Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230288

Sturdies advanced in chatbots based on deep learning

Dexiao Zhang

United International College, Beijing normal university&Hong kong Baptist university, Zhuhai, Guangdong province, 519000, China

1811031224@mail.sit.edu.cn

Abstract: Chat robots have always been a hot research topic in the field of natural language processing and human-computer interaction, aiming to build models to understand human input and provide logical answers. Early chat robots were mostly based on matching or rules, but due to the lack of manual feature representation capabilities and the complexity of rules, they could not meet practical application requirements. Thanks to the rapid development of deep learning, chat robots based on convolutional neural networks have made breakthroughs in accuracy and speed in recent years. According to the different design ideas of chat robots, the existing work mainly includes two categories: retrieval type chat robots and generation type robots. Focusing on the two main frameworks mentioned above, this article introduces the latest research progress of chat robots in detail, including the basic steps, design ideas, advantages and disadvantages of representative algorithms. We also introduced common chat robot modeling data sets and evaluation indicators. Finally, we discussed the existing problems and future development directions in the field of chat robots.

Keyword: chatbots, retrieval-based, deep learning.

1. Introduction

Chat robots have always been a hot research topic in the fields of natural language processing and human-computer interaction, which is an artificial intelligence program and a model of human-computer interaction (HCI) [1]. According to the dictionary, a chatbot is "a computer program designed to simulate a conversation with a human user, especially over the Internet" [2], which use natural language processing (NLP) and sentiment analysis to communicate with humans or other chatbots in human language through text or verbal speech [3-4]. By interacting with humans in natural language, chatbots can provide various services to users, such as customer service, entertainment, education, etc. Nowadays, the chatbots have attracted lots of attentions from academia and industry.

The research on chat robots has a long history, which can date back to Eliza, a chat robot developed in the 1960s. This chat robot mainly relies on templates, and its basic idea is to match the user's input with the pre written template and return the answer. If the matching fails, some vague responses will be fed back, such as "I don't know". However, due to the complexity and variety of human languages, relying solely on templates cannot enumerate all situations. Template based chat robots cannot meet the actual application requirements. Thanks to the rapid development of machine learning and even deep learning, as well as the construction of large-scale corpus, data-driven chat robots have gradually matured. According to different design ideas, existing research can be divided into two categories:

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230288

retrieval based chat robots and generation based chat robots.

(1) Retrieval based chat robots use information retrieval technology to match a previously stored conversation corpus as a response to a user's conversation request. The retrieval based model is relatively simple and mainly based on user input and contextual content, which use a knowledge base (storing predefined response content) and some heuristic methods to obtain an appropriate response. Heuristics include simple rule based expression matching, and complex machine learning classifiers. Identify which category the user's intention belongs to through classification, and then search for an answer in the corresponding category. These systems cannot generate any new content, just find the appropriate content from a fixed dataset as a response.

(2) Generation based chat robots refer to the use of natural language generation technology to automatically respond to user conversation requests. Generated chat robots are more complex, which do not rely on predefined reply content while use a generative method to generate new reply content word by word. Generative models are typically based on machine translation models. Unlike traditional machine translation models, the task of generative models is not to translate a sentence into a sentence in another language, but to translate the user's input into a response.

For models based on retrieval technology, due to the use of a knowledge base and the pre-defined data, the content to be responded to is syntactically smoother and less prone to syntax errors. The model based on retrieval technology does not have a conversation concept and cannot provide a more [intelligent] response in combination with context. This means the generative model may be more intelligent and more effectively to utilize contextual information to know what you are discussing. However, generative models are relatively difficult to train and require large-scale data. In addition, the output content often contains some grammatical errors (especially for long sentences).

Focusing on the two main frameworks mentioned above, this paper introduces in detail the latest research progress in the field of chat robots. Specifically, in Section 2, we introduce the representative methods of retrieval based chat robots and generation based chat robots, including their design ideas, key steps, advantages and disadvantages. Secondly, in Section 3, we introduce the commonly used chat robot modeling corpus and evaluation indicators. Finally, we also discussed the existing problems and future development directions in the field of chat robots, with the aim of providing some new insights for the development of this topic.

2. Method

Existing methods are either retrieval-based or generation-based. Retrieval-based approaches [5] retrieve response candidates from a pre-built index, rank the candidates and select responses from the top-ranked candidates, while generation-based approaches [6-7] use natural language generation (NLG) techniques to respond to messages. In general, matching algorithms must overcome the semantic gap between two objects [8]. In this work, we investigate the response selection of retrieval-based chatbots in the single-round case, since retrieval-based approaches always return fluent responses, and single rounds are the basis of conversations in chatbots.

2.1. Retrieval-based chatbots

A retrieval chat robot uses text matching and sorting learning techniques to find the most suitable response for the current input from a conversation corpus, whose design thinking and process can be summarised in the following steps: (1) identify the domain and text base. (2) identify the domain and relevant text base that the bot needs to cover. (3) collect and pre-process the text base that the bot needs to use, including processing the text base for word separation, deactivation processing, lexical annotation, entity recognition, etc. (4) designing the question classification module, which is responsible for classifying the questions entered by the user so that the bot can quickly search for answers from the relevant text base. (5) design the answer generation module, which is responsible for processing the text obtained from the search. (6) testing and optimisation: Conduct testing and optimisation of the bot, including evaluation and optimisation of the bot's answer accuracy, response time and interaction experience to improve the bot's performance and user satisfaction. On the basis of the above steps, a

DOI: 10.54254/2755-2721/8/20230288

complete retrieval chat robot is mainly composed of three modules: (1) candidate index retrieval module; (2) similarity feature calculation module; (3) sequencing learning module. The candidate index retrieval module collects large-scale human conversation corpus in advance and organizes the corpus into a question and answer format. Then, it uses indexing techniques in information retrieval to index these conversations for rapid retrieval by online modules. To realize the three key modules, the core algorithms of retrieval-based chatbots are based on a retrieval approach and consist of the following main algorithms: Natural Language Processing (NLP) algorithms: This algorithm is used to transform the natural language input by the user into machine-recognisable language. This includes techniques such as word separation, syntactic analysis and semantic analysis. Text retrieval algorithms: This algorithm is used to retrieve textual information that matches the user input, usually using algorithms such as vector space models, bag-of-words models to represent the text and techniques such as cosine similarity for matching. Semantic matching algorithms: This algorithm finds the best matching answer by matching the natural language entered by the user with a corpus pre-prepared by the system. This involves the use of techniques such as word vector models, topic models to represent the text and matching using similarity matching algorithms.

The advantages of retrieval chatbots are as follows: (1) simplicity of implementation. Retrieval chatbots are relatively simple to implement, do not require large amounts of training data and complex models, and can be deployed and brought online quickly. (2) Efficient. retrieval chatbots are fast to process and can give accurate answers in a short time. The retrieval chatbots also contains some disadvantages: (1) reliance on text libraries. Retrieval chatbots rely on existing text libraries for matching and cannot provide accurate answers to questions that do not appear in the library or to questions in new domains. (2) cannot automatically generate answers. As retrieval bots mainly match existing text, they cannot automatically generate answers from a corpus that does not already exist. In summary, although retrieval chatbots have some limitations, they have significant advantages in speed and efficiency when dealing with questions from existing text corpora. However, in the future, as natural language processing technology continues to advance, retrieval bots will be replaced by more intelligent and adaptive generative bots.

2.2. Generative chatbots

The generative chat robot differs from the original retrieval-based chatbots, which refers to the automatic organization of language responses when a person and a machine have a conversation. Generative chatbots typically use natural language generation algorithms based on neural networks, such as Recurrent Neural Networks (RNNs), Transformer Convolutional Networks (TCNs), GPTs and so on. These algorithms are able to generate natural language responses related to the input by learning a large corpus. The design thinking and design process is divided into several processes: data collection, data cleaning,, deployment of models, training of models, and feedback optimisation. The advantage of generative chatbots is that they can generate expressive, contextualised natural language responses that provide a more realistic conversational experience. However, generative chatbots also have some disadvantages; they are difficult to control the generated responses, they may produce responses that are not contextual or appropriate, they require large amounts of training data and computational resources, they are expensive to train and deploy models, and the generated responses may have incorrect, ambiguous or non-canonical syntax issues. Therefore, in practical applications, generative chatbots are often combined with retrieval chatbots to improve the interaction and usability of chatbots.

Currently, the key algorithms for generating chat robots mainly include RNN and seq2seq. RNN mainly solves the sequence problem, which means that the current output is not only related to the input at the current moment but also related to the output at the previous moment, which constitutes the association between the preceding and following words. However, RNN networks also have many problems that are difficult to solve, such as too large a dataset, increasing the difficulty of training, and prone to problems such as gradient disappearance and gradient explosion. The eq2seq model is also known as the sequence to sequence model, because it has achieved good results in translation systems, and is subsequently applied to chat robots. The seq2seq model consists of two parts: an encoder and a

DOI: 10.54254/2755-2721/8/20230288

decoder, which are introduced into the chat robot task. The encoder is defined as a question sentence, and the decoder is defined as a reply sentence.

3. Experiment

3.1. Metrics for evaluating performance

There are many different views on how to evaluate the performance of chatbots. From an information retrieval (IR) perspective, chatbots have many functions. Evaluators can ask questions and requests to the bot , and assess effectiveness by measuring accuracy, precision, recall and Fscore relative to the correct chatbot response [8].

The performance metrics we can evaluate for chatbots can be divided into: Accuracy: measures whether the responses generated by the chatbot are correct and accurate. This can often be done using a manual evaluation where a human reviews the responses and gives a score or feedback. Fluency: A measure of how fluent, natural, and linguistically appropriate the responses generated by the chatbot are. This can often be assessed using language model metrics such as Perplexity and BLEU. Diversity: Measures how diverse, creative and varied the responses generated by the chatbot are. This metric can often be assessed by evaluating how similar different responses are. Usability: Measures the usability of a chatbot, including aspects such as responsiveness, stability, scalability, and flexibility. This metric can often be evaluated by testing the response time, concurrency, and troubleshooting of chatbots. User Satisfaction: Measures user satisfaction and experience with the chatbot. This can often be assessed using user feedback, surveys, user retention rates, etc. Users are asked to complete a satisfaction survey, which can be statistically analysed or even simply rated on a response-by-response basis. ,

Evaluating the performance of a chatbot based on its interaction with the user can be very subjective. In different situations, the metrics may vary depending on the context, domain and type of interaction. Therefore, it is crucial to identify benchmarks or specific criteria to evaluate chatbots in a standardised way [9].

3.2. Corpus

Common existing corpora for chatbots are mainly as follows:

(1) Cornell Movie-Dialogs Corpus. Cornell Movie-Dialogs Corpus is a corpus of movie conversations collected by Cornell University, containing a large amount of text from movie conversations that can be used to train chatbot models.

(2) Persona-Chat. A corpus of persona conversations published by the Facebook AI Research team, containing conversations about different characters that can be used to train chatbot models with personalisation features.

(3) DailyDialog. DailyDialog is a conversation corpus published by the Chinese University of Hong Kong, containing everyday conversations from different domains and topics, which can be used to train chatbot models. et al. 2013). The latter is constructed from posts and replies on social networks. networks that are noisy, brief, and related to real conversations [10].

(4) Twitter dataset. Twitter dataset contains public data obtained from the Twitter social media platform, including a large number of conversations, opinions and sentiment expressions, which can be used to train chatbot models.

(5) Weibo dataset. Weibo dataset contains public data from China's largest microblogging social media platform, including a large number of conversations, opinions and sentiment expressions, etc., which can be used to train a chatbot model.

These corpora can be used to train chatbot models to help bots better understand and respond to user input. At the same time, these corpora have different characteristics and domains, and the appropriate corpus can be selected for training according to actual needs

3.3. Algorithms or methods for getting metrics on chatbot performance

User satisfaction surveys: Users are asked about their experience and satisfaction with the chatbot by

DOI: 10.54254/2755-2721/8/20230288

providing them with a questionnaire to understand how well the bot is performing. This is usually assessed using a Likert scale, such as a 1-5 scale, where 1 means 'very dissatisfied' and 5 means 'very satisfied'.

Conversation quality assessment: Human experts or natural language processing techniques (e.g. BLEU, ROUGE, etc.) are used to assess the conversation quality of chatbots. Conversation quality includes grammatical correctness, fluency, semantic accuracy and ability to convey information. v Success rate: The ability of a chatbot to solve a user's problem or execute a user's request, usually expressed as a percentage. For example, if a bot is able to successfully answer 80% of a user's questions, then it has a success rate of 80%.

Response time: The speed at which a chatbot responds to a user's message, usually measured in seconds. The shorter the response time, the higher the user satisfaction rate is typically.

Conversation length: The average number of messages a chatbot sends in a conversation. Shorter conversations usually mean that the bot is able to solve problems faster, but may also mean that it does not fully understand the user's needs.

Multi-conversation fluency: the ability of a chatbot to remember previous conversations and stay in tune with the user during multiple rounds of conversation. Fluency assessments typically involve multiple consecutive rounds of conversation. These metrics can be adapted and improved for specific application scenarios and goals to improve the performance and user experience of chatbots.

4. Discussion

Although chatbots have grown considerably in recent years, a number of issues remain, including: Limited understanding: chatbots often struggle to understand the nuances of human language, leading them to provide irrelevant or incorrect answers to users. Lack of contextual understanding: chatbots may not be able to understand the context and background of a conversation, resulting in genericised responses that do not meet the user's specific needs or questions. Inadequate learning: Some chatbots lack the ability to learn and continuously improve, resulting in them providing the same answers over and over again, even when they are not useful or relevant.

(1) Technical problems. Chatbots may experience technical problems such as slow response times or system crashes. Privacy and security issues: Skybots may require users to provide sensitive information, such as personally identifiable information, financial information, etc., which may raise privacy and security concerns. User acceptance: Some users may be less willing to engage in conversations with bots, preferring to communicate with real humans. These issues pose challenges for the development and use of chatbots and require further improvement and innovation by developers and designers.

(2) Can we make a graphical chatbot. Currently the most common type of chatbot is a text-based chatbot, could we make a graphical chatbot. This type of chatbot can interact with users through a combination of images and text, thus providing a more vivid and intuitive communication experience. For example, it can answer users' questions or provide information through images, GIFs, videos, etc. The development of graphical chatbots requires designers and developers to use a combination of image processing, natural language processing, machine learning and other techniques, combined with a domain- and scenario-specific corpus for training and optimisation. For example, in the e-commerce domain, a graphical chatbot that can recognise product images and provide product information can be developed to improve the user's shopping experience. However, compared to traditional text chatbots, graphical chatbots require more technical and resource investment, and are therefore relatively expensive to develop and maintain. Also, as image recognition and processing technologies are still evolving, graphical chatbots may face some technical challenges and limitations

(3) Lightweighting. Lightweighting of chatbots refers to reducing the size and complexity of a model while maintaining its performance, thereby reducing the storage and computational resource consumption of the model. The following are several commonly used techniques for lightweighting chatbot models: Parameter sharing: When training a model, the size of the model can be reduced by sharing parameters. For example, in a sequence-to-sequence model, the parameters of the encoder and decoder can be shared, thus reducing the number of parameters of the model. Pruning: After the model

DOI: 10.54254/2755-2721/8/20230288

has been trained, the model size can be reduced by removing some redundant parameters through pruning techniques. Pruning can be done based on weights, channels, neurons, etc. Quantization: Converting floating point parameters in a model into integer parameters can significantly reduce the size of the model and speed up the inference process of the model. Distillation: Transferring the knowledge of a complex model to a smaller model, thus allowing the smaller model to have high performance as well. Distillation can be based on soft labels or knowledge matrices. Network structure optimisation: The size and complexity of a model can be further reduced by adjusting the network structure, reducing the number of network layers, etc. For example, lightweight structures such as deeply separable convolutions can be used instead of traditional convolutional layers. Lightweighting of chatbots can be achieved by a variety of technical means. In practice, suitable techniques need to be selected, optimised and tuned according to the actual situation, in order to achieve the goal of reducing the size and complexity of the model while ensuring its performance.

5. Conclusion

A chatbot is an "online human-computer dialogue system with natural language". By comparing chatbots and generative chatbots we find that they differ in terms of accuracy, conversational fluency, knowledge coverage, conversational ability, and training difficulty. In terms of knowledge coverage retrieval chatbots look for answers in a given knowledge base and therefore have a wide range of knowledge coverage in these domains. In contrast, the knowledge of a generative chatbot depends on the data it is trained on and therefore may not cover as wide a range of knowledge domains as a retrieval chatbot. Generative chatbots are able to handle more complex multi-round conversations in terms of their conversational capabilities, as they can generate new responses and perform contextual understanding and reasoning during the conversation. Retrieval chatbots, on the other hand, can only respond based on a predefined library of responses, which may not be accurate enough. In terms of training difficulty, generative chatbots are typically more difficult to train because they require a larger amount of data and higher computational resources. Retrieval chatbots, on the other hand, only need to match on a pre-prepared knowledge base and are therefore relatively easy to train. In summary, both types of chatbots have their own advantages and limitations, and the choice of which type of chatbot to use depends on the application scenario and requirements. The algorithms represented by chatbots are generative models, retrieval models, and hybrid models. The algorithm design of a chatbot depends on the specific application scenario and requirements, and the different algorithms have their advantages and disadvantages. Therefore, when implementing a chatbot, it is necessary to choose the right algorithm for the situation, and to optimise and debug it. In the future we can continue to improve natural language processing techniques and enhance sentiment analysis capabilities to optimise chatbots.

References

[1] Ren F, Bao Y. A review on human-computer interaction and intelligent robots[J]. International Journal of Information Technology & Decision Making, 2020, 19(01): 5-47.
[2] Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications[J]. Machine Learning with Applications, 2020, 2: 100006.
[3] Dokukina I, Gumanova J. The rise of chatbots–new personal assistants in foreign language learning[J]. Procedia Computer Science, 2020, 169: 542-546.
[4] Ji Z, Lu Z, Li H. An information retrieval approach to short text conversation[J]. arXiv preprint arXiv:1408.6988, 2014.
[5] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation[J]. arXiv preprint arXiv:1503.02364, 2015.
[6] Vinyals O, Le Q. A neural conversational model[J]. arXiv preprint arXiv:1506.05869, 2015.
[7] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences[J]. Advances in neural information processing systems, 2014, 27.

DOI: 10.54254/2755-2721/8/20230288

[8] Sedoc J, Ippolito D, Kirubarajan A, et al. Chateval: A tool for chatbot evaluation[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations). 2019: 60-65.
[9] Vijayaraghavan V, Cooper J B. Algorithm inspection for chatbot performance evaluation[J]. Procedia Computer Science, 2020, 171: 2267-2274.
[10] Ritter A, Cherry C, Dolan B. Data-driven response generation in social media[C]//Empirical Methods in Natural Language Processing (EMNLP). 2011.