Learning to Control the Specificity in Neural Response Generation

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu and Xueqi Cheng
University of Chinese Academy of Sciences, Beijing, China
CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology
{zhangruqing,fanyixing}@software.ict.ac.cn
{guojiafeng,lanyanyan,junxu,cxq}@ict.ac.cn

Abstract

In conversation, a general response (e.g., "I don't know") could correspond to a large variety of input utterances. Previous generative conversational models usually employ a single model to learn the relationship between different utterance-response pairs, thus tend to favor general and trivial responses which appear frequently. To address this problem, we propose a novel controlled response generation mechanism to handle different utterance-response relationships in terms of specificity. Specifically, we introduce an explicit specificity control variable into a sequence-to-sequence model, which interacts with the usage representation of words through a Gaussian Kernel layer, to guide the model to generate responses at different specificity levels. We describe two ways to acquire distant labels for the specificity control variable in learning. Empirical studies show that our model can significantly outperform the state-of-the-art response generation models under both automatic and human evaluations.

1 Introduction

Human-computer conversation is a critical and challenging task in AI and NLP. There have been two major streams of research in this direction, namely task oriented dialog and general purpose dialog (i.e., chit-chat). Task oriented dialog aims to help people complete specific tasks such as buying tickets or shopping, while general purpose dialog attempts to produce natural and meaningful conversations with people regarding a wide range of topics in open domains (Perez-Marin, 2011; Sordoni et al.). In recent years, the latter has attracted much attention in both academia and industry as a way to explore the possibility in developing a general purpose AI system in language (e.g., chatbots).

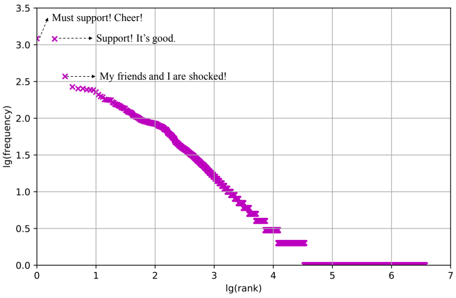

Figure 1: Rank-frequency distribution of the responses in the chit-chat corpus, with x and y axes being lg(rank order) and lg(frequency) respectively.

A widely adopted approach to general purpose dialog is learning a generative conversational model from large scale social conversation data. Most methods in this line are constructed within the statistical machine translation (SMT) framework, where a sequence-to-sequence (Seq2Seq) model is learned to "translate" an input utterance into a response. However, general purpose dialog is intrinsically different from machine translation. In machine translation, since every sentence and its translation are semantically equivalent, there exists a 1-to-1 relationship between them. However, in general purpose dialog, a general response (e.g., "I don't know") could correspond to a large variety of input utterances. For example, in the chit-chat corpus used in this study (as shown in Figure 1), the top three most frequently appeared responses are "Must support! Cheer!", "Support! It's good.", and "My friends and I are shocked!", where the response "Must support! Cheer!" is used for 1216 different input utterances. Previous Seq2Seq models, which treat all the utterance-response pairs uniformly and employ a single model to learn the relationship between them, will inevitably favor such general responses with high frequency. Although these responses are safe for replying different utterances, they are boring and trivial since they carry little information, and may quickly lead to an end of the conversation.

There have been a few efforts attempting to address this issue in literature. Li et al. (2016a) proposed to use the Maximum Mutual Information (MMI) as the objective to penalize general responses. It could be viewed as a post-processing approach which did not solve the generation of trivial responses fundamentally. Xing et al. (2017) pre-defined a set of topics from an external corpus to guide the generation of the Seq2Seq model. However, it is difficult to ensure that the topics learned from the external corpus are consistent with that in the conversation corpus, leading to the introduction of additional noises. Zhou et al. (2017) introduced latent responding factors to model multiple responding mechanisms. However, these latent factors are usually difficult in interpretation and it is hard to decide the number of the latent factors.

In our work, we propose a novel controlled response generation mechanism to handle different utterance-response relationships in terms of specificity. The key idea is inspired by our observation on everyday conversation between humans. In human-human conversation, people often actively control the specificity of responses depending on their own response purpose (which might be affected by a variety of underlying factors like their current mood, knowledge state and so on). For example, they may provide some interesting and specific responses if they like the conversation, or some general responses if they want to end it. They may provide very detailed responses if they are familiar with the topic, or just "I don't know" otherwise. Therefore, we propose to simulate the way people actively control the specificity of the response.

We employ a Seq2Seq framework and further introduce an explicit specificity control variable to represent the response purpose of the agent. Meanwhile, we assume that each word, beyond the semantic representation which relates to its meaning, also has another representation which relates to the usage preference under different response purpose. We name this representation as the usage representation of words. The specificity control variable then interacts with the usage representation of words through a Gaussian Kernel layer, and guides the Seq2Seq model to generate responses at different specificity levels. We refer to our model as Specificity Controlled Seq2Seq model (SC-Seq2Seq). Note that unlike the work by (Xing et al., 2017), we do not rely on any external corpus to learn our model. All the model parameters are learned on the same conversation corpus in an end-to-end way.

We employ distant supervision to train our SC-Seq2Seq model since the specificity control variable is unknown in the raw data. We describe two ways to acquire distant labels for the specificity control variable, namely Normalized Inverse Response Frequency (NIRF) and Normalized Inverse Word Frequency (NIWF). By using normalized values, we restrict the specificity control variable to be within a pre-defined continuous value range with each end has very clear meaning on the specificity. This is significantly different from the discrete latent factors in (Zhou et al., 2017) which are difficult in interpretation.

We conduct an empirical study on a large public dataset, and compare our model with several state-of-the-art response generation methods. Empirical results show that our model can generate either general or specific responses, and significantly outperform existing methods under both automatic and human evaluations.

2 Related Work

In this section, we briefly review the related work on conversational models and response specificity.

2.1 Conversational Models

Automatic conversation has attracted increasing attention over the past few years. At the very beginning, people started the research using handcrafted rules and templates (Walker et al., 2001; Williams et al., 2013; Henderson et al., 2014). These approaches required little data for training but huge manual effort to build the model, which is very time-consuming. For now, conversational models fall into two major categories: retrieval-based and generation-based. Retrieval-based conversational models search the most suitable response from candidate responses using different schemas (Kearns, 2000; Wang et al., 2013; Yan et al., 2016). These methods rely on preexisting responses, thus are difficult to be extended to open domains (Zhou et al., 2017). With the large amount of conversation data available on the Internet, generation-based conversational models developed within a SMT framework (Ritter et al., 2011; Cho et al., 2014; Bahdanau et al., 2015) show promising results. Shang et al. (2015) generated replies for short-text conversation by encoder-decoder-based neural network with local and global attentions. Serban et al. (2016) built an end-to-end dialogue system using generative hierarchical neural network. Gu et al. (2016) introduced copynet to simulate the repeating behavior of humans in conversation. Similarly, our model is also based on the encoder-decoder framework.

2.2 Response Specificity

Some recent studies began to focus on generating more specific or informative responses in conversation. It is also called a diversity problem since if each response is more specific, it would be more diverse between responses of different utterances. As an early work, Li et al. (2016a) used Maximum Mutual Information (MMI) as the objective to penalize general responses. Later, Li et al. (2017) proposed a data distillation method, which trains a series of generative models at different levels of specificity and uses a reinforcement learning model to choose the model best suited for decoding depending on the conversation context. These methods circumvented the general response issue by using either a post-processing approach or a data selection approach.

Besides, Li et al. (2016b) tried to build a personalized conversation engine by adding extra personal information. Xing et al. (2017) incorporated the topic information from an external corpus into the Seq2Seq framework to guide the generation. However, external dataset may not be always available or consistent with the conversation dataset in topics. Zhou et al. (2017) introduced latent responding factors to the Seq2Seq model to avoid generating safe responses. However, these latent factors are usually difficult in interpretation and hard to decide the number.

Moreover, Mou et al. (2016) proposed a content-introducing approach to generate a response based on a predicted keyword. Yao et al. (2016) attempted to improve the specificity with the reinforcement learning framework by using the averaged IDF score of the words in the response as a reward. Shen et al. (2017) presented a conditional variational framework for generating specific responses based on specific attributes. Unlike these existing methods, we introduce an explicit specificity control variable into a Seq2Seq model to handle different utterance-response relationships in terms of specificity.

3 Specificity Controlled Seq2Seq Model

In this section, we present the Specificity Controlled Seq2Seq model (SC-Seq2Seq), a novel Seq2Seq model designed for actively controlling the generated responses in terms of specificity.

3.1 Model Overview

The basic idea of a generative conversational model is to learn the mapping from an input utterance to its response, typically using an encoder-decoder framework. Formally, given an input utterance sequence $\mathbf{X} = (x_1, x_2, \ldots, x_T)$ and a target response sequence $\mathbf{Y} = (y_1, y_2, \ldots, y_{T'})$, a neural Seq2Seq model is employed to learn $p(\mathbf{Y}|\mathbf{X})$ based on the training corpus $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})|\mathbf{Y} \text{ is the response of } \mathbf{X}\}$. By maximizing the likelihood of all the utterance-response pairs with a single mapping mechanism, the learned Seq2Seq model will inevitably favor those general responses that can correspond to a large variety of input utterances.

To address this issue, we assume that there are different mapping mechanisms between utterance-response pairs with respect to their specificity relation. Rather than involving some latent factors, we propose to introduce an explicit variable s into a Seq2Seq model to handle different utterance-response mappings in terms of specificity. By doing so, we hope that (1) s would have explicit meaning on specificity, and (2) s could not only interpret but also actively control the generation of the response $\mathbf{Y}$ given the input utterance $\mathbf{X}$. The goal of our model becomes to learn $p(\mathbf{Y}|\mathbf{X}, s)$ over the corpus $\mathcal{D}$, where we acquire distant labels for s from the same corpus for learning. The overall architecture of SC-Seq2Seq is depicted in Figure 2, and we will detail our model as follows.

3.1.1 Encoder

The encoder is to map the input utterance $\mathbf{X}$ into a compact vector that can capture its essential topics. Specifically, we use a bi-directional GRU (Cho et al., 2014) as the utterance encoder, and each word x_i is firstly represented by its semantic representation $\mathbf{e}_i$ mapped by semantic embedding

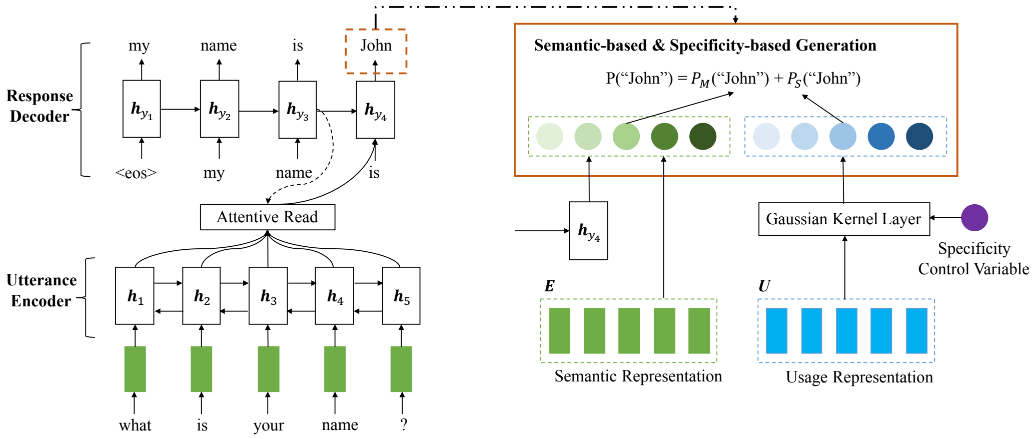

Figure 2: The overall architecture of SC-Seq2Seq model.

matrix $\mathbf{E}$ as the input of the encoder. Then, the encoder represents the utterance $\mathbf{X}$ as a series of hidden vectors $\{\mathbf{h}_t\}_{t=1}^{T}$ modeling the sequence from both forward and backward directions. Finally, we use the final backward hidden state as the initial hidden state of the decoder.

3.1.2 Decoder

The decoder is to generate a response $\mathbf{Y}$ given the hidden representations of the input utterance $\mathbf{X}$ under some specificity level denoted by the control variable s. Specifically, at step t, we define the probability of generating any target word y_t by a "mixture" of probabilities:

$$p(y_t) = \beta p_M(y_t) + \gamma p_S(y_t), \tag{1}$$

where $p_M(y_t)$ denotes the semantic-based generation probability, $p_S(y_t)$ denotes the specificity-based generation probability, β and γ are the coefficients.

Specifically, $p_M(y_t)$ is defined the same as that in traditional Seq2Seq model (Sutskever et al., 2014):

$$p_M(y_t = w) = \mathbf{w}^{\mathrm{T}}(\mathbf{W}_M^h \cdot \mathbf{h}_{y_t} + \mathbf{W}_M^e \cdot \mathbf{e}_{t-1} + \mathbf{b}_M), \tag{2}$$

where $\mathbf{w}$ is a one-hot indicator vector of the word w and $\mathbf{e}_{t-1}$ is the semantic representation of the $t-1$-th generated word in decoder. $\mathbf{W}_M^h$, $\mathbf{W}_M^e$ and $\mathbf{b}_M$ are parameters. $\mathbf{h}_{y_t}$ is the t-th hidden state in the decoder which is computed by:

$$\mathbf{h}_{y_t} = f(y_{t-1}, \mathbf{h}_{y_{t-1}}, \mathbf{c}_t), \tag{3}$$

where f is a GRU unit and $\mathbf{c}_t$ is the context vector to allow the decoder to pay different attention to different parts of input at different steps (Bahdanau et al., 2015).

$p_S(y_t)$ denotes the generation probability of the target word given the specificity control variable s. Here we introduce a Gaussian Kernel layer to define this probability. Specifically, we assume that each word, beyond its semantic representation $\mathbf{e}$, also has a usage representation $\mathbf{u}$ mapped by usage embedding matrix $\mathbf{U}$. The usage representation of a word denotes its usage preference under different specificity. The specificity control variable s then interacts with the usage representations through the Gaussian Kernel layer to produce the specificity-based generation probability $p_S(y_t)$:

$$\begin{aligned} p_S(y_t = w) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(\Psi_S(\mathbf{U}, \mathbf{w}) - s)^2}{2\sigma^2}), \\ \Psi_S(\mathbf{U}, \mathbf{w}) &= \sigma(\mathbf{w}^{\mathrm{T}}(\mathbf{U} \cdot \mathbf{W}_U + \mathbf{b}_U)), \end{aligned} \tag{4}$$

where σ^2 is the variance, and $\Psi_S(\cdot)$ maps the word usage representation into a real value with the specificity control variable s as the mean of the Gaussian distribution. $\mathbf{W}_U$ and $\mathbf{b}_U$ are parameters to be learned. Note here in general we can use any real-value function to define $\Psi_S(\mathbf{U}, \mathbf{w})$. In this work, we use the sigmoid function $\sigma(\cdot)$ for $\Psi_S(\mathbf{U}, \mathbf{w})$ since we want to define s within the range [0,1] so that each end has very clear meaning on the specificity, i.e., 0 denotes the most general response while 1 denotes the most specific response. In the next section, we will also keep this property when we define the distant label for the control variable.

3.2 Distant Supervision

We train our SC-Seq2Seq model by maximizing the log likelihood of generating responses over the training set $\mathcal{D}$:

$$\mathcal{L} = \sum_{(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}} \log P(\mathbf{Y}|\mathbf{X}, s; \theta). \tag{5}$$

where θ denotes all the model parameters. Note here since s is an explicit control variable in our model, we need the triples $(\mathbf{X}, \mathbf{Y}, s)$ for training. However, s is not directly available in the raw conversation corpus, thus we acquire distant labels for s to learn our model. We introduce two ways of distant supervision on the specificity control variable s, namely Normalized Inverse Response Frequency (NIRF) and Normalized Inverse Word Frequency (NIWF).

3.2.1 Normalized Inverse Response Frequency

Normalized Inverse Response Frequency (NIRF) is based on the assumption that a response is more general if it corresponds to more input utterances in the corpus. Therefore, we use the inverse frequency of a response in a conversation corpus to indicate its specificity level. Specifically, we first build the response collection $\mathcal{R}$ by extracting all the responses from $\mathcal{D}$. For a response $\mathbf{Y} \in \mathcal{R}$, let $f_{\mathbf{Y}}$ denote its corpus frequency in $\mathcal{R}$, we compute its Inverse Response Frequency (IRF) as:

$$\text{IRF}_{\mathbf{Y}} = \log(1 + |\mathcal{R}|)/f_{\mathbf{Y}}, \tag{6}$$

where $|\mathcal{R}|$ denotes the size of the response collection $\mathcal{R}$. Next, we use the min-max normalization method (Jain et al., 2005) to obtain the NIRF value. Namely,

$$\text{NIRF}_{\mathbf{Y}} = \frac{\text{IRF}_{\mathbf{Y}} - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'})}{\max_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'}) - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'})}. \tag{7}$$

where $\max(\text{IRF}_{\mathcal{R}})$ and $\min(\text{IRF}_{\mathcal{R}})$ denotes the maximal and minimum IRF value in $\mathcal{R}$ respectively. The NIRF value is then used as the distant label of s in training. Note here by using normalized values, we aim to constrain the specificity control variable s to be within the pre-defined continuous value range [0,1].

3.2.2 Normalized Inverse Word Frequency

Normalized Inverse Word Frequency (NIWF) is based on the assumption that the specificity level of a response depends on the collection of words it contains, and the sentence is more specific if it contains more specific words. Hence, we can use the inverse corpus frequency of the words to indicate the specificity level of a response. Specifically, for a word y in the response $\mathbf{Y}$, we first obtain its Inverse Word Frequency (IWF) by:

$$\text{IWF}_y = \log(1 + |\mathcal{R}|)/f_y, \tag{8}$$

where f_y denotes the number of responses in $\mathcal{R}$ containing the word y. Since a response usually contains a collection of words, there would be multiple ways to define the response-level IWF value, e.g., sum, average, minimum or maximum of the IWF values of all the words. In our work, we find that the best performance can be achieved by using the maximum of the IWF of all the words in $\mathbf{Y}$ to represent the response-level IWF by

$$\text{IWF}_{\mathbf{Y}} = \max_{y \in \mathbf{Y}}(\text{IWF}_y). \tag{9}$$

This is reasonable since a response is specific as long as it contains some specific words. We do not require all the words in a response to be specific, thus sum, average, and minimum would not be appropriate operators for computing the response-level IWF. Again, we use min-max normalization to obtain the NIWF value for the response $\mathbf{Y}$.

3.3 Specificity Controlled Response Generation

Given a new input utterance, we can employ the learned SC-Seq2Seq model to generate responses at different specificity levels by varying the control variable s. In this way, we can simulate human conversations where one can actively control the response specificity depending on his/her own mind. When we apply our model to a chatbot, there might be different ways to use the control variable for conversation in practice. If we want the agent to always generate informative responses, we can set s to 1 or some values close to 1. If we want the agent to be more dynamic, we can sample s within the range [0,1] to enrich the styles in the response. We may further employ some reinforcement learning technique to learn to adjust the control variable depending on users' feedbacks. This would make the agent even more vivid, and we leave this as our future work.

4 Experiment

In this section, we conduct experiments to verify the effectiveness of our proposed model.

4.1 Dataset Description

We conduct our experiments on the public Short Text Conversation (STC) dataset[1] released in NTCIR-13. STC maintains a large repository of post-comment pairs from the Sina Weibo which is one of the popular Chinese social sites.

[1] http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm

Utterance-response pairs	3,788,571
Utterance vocabulary #w	120,930
Response vocabulary #w	524,791
Utterance max #w	38
Utterance avg #w	13
Response max #w	74
Response avg #w	10

Table 1: Short Text Conversation (STC) data statistics: #w denotes the number of Chinese words.

STC dataset contains roughly 3.8 million post-comment pairs, which could be used to simulate the utterance-response pairs in conversation. We employ the Jieba Chinese word segmenter[2] to tokenize the utterances and responses into sequences of Chinese words, and the detailed dataset statistics are shown in Table 1. We randomly selected two subsets as the development and test dataset, each containing 10k pairs. The left pairs are used for training.

4.2 Baselines Methods

We compare our proposed SC-Seq2Seq model against several state-of-the-art baselines: (1) **Seq2Seq-att**: the standard Seq2Seq model with the attention mechanism (Bahdanau et al., 2015); (2) **MMI-bidi**: the Seq2Seq model using Maximum Mutual Information (MMI) as the objective function to reorder the generated responses (Li et al., 2016a); (3) **MARM**: the Seq2Seq model with a probabilistic framework to model the latent responding mechanisms (Zhou et al., 2017); (4) **Seq2Seq+IDF**: an extension of Seq2Seq-att by optimizing specificity under the reinforcement learning framework, where the reward is calculated as the sentence level IDF score of the generated response (Yao et al., 2016). We refer to our model trained using NIRF and NIWF as **SC-Seq2Seq**$_{\text{NIRF}}$ and **SC-Seq2Seq**$_{\text{NIWF}}$ respectively.

4.3 Implementation Details

As suggested in (Shang et al., 2015), we construct two separate vocabularies for utterances and responses by using 40,000 most frequent words on each side in the training data, covering 97.7% words in utterances and 96.1% words in responses respectively. All the remaining words are replaced by a special token <UNK> symbol.

We implemented our model in Tensorflow[3]. We tuned the hyper-parameters via the development set. Specifically, we use one layer of bi-directional GRU for encoder and another uni-directional GRU for decoder, with the GRU hidden unit size set as 300 in both the encoder and decoder. The dimension of semantic word embeddings in both utterances and responses is 300, while the dimension of usage word embeddings in responses is 50. We apply the Adam algorithm (Kingma and Ba, 2015) for optimization, where the parameters of Adam are set as in (Kingma and Ba, 2015). The variance σ^2 of the Gaussian Kernel layer is set as 1, and all other trainable parameters are randomly initialized by uniform distribution within [-0.08,0.08]. The mini-batch size for the update is set as 128. We clip the gradient when its norm exceeds 5.

Our model is trained on a Tesla K80 GPU card, and we run the training for up to 12 epochs, which takes approximately five days. We select the model that achieves the lowest perplexity on the development dataset, and we report results on the test dataset.

4.4 Evaluation Methodologies

For evaluation, we follow the existing work and employ both automatic and human evaluations: (1) **distinct-1 & distinct-2** (Li et al., 2016a): we count numbers of distinct unigrams and bigrams in the generated responses, and divide the numbers by total number of generated unigrams and bigrams. Distinct metrics (both the numbers and the ratios) can be used to evaluate the specificity/diversity of the responses. (2) **BLEU** (Papineni et al., 2002): BLEU has been proved strongly correlated with human evaluations. BLEU-n measures the average n-gram precision on a set of reference sentences. (3) **Average & Extrema** (Serban et al., 2017): Average and Extrema projects the generated response and the ground truth response into two separate vectors by taking the mean over the word embeddings or taking the extremum of each dimension respectively, and then computes the cosine similarity between them. (4) **Human evaluation**: Three labelers with rich Weibo experience were recruited to conduct evaluation. Responses from different models are randomly mixed for labeling. Labelers refer to 300 random sampled test utterances and score the quality of the responses with the following criteria: 1) **+2**: the response is not only semantically relevant and grammatical, but also informat-

[2] https://pypi.python.org/pypi/jieba

[3] https://www.tensorflow.org/

	Models	distinct-1	distinct-2	BLEU-1	BLEU-2	Average	Extrema
SC-Seq2Seq$_{\text{NIRF}}$	$s=1$	5258/0.064	16195/0.269	15.109	7.023	0.578	0.380
	$s=0.8$	5337/0.065	16105/0.271	15.112	7.003	0.578	0.381
	$s=0.5$	5318/0.065	16183/0.269	15.054	7.001	0.578	0.380
	$s=0.2$	5323/0.065	16087/0.270	15.168	7.032	0.580	0.380
	$s=0$	5397/0.066	16319/0.271	15.093	7.011	0.577	0.380
SC-Seq2Seq$_{\text{NIWF}}$	$s=1$	**11588/0.116**	**27144/0.347**	12.392	5.869	0.554	0.353
	$s=0.8$	6006/0.051	17843/0.257	11.492	5.703	0.553	0.350
	$s=0.5$	2835/0.050	9537/0.235	**16.122**	**7.674**	**0.609**	**0.399**
	$s=0.2$	1534/0.048	5117/0.218	8.313	4.058	0.542	0.335
	$s=0$	1038/0.046	3154/0.211	4.417	3.283	0.549	0.334

Table 2: Model analysis of our SC-Seq2Seq under the automatic evaluation.

Models	distinct-1	distinct-2	BLEU-1	BLEU-2	Average	Extrema
Seq2Seq-att	5048/0.060	15976/0.168	15.062	6.964	0.575	0.376
MMI-bidi	5074/0.082	12162/0.287	15.772	7.215	0.586	0.381
MARM	2566/0.096	3294/0.312	7.321	3.774	0.512	0.336
Seq2Seq+IDF	4722/0.052	15384/0.229	14.423	6.743	0.572	0.369
SC-Seq2Seq$_{\text{NIWF},s=1}$	**11588/0.116**	**27144/0.347**	12.392	5.869	0.554	0.353
SC-Seq2Seq$_{\text{NIWF},s=0.5}$	2835/0.050	9537/0.235	**16.122**	**7.674**	**0.609**	**0.399**

Table 3: Comparisons between our SC-Seq2Seq and the baselines under the automatic evaluation.

ive and interesting; 2) **+1**: the response is grammatical and can be used as a response to the utterance, but is too trivial (e.g., "I don't know"); 3) **+0**: the response is semantically irrelevant or ungrammatical (e.g., grammatical errors or UNK). Agreements to measure inter-rater consistency among three labelers are calculated with the Fleiss' kappa (Fleiss and Cohen, 1973).

4.5 Evaluation Results

Model Analysis: We first analyze our models trained with different distant supervision information. For each model, given a test utterance, we vary the control variable s by setting it to five different values (i.e., 0, 0.2, 0.5, 0.8, 1) to check whether the learned model can actually achieve different specificity levels. As shown in Table 2, we find that: (1) The SC-Seq2Seq model trained with NIRF cannot work well. The test performances are almost the same with different s value. This is surprising since the NIRF definition seems to be directly corresponding to the specificity of a response. By conducting further analysis, we find that even though the conversation dataset is large, it is still limited and a general response could appear very few times in this corpus. In other words, the inverse frequency of a response is very weakly correlated with its response specificity. (2) The SC-Seq2Seq model trained with NIWF can achieve our purpose. By varying the control variable s from 0 to 1, the generated responses turn from general to specific as measured by the distinct metrics. The results indicate that the max inverse word frequency in a response is a good distant label for the response specificity. (3) When we compare the generated responses against ground truth data, we find the SC-Seq2Seq$_{\text{NIWF}}$ model with the control variable s set to 0.5 can achieve the best performances. The results indicate that there are diverse responses in real data in terms of specificity, and it is necessary to take a balanced setting if we want to fit the ground truth.

Baseline Comparison: The performance comparisons between our model and the baselines are shown in Table 3. We have the following observations: (1) By using MMI as the objective, MMI-bidi can improve the specificity (in terms of distinct ratios) over the traditional Seq2Seq-att model. (2) MARM can achieve the best distinct ratios among the baseline methods, but the worst in terms of the distinct numbers. The results indicate that MARM tends to generate specific but very short responses. Meanwhile, its low BLEU scores also show that the responses generated by MARM deviate from the ground truth significantly. (3) By using the IDF information as the reward to train

	+2	+1	+0	kappa
Seq2Seq-att	29.32%	25.27%	45.41%	0.448
MMI-bidi	30.40%	24.85%	44.75%	0.471
MARM	20.11%	27.96%	51.93%	0.404
Seq2Seq+IDF	28.81%	23.87%	47.33%	0.418
SC-Seq2Seq$_{\text{NIWF,s=1}}$	42.47%	14.29%	43.24%	0.507
SC-Seq2Seq$_{\text{NIWF,s=0.5}}$	20.62%	40.16%	39.22%	0.451
SC-Seq2Seq$_{\text{NIWF,s=0}}$	14.34%	46.38%	39.28%	0.526

Table 4: Results on the human evaluation.

the Seq2Seq model, the Seq2Seq+IDF does not show much advantages, but only achieves comparable results as MMI-bidi. (4) By setting the control variable s to 1, our SC-Seq2Seq$_{\text{NIWF}}$ model can achieve the best specificity performance as evaluated by the distinct metrics. By setting the control variable s to 0.5, our SC-Seq2Seq$_{\text{NIWF}}$ model can best fit the ground truth data as evaluated by the BLEU scores, Average and Extrema. All the improvements over the baseline models are statistically significant (p-value < 0.01). These results demonstrate the effectiveness as well as the flexibility of our controlled generation model.

Table 4 shows the human evaluation results. We can observe that: (1) SC-Seq2Seq$_{\text{NIWF,s=1}}$ generates the most informative responses and interesting (labeled as "+2") and the least general responses than all the baseline models. Meanwhile, SC-Seq2Seq$_{\text{NIWF,s=0}}$ generates the most general responses (labeled as "+1"); (2) MARM generates the most bad responses (labeled as "+0"), which indicates the drawbacks of the unknown latent responding mechanisms; (3) The kappa values of our models are all larger than 0.4, considered as "moderate agreement" regarding quality of responses. The largest kappa value is achieved by SC-Seq2Seq$_{\text{NIWF,s=0}}$, which seems reasonable since it is easy to reach an agreement on general responses. Sign tests demonstrate the improvements of SC-Seq2Seq$_{\text{NIWF,s=1}}$ to the baseline models are statistically significant (p-value < 0.01). All the human judgement results again demonstrate the effectiveness of our controlled generation mechanism.

4.6 Case Study

To better understand how different models perform, we conduct some case studies. We randomly sample three utterances from the test dataset, and show the responses generated by different models.

Utterance: 三室一厅的帐篷，你值得拥有！A tent with three rooms and one hall, you're worth it!

Seq2Seq-att: 这个不错，很好！It looks good, very good!

MMI-bidi: 这个不错，我喜欢。It looks good, I like it.

MARM: 不错。Good.

Seq2Seq+IDF: 帐篷不错。The tent is good.

s=1: 我想知道这帐篷是什么牌子的？I want to know what the brand of this tent is?

s=0.8: 我想知道这是在哪买的？I want to know where to buy it?

s=0.5: 我也想买一个,可惜买不起！I want one too, but it's a pity that I can't afford it.

s=0.2: 不错不错！Good, good.

s=0: 真的吗？Really?

Utterance: 哎哟! 被发现了! 汤唯有必要那么美吗? Oh! Got caught! Why is Tang Wei so beautiful?

Seq2Seq-att: 我超喜欢她！I really like her!

MMI-bidi: 好美。我超喜欢她！So beautiful. I really like her!

MARM: 我不喜欢汤唯。I don't like Tang Wei.

Seq2Seq+IDF: 我超喜欢她的，我喜欢。I really like her, I like.

s=1: 我超喜欢她的眼妆，她的气质很美！I really like her eye make-up, and her temperament is so beautiful!

s=0.8: 我超喜欢她的，她的眼睛很雪亮！I really like her, and her eyes are so bright!

s=0.5: 我觉得她很美，很喜欢她！I think she is very beautiful, and I really like her!

s=0.2: 我觉得她很美！I think she is very beautiful.

s=0: 是的。Yes.

Utterance: 将来也要养一只这么乖的的萨摩，让它陪着孩子一起长大。I will keep such a sweet samoyed, and let it accompany my children to grow up.

Seq2Seq-att: 这不是萨摩吗？Is it a samoyed?

MMI-bidi: 好可爱的萨摩。What a lovely samoyed.

MARM: 好可爱啊。So lovely.

Seq2Seq+IDF: 好可爱的狗狗。What a lovely dog.

s=1: 萨摩耶好可爱，真想养一只! The samoyed is so lovely, and I really want to keep one!

s=0.8: 萨摩好可爱，好想掐掐。The samoyed is so lovely, and I really want to pinch it.

s=0.5: 好可爱的狗狗，好可爱的狗狗。What a lovely dog, what a lovely dog.

s=0.2: 好可爱！好可爱！So lovely, so lovely!

s=0: 好可爱！So lovely!

Table 5: Examples of response generation from the STC test data. $s = 1, 0.8, 0.5, 0.2, 0$ are the outputs of our SC-Seq2Seq$_{\text{NIWF}}$ with different s values.

As shown in Table 5, we can find that: (1) The responses generated by the four baselines are often quite general and short, which may quickly lead to an end of the conversation. (2) SC-Seq2Seq$_{\text{NIWF}}$ with large control variable values (i.e., $s > 0.5$) can generate very long and specific responses. In these responses, we can find many informative words. For example, in case 2 with s as 1 and 0.8, we can find words like "眼妆(eye make-up)", "气质(temperament)" and "雪亮(bright)" which are quite specific and strongly related to the conversation topic of "beauty". (3) When we decrease the control variable value, the generated responses become more and more general and shorter from our SC-Seq2Seq$_{\text{NIWF}}$ model.

爸爸(dad)		水果(fruits)		脂肪肝(fatty liver)		单反相机(DSLR)	
Usage	Semantic	Usage	Semantic	Usage	Semantic	Usage	Semantic
更好(better)	妈妈(mother)	尝试(attempt)	蔬菜(vegetables)	坐久(outsit)	胖(fat)	亚洲杯(Asian Cup)	照相机(camera)
睡觉(sleep)	哥哥(brother)	诱惑(tempt)	牛奶(milk)	素食主义(vegetarian)	减肥(diet)	读取(read)	摄影(photography)
快乐(happy)	老公(husband)	表现(express)	西瓜(watermelon)	散步(walk)	高血压(hypertension)	半球(hemispherical)	镜头(shot)
无聊(boring)	爷爷(grandfather)	拥有(own)	米饭(rice)	因果关系(causality)	亚健康(sub-health)	防辐射(anti-radiation)	影楼(studio)
电影(movie)	姑娘(girl)	梦想(dream)	巧克力(chocolate)	哑铃(dumbbell)	呕吐(emesis)	无人机(UAV)	写真(image)

Table 6: Target words and their top-5 similar words under usage and semantic representations respectively.

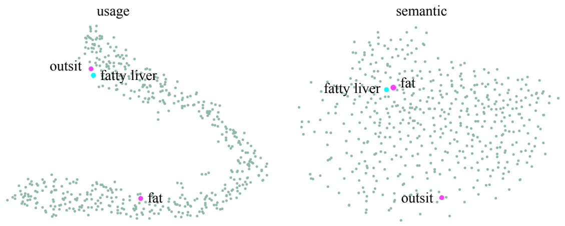

Figure 3: t-SNE embeddings of usage and semantic vectors.

4.7 Analysis on Usage Representations

We also conduct some analysis to understand the usage representations of words introduced in our model. We randomly sample 500 words from our SC-Seq2Seq$_{\text{NIWF}}$ and apply t-SNE (Maaten and Hinton, 2008) to visualize both usage and semantic embeddings. As shown in Figure 3, we can see that the two distributions are significantly different. In the usage space, words like "脂肪肝(fatty liver)" and "久坐(outsit)" lie closely which are both specific words, and both are far from the general words like "胖(fat)". On the contrary, in the semantic space, "脂肪肝(fatty liver)" is close to "胖(fat)" since they are semantically related, and both are far from the word "久坐(outsit)". Furthermore, given some sampled target words, we also show the top-5 similar words based on cosine similarity under both representations in Table 6. Again, we can see that the nearest neighbors of a same word are quite different under two representations. Neighbors based on semantic representations are semantically related, while neighbors based on usage representations are not so related but with similar specificity levels.

5 Conclusion

We propose a novel controlled response generation mechanism to handle different utterance-response relationships in terms of specificity. We introduce an explicit specificity control variable into the Seq2Seq model, which interacts with the usage representation of words to generate responses at different specificity levels. Empirical results showed that our model can generate either general or specific responses, and significantly outperform state-of-the-art generation methods.

6 Acknowledgments

This work was funded by the 973 Program of China under Grant No. 2014CB340401, the National Natural Science Foundation of China (NSFC) under Grants No. 61425016, 61472401, 61722211, and 20180290, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, and the National Key R&D Program of China under Grants No. 2016QY02D0405.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the conference on empirical methods in natural language processing*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285.

Michael Kearns. 2000. Cobot in lambdamoo: A social statistics agent.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data distillation for controlling specificity in dialogue generation. *arXiv preprint arXiv:1702.06703*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

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

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*, pages 196–205.

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.

Marilyn A Walker, Rebecca Passonneau, and Julie E Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 515–522. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the conference on empirical methods in natural language processing*.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

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

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–64. ACM.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *AAAI*, pages 3400–3407.