post as content vectors, and the second encoder compresses various topic words to topic vectors as representations of topic information. Both the content vectors and the topic vectors are averaged with weights that are calculated by the attention modules, and fed to the decoder as context vectors to jointly determine the response generation. In TAJA-Seq2Seq, information from the input post and information from the topic words interact with each other in the joint attention modules. Specifically, in the attention module of the content encoder, a topic vector obtained from a topic summarizer is taken as an extra input in the learning of the weights of the content vectors. In the attention module of the topic encoder, the final state of the content encoder is used to learn the weights of the topic vectors. By this means, the topic information acts as prior knowledge and helps calibrate the content emphasis in the input post which the response should focus on, and the content information helps select semantically relevant words from topic information that the response should be related to. We obtain the topic words from a pre-trained Twitter LDA model. Empirical study on large scale human judged generation results show that with topic information as prior knowledge, TAJA-Seq2Seq can generate more informative, diverse, and topically relevant responses than the traditional Seq2Seq model with attention.

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

## 2 Background: sequence-to-sequence model and attention mechanism

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

### 2.1 Sequence-to-sequence model

In Seq2Seq, given a source sequence (post) $\mathbf{X} = (x_1, x_2, \ldots, x_T)$ 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, the Seq2Seq model 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 $c$ as input. The objective function of Seq2Seq can be written as

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

The encoder RNN calculates the context vector $c$ by

$$\mathbf{h}_t = f(x_t, \mathbf{h}_{t-1}); \mathbf{c} = \mathbf{h}_T, \tag{2}$$

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

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W_{xi}}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1}) \\
\mathbf{f}_t &= \sigma(\mathbf{W_{xf}}\mathbf{x}_t + \mathbf{W_{hf}}\mathbf{h}_{t-1}) \\
\mathbf{o}_t &= \sigma(\mathbf{W_{xo}}\mathbf{x}_t + \mathbf{W_{ho}}\mathbf{h}_{t-1}) \\
\hat{\mathbf{c}}_t &= tanh(\mathbf{W_{xc}}\mathbf{x}_t + \mathbf{W_{hc}}\mathbf{h}_{t-1}) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot tanh(\mathbf{c}_t),
\end{aligned} \tag{3}$$

where $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ are the input gate, forget gate and output gate respectively and $\hat{\mathbf{c}}_t, \mathbf{c}_t$ are proposed and true cell values.

The decoder is a standard RNN language model(Mikolov et al., 2010) except conditioned on the context vector $\mathbf{c}$. The probability distribution $\mathbf{p}_t$ of candidate words at every time $t$ is calculated as

$$\mathbf{s}_t = f(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}); \mathbf{p}_t = softmax(\mathbf{s}_t, y_{t-1}) \tag{4}$$

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.

### 2.2 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 solve this problem, attention mechanism (Bahdanau et al., 2014) which is first proposed for MT, is added to Seq2Seq for response generation(Shang et al., 2015). 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 $\{\mathbf{h}_t\}_{t=1}^{T}$. Formally, $\mathbf{c}_i$ is defined as

$$\mathbf{c}_i = \Sigma_{j=1}^{T} \alpha_{ij} \mathbf{h}_j, \tag{5}$$

where $\alpha_{ij}$ is given by

$$\alpha_{ij} = \frac{exp(e_{ij})}{\Sigma_{k=1}^{T} exp(e_{ik})}; e_{ij} = \eta(\mathbf{s}_{i-1}, \mathbf{h}_j) \tag{6}$$

$\eta$ is usually implemented as a multi-layer perceptron (MLP).

## 3 Topic augmented joint attention based Seq2Seq

We propose a topic augmented joint attention based Seq2Seq (TAJA-Seq2Seq) model to incorporate topic information as prior knowledge into the process of response generation. In the following sections, we first show how we obtain the topic information related to the contents of posts, then we describe details of the network structure of TAJA-Seq2Seq and show how the topic information works in the structure.

### 3.1 Topic acquisition

We obtain a topic word list for each input post from a Twitter LDA model (Zhao et al., 2011).

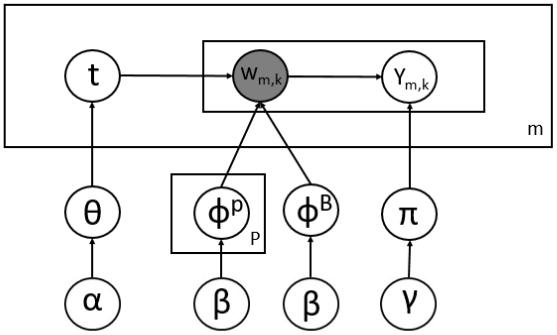


**Figure 1:** Graphical model of Twitter LDA

Twitter LDA belongs to the family of probabilistic topic models (Blei et al., 2003) and represents the state-of-the-art topic model for short texts (Zhao et al., 2011). We choose Twitter LDA among various probabilistic topic models because in conversation data, the input posts are short, informal, and contain quite a lot of scattered topics. These characteristics are similar with those of Twitter data and meet the assumptions of Twitter LDA quite well. The basic assumption of Twitter LDA is that each post corresponds to only one topic. Each word in the post is either a background word or a topic word under the topic of the post. Specifically, Twitter LDA first draws a mutlinomial distribution $\theta$ from a Dirichlet prior Dir($\alpha$) that represents the topic distribution of the whole data set. Second it draws $P$ multinomial distributions $\{\phi^p\}_{p=1}^{P}$ from Dir($\beta$). They model the word distributions for the $P$ topics. Finally a Bernoulli distribution $\pi$ from Dir($\gamma$) and another multinomial distribution $\phi^B$ from Dir($\beta$) are set to model the existence of background words. Given an input message $m$, the model then draws a topic $z_m$ based on $\theta$. For the $l$-th word $w_{m,l}$ in $m$, an indicator $Y_{m,l}$ is first sampled from $\pi$. If $Y_{m,l} = 1$, then $w_{m,l}$ is a topic word and is sampled from $\phi^{z_m}$; otherwise, $w_{m,l}$ would be treated as a background word and sampled from $\phi^B$. Figure 1 gives the graphical model of Twitter LDA.

In training, we concatenate the post and the response of each sample to form a short document and employ the collapsed Gibbs sampling algorithm (Zhao et al., 2011) to estimate the parameters. After we get the estimations of the parameters, we use them to assign a topic $p$ to each post $\mathbf{X}$ by the generation process described above. Then, we pick top $n$

words with the highest probabilities in $\phi^p$ for topic $p$ as the topic word list. Each word $w$ in the topic word list corresponds to a topic distribution which is calculated by Equation (7), where $C_{wp}$ is the number of times that $w$ is assigned to topic $p$ in training. These topic distributions will be used as topic vectors in TAJA-Seq2Seq.

$$p(p|w) \propto \frac{C_{wp}}{\sum_{p'} C_{wp'}}. \tag{7}$$

### 3.2 Structure of TAJA-Seq2Seq network

The data format of TAJA-Seq2Seq is $(\mathbf{K}^p, \mathbf{X}, \mathbf{Y})$, where $p$ is the topic assigned to $\mathbf{X}$ and $\mathbf{K}^p = (\mathbf{k}_1^p, \mathbf{k}_2^p, ..., \mathbf{k}_n^p)$ are the topic words of $\mathbf{X}$. $\mathbf{k}_j^p$ represents the topic vector of the $j$-th word calculated using Equation (7). Figure 2 gives the structure of our topic augmented joint attention based Seq2Seq model (TAJA-Seq2Seq) for response generation. TAJA-Seq2Seq has a content encoder and a topic encoder. The content encoder embeds an input post from both ends into vector space with a bidirectional LSTM-RNN. The topic encoder obtains the vectors of the topic words of the input post by looking up the topic vector table. In addition to the two encoders, TAJA-Seq2Seq also contains a topic summarizer which transforms the vectors of topic words to a single vector $q$ with an MLP. On top of each encoder, there is an attention module. For the $i$-th word in the decoder, the content attention module takes the former hidden state of the decoder $\mathbf{s}_{i-1}$, the hidden states of the content encoder $\{\mathbf{h}_t\}_{t=1}^T$, and $q$ from the topic summarizer as input, and calculate combination weights of $\{\mathbf{h}_t\}_{t=1}^T$. Similarly, in the topic attention module, for the $i$-th word in the decoder, the weights for topic vectors $\{\mathbf{k}_j^p\}_{j=1}^n$ are calculated according to the former hidden state of the decoder $\mathbf{s}_{i-1}$, the topic vector themselves, and the final state of the content encoder $\mathbf{h}_T$. Here, $q$ encodes the topic information and helps calibrate the content emphasis in the input post, and $\mathbf{h}_T$ encodes the content information and helps select topic words that are more relevant to the $i$-th word in the generated response. Different from the attention module in Seq2Seq, topic information and content information determine the combination weights of each other in the joint attention mechanism and make more accurate decisions on the semantic focus of every response word.

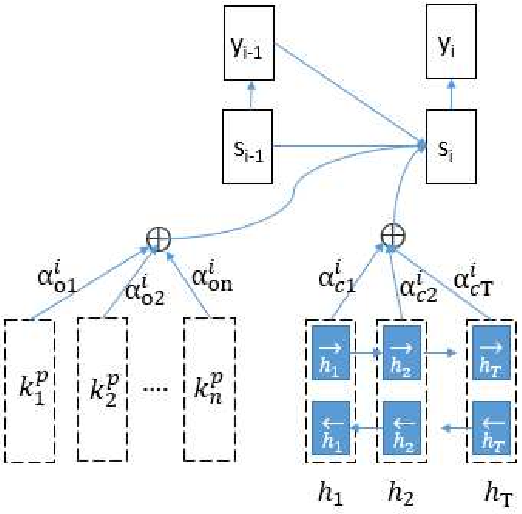


**Figure 3:** Connections of the decoder and 2 attention modules

Specifically, $\forall i$ in the decoder, the content attention module calculates unnormalized weights $\mathbf{E}_c^i = (e_{c1}^i, e_{c2}^i, \ldots, e_{cT}^i)$ for $\{\mathbf{h}_t\}_{t=1}^T$. $e_{ct}^i$ corresponds to $\mathbf{h}_t$ and is calculated as

$$e_{ct}^i = \eta_c(\mathbf{s}_{i-1}, \mathbf{h}_t, q), \tag{8}$$

where $\eta_c$ is an MLP with tanh as the activation function and $q$ is the output of the topic summarizer. The topic summarizer concatenates $(k_1^p, k_2^p, \ldots, k_n^p)$ as a vector $q'$, then it calculates $q$ as

$$q = \tanh(\mathbf{W}_s \cdot q' + \mathbf{b}_s), \tag{9}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are parameters. The topic attention module calculates unnormalized weights $\mathbf{E}_o^i = (e_{o1}^i, e_{o2}^i, ..., e_{on}^i)$ for $(k_1^p, k_2^p, ..., k_n^p)$. $\forall j$, $e_{oj}^i$ is represented as

$$e_{oj}^i = \eta_o(\mathbf{s}_{i-1}, \mathbf{k}_j^p, \mathbf{h}_T). \tag{10}$$

Both $\mathbf{E}_c^i$ and $\mathbf{E}_o^i$ are further normalized with a softmax function as described in Equation (6).

The decoder of TAJA-Seq2Seq is an RNN language model with the output of the content attention module and the output of the topic attention module as input. Specifically, for step $i$, the calculation details of the decoder are given by

$$\begin{aligned} \mathbf{s}_i &= f(y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i, \mathbf{o}_i) \\ \mathbf{p}_i &= softmax(\mathbf{s}_i, y_{i-1}), \end{aligned} \tag{11}$$

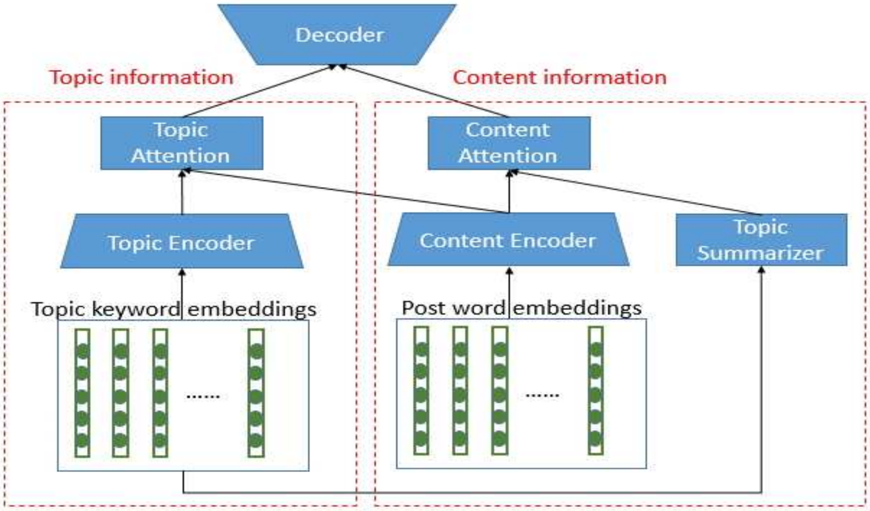


**Figure 2:** Structure of TAJA-Seq2Seq

where $\mathbf{c}_i$ and $\mathbf{o}_i$ are the output of the content attention module and the topic attention module respectively, and $\mathbf{p}_i$ is the probability distribution of the $i$-th word in response generation. Figure 3 shows how the two attention modules work with the decoder. We use beam search(Steinbiss et al., 1994) to generate responses using the trained model. We publish the Python code of TAJA-Seq2Seq in Github[2].

TAJA-Seq2Seq makes use of the topic information provided by Twitter LDA as prior knowledge to enhance the performance of response generation. Since topic keywords of a post connect it with other semantic related responses under the same topic during training, the topic keywords mainly play two roles, classification and association, during the generation of the response. On the one hand, some of the topic keywords, like “thriller” for post “I’ve watched The Orphan” , classify the post to a specific semantic group (such as descriptions of thrillers ) and relate it to many candidate responses in this topic group, hence makes the model to generate more targeted and well-focused responses such as “This movie is so disgusting”. On the other hand some topic words like “high-heels” for post “You are so tall”, extend the semantic meaning of posts and can lead to more diverse and novel responses like “I am not tall. This is because of my high-heels”.

Moreover, in TAJA-Seq2Seq, content vectors and topical vectors jointly affect the weight learning of each other in the joint attention module during the generation of each word in the response sequence. This simulates how people behave in conversation. When reacting to an input post, people find relevant information from their prior knowledge according to the content of the post, let the information help them determine the important parts in the post that they want to focus, and use both the information from prior knowledge and the main content focus from the post to create their reply. In our model, topic attention simulates the process of relevant information finding and content attention simulates the determination of the content focus.

TAJA-Seq2Seq also helps make better choice on the first word in the response. First words matter much because the decoder contains a language model part to guarantee the fluency of the generated responses. If the first word is wrongly chosen, then the whole sentence will never have a chance to go back to a proper meaning. In traditional attention based Seq2Seq, the weights of $\{\mathbf{h}_t\}_{t=1}^T$ of the first word in response only depend on themselves since there is no $\mathbf{s}_{i-1}$ when $i$ equals to 0. In TAJA-Seq2Seq, for the two attention modules, the weights of $\{\mathbf{h}_t\}_{t=1}^T$ and $\{\mathbf{k}_j^p\}_{j=1}^n$ for the first decoder state depend not only on themselves, but also on the topic vector $q$ and the content vector $\mathbf{h}_T$. Topics provide extra information for a better choice of the first word in response generation.

[2] https://github.com/LynetteXing1991/TAJA-Seq2Seq

## 4 Experiments

We evaluate TAJA-Seq2Seq from different perspectives and compare the model with a series of encoder-decoder based models using real-world data.

### 4.1 Experiment setup

**Data sets** we crawled $5,671,846$ post-response pairs from Sina Weibo[3]. We then use Stanford Chinese word segmenter[4] to tokenize post-response pairs as inputs of all models. Pairs with posts or responses longer than $50$ tokens are removed and we also remove responses with frequency higher than 10 since they would dominate the generated results if they are included in training data. $1,147,758$ distinct post-response pairs are left, and we use them to build the training data set $\mathcal{D}$. We then separately construct vocabulary sets for posts and responses. The sizes of two vocabulary sets are both $20,000$. The post vocabulary covers $98.8\%$ words that appear in posts and the response vocabulary covers $98.3\%$ in responses. Other words are treated as "UNK" in training. Our test data contains $154$ randomly sampled posts which are not in $\mathcal{D}$. We follow the same procedure as the training set $\mathcal{D}$ to construct the test set.

**Parameter tuning** in Twitter LDA, we set the number of topics $T$ as 200 and $\alpha = 1/T$, $\beta = 0.01$, $\gamma = 0.01$. For each topic, we set the number of topic words $n$ in the topic word list as 10. In all Seq2Seq based models, including our TAJA-Seq2Seq and all other baseline models, we set the dimensions of the hidden states of the encoder and the decoder as $1000$, and the dimensions of all word embeddings as $620$. We initialize the parameters of the Seq2Seq based models using isotropic Gaussian distributions and then use AdaDelta algorithm (Zeiler, 2012) to train the models on a NVIDIA Tesla K40 GPU.

### 4.2 Baselines

We compare our model with three baselines. All of the models are implemented with an open source deep learning tool, Blocks[5]. The three baselines are as follows,

**TA-Seq2Seq**: this is a simplified version of TAJA-Seq2Seq in which we remove the topic encoder and topic attention module, and only leave the content encoder, content attention and the topic summarizer.
**S2SA**: standard Seq2Seq model with attention mechanism which is firstly proposed in (Bahdanau et al., 2014) for machine translation. We implement it with a standard bi-directional encoder and an MLP as the attention network.
**S2SA-NMI**: Seq2Seq model with a NMI objective function proposed in (Li et al., 2015) to improve the diversity of the generated results. Since NMI-bidi performs better among the two models in (Li et al., 2015) according to the existing work, we select NMI-bidi as a baseline.

We do not compare our model with Seq2Seq without attention and the traditional retrieval-based methods because it has been proven in (Shang et al., 2015) that S2SA outperforms these two methods and represents the state-of-the-art in response generation.

### 4.3 Evaluation Methods

**Human Annotation**: we recruit human annotators to evaluate the quality of the generated responses of different models. For every post, we show top $10$ generated responses from every model. 4 labelers with Weibo experience are asked to judge the quality of the responses. Query-response pairs from different models are randomly shuffled and split to $4$ parts. The labeling criteria is

**+2**: The response is relevant, natural, well-focused and informative.

**+1**: The response is fair to answer this post, but its relevance with the post is not very strong. For example, some responses are general and can be used to respond to many posts, like "Yes, I see" , "Wow" and "No I don't know".

**0**: The response cannot be used to answer this post, semantically irrelevant or disfluent.

**-1**: The post is hard to understand so that the quality of the response cannot be judged.

We set the $-1$ score because some posts in Sina Weibo need special background knowledge to understand.

**D-score**: we use D-score to evaluate the diversity of the generated responses. For every post $p$, we pick responses labeled +1 and +2 from the judged

[3] www.sina.com

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

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

| Models | +2 | +1 | 0 | -1 | D-Score |
|---|---|---|---|---|---|
| TAJA-Seq2Seq | 43.5% | 25.8% | 30.5% | 0.3% | 30.34 |
| TA-Seq2Seq | 36.1% | 30.2% | 33.4% | 0.3% | 27.45 |
| S2SA-NMI | 30.1% | 37.2% | 32.5% | 0.3% | 25.48 |
| S2SA | 29.9% | 39.0% | 30.9% | 0.2% | 25.42 |

**Table 1:** Human Annotation Results and D-score

10 responses to construct a candidate list $\mathbf{R}^p = (r_1^p, r_2^p, ..., r_y^p)$, in which $y$ is the total number of proper generated responses of post $p$. Given the candidates of all posts, we calculate D-score as follows,

$$Dscore = \frac{1}{P * y} \Sigma_p \Sigma_{i=1}^{y} d_{r_i^p} \tag{12}$$

where $d_{r_i^p}$ is the number of distinct words in $r_i^p$ and $P$ is the number of posts. D-score represents the average number of distinct words in generated responses with good quality, and thus to some extent reflects the diversity of the generated responses of every model.

### 4.4 Results

For human annotation results, it is clear from Table 1 that TAJA-Seq2Seq and TA-Seq2Seq generate much more high-quality responses (responses labeled as +2) than other baselines, outperform 6% and 13.4% respectively. We conducted sign test, and the results show that the improvement is statistically significant (with $p$-value $< 0.01$). The total numbers of positive responses (responses labeled as +1 or +2) for all 4 models are relatively consistent and are all around 70%, while TAJA-Seq2Seq and TA-Seq2Seq decrease the number of fair and plain responses and increase the number of natural and interesting ones. This supports our claim on the advantage of our model, that is by leveraging topic information as prior knowledge, our model could involve various relevant information to enrich generated responses and make them more informative and diverse. An interesting phenomenon observed from the human annotation results is that the TA-Seq2Seq model contributes half of the improvement of the whole model. More precisely, the topic summarizer that helps calculate the weights of the context vectors can make considerable improvement itself. It indicates that topic information is really useful for response generation even in a simple way.

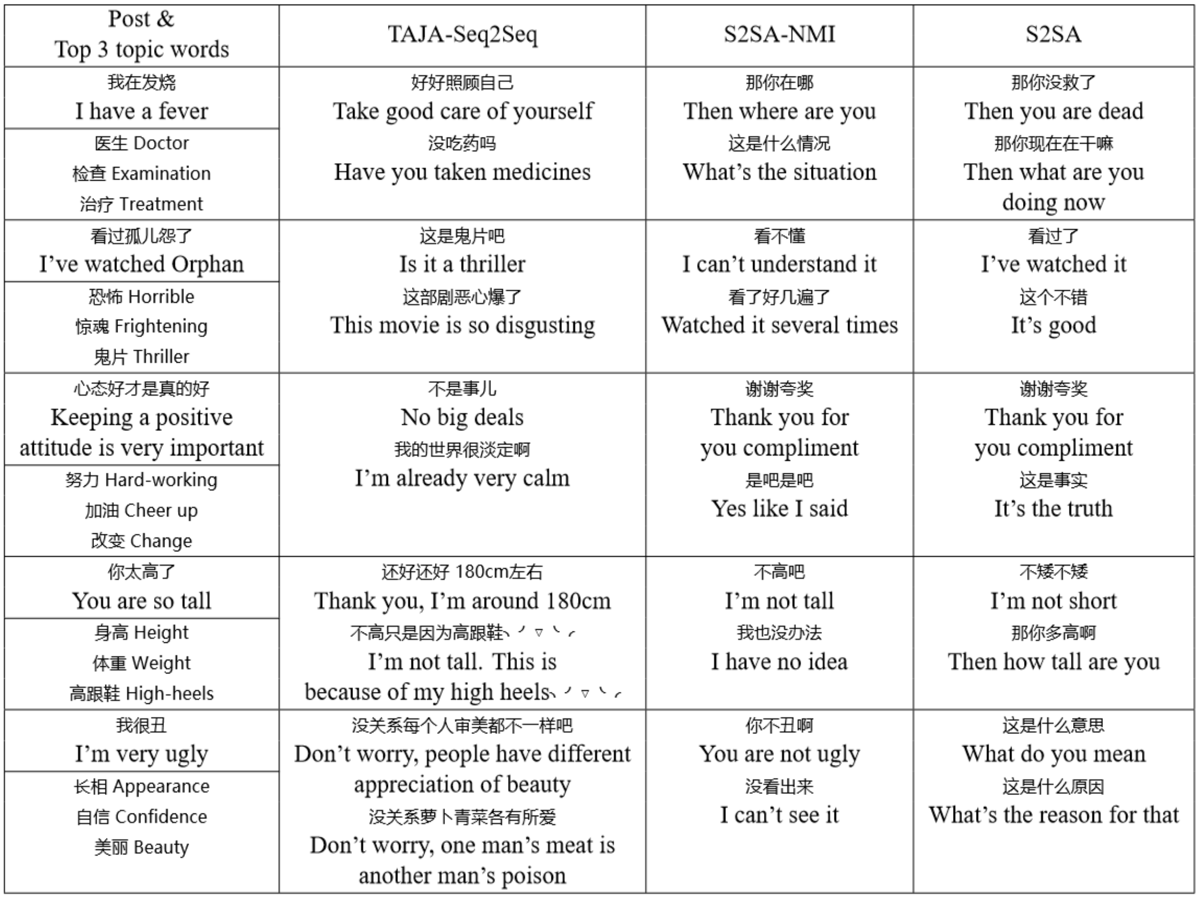
For D-score, we can see from Table 1 that TAJA-Seq2Seq has more distinct words in its good responses and outperforms traditional S2SA models to 4 points. The performance of models on diversity is consistent with that on quality in Table 1.

### 4.5 Case Study

In Table 2, we list some cases to show the quality of the generated responses and the effectiveness of topic information as prior knowledge. For every post, we list the top 3 topic words as showcase of topic information. From the table, the advantages of TAJA-Seq2Seq over the traditional S2SA can be summarized as: 1) TAJA-Seq2Seq connects the post with a topic that generally and accurately describes the characteristics of the post, like case 2,3 and 5. In case 2 TAJA-Seq2Seq categorizes the movie “The Orphan” directly into the “thriller” topic thus the generated responses are very targeted and well-focused on the characteristics of thrillers. On the other hand, responses from traditional S2SA fails to capture these characteristics and gives too general responses related to “watch movies”; 2) TAJA-Seq2Seq associates the post with other relevant elements, and makes the responses more informative and diverse. For example, in case 4, TAJA-Seq2Seq associates “tall” with relevant elements like “high heels”, and can generate interesting responses like “No this is because of my high heels”.

## 5 Related work

Besides traditional attention module introduced in background knowledge section, there are also other encoder-decoder based models aiming at improving the quality of generated responses from various perspectives. Similar with TAJA-Seq2Seq, there are encoder-decoder models that take the benefits of extra informations. For involving context information into one-round chat, A. Sordoni(Sordoni et al., 2015) proposed a encoder-



| Post &<br>Top 3 topic words | TAJA-Seq2Seq | S2SA-NMI | S2SA |
|---|---|---|---|
| 我在发烧<br>I have a fever | 好好照顾自己<br>Take good care of yourself<br>没吃药吗<br>Have you taken medicines | 那你在哪<br>Then where are you<br>这是什么情况<br>What's the situation | 那你没救了<br>Then you are dead<br>那你现在在干嘛<br>Then what are you doing now |
| 医生 Doctor<br>检查 Examination<br>治疗 Treatment | | | |
| 看过孤儿怨了<br>I've watched Orphan | 这是鬼片吧<br>Is it a thriller<br>这部剧恶心爆了<br>This movie is so disgusting | 看不懂<br>I can't understand it<br>看了好几遍了<br>Watched it several times | 看过了<br>I've watched it<br>这个不错<br>It's good |
| 恐怖 Horrible<br>惊魂 Frightening<br>鬼片 Thriller | | | |
| 心态好才是真的好<br>Keeping a positive attitude is very important | 不是事儿<br>No big deals<br>我的世界很淡定啊<br>I'm already very calm | 谢谢夸奖<br>Thank you for you compliment<br>是吧是吧<br>Yes like I said | 谢谢夸奖<br>Thank you for you compliment<br>这是事实<br>It's the truth |
| 努力 Hard-working<br>加油 Cheer up<br>改变 Change | | | |
| 你太高了<br>You are so tall | 还好还好 180cm左右<br>Thank you, I'm around 180cm<br>不高只是因为高跟鞋ヽ╯▽╰╭<br>I'm not tall. This is because of my high heelsヽ╯▽╰╭ | 不高吧<br>I'm not tall<br>我也没办法<br>I have no idea | 不矮不矮<br>I'm not short<br>那你多高啊<br>Then how tall are you |
| 身高 Height<br>体重 Weight<br>高跟鞋 High-heels | | | |
| 我很丑<br>I'm very ugly | 没关系每个人审美都不一样吧<br>Don't worry, people have different appreciation of beauty<br>没关系萝卜青菜各有所爱<br>Don't worry, one man's meat is another man's poison | 你不丑啊<br>You are not ugly<br>没看出来<br>I can't see it | 这是什么意思<br>What do you mean<br>这是什么原因<br>What's the reason for that |
| 长相 Appearance<br>自信 Confidence<br>美丽 Beauty | | | |

**Figure 4:** Cases of the generation responses from TAJA-Seq2Seq, S2SA-NMI and S2SA

decoder based model, the encoder of which is a two-part feed-forward network that concludes context and post information separately and fed them both to the decoder for generating responses. To keep speaker information consistent during chat, J. Li (Li et al., 2016) added personal information from knowledge base as an extra input of Seq2Seq and concatenate it with the word embedding for every time step. Before TAJA-Seq2Seq, there are also works aiming at simulating humans' behaviour during chat to improve chat machine. J. Gu(Gu et al., 2016) proposed copying mechanism to mimic the repeating phenomenon during human conversation. K. Yao (Yao et al., 2015) added an extra RNN between the encoder and the decoder with attention to conclude and explain intentions from posts. Moreover, topic information has been used in similar tasks and has been proven effective. S. Ghosh(Ghosh et al., 2016) proposed Contextual LSTM that involved topic information to represent contexts and concatenated topic information with the word embedding for every time step. CLSTM improved 3 NLP tasks, word prediction, next sentence selection and sentence topic prediction. These success of topical information in similar NLP tasks provokes our trial of involving it into neural response generation.

## 6 Conclusion and Future work

We proposed TAJA-Seq2Seq to naturally incorporate topic information as prior knowledge into Seq2Seq network and improved the quality and diversity of response generation. In future work, we will implement this structure for other kinds of extra information, such as contexts and knowledge from knowledge base, to enhance response generation quality in different aspects.

## References

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

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Cho et al.2014] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

[Cho et al.2015] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *Multimedia, IEEE Transactions on*, 17(11):1875–1886.

[Ghosh et al.2016] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.

[Gu et al.2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Li et al.2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

[Li et al.2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model.

[Mikolov et al.2010] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, volume 2, page 3.

[Shang et al.2015] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

[Sordoni et al.2015] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses.

[Steinbiss et al.1994] Volker Steinbiss, Bach-Hiep Tran, and Hermann Ney. 1994. Improvements in beam search. In *ICSLP*, volume 94, pages 2143–2146.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Yao et al.2015] Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. *Computer Science*.

[Zeiler2012] Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

[Zhao et al.2011] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.