Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism

Chongyang Tao[†], **Shen Gao**[†], **Mingyue Shang**[†], **Wei Wu**[◇], **Dongyan Zhao**[†,‡], **Rui Yan**[†,‡*]
[†] Institute of Computer Science and Technology, Peking University, Beijing, China
[◇] Microsoft Corporation, Beijing, China
[‡] Beijing Institute of Big Data Research, Beijing, China
{chongyangtao,shangmy,zhaody,ruiyan}@pku.edu.cn, 63388@qq.com, wuwei@microsoft.com

Abstract

Attention mechanism has become a popular and widely used component in sequence-to-sequence models. However, previous research on neural generative dialogue systems always generates universal responses, and the attention distribution learned by the model always attends to the same semantic aspect. To solve this problem, in this paper, we propose a novel Multi-Head Attention Mechanism (MHAM) for generative dialog systems, which aims at capturing multiple semantic aspects from the user utterance. Further, a regularizer is formulated to force different attention heads to concentrate on certain aspects. The proposed mechanism leads to more informative, diverse, and relevant response generated. Experimental results show that our proposed model outperforms several strong baselines.

1 Introduction

To build an intelligent human-computer dialogue system is of growing interest recently. Dialogue systems are generally categorized into task-oriented systems (e.g., agents or virtual assistants) and non-task-oriented systems in the open-domain (e.g., chatbots). Task-oriented dialogue systems aim at helping users to accomplish particular tasks, such as booking a restaurant and vacation scheduling, while chatbot systems are designed to freely converse with humans without any hard limits or domain constraints. As the amount of human-to-human conversational data on social media increases, such data from the open domain drives the development of dialogue systems which are regarded as hot applications in the spotlight.

In this paper, we focus on the generative conversational model for open domain dialogue systems. Most generative conversational models are implemented based on the classic sequence-to-sequence (Seq2Seq) neural network model [Sutskever *et al.*, 2014]. The Seq2Seq model is originally proposed for machine translation and later adapted to various natural language generation tasks, such as text summarization [Rush *et al.*, 2015] and dialogue generation [Mou *et al.*, 2016; Yan *et al.*, 2016; Tian *et al.*, 2017]. However, the standard Seq2Seq model compresses all the necessary information of an input sequence into a fixed-length vector. Its performance drops rapidly as the length of an input sequence increases [Cho *et al.*, 2014]. To address this issue, Bahdanau *et al.* (2015) proposed an attention mechanism applied onto the decoding sequence, which learns to align the input sequence and the output sequence jointly. Such a mechanism outperforms the basic Seq2Seq model significantly in natural machine translation. Researchers thereafter apply it to response generation for conversations in the open domain, which also yields impressive advances [Shang *et al.*, 2015].

Query:	今天下雨，我们一起去吃火锅吧！ It's rainy today, let's go to eat hot pot!
Candidate 1:	我觉得不应该出门，还是在家做饭吧！ I think we shouldn't go out, let's cook at home!
Candidate 2:	好啊，我很久没吃火锅了。 Ok, I have not eaten the hotpot for a long time.
Candidate 3:	开车去还是坐地铁? Drive or take the subway?

Table 1: A query and three reasonable candidate replies.

The Seq2Seq model with the attention mechanism seems to be a great success in dialogue systems, but it still has insufficiency. Previous research has revealed that Seq2Seq with attention mechanism based dialogue system tends to suffer from generating trivial and universal responses [Li *et al.*, 2016a]. Recall that the Seq2Seq model with attention mechanism is originally designed for machine translation. Although neural-based machine translation and conversation generation can both be treated as a translation from an input sentence to an output sentence, the impact of attention mechanism in the decoding stage is disparate. In machine translation, the attention mechanism helps to correctly align each target word with the relevant words, which agrees well with human's intuition. But it is less interpretable when it comes to conversation generation. Compared with machine translation, there is few one-to-one word alignments between the input sentence and the output sentence in conversation generation task. Thus it is not an appropriate strategy for conversation genera-

* Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

tion that the decoder focuses on different information at every time-step. Cho *et al.* (2014) has indicated that the generation process of a better response has a relatively more centralized attention distribution.

Given an utterance, humans tend to focus on certain aspects to respond, rather than disperse attention to every word. We note that an utterance could have many valid responses that focus on different aspects. As shown in Table 1, candidate 1 emphasizes on "go out", while the attention of candidate 2 is on "eat hotpot". Thus it is intuitive to guide the decoder, when generating different response, to pay attention to different aspects.

The standard attention mechanism for the Seq2Seq model fails to revolve around different aspects of the input query for conversations. It tries to align different responses with the same aspect of the input utterance and results in a severe problem by aligning universal terms among input and output utterances. What's more, such attention mechanism is unconstrained and tends to generate more dispersed attention distribution over the input utterance. Researchers have revealed that Seq2Seq model with standard mechanism suffers from generating trivial, universal and informativeness responses [Li *et al.*, 2016a], which concur with our observations and explanations.

To address the problem of universal responses and more importantly, to bring diversity into conversations, we propose a *Multi-Head Attention Mechanism* (MHAM) for Seq2Seq model. To be more specific, we first project hidden states of the encoder to different semantic spaces through learnable projection matrices. Then the standard attention mechanism is applied by the decoder for all semantic spaces to jointly attend to information from multiple aspects. In addition, we introduce a penalization term to avoid the attention signal vectors suffering from a redundancy problem when the attention mechanism may provide similar attention weights for all aspects. The penalty is designed to penalize the redundancy of attention weight across different aspects, but also to push the decoder to attend to a certain aspect consistently. The constrained model is name as *Constrained Multi-Head Attention Mechanism* (CMHAM).

Our contributions are main-fold:

- We propose a Seq2Seq model with multi-head attention mechanism for dialogue system to generate the response with diverse attention.
- We incorporate an elaborative penalization term to force every head of the multi-head attentions to concentrate on a certain aspect, as well as to control the diversity of these attentions.

Experiments show that our model outperforms several existing Seq2Seq based conversational models in terms of both automatic evaluation metrics and human judgement.

2 Model

2.1 Seq2Seq Model and Attention Mechanism

Sequence-to-sequence model (Seq2Seq) was first proposed in machine translation. The idea was to translate one sequence to another sequence through an encoder-decoder neural architecture. Recently, dialog generation has been treated as sequence translation from a query to a reply [Mou *et al.*, 2016; Xing *et al.*, 2017; Zhou *et al.*, 2017; Yao *et al.*, 2017].

Formally, given an input query $X = (x_1, x_2, \cdots, x_n)$, the encoder network sequentially reads the word in X and encodes it as a context vector c through a recurrent neural network (RNN). The decoder network sequentially generates a reply $Y = (y_1, y_2, \cdots, y_m)$ with context vector c as input. The Seq2Seq models are typically trained with the objection function:

$$p(Y|X) = \prod_{t=1}^{M} p(y_t|c, y_1, \cdots, y_{t-1}) \quad (1)$$

The encoder RNN computes the context vector c as follows:

$$h_t = f(x_t, h_{t-1}); \; c = h_N \quad (2)$$

where h_t is the hidden state at time t. $f(\cdot)$ is a non-linear activation function, which can be a logistic function, the sophisticated long short-term memory (LSTM) unit [Hochreiter and Schmidhuber, 1997], or the recently proposed gated recurrent unit (GRU) [Chung *et al.*, 2014]. We employ LSTM as $f(\cdot)$ in our paper.

The decoder RNN generates word by word conditioned on the context vector c and the decoder hidden state s_t. The output probability distribution $o_t \in \mathbb{R}^{D_v}$ (D_v denotes the vocabulary size.) over vocabulary at time t can be calculated as:

$$s_t = f(c, y_{t-1}, s_{t-1}) \quad (3)$$

$$o_t = softmax(y_{t-1}, s_t) \quad (4)$$

In primitive Seq2Seq model, c is the same for generating all output words. To alleviate this problem, the attention mechanism [Bahdanau *et al.*, 2015] is usually adopted to allow the decoder to pay different attention to each part of input at every timestep. The attention mechanism computes a different c_t which is the wighted sum of hidden states of the encoder. $c_t = \sum_{i=1}^{n} a_{t,i} h_i$, where $a_{t,i}$ is the attention weight over h_i at time t and indicates how much the i-th word contributes to generating the j-th word. $a_{t,i}$ is usually defined as:

$$e_{t,i} = g(s_t, h_i); \; a_t = softmax(e_t) \quad (5)$$

where g is a function that calculates the similarity between h_i and s_t. In this paper we use bilinear function as $g(s_t, h_i) = v^T tanh(W_h h_i + W_s s_t)$, where v, W_h and W_s are parameter matrices.

2.2 Multi-head Attention Mechanism

The context vector obtained by traditional attention mechanism focuses on a specific representation subspace of the input sequence. Such context vector is expected to reflect one aspect of the semantics in the input. However, a sentence usually involves multiple semantics spaces, especially for a long sentence. In this paper, we propose a multi-head attention mechanism for Seq2Seq model to allow the decoder RNN to jointly attend to information from different representation subspaces of the encoder hidden states at the decoding

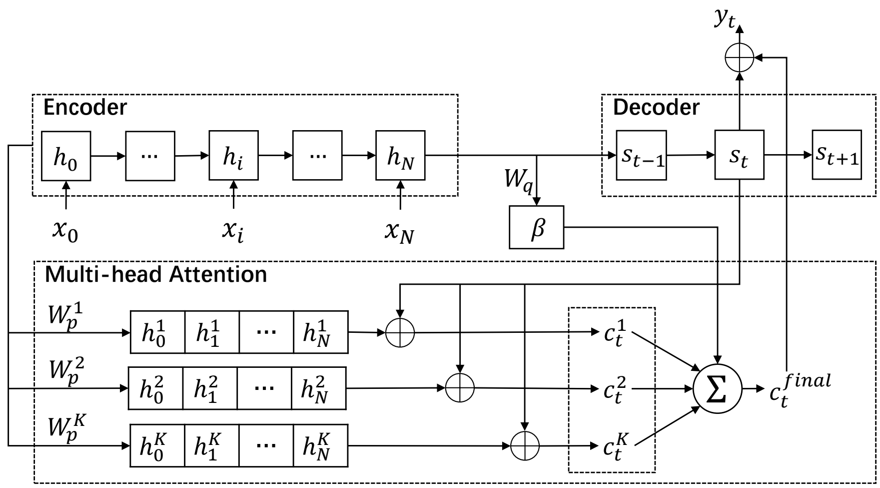


Figure 1: An overview of the proposed Seq2Seq model with multi-head attention mechanism.

process. The idea of multi-head has been applied to learn the sentence representation in self-attention [Lin *et al.*, 2017; Vaswani *et al.*, 2017].

Formally, we first project hidden states of the encoder to K different semantic spaces through different learnable projection matrices as follows:

$$h_i^k = W_p^k \cdot h_i \quad k \in (1, \cdots, K); i \in (1, \cdots, N) \tag{6}$$

where $W_p^k \in \mathbb{R}^{d \times d}$ is the learnable projection matrix for the k-th semantic space, and d is the dimension of hidden units of the encoder and decoder RNN.

We perform standard attention mechanism for all semantic spaces to obtain multiple attention probability distributions over the input words, and then those distributions are used to generate K different context vectors $\{c_t^1, \cdots, c_t^K\}$ that focus on different components of the input sentence. To be specific, for k-th semantic space, $\alpha_{t,i}^k$ denotes the attention weight over the i-th encoder hidden state h_i^k at time t, which is defined in Eq. 7. Then we can compute the context vector $c_t^k \in \mathbb{R}^d$ through a weighted sum of the encoder hidden states.

$$\alpha_{t,i}^k = \frac{e^{g(h_i^k, s_t)}}{\sum_i e^{g(h_i^k, s_t)}}; \; c_t^k = \sum_{i=1}^{T} \alpha_{t,i}^k \cdot h_i^k \tag{7}$$

For each decoding timestep, we need to combine all context vectors for word generation. Simply, we can employ concatenation strategy or pooling strategy. In this paper, we adopt a soft-attention approach to combine K context vectors. Specifically, we first calculate a weight vector for context vectors conditioned on the final hidden state h_N of encoder. Then the final context vector c_t^{final} is obtained by a weighted sum of all context vectors from different semantic spaces.

$$r = softmax(W_q \cdot h_N) \tag{8}$$

$$c_t^{final} = \sum_{k=1}^{K} r_k \cdot c_t^k \tag{9}$$

where $W_q \in \mathbb{R}^{K \times d}$ is a trainable parameter and softmax function ensures all weights sum up to 1. $r \in \mathbb{R}^{K \times 1}$ and r_k is the weight for k-th head. We can directly replace c with c_t^{final} in Eq. 3 for generating words in the decoder RNN.

2.3 Penalty Term

We can notice that there is a potential drawback of the multi-head attention mechanism, i.e., the context vectors can suffer from redundancy problem if the attention mechanism always provides similar attention weight for all semantic spaces. Motivated by recently work [Bousmalis *et al.*, 2016; Lin *et al.*, 2017], we introduce a penalty term, which can not only penalize the redundancy of attention weight vectors across different aspects of the source sentence but also encourage the decoder to attend to a specific aspect consistently.

For each head, we first calculate the average accumulated attention weight on each source word. Formally, the average accumulated attention weight on i-th input word for k-th head can be calculated as:

$$\delta_i^k = \frac{1}{M} \sum_{t=1}^{M} \alpha_{t,i}^k \tag{10}$$

here M is the length of the decoder. The above computation is equivalent to performing a mean pooling across different decoding time (input words) and over different semantic spaces.

We can obtain K average accumulated attention weight vectors for all semantic spaces. δ^k represents the attention vector for k-th semantic space and $\sum \delta^k = 1$. We concatenate all those vectors into a matrix $\Delta = \{\delta^1 \oplus \delta^2 \oplus \cdots \oplus \delta^K\} \in \mathbb{R}^{K \times N}$. We define the loss via a soft subspace orthogonality constraint between the attention weight vector of each space (head) as follows:

$$\mathcal{L}_{penalization} = \|\Delta \cdot \Delta^T - I\|_F^2 \tag{11}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm and $I \in \mathbb{R}^{K \times K}$ is an identity matrix. For any non-diagonal elements $(\Delta \cdot \Delta^T)_{mn} (m \neq n)$, it represents a summation of element-wise product of δ^m and δ^n. In the extreme case, when the two

Model	*Embedding Average*	*Embedding Greedy*	*Embedding Extrema*	*Human Score*
RNNatt [Bahdanau *et al.*, 2015]	0.5469	0.3385	0.4491	0.73
MMI-anti [Li *et al.*, 2016a]	0.5158	0.3112	0.4149	0.75
HGFU [Yao *et al.*, 2017]	0.5651	0.3404	0.4639	0.79
MHAM	0.5705	0.3460	0.4688	0.98
CMHAM	0.5889	0.3608	0.4830	1.23

Table 2: Reply evaluation using embedding-based metrics as well as human evaluation.

attention vectors are orthogonal, $(\Delta \cdot \Delta^T)_{mn}$ will be 0. Otherwise, $(\Delta \cdot \Delta^T)_{mn}$ will be a positive value. We can notice that the elements on the diagonal of $(\Delta \cdot \Delta^T)$ will be forced to approximate 1 since we subtract an identity matrix from $(\Delta \cdot \Delta^T)$. Such a penalty term will encourage attention vector for each head to focus on as few input words as possible. In the most extreme case, the attention vectors for each head all concentrate on a single word and different heads attend to different words.

Finally, the loss function of our model can be defined as:

$$\mathcal{L} = \lambda \mathcal{L}_{task} + \gamma \mathcal{L}_{penalization} \tag{12}$$

where $\mathcal{L}_{task} = -\log p(y|x)$ is the negative log-likelihood loss function for sequence generation. λ and γ are hyper-parameters that control the interaction of the loss terms. $\lambda + \gamma = 1$. In our model, all parameters are randomly initialized and automatically updated through back-propagation algorithm.

3 Experiment

3.1 Dataset

We evaluated our model on a massive Chinese conversation corpus crawled from an online forum Douban[1]. After we removed low-quality query-reply pairs, there remain 1,600,243 pairs for model training, 10000 pairs for model validation, and 2000 pairs for model testing. We performed Chinese text segmentation by the Jieba word tokenizer [2]. The query contains on average 13 words and reply contains on average 11 words.

3.2 Implementation Details

In our model, the vocabulary size is 63,000. We employ 610-dimensional word embeddings, which are randomly initialized in the beginning and trained during the training process, and 1000-dimensional hidden units in the encoder and decoder RNN, following [Yao *et al.*, 2017]. We utilize Adaptive Sub-gradient Methods (AdaGrad) [Duchi *et al.*, 2011] optimizer on mini-batches of size 32, with learning rate 0.15 and gradient clipping 2. The number of head is $K = 5$. As the number of heads increases, the semantic subspace of some heads become similar. 5 heads can attend most semantic parts in a query. We train our model in 300K iterations (about 10 epochs) and keep the best model on the validation set. For decoding process, we use beam search with a beam size of 5 and select the top-1 generated reply for evaluation. For the coefficient of penalty term, we take the hyper-parameters which achieve the best performance on the validation set via a small grid search. Finally, we choose λ as 0.95 and γ as 0.05.

Model	*distinct-1*	*distinct-2*
RNNatt [Bahdanau *et al.*, 2015]	0.0201	0.0812
MMI-anti [Li *et al.*, 2016a]	0.0304	0.1484
HGFU [Yao *et al.*, 2017]	0.0101	0.0537
MHAM	0.0406	0.1561
CMHAM	0.0502	0.1749

Table 3: Results of diversity evaluation in terms of system-level diversity.

3.3 Evaluation

Researchers usually employ BLEU [Papineni *et al.*, 2002] as an evaluation metric for generative dialogue systems. However, BLEU measures word overlap between the generated reply and the ground truth, which is too strict for evaluating dialogue systems due to significant diversity in the space of valid replies to a given context. Besides, [Liu *et al.*, 2016; Tao *et al.*, 2018] conduct empirical experiments and show weak correlation between BLEU and human annotation. In this paper, we consider three embedding-based metrics (including *Embedding Average*, *Embedding Greedy* and *Embedding Extreme*) to evaluate our model, following several recently studies on dialog systems [Serban *et al.*, 2017; Xu *et al.*, 2017]. The three metrics compute the similarity between the generated reply and reference reply according to the word embedding.

We also use human evaluation in our experiment since automatic evaluation metrics may not always consistent with human perception [Stent *et al.*, 2005]. Three educated annotators are invited to judge the quality of 200 randomly sampled replies generated by different models. We show human annotators a dialogue query along with replies generated from each model. Annotators judge the quality of the replies by rating an integer score among 0, 1, and 2. A score of 2 indicates a relevant, natural and informative reply; 1 indicates that the reply is relevant and natural, but is too universal; 0 indicates a bad reply that is either dis-fluent or semantically irrelevant. In our experiment, the average Cohen's kappa score is 0.3366, indicating that annotators reach good agreement.

3.4 Baselines

We compare our model with several state-of-the-art neural conversation models. **RNNatt** is a Seq2Seq architecture with soft attention mechanism [Bahdanau *et al.*, 2015], which has been widely adopted as a baseline for comparison. Note that

[1]http://www.douban.com

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

ID	Query	Reply	
1	你很活跃耶，哈哈 (You're very active, haha)	**RNNatt:**	是的，你也是？(Yes, are you the same?)
		HGFU:	就是开心(I am just happy)
		MMI-anti:	还好吧，哈哈 (Just ok, haha.)
		CMHAM:	你也活跃起来嘛。(You can also be active.)
2	你情绪怎么了 (what's your mood?)	**RNNatt:**	我也不知道怎么了 (I don't know why)
		HGFU:	不知道该怎么办 (I don't know what to do)
		MMI-anti:	谢谢你的回复 (Thanks for your reply)
		CMHAM:	我情绪不好 (I am in a bad mood)
3	我想知道长脸适合中分么？ (I wonder whether the long face is suitable for central parting?)	**RNNatt:**	我也不知道 (I don't know why.)
		HGFU:	可以啊，可以试试。(Sure, you can have a try.)
		MMI-anti:	哈哈哈哈哈 (Haha, haha.)
		CMHAM:	那要看你的脸型对不对！(It depends on the shape of your face.)

Table 4: Generation examples for different dialog models.

this model is a special case of our model when the head $K = 1$ without the penalty term. **MMI-anti** [Li *et al.*, 2016a] is a Seq2Seq model with Maximum Mutual Information (MMI) between inputs and outputs as the objective function, which aims at generating more diverse responses. **HGFU** [Yao *et al.*, 2017] is also a neural generative dialogue model, which incorporates an additional pre-trained cue word into the decoding process in a "soft" manner to generate a more meaningful response.

3.5 Experiment Results

Table 2 shows the performance of our model and the baselines in terms of embedding-based evaluation metrics as well as human evaluation. We can see that our proposed models outperform baseline models, which indicates that coupling the Seq2Seq model with a multi-head attention mechanism is a better method for response generation tasks.

We also notice that the performance of our model is better than HGFU [Yao *et al.*, 2017]. This is due to the different content-introducing mechanism. Concretely, a predicted cue word is incorporated into the decoding process in HGFU. Differently, our model introduces auxiliary information from the query itself, which makes our generated reply more relevant to the given query. This can be proved in our case study (see Table 4). Besides, we find that the predicted cue words in HGFU are not always pertinent or appropriate, which further has a direct impact on reply generation.

Furthermore, *CMHAM* achieves better performance than *MHAM* in terms of embedding-based metrics, which demonstrates the effectiveness of the penalty term. As mentioned above, the penalty term encourages the decoder to attend different aspects of the source query, which introduce much more information during the reply generation. The results of human evaluation can also demonstrate the strength of *CMHAM*.

Our conclusions above can also be supported by a case study. As the examples shown in Table 4, we can see that our model can usually generate more meaningful and informative replies. Furthermore, our model appears to be better at generating more relevant replies compared with baseline models. We attribute the improvement of our model to the multi-head attention mechanism which allows the model to attend to information jointly from different representation spaces, so as to better understand the utterance.

3.6 Further Analysis

To further investigate the effectiveness of the obtained context vector for each head, we generate K replies from K different heads of attention from model. Concretely, the context vector for i-th head c_t^k is concatenated with the hidden state s_t of decoder, and then we use the concatenated vector to predict a word distribution at time t. We apply beam search for each head with a beam size of 5 and select the top-1 generated reply as the final output of this head.

We adopt the *distinct-1* and *distinct-2* metrics proposed by Li *et al.* (2016a) to measure the informativeness and diversity of the generated replies. The distinct-1 (distinct-2) measures the ratio of distinct unigrams (bigrams for distinct-2). The results are shown in Table 3. It can been seen that our model has the best performance both in *distinct-1* and *distinct-2*. Besides, we can notice that *CMHAM* is better than *MHAM*. This implies our proposed penalty term in objective function can improve the diversity of generated replies.

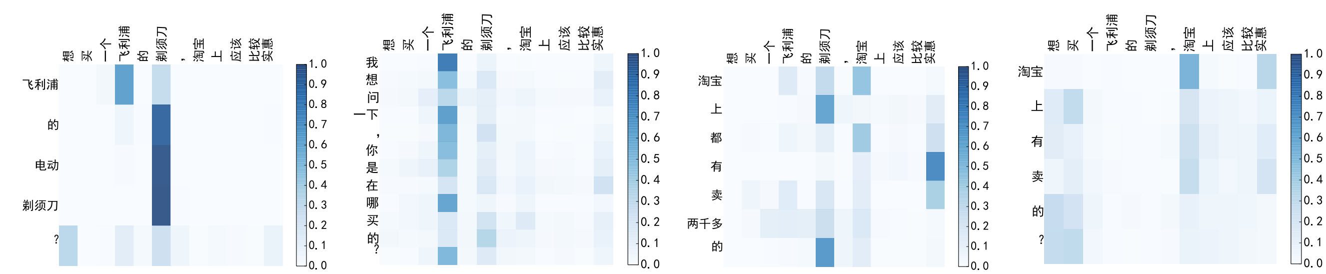
Table 5 shows an example and its generated replies for different heads. We can observe that five heads generate reasonable replies, and each reply attends to specific semantic aspects of the input query. Figure 2 exhibits the visualization of attention weight of the first four heads. We can see that different attention heads can attend to different semantic parts of the input query. For instance, the generated replies of *Head-1* and *Head-2* mainly focus on "Philips" and "razor" in the query, while *Head-3* and *Head-4* mainly attend to "Taobao" and "affordable". As the case shown, our model has the capability to generate different replies that focus on specific semantic parts of the input query.

4 Related Work

As the general Seq2Seq dialogue model with attention mechanism suffers from generating trivial and universal responses, various attempts have been made to tackle this problem. One line of research has focused on improving the output diversity. Li *et al.* (2016a) proposed a method that uses Maximum Mutual Information (MMI) between inputs and outputs as the objective function. This approach penalized unconditionally high frequency response to reduce generic responses. Li *et al.* (2016b) introduced a diversity-promoting decoding algorithm by adding an intra-sibling ranking term to the standard beam search algorithm, which favors choosing hypotheses

Query		Reply
想买一个飞利浦的剃须刀，淘宝上应该比较实惠 (I want to buy a Philips razor, and it should be more affordable in Taobao.)	**Head-1:**	飞利浦的电动剃须刀？ (Is it a Philips electric razor?)
	Head-2:	我想问一下，你是在哪买的？ (I want to know where you bought it.)
	Head-3:	淘宝上都有卖两千多的 (Some razors on Taobao are sold for more than 2,000 yuan.)
	Head-4:	淘宝上有卖的？ (Does Taobao sell it?)
	Head-5:	好的，我去看看。 (OK, I will take a look.)

Table 5: Generation examples for each head in our model.



(a) Head-1 (b) Head-2 (c) Head-3 (d) Head-4

Figure 2: Visualization of attention weight for different attention heads. In each subplot, the horizontal axis represents the input query, and the vertical axis represents the output reply for each head. Darker squares refer to larger weights. The corresponding translation for each head can be seen in Table 5.

from diverse parents. Such method focuses on diversifying the output of the decoder at word-level. Zhou *et al.* (2017) proposed a mechanism-aware machine with the framework of encoder-diverter-decoder that models the responding mechanisms as latent embeddings. Recent work from Zhao *et al.* (2017) combined general Seq2Seq model with conditional variational auto-encoders, which introduces a latent variable to capture discourse-level variations. Different from previous research, our work addresses the universal responses issue by exploring the multi-aspects information in conversations. It is inspired by the intuition that, given an utterance, people are likely to partially focus on a certain aspect.

On the other hand, some researchers have adapted content introducing to alleviate the problem. Xing *et al.* (2017) incorporated topic information as prior knowledge into the Seq2Seq framework with attention mechanism to encourage the model to generate more topic coherent responses. Mou *et al.* (2016) presented a method that uses the point-wise mutual information to predict a keyword and makes the word explicitly occur in the generated response. This method is to some extent rigid. Yao *et al.* (2017) also adapted the approach of predicting a cue word from the query, but it proposed an implicit method to utilize the cue word. However, since the cue word is predicted only by the query, it has the risk of having low relatedness with the whole conversation. Instead of predicting a word, our approach utilizes attention mechanism with multi-head structure to partially focus on words in a previous utterance, which is more intuitive.

Attention mechanisms have become an integral part of the Seq2Seq framework, thus many efforts have been made to improve the attention architecture. Lin *et al.* (2017) presented a self-attention mechanism to extract different aspects of a sentence into multiple vector-representations. Such method could also be referred as multi-head attention mechanism. Vaswani *et al.* (2017) also adapted this method in neural machine translation task, as it allows the model to jointly attend to information from different representation subspaces at different positions. To enhance the performance of multi-head mechanism in dialogue systems, our approach incorporates a penalty term to force the attention of each head to concentrate on a certain aspect, and to control the multiplicity.

5 Conclusion

In this paper, we propose a Seq2Seq model with multi-head attention mechanism, which can attend to different semantic parts of an input query for the decoder to explicitly generate reply. We call it *Multi-head Attention Aware Dialog System* (MHAM). Experiments show that our model outperforms the existing neural-based dialogue models in terms of both automatic evaluation metrics and human judgement.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001) and the National Science Foundation of China (No. 61672058). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Microsoft Collaborative Research Program.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Bousmalis *et al.*, 2016] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, pages 343–351, 2016.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[Duchi *et al.*, 2011] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

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

[Li *et al.*, 2016a] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119, 2016.

[Li *et al.*, 2016b] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

[Lin *et al.*, 2017] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132, 2016.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*, pages 3349–3358, 2016.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[Rush *et al.*, 2015] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[Serban *et al.*, 2017] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586, 2015.

[Stent *et al.*, 2005] Amanda Stent, Matthew Marge, and Mohit Singhai. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351, 2005.

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

[Tao *et al.*, 2018] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*, pages 722–729, 2018.

[Tian *et al.*, 2017] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make context more useful? An empirical study on context-aware neural conversational models. In *ACL*, volume 2, pages 231–236, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

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

[Xu *et al.*, 2017] Zhen Xu, Bingquan Liu, Baoxun Wang, SUN Chengjie, Xiaolong Wang, Zhuoran Wang, and Chao Qi. Neural response generation via gan with an approximate embedding layer. In *EMNLP*, pages 628–637, 2017.

[Yan *et al.*, 2016] Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. Shall i be your chat companion?: Towards an online human-computer conversation system. In *CIKM*, pages 649–658, 2016.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, pages 2180–2189, 2017.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664, 2017.

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