A Summary of Research on Intelligent Dialogue Systems

Bangjing Wang1 a, Hui Cao1 *
1Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

Abstract. Dialogue systems have always been an important research direction of natural language processing. In recent years, the potential and commercial value of dialogue systems have attracted people's attention, and technologies such as intelligent customer service and chatbots have gradually become popular. First, it outlines the concept and classification of the dialog system; then introduces the constituent modules of the dialog system, and mainly studies the key technologies and mainstream ideas; finally, summarizes the problems encountered at this stage and looks forward to the future research direction.

1. Introduction

Science fiction movies often have virtual assistants with high intelligence, they can freely talk to the host, and provide help to the host, with a full sense of technology. With the advent of deep learning and other technologies, creating a robot that can chat with us is no longer a fantasy. First of all, in terms of data sources, there are a lot of dialogue data on the Internet, including some film and television works, articles, and novels. Secondly, deep learning technology has been verified in many natural language processing tasks, and it has performed very well in many complex tasks. A large number of documents using deep learning technology to construct dialogue system have emerged.

2. Classification of dialogue systems

The dialogue system can be understood as the use of machine learning, deep learning and other advanced technologies, so that the machine understands human language, correctly recognizes the intention of human language, and can communicate with humans or respond freely[1]. The application value of this technology is to replace manual work in highly repetitive operations, and to perform recovery work with high efficiency and high standards.

Dialogue systems can be divided into two types according to the number of dialogue rounds[1]: single-on dialogue systems and multi-round dialogue systems. According to the purpose of the dialogue, it can be divided into two types: task-oriented dialogue system, non-task-oriented dialogue system, and it can also be generalized as a task-oriented and small-talk dialogue system. This article mainly studies the non-task-oriented dialogue system. The non-task-oriented dialogue system can be divided into two types according to the mode: retrieval type and generative type[2].

2.1. Retrieval dialogue model

The retrieval model needs to construct a response database, and retrieve the most suitable response in the database as output based on heuristic inference rules, input, and context characteristics. The advantages of this method are obvious. Because the response database is built by hand, there is no grammatical error in the sentence that is answered while ensuring the quality of the database[2].
there are also certain flaws. For knowledge not covered in the database, the retrieval model cannot respond, nor can it generate any new text, and cannot be applied to scenarios where there is no predefined response.

2.2. Generative dialogue model

Compared with the retrieval model, the generative model does not depend on the constructed response data, but the model learns from the database to generate a new response. The advantage of this method is that it answers smartly and is closer to human conversation[3]. However, the model training process is more difficult. First, the model training requires high hardware resources. Second, the model learning features require a large amount of data. Finally, there may be grammatical errors when the output is long.

3. Research methods

3.1. Retrieval dialogue model

Retrieval dialogue refers to selecting an appropriate answer from the response database as an output by searching and matching. The general construction method is: first construct a response database that contains a large amount of data, then match the input data through information retrieval and other methods[4], and finally use the matching candidate answer scores as output.

To solve the problem of low utilization rate of dialogue interactive information in the retrieval dialogue model, Tao[4] proposed a deep retrieval model based on retrieval in 2019. An interaction-over-interaction network is defined in this article. The network is composed of interaction blocks. Each block contains a self-attention module, an interaction module, and a compression module. The self-attention module is used to extract the dependency between the question or reply; the interaction module models the dependency between the question and reply[4]; the compression module combines the results of the first two modules.
Let \( U_{k-1} \) and \( R_{k-1} \) be the \( K \)th input, \( U_0 = E_u, R_0 = E_r \), and the self-attention module is defined as:

\[
U^* k = f_{ATT}(U_{k-1}, U_{k-1}) \\
R^* k = f_{ATT}(R_{k-1}, R_{k-1})
\]

(1)

(2)

The interactive module first let \( U_{k-1} \) and \( R_{k-1} \) pass:

\[
U_k' = f_{ATT}(U_{k-1}, R_{k-1}) \\
R_k' = f_{ATT}(R_{k-1}, U_{k-1})
\]

(3)

(4)

After that, \( U_{k-1} \) and \( R_{k-1} \) interact with \( U_k \) and \( R_k \) respectively:

\[
U_k'' = (U_k \odot U_k') \\
R_k'' = (R_k \odot R_k')
\]

(5)

(6)

Finally, the \( U_{k-1} \) and \( R_{k-1} \) are updated to \( U_k \) and \( R_k \) through the compression module as the output of the module. After passing the ioi network, in the \( K \)th interaction module, three similarity matrices are constructed:

\[
M_{i,1}^k = \frac{u_{i-1,k}^k(r_{k-1})^T}{\sqrt{d}} \\
M_{i,2}^k = \frac{u_{i,k}^k(r_{k}^k)^T}{\sqrt{d}} \\
M_{i,3}^k = \frac{u_{i,k}^k(r_{k}^k)^T}{\sqrt{d}}
\]

(7)

(8)

(9)

Then connect these three matrices into a three-dimensional vector, and use the convolutional neural network to extract features from \( T_{k+1} \) and match. The final feature output is mapped to the \( d \)-dimensional vector through the hidden layer to obtain the final matching score.

Finally, the model is validated on three data sets. The ioi model performs better when interacting with medium depth. Compared with the existing model, the accuracy rate is improved by 2%.

### 3.2 Generative dialogue model

Generative dialogue can be regarded as a seq2seq problem. However, the main reason why the seq2seq paradigm is successful in solving problems in the field of machine translation is that the output interval corresponding to the input problem is limited[5]. If the same technology is used to build the dialogue model, the first problem is: unable to pay attention. When it comes to context, the model cannot pay attention to the contextual connection. For example, the above input "I am sick today", and then one of the inputs "what to do next", the reply "go to play basketball together", this reply is for the current input is correct, but the above is not taken into account at all. In fact, the reply of playing basketball should not appear in the output space; the second question is: the gender, age, work, status, income and so on have a certain impact on the output space, these conditions have a certain limit on the recovery space.

In response to the above problems, Li[6] proposed an incremental generating dialogue model based on Transformer in 2019. The input of this model contains multiple rounds of dialogue history and a document. The goal is to output a reply that integrates the knowledge of the document. The model updates the Transformer model structure and incrementally updates the dialogue sentences and associated documents to the model. The model uses a two-stage decoding idea. The first stage focuses on contextual coherence, and the second stage introduces relevant documents to modify the results of
the first stage.

In Figure 4, the upper part is the encoder frame and the lower part is the decoder frame. The overall framework consists of three parts: self-attention encoder, incremental encoder, and collaborative decoder. The self-attention encoder independently encodes document knowledge and current dialogue; the incremental encoder uses incremental encoding to encode multiple rounds of dialogue, and uses the attention mechanism to establish context-sensitive document knowledge representation[7]; the collaborative decoder is a two-way conversion. For the decoder, the first pass of decoding relies on the dialogue context to generate a response, and the second pass of decoding uses document knowledge for further correction.

Encoder construction steps:

First, the multi-head attention method is used to merge and encode the document knowledge and the contextual dialogue:

\[ C^{(k)} = ITE(C^{(k-1)}, d^{(k)}, Inu^{(k)}) \]  

The coding layer of the Encoder structure contains four sublayers:

a. multihead self-attention:

\[ B^{(n)} = MultiHead(C^{(n-1)}, C^{(n-1)}, C^{(n-1)}) \]  

b. multi-head knowledge attention:

\[ E^{(n)} = MultiHead(B^{(n)}, d^{(k)}, d^{(k)}) \]  

c. multi-head context attention:

\[ F^{(n)} = MultiHead(E^{(n)}, c^{(k-1)}, c^{(k-1)}) \]  

d. position-wise fully connected feed-forward network:

\[ C^{(n)} = FFN(F^{(n)}) \]  

Decoder construction steps are as follows:

The network structure of the first decoding and the second decoding is the same, which are composed of four sublayers[8]:

a. multi-head self-attention:

\[ G_{1}^{(n)} = MultiHead(R_{1}^{(n-1)}, R_{1}^{(n-1)}, R_{1}^{(n-1)}) \]  

b. multi-head context attention:

\[ H_{1}^{(n)} = MultiHead(G_{1}^{(n)}, c^{(k)}, c^{(k)}) \]  

c. multi-head utterance attention:

\[ M_{1}^{(n)} = MultiHead(H_{1}^{(n)}, SA_u(U^{(k)}), SA_u(U^{(k)})) \]  

d. position-wise fully connected feed-forward network:

\[ R_{1}^{(n)} = FFN(M_{1}^{(n)}) \]  

Then use the softmax function to obtain the probability of decoding the word for the first time. In the second decoding process, you only need to replace the historical dialogue information of the second sublayer with the current document representation, and the current round of sentence information of the third sublayer with the output of the first decoding[9], and finally output after correcting with the knowledge of the document word.

The change in the decoder part of the model reduces the perplexity of the results and improves the performance of the overall model[10].
4. Conclusion and Future Work
Judging from the current technical level, the realization of the retrieval dialogue system is better, and the overall dialogue system is less effective due to the contextual problems and the like. As far as the application in the commercial field is concerned, the prospect of the generative dialogue system is very broad, and it is also the current development trend of the dialogue system. The follow-up plan is based on the above models, research and integration of ELMo, BERT and other models to achieve a high-performance dialogue system.

References
[1] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In Advances in Neural Information Processing Systems, pages 2042–2050.
[2] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017b. FusionNet: Fusing via fullyaware attention with application to machine comprehension. In International Conference on Learning Representations.
[3] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1468–1478.
[4] Tao, C., Wu, W., Xu, C., Hu, W., & Yan, R.. (2019). One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
[5] Guo, X., Zheng, L. H., Li, M. Z., & Zhang, Y.. (2014). Intelligent data acquisition and cloud services for apple orchard. International Journal of Agricultural and Biological Engineering, 7(2), 146-153.
[6] Li, Z., Niu, C., Meng, F., Feng, Y., Li, Q., & Zhou, J. (2019). Incremental transformer with deliberation decoder for document grounded conversations.

[7] Shen, Y., Li, Y., Sun, J., Ding, W., & He, J. (2019). Hashtag recommendation using lstm networks with self-attention. Computers, Materials and Continua, 61(3), 1261-1269.

[8] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In Thirtieth AAAI Conference on Artificial Intelligence.

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

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008.