Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230945

Performance analysis and comparison of representative chatbots based on deep learning

Runpu Wang

Shanghai Jiao Tong University, Shanghai, Minhang District, 201100, China

sjtu-walter@sjtu.edu.cn

Abstract. In today's society, chat robots have entered people's vision. They can mainly carry out corresponding human-computer interaction, and are also research hotspots in the field of science and technology. Chat robots can build models to understand the input content, and then output relatively natural answers. In recent years, their birth has also produced a certain range of applications. However, the most representative chat robots are mainly based on retrieval and generation. Their patterns are different, so their behavior is also different. At the same time, in order to enable users to choose better chat robots, this paper compares the performance of retrieval chat robots based on SMN model, retrieval chat robots based on DAM model and seq2seq generation chat robots respectively. Their performance was analyzed and evaluated, mainly from the appropriateness and diversity of their responses and the difficulty of the training model.

Keywords: Chatbot, Seq2Seq, SMN, DAM, Natural language processing.

1. Introduction

Recently, deep learning has been developing continuously in the fields of automatic driving, video and so on. Thus, deep learning can also be applied in the fields related to natural language, and it is also very effective. However, natural language processing and human-computer interaction are also hot topics in today's society. Chat robots aim to build models to understand the input content and output natural and logical answers. They have been applied in e-commerce, medical care, etc. and have also been recognized by the society..

However, for chat robots, their main purpose is to automatically find answers to some questions. For early chat robots, templating is too serious. If the user's content matches the template that has been written, then a better answer can be obtained. Otherwise, some vague answers will be obtained, such as "I don't know this question". Previously, due to the Loebner Award and Chatter Gbox Challenge, regular chat robots were generally popular. After. A variety of chat robots are emerging, such as Parry and Alicebot. In this way, the development of chat robots is slowly limited. It is not easy to develop. It is a bit like the market is saturated.

With the continuous development of machine learning, data-driven chat robots are gradually being studied. They are more high-end robots, like retrieval based chat robots and generation based chat machines, which use more advanced technologies. The retrieval based chat robot uses information retrieval technology as the main carrier, which can store dialogue materials first, and then respond to requests to meet users. The retrieval based chat robot is more simple, because it mainly uses a database of predefined responses and some heuristic reasoning to select appropriate responses according to input and context [1]. In addition, they can also be obtained by grabbing the online conversation materials before human beings, so the computational load is relatively small, and the response can be given quickly, so the efficiency is relatively high. For the chat robot based on the 2nd generation, they refer to the use of natural language generation technology to automatically respond to users' conversation requests. The generative model is mainly generated with new responses, rather than relying on predefined responses. However, the current mainstream algorithm is to sequence model.

However, these two design ideas have their own advantages and disadvantages. After all, nothing is perfect. Although retrieval based methods have more accurate calculation results, they cannot operate well on the scene. If you want to generate a more efficient model, you need a lot of training data. However, it is easy to make mistakes, which is their disadvantage. Also, we need to know that the models used in their design are different, so their behaviors are more different, which is also easy to understand. In this paper, we will carry out a detailed study of various aspects of the performance of chat robots, so as to greatly explore their convenience. Specifically, we studied retrieval chat robots based on SMN model, DAM model and seq2seq respectively. The performance of chat robots with different design methods is evaluated in detail. We will discuss the appropriateness, diversity and difficulty of training models of chat robots.

The structure design of this paper is mainly like this. In section 2, we introduce retrieval chat robots based on SMN model, DAM model and seq2seq, and discuss their respective theories and main implementation processes. In section 3, we compared the application performance of different chat robots. Finally, in the fourth section, we summarized some problems in the development of different chat robots, and looked forward to their development direction and situation.

2. Methods

2.1. Chatbot based on SMN

We will test the retrieval based chat robot based on SMN [2] according to our standards, and use the Ubuntu corpus [3] dataset. However, for SMN, it is mainly to match the responses of each discourse, and then collect and complete the next step. For matching, another method is to obtain two granularity matching representations by performing correlation matching at the sub sequence level. Then we need to perform convolution related merging operation to extract the matching information. However, after the operations mentioned above, the matching information of each utterance is obtained. GRU [4] is used to accumulate all matching information in chronological order. Finally, the corresponding matching degree is calculated by using these hidden states.

However, in the case of a discourse response match, after the word is embedded, the response and the discourse become two arrays. The matrix representation of each discourse will be obtained, and each word will be encoded into a vector with the length of d. Then the final data size of this discourse is $R^{d \times n_u}$, Similarly, the response representation we chose is $R^{d \times n_r}$. This gives us a similarity matrix representation of word-level on a scale of $R^{n_d \times n_r}$. Finally, we will extract the features that have a large impact in the text [4]. We now have the key characteristics of each utterance and that candidate response, first pulling them into a one-dimensional form. Then, we put it in the GRU for encoding. The purpose of this is also so that each utterance's feature affects all other utterances, and to preserve timing information between utterances. The last step is to merge the output of the GRU above.

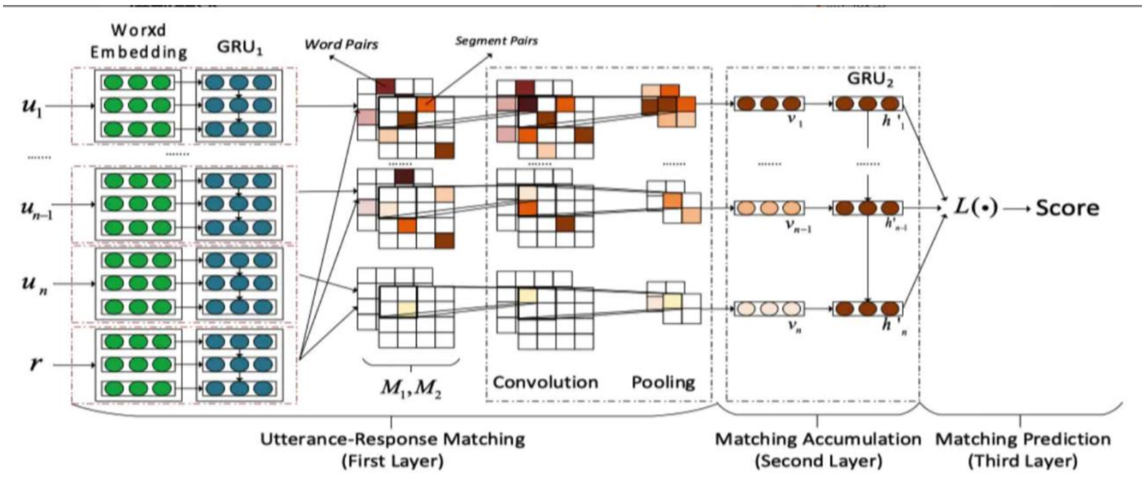

Figure 1. Basic structure of Seq2Seq model.

2.2. Chatbot based on DAM

We will test our retrieval-based chatbot based on DAM[5] against our criteria, and using the Ubuntu Corpus [3] dataset. First, each utterance in the response and context obtains different representations through the self-attention layer. Attention is a important part of this model and it is very effective [6-7]. Two layers of self-attention layers are stacked, that is, each sentence has two different granularity representations, plus the word embedding representation at the beginning, each There are three expressions in a sentence. The response is matched with each utterance to form a 2D matching matrix. The next step is to combine all the obtained matrices, then perform 3D convolution and pooling operations, and get a final matching score in the connection linear layer.

However, for DAM, they can mainly express the meaning of the head word in combination with the context, and then generate some relatively complex fragment meanings according to the head word. Given text dependency and dependency information, each statement in the context and response will be matched according to fragments of different granularity. Then the DAM can match information from the context, then perform matching feature operations, and finally perform content fusion. In addition, since most attention computing can be completely parallelized, DAM can continue to carry out relevant research in the subsequent design.

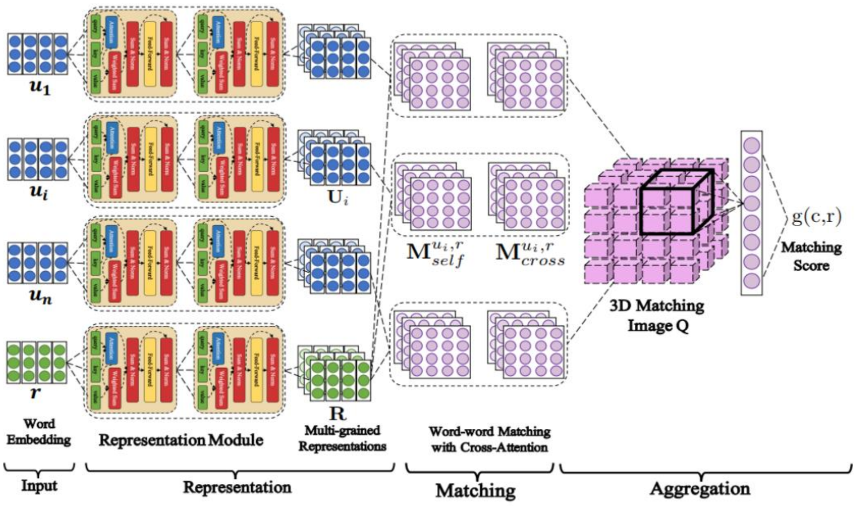

Figure 2. Overview of DAM model.

2.3. Chatbot based on Seq2seq

But for Seq2Seq [8], it is mainly an architecture, mainly used for the role of codec, which is also a hot research field. When talking about Seq2Seq (Sequence to Sequence), we should know that it is based on a given sequence and has specificity. Then comes Seq2Seq model, which is an important variant of RNN. Seq2Seq is mainly composed of encoder, decoder and state vector. The basic architecture of Seq2Seq is shown in Figure 1. The function of the encoder is to convert the input variable length sequence into a fixed length state vector. However, the state vector has information about the input sequence, and the related information can be sent to the decoder. Therefore, the decoder can generate a sequence of variable length.

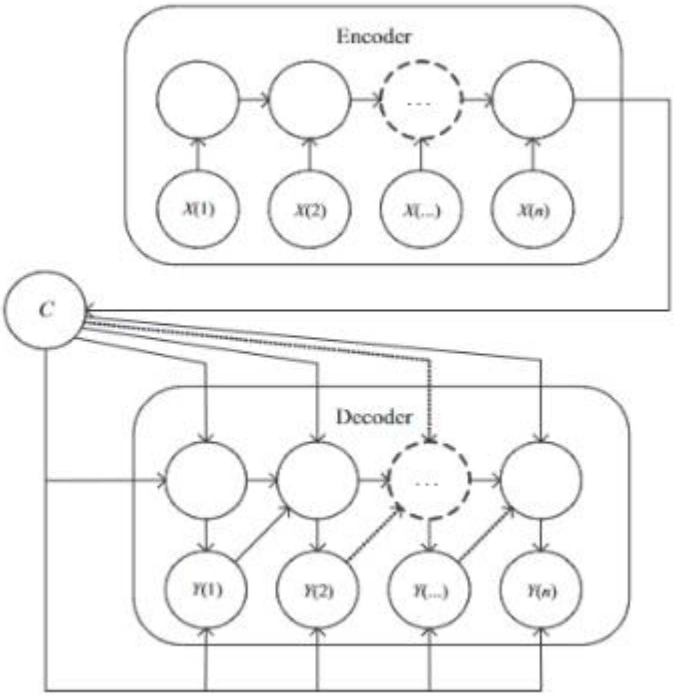

Figure 3. Basic structure of Seq2Seq model.

There is also the encoder decoder [9]. Speaking of it, it is mainly its architecture, which simplifies the data set to a certain extent. The data set is converted into a fixed length vector and represented by the encoder. Then the vector is introduced into the network as a data set for training. However, the decoder can decode the output of the algorithm to obtain a series of uncertain answers. In this way, the efficiency and accuracy of the system can be fully improved.

In the Seq2Seq model, the long and short term memory (LSTM [10]) structure is mainly used to extract the unit structure of sentence information. However, LSTM is a cyclic neural network model, which can form a dynamic network. The recurrent neural network takes the feedback of the previous node as the input, and then there are relevant operations that can calculate the output and loss. It should be noted that the original Seq2Seq model only uses one intermediate semantic vector to represent the semantics of all input sequences, which will result in the loss of more information of the input sequences. At this time, the attention mechanism mainly allows the model to automatically search for relevant information, so that different output results can focus on different inputs. This method is widely used.

However, it is the original Seq2Seq model, which uses greedy search in the decoding process. This algorithm is relatively complex. In the decoding process, the output of each word selects the value with the highest probability, and the input at the next moment is this value. The final result is often repeated security responses, because there are more corpora in the training set, such as "I don't know". In addition, in the decoding process of Seq2Seq model, Beam Search algorithm should also be considered. In the decoding process, the Beam Search algorithm first outputs the first K words with the highest probability, and then takes these K words as the input. The output at the next time needs to be calculated, and then there will be K results. In this way, the responses generated by hungry words have diversified characteristics.

3. Experiment and performance analysis

3.1. Datasets processing

However, we have made relevant use of the Ubuntu Dialogue Corpus, which contains nearly 1 million conversations between two people extracted from the Ubuntu chat logs, to obtain technical support for various problems related to Ubuntu. The datasets contain the training set, test set and validation set. It contains the folder where the dialog appears, the ID number of the specific dialog, the timestamp of sending the dialog, the user who sent the dialog, the user who responded, and the text of the dialog. What we need to do is to delete the unanswered questions in the corpus. Since our research only focuses on whether the answers to the questions in one round are accurate, we only need to extract the first round of questions and answers from the corpus.

3.2. Criterion

Then we will discuss and study the appropriateness, diversity and difficulty of the training model of the chat robot reply. After model training, we also input the same set of data sets, make corresponding records, and finally judge the complexity of their respective models.

3.3. Performance analysis

Using our processed datasets, train the three models mentioned above and observe their loss function. Then we tested each of the three models using a test set containing 0.5 million pairs. In order to judge the accuracy of different models, We count the number of occur-rences of each model output matching the input (Table 1) and plot it into an image (See Figure 4).

Table 1. Relevant statistics of input and output matching of different models.

Method	SMN	DAM	Seq2seq
number	493324	471546	488641

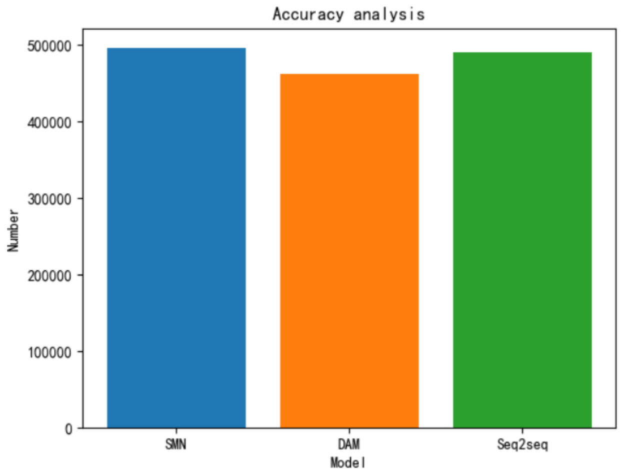

Figure 4. Correlation analysis on the accuracy of different chat robots.

Then, in order to verify the rationality of the answers generated by the three models, we selected the correct combination in all three models. Ask 30 test participants to choose which model produces the most reasonable and life-like answer to the same question. There are a total of 10 questions, whose the results are shown in Figure 5 and Table 2.

Table 2. Display of relevant forecast results.

	SMN	DAM	Seq2seq
number	116	121	63

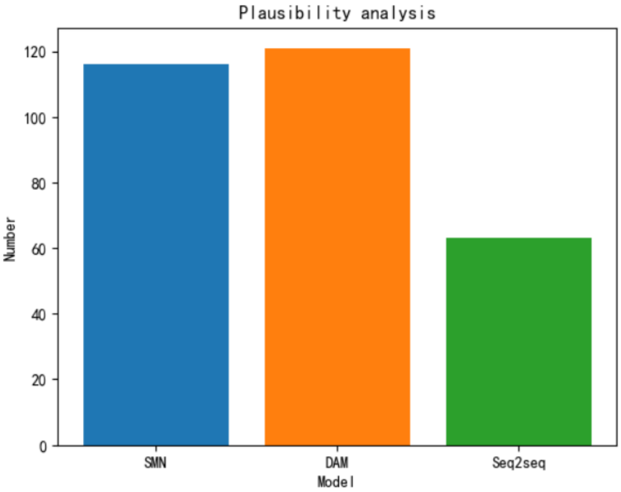

Figure 5. Analysis and comparison of the accuracy of some relevant results.

However, from the data analysis, we can also get some relevant research conclusions. The accuracy of the three models is not different from each other. And from the perspective of rationality, it can be seen that the output of the Seq2seq model is quite different from the dialogue in real life, which is mainly caused by different design models. It should also be noted that SAM model is capable of conducting multiple rounds of dialogue, DAM model is capable of semantic matching, and Seq2seq has more diversified answers due to its generation characteristics. However, this paper does not carry out relevant research on these, which can provide ideas for future generations to discuss.

4. Discussion

In recent years, with the continuous progress of science and technology, in-depth learning and a series of software science and technology operations, chat robot technology is a major technological product. However, at the current level of development, there are still many difficulties:

(1) Open field chat robot. As the environment is in a closed area, the chat robots can be used well and their performance is relatively good. For example, in the ticketing system, but in an open outdoor environment, their performance is not very good. At this time, the data collected by the chat robots is not accurate and fuzzy enough to respond well to users.

(2) Lightweight model. Chat robots need a lot of computing operations. Deep learning plays a major role in the development of chat robots, but it also has a huge demand for computing resources. For the chat function that is widely used at present, the edge 8 device does not have much computing resources and needs the background server to assist in computing. For this problem, the algorithm of lightweight chat robot and the rectification of algorithm still need more research and related application operations.

(3) Build a large-scale, high-quality corpus. The training data is the source of the robot's learning to speak, which is critical to the quality of the answers. An ideal training dataset is large and of high

quality. Specifically, the content of the corpus touches on all aspects, allowing the robot to know everything. The content of the corpus must also be reliable, where there is no bad information or content that does not answer the question. The largest corpus available is mixed, which prompts us to pay more attention to the selection of high-quality corpora.

In addition, for future chat robots, they should be more diversified, so their development direction is relatively broad, and their models will be lighter. However, there are still some problems in the answers of the current chat robots, and their response operations to users are unreasonable. On the other hand, the demand for lightweight computing is growing. They can be applied to robots with weak computing power, which is also a necessary part of robot design.

5. Conclusion

This paper mainly discusses the chat robot, and introduces and analyzes the performance of robots with different design ideas in detail. The purpose of this is to provide some useful suggestions for the selection of chat robots in different application environments. To sum up, we have studied retrieval chat robots based on DAM model, smnmodel and seq2seq respectively, evaluated their respective performance, advantages and disadvantages, and mainly studied and discussed their reply appropriateness, reply diversity and training model difficulty.

References

[1] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. arXiv preprint arXiv:1408.6988

[2] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Match- ing Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

[3] Wang Lubao Research on end-to-end task-based dialogue system based on deep learning [D]. Changchun University of Technology, 2022. DOI: 10.27805/d.cnki.gccgy.2022.000447.

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

[5] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classifification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[6] Wang Kexin Research and Implementation of Intelligent Chat Robot Based on Deep Learning [D]. Heilongjiang University, 2021. DOI: 10.27123/d.cnki.ghlju.2021.000387.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning, pages 2048–2057.

[8] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, Wei-Ying Ma. 2016. Topic Aware Neural Response Generation. arXiv preprint arXiv:1606.08340

[9] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078

[10] Xingjian Shi Zhourong Chen Hao Wang Dit-Yan Yeung. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv preprint arXiv:1506.04214