# Mechanism-Aware Neural Machine for Dialogue Response Generation

**Ganbin Zhou**[1,2], **Ping Luo**[1], **Rongyu Cao**[1,2], **Fen Lin**[3], **Bo Chen**[3], **Qing He**[1]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China.
{zhouganbin, caory}@ics.ict.ac.cn, {luop, heqing}@ict.ac.cn
[2]University of Chinese Academy of Sciences, Beijing 100049, China.
[3]Pattern Recognition Center, WeChat Technical Architecture Department, Tencent, China.

## Abstract

To the same utterance, people's responses in everyday dialogue may be diverse largely in terms of content semantics, speaking styles, communication intentions and so on. Previous generative conversational models ignore these 1-to-$n$ relationships between a post to its diverse responses, and tend to return high-frequency but meaningless responses.

In this study we propose a mechanism-aware neural machine for dialogue response generation. It assumes that there exists some *latent responding mechanisms*, each of which can generate different responses for a single input post. With this assumption we model different responding mechanisms as latent embeddings, and develop a *encoder-diverter-decoder* framework to train its modules in an end-to-end fashion. With the learned latent mechanisms, for the first time these decomposed modules can be used to encode the input into mechanism-aware context, and decode the responses with the controlled generation styles and topics. Finally, the experiments with human judgements, intuitive examples, detailed discussions demonstrate the quality and diversity of the generated responses with 9.80% increase of acceptable ratio over the best of six baseline methods.

## Introduction

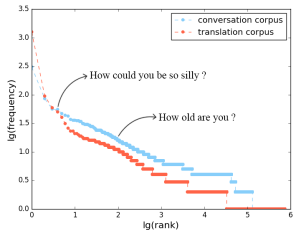
Conversational models, aiming at generating relevant and fluent responses in free-form natural language, have attracted increasing studies for the dialogue-based interface with its wide application fields from customer service to user entertainment (Abu Shawar and Atwell 2007; Grosz 2016). Previous rule-based (Williams and Young 2007; Misu et al. 2012; Young et al. 2013) and retrieval-based (Ji, Lu, and Li 2014) conversational models requires manual efforts in rule developing and feature engineering, or can only response the posts in pre-existing cases, thus are difficult to be extended to open domains. Recently, the vast amount of dialogue text generated by social media provides the data basis for generative models of dialogue systems, which are promising to outperform the conventional ones (Shang, Lu, and Li 2015).

Generative conversational models, which learn the mapping from an input post $\boldsymbol{x}$ to its response $\boldsymbol{y}$, are typically motivated by the previous studies in statistic machine translation (SMT). Instead of translating from one language to another, they "translate" an input post $\boldsymbol{x}$ to a response $\boldsymbol{y}$ via maximizing the probability of $p(\boldsymbol{y}|\boldsymbol{x})$. However, since the generic responses such as "*I see*", "*that's OK*" and "*that's great*" appear quite frequently in the corpus, the training objective of maximum likelihood tends to produce high-frequency responses, which might be meaningless and lack of diversity (Li et al. 2016).

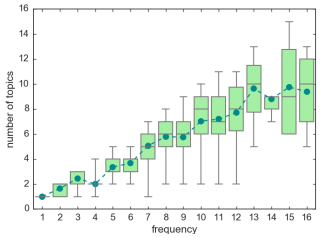
In this study, we find that the training corpus for conversational models is intrinsically different from the one for translation models in terms of *output diversity*. In translation corpus, since every sentence in a language and its translation in another language are semantically equivalent, there exists a *1-to-1* relationship between them. However, in conversation corpus, an input post might correspond to multiple responses with different semantics and speaking styles. For example, in free-chat corpus used in this study the input sentence "*how could you be so silly*" includes 62 different responses. It means that a *1-to-n* relationship between a post to its responses actually exists in open-domain conversation.

Furthermore, we argue that this issue of response diversity mainly comes from the different language mechanisms people use in responding the same utterance. For example, considering the input "*have you eaten yet?*" (a widely-used sentence in Chinese for greeting), the respondent who prefers rhetorical questions could response with "*how about you?*". In contrary, the respondent who prefers declarative sentences could response affirmatively with "*yes, I have*". Hence, even for the same input the responses generated by different mechanisms may be largely dissimilar in terms of language style and response content.

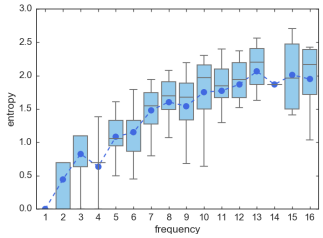
To address this issue of response diversity, we explicitly consider the multiplicity of responding mechanisms in modeling dialogues and propose a probabilistic framework of Mechanism-Aware Responding Machine (MARM). Specifically, we model the responding mechanisms as latent embeddings, and represent the mapping from an post $\boldsymbol{x}$ to its response $\boldsymbol{y}$ as a mixture of these responding mechanisms. Different from the conventional neural *encoder-decoder* (Cho et al. 2014; Sutskever, Vinyals, and Le 2014) for response generation (Shang, Lu, and Li 2015; Yin et al. 2016), a framework of *encoder-diverter-decoder* is developed, where the module of *diverter* is used to generate mechanism-aware context. After the model parameters are learned, the most likely mechanisms to an input post $\boldsymbol{x}$ are selected to encode




(a) Rank-frequency distribution.




(b) Topic number in the responses of conversation corpus.




(c) Topic entropy in the responses of conversation corpus.

Figure 1: Quantitative study of response diversity. The figures of topic number and entropy of translation corpus are not shown, since the translation sentences to a specific input always belong to the same topic.

the input into mechanism-aware context, and decode the responses with the controlled generation styles and topics.

To our best knowledge, this is the first time to generate mechanism-dependent responses, which helps to distinguish ungrammatical output from the meaningful but infrequent ones. To an input post $\boldsymbol{x}$, assume $\boldsymbol{y}'$ be an ungrammatical output while $\boldsymbol{y}^*$ be a meaningful but infrequent one. The principle of likelihood maximization assigns low probabilities to both $p(\boldsymbol{y}'|\boldsymbol{x})$ and $p(\boldsymbol{y}^*|\boldsymbol{x})$, thus it cannot distinguish them. Nevertheless, with the learned responding mechanisms we experimentally discover that upon some mechanism $m$ $p(\boldsymbol{y}^*|\boldsymbol{x}, m)$ might reach a big value, indicating that $\boldsymbol{y}^*$ is more likely to be generated by $\boldsymbol{x}$ using the mechanism $m$. However, $p(\boldsymbol{y}'|\boldsymbol{x}, m)$ still remains in a low value for all the mechanisms. Furthermore, we also observe that the generated responses upon different mechanisms vary in styles and topics. Therefore, we believe that responses generated by MARM are more likely to be meaningful and diverse.

In summary, our main contributions are three folds: 1) for the first time we quantitatively measured the response diversity in the conversation corpus; 2) we propose an encoder-diverter-decoder framework, in which the response mechanisms on the responses can be explicitly modeled to produce the mechanism-aware context for mechanism-dependent response generation; 3) we empirically show that the proposed method can yield performance visibly better than the other six neural-based generative models in terms of both response relevance and diversity by the human judgement. Specifically, it generates 9.80% increase of acceptable ratio over the best baseline method (see Table 1).

Finally, we would like to emphasize that the proposed mechanism-aware responding machine is task-agnostic and can be applied to any generative conversational model. In this study we apply it to the conversation model (Cho et al. 2014), where only one-round dialogue is considered. It can be easily extended to other one-round dialogue models, and also the model (Serban et al. 2015), where the context from previous multiple rounds of dialogues are considered.

## Quantitative Study of Response Diversity

Here, we quantitatively study the issue of response diversity in the training corpus for conversation compared with the one for machine translation. Specifically, we are given two corpora, namely $\mathcal{D}^c = \{(\boldsymbol{x}, \boldsymbol{y})|\boldsymbol{y}$ is the response of input $\boldsymbol{x}\}$ and $\mathcal{D}^t = \{(\boldsymbol{x}, \boldsymbol{y})|\boldsymbol{y}$ is translation of input $\boldsymbol{x}$ in different language$\}$. We collected $\mathcal{D}^c$ similar as suggested in (Wang et al. 2013) from the open social platform Tencent Weibo (detailed in the experiment section). We used a public corpus for machine translation (CWMT 2013) as $\mathcal{D}^t$. Originally, the size of $\mathcal{D}^t$ is bigger than that of $\mathcal{D}^c$. For fair comparison, we randomly sampled the original $\mathcal{D}^t$ so that the size of these two data sets is equal, namely $|\mathcal{D}^c| = |\mathcal{D}^t| = 780,852$.

For each corpus we first check how frequent each input sentence $\boldsymbol{x}$ occurs in it. Fig. 1(a) shows the rank-frequency distribution of the input sentences in the two corpora, with the $x$ and $y$ axes being $\lg$(rank order) and $\lg$(frequency) respectively. It shows that this sentence-level frequency exhibits the similar Zipf pattern, which widely occurs for the word frequency in the natural language corpus. For example, the sentences "how could you be so silly?" and "how old are you?" appear 62 and 13 times in the conversation corpus. Also, it shows that the curve for the conversation corpus is far above the one for the translation corpus excepts the starting points. There is a fact that 24.23% of the input sentences in the conversation corpus occur at least twice, however, this number decreases to 4.81% for the translation corpus.

With these frequent-occurred input posts in the conversation corpus, we further explore the diversity of their responses in terms of content semantics. Here, we firstly train Biterm topic model (Yan et al. 2013) on about 20 million sentences, apply it (with 500 topics) to inferring the topics on the whole corpus, and assign to each response only one topic with the maximal value of topic distribution. Then, for each input post, we count the number of different topics of its corresponding responses and calculate the entropy of their topic distribution. We observe that the diversity of corresponding responses increases with the number of covered topics and the entropy value of the topic distribution.

Fig. 1(b) and 1(c) show the response diversity in the conversation corpus. In these two figures, the $x$ axis represents the frequency of an input sentence, and the $y$ axis represents the number of covered topics (in Fig. 1(b)) and the entropy of the topic distribution (in Fig. 1(c)) for its responses, respectively. These two figures show that the average values of the corresponding $y$ axis (the dotted line) increase along

the frequency of the input post. It means that when an input post occurs more frequent in the conversation corpus, it is more likely that it leads to more diverse responses.

We also conduct the similar analysis on the translation corpus and obtain totally different results (the corresponding figures are omitted due to the space limitation). Though some sentences occur multiple times in the source language, the multiple translation sentences to a specific input always belong to the same topic. Furthermore, we calculate the edit distances among the translation sentences to a single input and find that they are different slightly. This result is not surprising since the two parallel sentences in the source and target language should have the same content semantics. Nevertheless, as to the natural language conversation (especially for free chat), the responses become quite divergent especially for the widely-occurred input. Therefore, we need to explicitly model the multiplicity of response mechanism in open environment conversation.

## Mechanism-Aware Response Machine

### Modeling of Response Mechanisms

Given an input sequence $\boldsymbol{x} = (x_1, x_2, \cdots, x_T)$ and a response sequence $\boldsymbol{y} = (y_1, y_2, \cdots, y_{T'})$, the generative conversation model aims to learn $p(\boldsymbol{y}|\boldsymbol{x})$ based on the training corpus $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})|\boldsymbol{y}$ is the response of input $\boldsymbol{x}\}$. The conventional neural encoder-decoder model first summarizes the post as a vector representation, then feeds this representation to a decoder to generate responses. Similar framework has been applied in machine translation with remarkable success (Cho et al. 2014; Sutskever, Vinyals, and Le 2014). However, the task of machine translation, which estimates the probability of a target language sentence conditioned on the source language sentence with the same meaning, is much easier than the task of conversation modeling with large degree of response diversity (as shown in the previous section). Thus, the modeling of $p(\boldsymbol{y}|\boldsymbol{x})$ for natural language conversation should be complex enough to represent all the suitable and diverse responses.

To this end, in this study we assume that there are $M$ latent mechanisms $\{m_i\}_{i=1}^M$ for response generation. Then, $p(\boldsymbol{y}|\boldsymbol{x})$ can be expanded as follows,

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} p(\boldsymbol{y}, m_i|\boldsymbol{x}) = \sum_{i=1}^{M} p(m_i|\boldsymbol{x}) p(\boldsymbol{y}|m_i, \boldsymbol{x}) \quad (1)$$

In Equ. (1), $p(m_i|\boldsymbol{x})$ represents the probability of the mechanism $m_i$ conditioned on $\boldsymbol{x}$. This probability actually measures the degree that $m_i$ can generate the response for $\boldsymbol{x}$. The bigger of this value is, the more degree that the mechanism $m_i$ can be used to generate the responses for $\boldsymbol{x}$. Additionally, $p(\boldsymbol{y}|m_i, \boldsymbol{x})$ measures the probability that the response $\boldsymbol{y}$ is generated by the mechanism $m_i$ for $\boldsymbol{x}$.

Now, the question is how to model $p(m_i|\boldsymbol{x})$ and $p(\boldsymbol{y}|m_i, \boldsymbol{x})$ in the framework of encoder-diverter-decoder. As shown in Fig. 2, a module of diverter is developed to bridge encoder and decoder. The diverter takes the hidden states of the encoder as input, which forms the summary

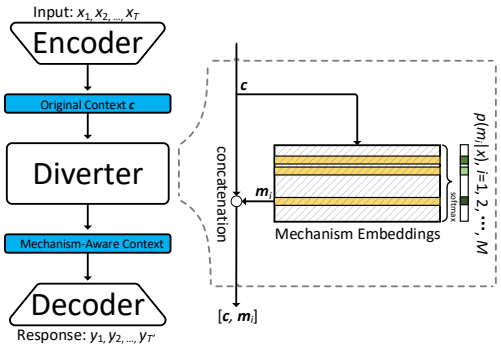



Figure 2: Structure of encoder-diverter-decoder model.

context $\boldsymbol{c}$ for the input post $\boldsymbol{x}$. Then, $p(m_i|\boldsymbol{x})$ can be modeled as follows (shown in the right part of Fig. 2),

$$p(m_i|\boldsymbol{x}) = \frac{\exp g(\boldsymbol{m}_i, \boldsymbol{c})}{\sum_{k=1}^{M} \exp g(\boldsymbol{m}_k, \boldsymbol{c})} \quad (2)$$

where $g$ can be nonlinear, potentially multi-layered function and $\boldsymbol{m}_i$ represents the embedding of the $i$-th mechanism. Here, $\{\boldsymbol{m}_i\}_{i=1}^M$ are trained as model parameters. Additionally, to avoid overfitting $g$ is defined with the *maxout* activation function (Goodfellow et al. 2013):

$$\begin{aligned} g(\boldsymbol{m}_i, \boldsymbol{c}) &= \boldsymbol{m}_i^{\mathrm{T}} \boldsymbol{W}_t \boldsymbol{t} \\ \boldsymbol{t} &= [\max\{\widetilde{t}_{2j-1}, \widetilde{t}_{2j}\}]_{j=1,2,\cdots,l_c}^{\mathrm{T}} \\ \widetilde{\boldsymbol{t}} &= \boldsymbol{W}_c \boldsymbol{c} \end{aligned} \quad (3)$$

where $\widetilde{t}_j$ is the $j$-th element of the vector $\widetilde{\boldsymbol{t}}$, $\boldsymbol{W}_t \in \mathbb{R}^{l_m \times l_c}$, and $\boldsymbol{W}_c \in \mathbb{R}^{2l_c \times l_c}$. Here, $l_c$ and $l_m$ denote the dimensions of $\boldsymbol{c}$ and $\boldsymbol{m}_i$ respectively.

Next, to model $p(\boldsymbol{y}|m_i, \boldsymbol{x})$ we must consider how an input $\boldsymbol{x}$ and a mechanism $m_i$ jointly determine the response $\boldsymbol{y}$. Since the hidden context $\boldsymbol{c}$ give a representation of the input $\boldsymbol{x}$, $\boldsymbol{c}$ can be combined with $\boldsymbol{m}_i$ to form a mechanism-aware context. For model simplicity, the concatenation of $[\boldsymbol{c}; \boldsymbol{m}_i]$ is utilized to form this mechanism-aware context. With this adapted context as input, the decoder is expected to generate mechanism-aware response for $p(\boldsymbol{y}|m_i, \boldsymbol{x})$.

It is worth mentioning that the proposed diverter model is independent of the concrete methods on how the decoder use the context for response generation. The mechanism-aware context can be fed to only the first hidden state unit (Sutskever, Vinyals, and Le 2014) or every hidden state unit in the decoder (Cho et al. 2014). The recent attention-based decoder (Bahdanau, Cho, and Bengio 2015) can also be applied to this mechanism-aware context to generate different context for every hidden state unit in the decoder. In this paper, the method in (Cho et al. 2014) is adopted. The details on the decoder is omitted due to the space limitation.

With the modeling of $p(m_i|\boldsymbol{x})$ and $p(\boldsymbol{y}|m_i, \boldsymbol{x})$ the objective of likelihood maximization, namely

$$\sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}^c} \log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}^c} \log \sum_{i=1}^{M} p(m_i|\boldsymbol{x}) p(\boldsymbol{y}|m_i, \boldsymbol{x}) \quad (4)$$

is used to learn the mechanism embeddings $\{\boldsymbol{m}_i\}_{i=1}^M$ and other model parameters. Observed from Equ.(4), the computational complexity is linear to the number of mechanisms. Thus, it is not difficult to capture complex phenomena in natural language if large number of mechanisms needed.

Note that calculating the logarithm of the total probability (required here) may cause overflow or underflow problems, which can be avoided by the technique of numerical computation in (LogSumExp 2016).

### Mechanism-Aware Response Generation

With all the responding mechanisms $\{\boldsymbol{m}_i\}_{i=1}^M$ obtained, we develop the following mechanism-aware method to generate responses for a new input $\boldsymbol{x}$. First, with the probabilities of $p(m_i|\boldsymbol{x})$ we select the $L$ ($L < M$) mechanisms with the maximal values of $p(m_i|\boldsymbol{x})$. These $L$ mechanisms are most likely to generate the appropriate responses for $\boldsymbol{x}$. Then, for each selected mechanism $m_l$, we utilize beam search to generate $K$ responses candidates by maximizing $p(\boldsymbol{y}|m_l, \boldsymbol{x})$. Finally, all the $L \times K$ generated response candidates are re-ranked using following score:

$$p(\boldsymbol{y}, m_l|\boldsymbol{x}) = p(m_l|\boldsymbol{x})p(\boldsymbol{y}|m_l, \boldsymbol{x}) \tag{5}$$

where the candidate $\boldsymbol{y}$ is generated by mechanism $m_l$. Only the top $K$ candidates are returned as the final responses.

The ranking measure in Equ. (5) contains two folds. First, maximizing $p(m_l|\boldsymbol{x})$ guarantees that the responses generated by $m_l$ are relevant to the input $\boldsymbol{x}$. Second, maximizing $p(\boldsymbol{y}|m_l, \boldsymbol{x})$ guarantees that the response $\boldsymbol{y}$ generated by $m_l$ is fluent and grammatical for $\boldsymbol{x}$. Thus, this mechanism-aware method is expected to generate the appropriate responses, which are both relevant and fluent to the input.

Additionally, instead of using $p(\boldsymbol{y}|\boldsymbol{x})$, we use $p(\boldsymbol{y}, m_l|\boldsymbol{x})$ to rank the response candidate $\boldsymbol{y}$ generated by $m_l$. With this new measure, a response $\boldsymbol{y}$ which has a relatively low value of $p(\boldsymbol{y}|\boldsymbol{x})$ may be promoted to higher rank if $\boldsymbol{y}$ has higher probability $p(\boldsymbol{y}, m_l|\boldsymbol{x})$. It means that the grammatical but infrequent responses, whose values of $p(\boldsymbol{y}|\boldsymbol{x})$ are relatively low, may be ranked higher by some mechanism, and then more likely to be chosen in the final responses.

Furthermore, since the mechanism-aware contexts are different for different mechanisms, the responses vary among mechanisms. The experiments also show that different mechanisms has different influences on the wording and speaking styles in responding. Since the MARM generates responses from $L$ different mechanisms, these responses tend to be more diverse. Hence, we argue that the method of mechanism-aware response generation will produce meaningful and diverse responses, which will be further demonstrated in the experimental section.

## Experiment Process

### Dataset Details

To obtain the conversation corpus, we collected nearly 14 million post-response pairs from Tencent Weibo[1]. Then, we remove spams and advertisements from dataset, and only retain high-quality post-response tuples. Totally, we have $815,852$ pairs left, among which $775,852$ ones are for training, and $40,000$ for model validation.

### Benchmark Methods

We implemented six conversation models for comparison:

1) RNNs2s (Sutskever, Vinyals, and Le 2014): The one-layer seq2seq model, which uses the last hidden state of the encoder as the initial hidden state of the decoder. 2) RNNencdec (Cho et al. 2014): The one-layer encoder-decoder model, which feeds the last hidden state of encoder to every cell and softmax unit of the decoder. 3) RNNatt (Bahdanau, Cho, and Bengio 2015): The model based on the encoder-decoder framework with attention signal. 4) NRM (Shang, Lu, and Li 2015): The neural responding machine with both the global and local scheme for attention modeling. 5) MMMI-bidi and MMI-antiLM (Li et al. 2016): The one-layer encoder-decoder model using Maximum Mutual Information (MMI) as the objective function to reorder generated responses. It obtains two variants $\lambda = 0.5$ and $\gamma = 1$.

Note that the benchmarks are the state-of-the-art for dialogue generation based on one-round input. They differ in how the context of the input post from the encoder is fed to the decoder for response generation. The proposed mechanism-aware model with the diverter module can be applied to any of these models, to improve the responding performance in another vertical direction. In this study the mechanism-aware model based on RNNencdec is implemented for evaluation. Again, we stress that the mechanism-aware model can be easily applied to the models (Serban et al. 2015), where the context is summarized from previous multiple rounds of dialogues. In the future MARM for multi-round dialogue systems will be evaluated.

### Implementation Details

Note that segmentation granularity and vocabulary size have an impact on model performances, for fair comparison, we used the vocabulary of 8,000 words (a mixture of Chinese words and characters) for all models. This vocabulary covers $99.93\%$ of the words in the corpus. All the other characters are replaced with a special token "UNK".

As suggested in (Shang, Lu, and Li 2015), the word embeddings for the encoders and decoders are treated respectively. Some initial experiments demonstrated that the two separate sets of word embeddings can improve the performance. For fair comparison, the dimension of the word embedding is set to 128 for all the models. As suggested in (Cho et al. 2014), the GRU unit is simpler and faster to converge than LSTM. Thus, we applied the one-layer GRU units (each with 1024 cells) to all the models in experiment. For MARM, the number of mechanisms is $M = 4$. As the mechanism increases, the language styles of some mechanisms become similar. While more mechanisms yield better performance in terms of objective function, 4 mechanisms are suitable to generate responses with distinctive wording clusters and satisfactory quality in experiments. The mechanism embeddings with 128 dimensions are initialized by a uniform distribution between -0.2 and 0.2. For response generation, we select top $L = 2$ mechanisms for beam search,

[1] http://t.qq.com/?lang=en_US

Table 1: The results from human judgement.

| Models | %Acceptable* | | | | | %Bad | %Normal | %Good | Diversity |
|---|---|---|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 | | | | |
| RNNencdec | 59.67 | 57.17 | 55.78 | 54.50 | 53.13 | 46.83 | 36.76 | 16.41 | 1.183 |
| NRM | 57.33 | 55.17 | 54.00 | 52.83 | 51.40 | 48.57 | 37.49 | 13.94 | 1.457 |
| RNNatt | 52.00 | 50.50 | 47.89 | 46.25 | 45.13 | 54.87 | 36.93 | 8.20 | 1.513 |
| RNNs2s | 55.00 | 50.00 | 47.33 | 46.00 | 44.33 | 55.64 | 39.09 | 5.27 | 1.553 |
| MMI-antiLM | 49.00 | 45.67 | 44.00 | 43.25 | 43.40 | 56.60 | 36.47 | 6.93 | 1.053 |
| MMI-bidi | 58.67 | 58.33 | 55.89 | 54.67 | 54.60 | 45.40 | 45.33 | 9.27 | 1.523 |
| MARM | 64.67 | 66.67 | 65.22 | 64.33 | 64.40 | 35.60 | 41.13 | 23.27 | 2.687 |

*The"Top-$k$" denotes the responses with top-$k$ probabilities in each group.

the number of response candidates from each mechanism is $K = 5$. The beam size is 200 for all models.

All the other parameters are initialized by a uniform distribution between -0.01 and 0.01. In training, we divided the corpus into mini-batches whose size is 128, and used the RMSProp (Graves 2013) algorithm for optimization. The training stops if the perplexity of the validation set does not decrease in 7 consecutive epochs. For each model, we selected the parameters updated after epoch with the least perplexity for the further evaluation. All the models were implemented using Theano (Theano Development Team 2016).

### Human Judgement

Due to response diversity, it is practically impossible to establish a data set which adequately cover all the responses for given posts. Thus, the evaluation measures, such as the BLEU score (Papineni et al. 2002), are not appropriate. Additionally, although the perplexity and BLEU are widely used in translation evaluation the lower values of these measures do not lead to better responses evaluated by human judgement (Liu et al. 2016). Hence, we only use the careful human judgement in our experiment.

Several labelers from a professional company were invited to evaluate the quality of the responses for 300 random sampled posts. For each test post, every model generated 5 responses as a group. For each response the labelers were asked to score its quality with one of the following 3 levels:

1) Bad: the response is NOT grammatical or relevant. 2) Normal: the response is grammatical and relevant to the input post. 3) Good: beyond the level of Normal, the response is interesting and meaningful.

If a response is scored Normal or Good, we call this response is *acceptable*. If the scores for a single response are different, it will be considered again in a group discussion for a consensus. Additionally, to evaluate the diversity of the responses, for the 5 responses generated by a given model for a post, the labelers were asked to annotate the number of different meanings among the accepted responses.

Hence, with the labeling results we calculated the percentage of bad, normal, good, and acceptable responses, and also the average values of the different meanings in the responses. Note that previous studies (Ritter, Cherry, and Dolan 2011; Shang, Lu, and Li 2015) only used the top-1 response for human judgement, and did not consider the diversity of the generated responses.

## Experimental Results and Analysis

### Experimental Results

The experimental results are summarized in Table 1. It shows that the Acceptable ratio and Diversity of MARM visibly outperform other models. Consider the Top-5 Acceptable ratio. The best baseline method MMI-bidi obtain 54.60% Acceptable ratio, while MARM reaches 64.40% with the increase percentage of 17.95%. We observe that this improvement is mainly from more Good responses generated (16.41% vs. 23.27%), indicating that MARM outputs more meaningful responses in the experiments.

Also, it is interesting to see that for the 4 algorithms of RNNencdec, NRM, RNNatt, and RNNs2s their Acceptable ratios at Top $k$ level has a sharp decrease tendency with the increase of $k$. For example, the Acceptable ratio of RNNencdec decreases from 59.67% at Top-1 level to 53.13% at Top-5 level. It indicates that the responses ranked at lower positions generated by these models obtain more unsatisfactory quality. However, the Acceptable ratio of MARM decreases from 64.67% at Top-1 level to 64.40% at Top-5 level, only 0.27% decrease, indicating that MARM generates high-quality responses even when they are ranked at the lower positions. We believe that this stable performance is due to the fact that the 5 selected responses may be generated from different responding mechanisms, and each of them is one of the most probable responses generated by a selected mechanism.

Meanwhile, we observe that the diversity measure of MARM reaches 2.687, much bigger that 1.523 from the best baseline of MMI-bidi. Thus, compared with the other models, MARM not only promotes to generate relevant and grammatical responses, but also generates diverse ones with the support of multiple mechanisms in the experiments.

Furthermore, we find the following two interesting observations. 1) For the four methods (RNNencdec, NRM, RNNatt and RNNs2s) which directly model $p(\boldsymbol{y}|\boldsymbol{x})$, we observe that their Acceptable ratio decreases as the Diversity measure increases. In other words, fitting the data better leads to more similar responses for a given post. 2) The models of RNNatt and NRM using the attention technique obtain unsatisfactory performance. The attention technique was developed to consider the word alignment between the long sentences in the source and target languages for machine translation. However, this alignment may not occur frequently especially in the experimental training cor-

pus with response diversity. Thus, the models with attention technique may overfit the training data but achieve poor performance on the test data in the experiments.

## Analysis on Responding Mechanisms

In order to intuitively understand what the learned responding mechanisms are and how they influence the process of response generation, we identified the keywords in the responses generated by different responding mechanisms. With each mechanism $m_i$ we used it to generate 5 responses for each of the 300 posts in the test set. We put all the 1,500 responses from a mechanism $m_i$ together to form a pseudo document $D_i$. Then, we calculated the following measure with the $j$-th word and the $i$-th mechanism, namely $p(word_j|m_i) = \frac{n_i^j}{\sum_{k=1}^{M} n_k^j}$, where $n_i^j$ is the number of times that the $j$-th word occurs in $D_i$. Clearly, the bigger of this value, the more likely that the mechanism $m_i$ utilizes it for response generation. Finally, for each mechanism we listed the top-10 keywords ranked by this measure, as shown in Table 2. Note that in this table we only considered the keywords which occur enough times in $D_i$, namely $n_i^j > 100$.

Table 2: Keywords from different responding mechanisms

| Keyword in $m_1$ | | Keyword in $m_2$ | | Keyword in $m_3$ | | Keyword in $m_4$ | |
|---|---|---|---|---|---|---|---|
| Chinese | English | Chinese | English | Chinese | English | Chinese | English |
| 看 | look | 个 | a | 哪里 | where | 不会 | cannot |
| 好好 | ok | 看来 | seem | ? | ? | 一直 | always |
| 还是 | still | 不错 | nice | 呀 | oh | 可是 | however |
| 那 | that | 小 | little | 么 | why | 一定 | must |
| 注意 | attention | 还 | yet | 怎么 | how | 都 | all |
| 自己 | self | 这样 | this | 什么 | what | 陪 | company |
| 知道 | know | 人 | person | 在 | stay | 也 | also |
| 就 | at once | 可 | may | 样子 | appearance | 会 | can |
| 可以 | can | 微 | tiny | 想 | think | 认错 | admit |
| 去 | go to | 对 | yes | 说 | say | 很 | very |

Table 2 shows the keywords for each responding mechanisms. For $m_1$, conjunction words, such as *that*, *still* and *at once*, occupy a large proportion. For $m_2$, modifier words, such as *nice*, *little* and *tiny*, occupy a large proportion. For $m_3$, the words about questions, such as *where*, *why*, *how* and the question mark, occur frequently. Thus, this mechanism is more likely to generate interrogative or rhetorical sentences. For $m_4$, most of the keywords, such as *cannot*, *always* and *must*, have the affirmative or negative tones. Thus, it may generate the responses in the form of emphatic sentences. These observations indicate that the obtained mechanisms have certain influence on the wording and speaking styles in responding. Hence, for the same input, we can utilize different mechanisms to increase the diversity of responses.

Additionally, some examples in Table 3 empirically shows how the different mechanisms generate relevent, fluent and diverse responses. These examples are consistent with the analysis in Table 2.

Table 3: The responses from MARM for input examples

| Questions | Mechanism 1 | Mechanism 2 | Mechanism 3 | Mechanism 4 |
|---|---|---|---|---|
| 我要坐地铁回家<br>I take subway back home | 好的，路上注意安全<br>Ok, be safe on the subway | 真是个不错的选择<br>That's a good choice | 坐车去哪里<br>Where are you going on the subway | 坐车一定要注意安全<br>Be careful on the subway |
| 生活中怎样感到开心快乐<br>How do I feel happy in my life | 开心就来和我聊天<br>Have a talk with me if you are happy | 看来你今天心情很不错<br>You seem to be happy today | 有什么开心的事情<br>What are the happy things you have experieced | 和我聊天就很开心<br>Talk with me and you will be happy |
| 明天考试、求安慰<br>I have a test tomorrw, and I need some solace | 好好复习就好了<br>Just work hard on the reviews | 我会为你加油的<br>I will stand by you | 你要去哪里试试<br>Where are you taking the test | 我会一直安慰你的<br>I will always be on your side |
| 上班好累啊<br>The work is really hard | 那就好好休息<br>Take a good rest | 要懂得劳逸结合<br>You need to balance work and life | 你是不是累了<br>Are you tired | 上班就是这样<br>That is work |

## Related Work

The related works of conversation models are five-folds.

**Statistic Machine Translation.** The basic neural-based encoder-decoder framework for generative conversation models is actually from the studies of statistic machine translation. Sutskever et al. (2014) used a multilayered LSTM as the encoder and another deep LSTM as the decoder for machine translation. Later, Cho et al. (2014) proposed the RNN encoder-decoder framework, where the generated context from the input is fed to every unit in the decoder. Bahdana et al. (2015) extended the encoder-decoder framework with the attention technique to improve the performance of SMT for long input sentences. However, all these SMT studies do not consider the issue of response diversity.

**Conversation Models.** Along the way of neural SMT, many recent studies showed that these models can also be successfully used in conversation modeling, another sequence-to-sequence learning problem. Specifically, Shang et al. (2015) further extended the attention technique with both global and local schemes for generating short conversation. Their study qualitatively analyzed the issue of response diversity, but lacked the quantitative study on it. Most recently, researchers begun to investigate models for multiple-round conversation. Serban et al. (2015) built an end-to-end dialogue system using generative hierarchical neural network. A related model proposed by Sordoni et al. (2015) applied a hierarchical recurrent encoder-decoder model for query suggestion. The basic idea for multiple-round conversation is to extend the context generation from the immediate previous sentence to several previous ones.

**Response Diversity.** Some recent studies began to tackle the issue of response diversity from both SMT and conversation sides. Gimpel et al. (2013) proposed the methods, namely system combination and discriminating re-ranking, to produce a diverse set of plausible translations. For conversation modeling, Li et al. (2016) argued that the traditional objective function is unsuited, and used Maximum Mutual Information (MMI) as the objective. They also mentioned that the MMI measure penalizes not only high-frequency responses but also fluent ones, and may lead to ungrammatical outputs. Thus, they reduced the MMI measure to a simple version. Our work addresses the response diversity issue by directly modeling the different responding mechanisms. The proposed mechanism-aware ranking method helps to promote the infrequent but meaningful responses.

**Discourse Relation.** Some recent studies focus on automatically recognizing the internal structure and logical relationship between adjacent sentences. Ji et al. (2016) proposed a RNN-based model for jointly modeling sequences of words and discourse relations. However, this work explic-

itly models the discourse relations between two sentences in a supervised manner with the manual annotations, while our work learns the latent mechanisms in an unsupervised way.

**Concept-to-text Generation.** The Concept-to-text generation models handle non-linguistic input and generate textual output. Konstas et al. (2012) proposed a joint model for content selection and surface realization with a probabilistic context-free grammar. Wen et al. (2015) proposed a RNN model for spoken dialogue systems which generate responses for given structured data as input. These models handle non-linguistic input, while generative conversational models handle linguistic input like sentences.

## Conclusion and Future Work

In this study, to address the issue of response diversity we propose a framework of encoder-diverter-decoder for conversation modeling, aiming to explicitly model the latent responding mechanisms in free chat. The learned mechanisms helps to generate mechanism-aware responses, which are empirically shown to be diverse, relevant, and fluent. Incorporating auxiliary information, including topic distribution, demographic information of the respondents and so on, into the conversation model will be promising to provide personalized responses in accordance with a specific demographic group. We will explore towards this direction in future work.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 91546122, 61573335, 61473273, 61473274), National High-tech R&D Program of China (863 Program) (No.2014AA015105), Guangdong provincial science and technology plan projects (No. 2015 B010109005). This work was also supported by the funding of WeChat cooperation project. We thank Hao Ye, Ming Bai and WeChat Chatbot Team for their constructive advices. We also thank the anonymous AAAI reviewers for their helpful feedback.

## References

Abu Shawar, B., and Atwell, E. 2007. Chatbots: Are They Really Useful? *LDV-Forum: Zeitschrift für Computerlinguistik und Sprachtechnologie*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation. In *EMNLP*.

CWMT. 2013. `http://www.liip.cn/CWMT2013/evaluation.html`.

Gimpel, K.; Batra, D.; Dyer, C.; and Shakhnarovich, G. 2013. A Systematic Exploration of Diversity in Machine Translation. In *EMNLP*.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; and Bengio, Y. 2013. Maxout Networks. In *ICML*.

Graves, A. 2013. Generating Sequences With Recurrent Neural Networks. *arXiv*.

Grosz, B. J. 2016. Ai100 report. `https://ai100.stanford.edu/2016-report`.

Ji, Y.; Haffari, G.; and Eisenstein, J. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. In *NAACL*.

Ji, Z.; Lu, Z.; and Li, H. 2014. An Information Retrieval Approach to Short Text Conversation. *arXiv*.

Konstas, I., and Lapata, M. 2012. Unsupervised concept-to-text generation with hypergraphs. In *NAACL*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NACCL*.

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. In *ACL*.

LogSumExp. 2016. `https://en.wikipedia.org/wiki/LogSumExp`.

Misu, T.; Georgila, K.; Leuski, A.; and Traum, D. 2012. Reinforcement Learning of Question-answering Dialogue Policies for Virtual Museum Guides. In *SIGDIAL*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven Response Generation in Social Media. In *EMNLP*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL*.

Sordoni, A.; Bengio, Y.; Vahabi, H.; Lioma, C.; Simonsen, J. G.; and Nie, J.-Y. 2015. A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion. In *CIKM*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural networks. In *NIPS*.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv*.

Wang, H.; Lu, Z.; Li, H.; and Chen, E. 2013. A Dataset for Research on Short-Text Conversation. In *EMNLP*.

Wen, T. H.; Gasic, M.; Kim, D.; Mrksic, N.; Su, P. H.; Vandyke, D.; and Young, S. 2015. Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking. In *SIGDIAL*.

Williams, J. D., and Young, S. 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech & Language*.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A Biterm Topic Model for Short Texts. In *WWW*.

Yin, J.; Jiang, X.; Lu, Z.; Shang, L.; Li, H.; and Li, X. 2016. Neural Generative Question Answering. In *IJCAI*.

Young, S.; Gasic, M.; Thomson, B.; and Williams, J. D. 2013. POMDP-based Statistical Spoken Dialog Systems: a Review. *Proceedings of the IEEE*.