Knowledge dialogue generation with multi-head attention mechanism

Fengrui Yang, Xiaosi Wu* and Luwei Zhao

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing,400065, China

*Corresponding author’s e-mail: S190131024@stu.cqupt.edu.cn

Abstract: An increasing number of research efforts are focusing on knowledge dialogue generation. Less attention is focused on increasing knowledge diversity in generated responses. A model of knowledge selection guided by a multi-head attention mechanism is proposed. First, the current input discourse and knowledge content are input into the Bi-GRU module to obtain the coding vector, and then obtain multiple aspects of semantics from the user input discourse coding vector based on the multi-head attention mechanism, so as to select different knowledge. A punishment item method is proposed to force different attention to focus on different aspects, and finally, use the user input and selected knowledge for the decoding stage. Experiments with manual and automated evaluations have proven that the model is superior to the baseline model compared to previous work.

1. Introduction

Generative dialogue systems are receiving increasing attention. The Seq2Seq model can generate fluent responses. And the responses generated by the generative dialogue systems based on the seq2seq framework have some issues, generally generating some safe but boring responses, such as "I don't know", "OK". Therefore, there has been a lot of research to enhance the quality of dialogue content. One approach is to help the model generate fuller responses by leveraging external knowledge.

Researchers have observed that humans will incorporate some knowledge in their daily conversations, so research has begun to add external knowledge to dialogue systems. The combination of an external knowledge base and a dialogue training corpus is what can bridge the gap between a dialogue system and a human with background knowledge. Recently, several studies have demonstrated that combining external knowledge can improve the informativeness of the generated content, increase user satisfaction, and help facilitate the development of open-domain dialogue systems. Ghazvininejad [1] stored texts as knowledge in memory networks and used them to generate more informative responses. Liu [2] proposed a neural knowledge diffusion model, while enhancing the model by convergence and divergent thinking on the knowledge base. Lian [3] proposed the use of posteriori knowledge to help guide the selection of appropriate knowledge during the training process. Zhang [4] improved generation by introducing a large amount of continuous background knowledge. Zheng [5] used the differences between the knowledge selected in different dialogue rounds for knowledge selection, which promoted accuracy in knowledge selection and produced informative responses.

However, current research on generative knowledge dialogues ignores the fact that a sentence, especially one that contains a long history dialogue, may involve multiple aspects of knowledge. Given a query, humans may focus on a specific aspect and then respond to it, or they may focus on multiple
aspects and respond. A single sentence can have many different valid responses, as shown in Table 1, where the first response focuses on "raining", the second response on "Fast and Furious", and the third response on "go out" for the same query. This shows that different aspects of the same query may involve different knowledge information, and thus have different responses.

Based on the above research, in response to the current Seq2Seq model based on combining external knowledge bases without considering that different aspects of the current user input may involve different aspects of knowledge information. The standard attention-based mechanism selects external knowledge information that will only be related to a specific aspect of the input discourse, which may lead to incomplete and not rich knowledge selection. Therefore, this paper proposes a model of knowledge dialogue system based on multiple attention mechanisms.

Table 1. A query and three reasonable candidate responses.

| Query | It's raining, let's go out and watch the Fast and the Furious! |
|-------|---------------------------------------------------------------|
| Response 1 | I don't think we should go out |
| Response 2 | Yeah, I really like Dawn Johnson's movies. |
| Response 3 | Do we drive there or take the subway? |

2. Model

The model framework proposed in this paper is an improvement on the Seq2Seq model. First, give the source sequence (query) \( X = \{x_1, x_2, ..., x_N\} \), \( N \) denotes the total number of words in the input sequence, and external knowledge \( K = \{k_1, k_2, ..., k_T\} \). The purpose of this paper is to generate the response \( Y = \{y_1, y_2, ..., y_T\} \) of the fused knowledge by selecting the appropriate knowledge from the external knowledge. The model consists of four modules: the utterance encoder, the knowledge encoder, the knowledge manager and the decoder. The overall scheme of the proposed knowledge generation dialogue system is shown in Figure 1.

![Figure 1. Framework of knowledge enhanced dialogue generation system based on multi-head attention mechanism.](image)

2.1. Encoder

In this paper, we use a bi-directional gated recurrent unit (Bi-GRU) as an encoder. First map each word in the input sequence \( X \) to a word vector \( \tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_{N_t}\} \). The forward RNN reads \( \tilde{X} \) from left to right to obtain the forward hidden state \( h_t^+ \) for each \( x_t \), and the backward RNN reads \( \tilde{X} \) in a reverse order in the same way to obtain the backward hidden state \( h_t^- \) of \( x_t \). The two direction hidden states are concatenated to obtain the final hidden state \( h_t \) for \( x_t \) as shown in equation (1):

\[
\tilde{h}_t = [\tilde{h}_t^+; \tilde{h}_t^-]
\] (1)

Where, \([;]\) denotes concatenation. The hidden state of the final utterance \( X \) is \( h = \{h_1, h_2, ..., h_N\} \), and this vector is fed to the knowledge manager for selecting knowledge and providing a decoding basis for decoding as a decoder.
The knowledge encoder has the same structure as the utterance encoder, but they do not share any parameters. The knowledge vector \( k = \{ k_1, k_2, \cdots, k_T \} \) is obtained and used by the knowledge manager.

2.2. Knowledge Manager

The purpose of the knowledge selector is to select the appropriate knowledge \( k_i \). To obtain more comprehensive and appropriate knowledge, this paper makes full use of the semantic content of the input utterance to select the knowledge selection. The knowledge manager based on the multi-head attention mechanism in this section is shown in figure 2.

Firstly, the discourse hidden states \( h = \{ h_1, h_2, \cdots, h_N \} \) are projected to \( M \) different semantic spaces by different learnable projection matrices as follows.

\[
    h^q_i = W^q_p \cdot h^p
\]

Where \( i \in (1, \cdots, N), q \in (1, \cdots, M), W^q_p \in \mathbb{R}^{d \times d} \) denotes the learnable projection matrix of the \( q \)-th semantic space, and \( d \) is the dimension of the hidden unit of the decoder.

The standard attention mechanism is implemented for all the semantic spaces and knowledge vectors to obtain multiple attention probability distributions on the knowledge vectors. And then these distributions are used to generate \( M \) different knowledge vectors \( \{ z^1, z^2, \cdots, z^M \} \), respectively. In the \( k \)-th semantic space, the calculation process is shown in equation (3).

\[
    \alpha^q_{ij} = \text{softmax}(h^T_i \cdot k): \quad z^q = \sum_{i=1}^{T} \alpha^q_{ij} \cdot k
\]

Where, \( \alpha^q_{ij} \) represents the attention weight of the \( i \)-th utterance encoder hidden state \( h^p_i \) to the \( j \)-th knowledge; then the related knowledge vector \( z^q \) is obtained by weighted summation of the knowledge vectors.

The final knowledge vectors selected from different semantic spaces are weighted and the final knowledge vector is obtained as shown in equation (4).

\[
    r = \text{soft max}(W_q \cdot h_N): \quad k_{\text{final}} = \sum_{n=1}^{M} r_k \cdot z^q
\]

\( r \in \mathbb{R}^{M \times 1} \), \( r_k \) is the weight of the \( k \)-th head. \( k_{\text{final}} \) is used directly in the decoding stage. In this paper, \( M \) is set to 3.

2.3. Decoder

The decoder in this paper adopts unidirectional GRU and adds an attention mechanism in the decoding stage. At moment \( t \), the context encoding vector \( c_t \) is jointly determined by the current input hidden state \( h_t \), the hidden state \( k_{\text{final}} \) of the selected knowledge, and the hidden state \( s_{t-1} \) of the decoder at the last moment, which is computed as shown in (5) and (6) below.

\[
    c_t = \sum_{i=1}^{N} \alpha_{ij} h^p_i + \sum_{i=N+1}^{N+T} \alpha_{ij} k_{\text{final}}^q
\]
\[ \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{N} \exp(e_{tj})} \]  \hspace{1cm} (6)

\[ e_{ti} = \tanh(s_{t-i}, h_i) \quad i = 1, \ldots, N \]

\[ e_{tN} = \tanh(s_{t-i}, k_i^{\text{final}}) \quad i = N+1, \ldots, N+T \]

Where \( \alpha_{ti} \) is the weighting factor.

The hidden state \( s_t \) of the decoder at moment \( t \) is updated as shown in equation (7):

\[ s_t = GRU(c_t, e(y_{t-1}), s_{t-1}) \]  \hspace{1cm} (7)

Where \( e(y_{t-1}) \) is the embedding vector of the word \( y_{t-1} \) generated in the last time step.

After obtaining the hidden state \( s_t \). The decoder generates a word-by-word response, which is calculated as shown in (8).

\[ p(y_t | c_t, X, K, y_{t-1}) = \text{softmax}(W_y s_t + b_y) \]  \hspace{1cm} (8)

where \( W_y \) and \( b_y \) are trainable parameters. The output of all moments is the final predicted response sequence \( Y = \{y_1, y_2, \ldots, y_L\} \).

2.4. Penalty Item

It is possible that the multi-head attention mechanism focuses on the same aspect of semantic information, which can cause redundancy of information. Inspired by literature [6] and literature [7], a penalty term is introduced, which not only penalizes the redundancy of the attention vector for different aspects of the input sentence but also forces the multi-head attention mechanism to focus on specific aspects of the input sentence.

The attention weight \( \alpha_{i,j} \) of the \( i \)-th utterance encoder hidden state \( h^n_i \) on the \( j \)-th knowledge is obtained from equation (3), and the weighting matrix \( A \) is calculated as shown in equation (9).

\[ L_{\text{penalization}} = \| A \cdot A^T - I \|^2_F \]  \hspace{1cm} (9)

Here \( \| \cdot \|^2_F \) denotes the squared Frobenius norm, and \( I \) is the unit matrix. To ensure that the weights of each row of the weighting matrix \( A \) are not the same as possible, \( \| A \cdot A^T - I \|^2_F \) can make \( A \cdot A^T \) approximates the unit matrix \( I \) as much as possible, the attention weights \( \alpha^n \) focus on specific words as much as possible, reduce the redundancy between weights, and finally add the penalty term to the final loss function.

In this paper, a negative log-likelihood loss function is used with the aim of measuring the difference between the true response and the response generated by the model, calculated as shown in equation (10).

\[ L_{\text{NLL}} = - \log p(y_t) \]  \hspace{1cm} (10)

The final loss is defined as shown in equation (11) below.

\[ L = L_{\text{NLL}} + L_{\text{penalization}} \]  \hspace{1cm} (11)

3. Experiment

3.1. Experimental design

In this paper, we use data from Baidu's knowledge graph-based Chinese chat conversation task dataset [8]. The data is based on real conversations, including "conversation", "goal" and "knowledge ", where the knowledge information is real events in the field of movies and entertainment characters, such as box offices of movies, directors, reviews of movies, ancestry and representative works of related characters, etc. Each knowledge is presented in the form of triples, and the conversation topics are around movies and entertainment characters. The dataset contains a total of 30,000 multi-round
conversations with about 120,000 dialogues. The experiment divides the dataset into training, validation and test sets with sizes of 19858, 1930 and 1205 respectively.

The model proposed in this paper is implemented on the PyTorch [9] platform. The utterance encoder and knowledge encoder use a two-layer GRU structure, and the number of hidden nodes in each GRU layer is set to 300. The decoder uses a single-layer GRU with 600 hidden nodes. The word embedding size is set to 300; the vocabulary size is within 30,000 words. In this paper, we use Adam [10] to optimize the objective function with batch size and learning rate set to 8 and 0.0005, respectively.

The following models were selected as comparison models for this paper.

a) Seq2Seq model [11], which uses only historical conversations as a priori information to generate responses in order to verify the necessity of introducing knowledge.

b) MemNet model [1], which is based on the classical Seq2Seq model in addition to a fact encoder that stores external knowledge in its memory unit.

c) The PostKS [3] model proposes the use of a posteriori knowledge to help guide the selection of appropriate knowledge during the training process.

The hyperparameter settings used in the baseline model are consistent with those of the model proposed in this paper.

3.2. Experimental results and analysis

In this paper, both manual and automatic evaluations are used. In the manual evaluation, 100 responses generated by different models were randomly selected for evaluation in this paper. Manual scoring from two metrics, content and knowledge, each evaluation metric contains three scores: 0, 1 and 2. A score of 0 for content means that the response is irrelevant to the current input or does not make sense. A score of 1 means the reply is grammatically correct but makes little sense or has low relevance to the current input. A score of 2 means that the reply is grammatically correct, logical and has high relevance to the current input. Automatic evaluation is the comparison of real responses with the responses generated by the model, and the automatic evaluation used in this paper is BLUE-1/2 and distinct-1/2 [11][12].

3.2.1 Experimental results. As can be seen from Table 2, the model in this paper obtained the highest scores in both automatic assessment and manual evaluation, outperforming all baselines except for the BLUE score metric. The Seq2Seq model does not incorporate external knowledge, so the seq2seq model will not be evaluated in terms of knowledge in manual evaluation. Compared with MemNet, the model in this paper improves 33.3% and 12.4% on diversity distnct-1 and distnct-2, respectively, and 7.66% and 6.18% on BLUE-1 and BLUE-2. Compared with PostKS, the model in this paper improved 24.4% and 10% on diversity distnct-1 and distnct-2, respectively. However, the PostKS model scores higher on the BLUE metric, probably because the PostKS model incorporates the use of input utterances and true responses as posterior distributions during training. Under the guidance of real response, the model tends to generate responses closer to the real response. Compared to Seq2Seq, the other three model's introduce knowledge entities, so the n-grams in the responses are more diverse. This model has the highest diversity score, which indicates that the multi-head knowledge selection mechanism in this paper can improve the diversity of responses.

| Model Comparison | distinct-1 | distinct-2 | BLUE-1 | BLUE-2 |
|------------------|------------|------------|--------|--------|
| Seq2Seq          | 0.038      | 0.083      | 0.190  | 0.171  |
| MemNet           | 0.042      | 0.137      | 0.235  | 0.178  |
| PostKS           | 0.045      | 0.140      | 0.258  | 0.190  |
| Model of this paper | **0.056** | **0.154** | 0.253  | 0.189  |

This paper makes statistics on the proportion of scores scored by raters for different models. The statistical results are shown in Table 3. Compared with the other three models, in terms of content, the "1" score of the seq2seq model is the highest. Most of the "1" scores are meaningless but safe replies, such as "I don't know" and "me too". In contrast, the "1" score of the reply generated by the model of
introducing external knowledge decreased, and the proportion of the "2" score increased, which verified the necessity of introducing external knowledge into the dialogue system.

In terms of knowledge, compared with MemNet and PostKS, the response generated by this model has the highest proportion of "2" scores, but the proportion of "1" is also the highest. It may be because this model obtains more knowledge by paying attention to the various semantic spaces of the current input, and introduces knowledge with low relevance, resulting in meaningless response generated by the model.

| Model                  | Content           | Knowledge          |
|------------------------|-------------------|--------------------|
|                        | 0     | 1      | 2      | 0      | 1      | 2      |
| Seq2Seq                | 35.47% | 49.12% | 15.41% | -      | -      | -      |
| MemNet                 | 35.66% | 44.21% | 20.13% | 34.24% | 30.21% | 35.55% |
| PostKS                 | 34.45% | 39.81% | 25.74% | 31.8%  | 31.52% | 36.68% |
| Model of this paper    | 38.64% | 34.70% | 26.66% | 26.91% | 33.31% | 39.78% |

3.2.2 Result Analysis. A sample is selected from the test set. Table 4 (a) shows the current input and related knowledge, and table 4 (b) shows the responses of the benchmark model and this model based on the current input and related knowledge. Different aspects of a sentence shown in Table 4 (a) can involve different knowledge. It can be seen that the responses generated by the model in this paper are more coherent and diverse. It is shown that the knowledge dialogue system based on multi-head attention mechanism can generate content-rich responses, enhance the experience of human-computer interaction, and introduce knowledge-generated responses in line with human communication habits.

Table 3. The proportion of the score in the artificial score

Table 4(a). Current input and related knowledge.

| Current Input                  | Knowledge                                      |
|--------------------------------|------------------------------------------------|
| Do you like Lin Feng's "The Apostle"? | [The Apostle  Time.com short comment's  The end song is very good ah] |
|                                | [The Apostle  Genre  Crime]                    |
|                                | [Lin Feng  Review  a perfect man]              |
|                                | [Lin Feng  Domain  Star]                       |
|                                | [Lin Feng  Main Achievements  The most popular male singer in Asia Pacific] |
|                                | .......                                         |

Table 4(b). A sample of each model generating a reply

| Model               | Reply                                      |
|---------------------|--------------------------------------------|
| Sq2Seq              | Yes, I like it.                            |
| MemNet              | The end credits of "The Apostles" are great.|
| PostKS              | The Apostle is a crime film.               |
| Model of this paper | I like Lin Feng. "The Apostle" is a crime movie. |

To further verify that the knowledge selected by each head is different, the knowledge vector selected by each head is used as the initial state of the decoder along with the current input vector, respectively. It can be observed that each header pays attention to a specific semantic aspect and selects the corresponding knowledge. The head 1 pays attention to "like" and gives a response, but there is no
relevant knowledge, indicating that not every semantic aspect will have relevant knowledge. In Head 2, the focus was on "McDull - The Rice Bowl Corps". The head 3 focus is on "Bloomberg", and the corresponding responses are related to external knowledge.

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

| Current input | Response |
|---------------|----------|
| Do you like watching "McDull - Rice Bowl Corps" starring Bloomberg? | Head 1: I don't like it.  
Head 2: "McDull - Rice Bowl Corps" of rating 6.9 points.  
Head 3: I really like Bloomberg very much! |

4 Conclusion
In this paper, we propose a knowledge selection mechanism based on multi-head attention, which maps the input utterance to multiple semantic spaces and makes full use of the multi-aspect information of the input utterance to guide knowledge selection. A penalty term is also proposed in order to force the multi-head attention mechanism to focus on the different contents of the input utterance. The attention mechanism is added in the decoding phase to dynamically attend to both the input utterances and the selected knowledge. Both automatic evaluation and manual evaluation prove the effectiveness of this model in knowledge selection. In the future work, we will further explore how to avoid choosing duplicate knowledge in multiple rounds of dialogue.

Reference
[1] Ghazvininejad M, Brockett C, M.W. Chang, et al. (2018) A knowledge-grounded neural conversation model[C]. Thirty-Second AAAI Conference on Artificial Intelligence, 19
[2] Liu S, Chen H, Ren Z, et al. (2018) Knowledge diffusion for neural dialogue generation[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 1489-1498
[3] Lian R, Min X, Fan W, et al. (2019) Learning to select knowledge for response generation in dialog systems. J.
[4] Zhang Y, Ren P, MD Rijke. (2019) Improving background based conversation with context-aware knowledge pre-selection.
[5] Zheng C, Cao Y, Jiang D, et al. (2020) Difference-aware knowledge selection for knowledge-grounded conversation generation. J.
[6] Tao C, Gao S, Shang M, et al. (2018) Get the point of my utterance! Learning towards effective responses with multi-head attention mechanism[C]/ Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18.
[7] Lin Z, Feng M, Santos C, et al. (2017) A structured self-attentive sentence embedding. J.
[8] Wu W, Guo Z, Zhou X, et al. (2019) Proactive human-machine conversation with explicit conversation goal. J. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
[9] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch.
[10] KINGMA D, ADAM B. (2015) A method for stochastic optimization [C]/ Proceedings of the 3rd International Conference on Learning Representations. San Diego, California, USA: International Conference on Learning Representations: 789-795.
[11] Sutskever I, Vinyals O, Le Q V. (2014) Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems:3104-3112.
[12] Papineni K, Roukos S, Ward T, et al. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation.
[13] Li J, Galley M, Brockett C, et al. (2015) A diversity-promoting objective function for neural conversation models. J. Computer Science.