A Survey on Intelligent Question and Answer Systems

Xuechao Guo, Bin Zhao, and Bo Ning(✉)

School of Information Science and Technology, Dalian Maritime University, Dalian, China
{xuechao,zhaobin,ningbo}@dlmu.edu.cn

Abstract. With the rapid development of technology in society, people are surrounded by all kinds of data in the information age. So the means to access external information accurately and quickly have become particularly important. Currently, intelligent question and answer systems are a promising area of research in the field of artificial intelligence and natural language processing. As an interactive system, it significantly differs from traditional search engines. When people perform advanced information retrieval, it can accurately understand the natural language questions asked by the user and give the user a corresponding answer using natural language, which meets people's needs for quickly, easily and accurately accessing information. Intelligent question and answer systems are already widely used in people's daily life, such as common intelligent voice interaction, online customer service, knowledge acquisition, emotional chat, etc. Deep learning, a branch of machine learning, is now widely used for various tasks in natural language processing. This paper focuses on the applications of deep learning to intelligent question and answer systems.

Keywords: Intelligent question and answer systems · Information retrieval · Natural language processing · Deep learning

1 Introduction

Traditional search engine systems have many drawbacks, such as a lack of accuracy in the representation of requirements when searching. The search needs of users are often very complex and specific, which cannot be expressed by a simple logical combination of a few keywords. Traditional search engines are not concise enough and return too many results making it extremely difficult for users to locate the information they need quickly and accurately. Lacking the support of semantic processing technology, traditional keyword-based information retrieval remains on the surface of language without containing semantics, making the retrieval effect mediocre. Intelligent question and answer systems based on natural language processing technology are one of the directions in which traditional

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 81–88, 2022.
https://doi.org/10.1007/978-3-031-23902-1_7

search engines are being improved. The query information is used to parse out the user's query intent, then the location of the answer is pinpointed from the document based on the intent, and finally the answer is extracted and returned to the user. Rather than just returning a documented distribution of answers to questions to the user, which is a great improvement both in terms of accuracy and meeting the user's search needs. Deep learning [10], a branch of machine learning, is a class of algorithms that attempt to perform high-level abstractions of data using multi-processing layer computational models that contain complex structures or consist of multiple non-linear transformations. It gradually transforms the initial low-level feature representation into a high-level feature representation through multi-layer processing, and allows complex learning tasks such as classification to be performed with simple models. Deep learning techniques have been used in a wide range of fields such as image recognition, speech recognition and natural language processing. A number of scholars have already applied deep learning to intelligent question and answer systems.

Classified according to the type of domain, intelligent question and answer systems can be divided into limited domain-oriented question and answer systems and open domain oriented question and answer systems. Depending on the implementation method, domain-limited QA systems are divided into Pipeline QA systems and end-to-end QA systems. Open domain QA systems are divided into search-based QA systems, generative-based QA systems and hybrid-based QA systems. The following article will focus on each of these types of QA systems.

2 Qualified Domain Oriented QA Systems

Domain-oriented QA systems, also known as task-based QA systems, are oriented towards vertical domains and aim to help users complete predetermined tasks or actions using as few dialogue rounds as possible. For example, booking tickets, restaurants, hotels, etc. The current mainstream research approach is to study task-based QA systems in three modules: Natural Language Understanding (NLU), Dialogue Management (DM) and Natural Language Generation (NLG) [2]. There are two broad approaches to the study of task-based QA systems, one is the Pipeline task-based QA systems and the other is the end-to-end task-based QA systems.

2.1 Pipelined Task-Based QA Systems

Pipelined task-based QA systems including Natural Language Understanding (NLU) modules, where Kim [8] and Lee [11], et al. use the user's dialogue to perform intention analysis and translate it into pre-defined semantic slots. Next comes the Dialogue State Tracking (DST) module, Williams et al. [26] assess the status of each round of dialogue based on information about the current dialogue and the history of the dialogue. There follows the Policy Learning (POL) module, it will make the next reaction based on the current state of

the conversation, i.e. it determines the action of the system. And finally, there is the Natural Language Generation (NLG) module. Wen et al. [25] convert the responses made by the policy learning module into the corresponding natural language responses provided to the user. The Dialogue State Tracking module and the Dialogue Policy Learning module make up the Dialogue Manager (DM), which is the core controller of the task-based QA systems. In some situations, several of these modules can be used in combination, depending on the needs of the task. Budzianowski et al. [1] combine NLU and DST modules and map the user's historical conversation information to the corresponding conversation states. Chen et al. [3] combine POL and NLG modules and map the user's historical discourse and conversation state to the system response. The first joint modelling of intention recognition and slot filling tasks using a GRU-based approach by Zhang et al. [31] After the GRU encodes the sentence the representation of the sentence is obtained by the max- pooling layer for intent recognition. Finally the joint learning of the two tasks is performed through a shared GRU layer to obtain an implicit relationship between them. Liu et al. [15] first used a sequence-to-sequence fusion attention mechanism approach to jointly model intention recognition and slot-filling tasks. Goo et al. [6] first used the slot-gate mechanism to explicitly focus on learning the relationship between slot-filling and intention recognition tasks and to obtain better semantic information through global optimization. Li et al. [12] use the Gate mechanism to use information to guide the slot-filling task, explicitly exploiting information about intentions. They explored the role of self-attention mechanisms on this task for the first time, which achieved good performance.

2.2 End-to-End Task-Based QA Systems

Although pipelined task-based QA systems can achieve good performance on individual tasks with a combination of modules, this modular system has significant drawbacks in deep learning training. When using a multi-step, multi-model to solve a complex task, inconsistency in the training objectives of each module makes it difficult to achieve optimal performance of the trained model. The end-to-end model uses only one model, one objective function and uses back propagation to optimise the parameters, the inherent pitfalls of multiple models are avoided. Recurrent neural networks are mostly used as encoder and decoder modules in end-to-end task-based dialogue systems. Madotto et al. [17] used an end-to-end memory network (MemNN) to address the problem of unstable performance in long sequences of recurrent neural networks as well as high temporal computational overhead. It uses several embedding matrices as external memory and reads the memory repeatedly using query vectors, it is therefore more suitable for storing information from external knowledge bases in task-based conversations. Eric et al. [5] propose a complete end-to-end model where they model conversation context and conversation generation based on an existing sequence-to-sequence architecture. They added an attention-based key-value pair mechanism to the retrieval of knowledge base entries, addressing the issue of how the model could be more smoothly interfaced with the knowledge base. Wu

et al. [27] fused memory networks and copy mechanisms to more smoothly embed the knowledge base during sequence generation, allowing for better integration of the knowledge base into task-based dialogue systems. Luo et al. [16] proposed the Profile Model and Preference Model based on the end-to-end memory network model. The former learns personalisation by embedding user profiles and uses global memory to store the conversation context of similar users. The latter is done by establishing a link between the portrait and the knowledge base, thus the best result among the candidate answers can be selected.

3 Open Domain Oriented QA Systems

The Open Domain QA systems differs significantly from the Limited Domain QA system in that it does not target domain-specific questions and complete any task. Instead, they are data-driven [20] and use natural language to mimic human discourse for everyday interactions. Three broad categories of open domain QA systems have been studied, including generative-based QA systems, retrieval-based QA systems and hybrid-based QA systems. The widespread use of deep learning techniques is now leading to breakthroughs in open domain QA systems.

3.1 Retrieval-Based QA Systems

Retrieval-based QA systems usually begin with the construction of a large corpus that can be retrieved by the system. The system identifies the responses from the conversation corpus that most closely resemble the input utterance. For each input statement, the retrieval model selects the statement with the greatest semantic match from the candidate statements as its response. In recent years, some models have taken into account not only current conversations but also rich historical conversations in the selection of responses. The core modules of a retrieval-based QA systems include a candidate answer retrieval module, a question-answer similarity calculation module and a ranking module. The overall step can be considered as a classification prediction of candidate responses. In another way, the question and the candidate responses are fed into a neural network model, and all candidate responses are classified or ranked.

The calculation of question-answer similarity is a key aspect of retrieval-based QA systems. Traditional text similarity calculation methods are based on surface text similarity calculation. The main method is to directly target the unprocessed text and use the degree of match and distance of the characters or strings in it as a criterion of similarity. This method is simple to implement and easy to understand. Kondrak et al. [9] used an n-gram model to calculate the similarity between characters, the main idea being to calculate the ratio of the number of identical N tuples to the total number of N tuples for two texts. Surface text similarity calculations do not take into account the semantic relationship of the text very well. In response to this problem, semantic similarity calculation methods have been proposed, of which neural network-based methods are currently the mainstream research direction. Mikolov et al. [18] proposed

the word vector embedding (Word2Vec) approach, which includes the CBOW model that uses context to predict central words and the Skip-gram model that uses central words to predict context. It uses contextual information to transform words into high-dimensional vectors. Shen et al. [22] based on convolutional neural networks to obtain a semantic representation vector of the user's query and the word vector of candidate answers with its fixed length by convolution and pooling operations, and then use the cosine similarity function to measure how well the queries and responses match. The task of simulating image recognition by Pang et al. [14] The matrix is first constructed based on the similarity between word vectors, then the fused information is gradually captured using convolutional and pooling layers, and finally the matching scores are calculated for answer ranking. Kim et al. [7] propose a DRCN model based on the generic framework of DenseRNN, which contains a multilayer recurrent neural network and an attention mechanism. The input to each layer of the neural network is a fusion of the outputs of all previous layers, and the output vector is used as input directly by the pooling layer before semantic fusion, thus alleviating the problem of gradient disappearance caused by the increase in the number of layers of the model. With the creation of the Transformer [24] model, more and more pre-trained models are emerging and performing well on various tasks in natural language processing. For example, the Bidirectional Encoder Representation from Transformers (BERT) [4] model addresses the case where the word2vec model is unable to determine multiple meanings of a word. The Bert model is trained using a large-scale unlabeled corpus, and the text obtained contains rich semantic information.

3.2 Generative-Based QA Systems

The main goal of a generative QA system is to generate responses based on the contextual information of the current conversation. This is usually done by using a deep learning based encoder-decoder architecture. Deep learning-based techniques do not usually rely on specific answer banks or templates, but rather on linguistic competence acquired from a large corpus. Specifically, a recurrent neural network is used to encode the input utterance as a vector representation, while another recurrent neural network is used at the decoding end and an attention mechanism is used to generate the replies one by one.

People do not simply ask one question and answer another in general chat. Answers often refer to the content of contextual chat messages, so contextual information should be introduced into the encoder, which helps the encoder to generate better session response content. Sordoni et al. [23] proposed the use of multilayer feedforward neural networks instead of recurrent neural networks in the encoder part. In this way, both the contextual information and the current conversation to be encoded by a multi-layer feed-forward neural network to generate an intermediate semantic representation of the encoder-decoder architecture, while avoiding the problem of recurrent neural networks being sensitive to excessively long inputs. Li et al. [13] used the idea of adversarial learning to train both a response generator and a response discriminator. Generators

are sequence-to-sequence based generative dialogue models. The discriminator is a classification model that completes a binary classification task which classifies the generated responses into two categories, human responses and machine responses, and is used to assess the quality of the responses. The core idea of the model is to motivate generators that can generate discourse in place of human replies. Xu et al. [29] used generative adversarial networks to the task of dialogue generation. They propose to replace the sampled decoding results in the decoder with an approximate embedding layer. It is an end-to-end model overall, allowing discriminators and generators to be trained with simultaneous parameter tuning by back-propagation. Shao et al. [21] connect the decoder head to tail and then use the generated part as part of an attention mechanism with further additions to the existing informativeness for generating long replies with high information content. Wu et al. [28] improved the word mapping at decoding and proposed a sequence-to-sequence model based on a dynamically decoded lexicon. The model makes it possible to have a different lexicon for each decoding step depending on the actual current conversation, in order to remove the interference of irrelevant words, narrow the mapping and speed up decoding.

3.3 Hybrid-Based QA Systems

QA systems based on retrieval-based approaches often respond to discourses that are limited by constructed corpus, and sometimes the retrieved content does not fit well with the contextual content. QA systems based on generative approaches may generate discourse that is more generic, or even output answers that are not relevant to the user's input question. The hybrid-based QA systems integrates both retrieval and generative approaches, providing a clever blend of the two. Qiu et al. [19] generated responses using a sequence-to-sequence model and then scored the responses obtained by the retrieval method using the rerank model. If the score is below the initially set threshold, the response generated by the seq2seq model is used directly as the answer to the user, otherwise the response obtained by retrieval is used as the answer to the user. Liu et al. [30] built a neural network dialogue model that mixes retrieval-based ranking models and generative models, and combines the advantages of retrieval-based and generative QA systems, providing new insights into how to integrate retrieval and generative models to build QA systems.

4 Conclusion

Current intelligent question and answer systems have significant shortcomings in terms of semantics, consistency and interactivity. In terms of semantics, deep learning-based generative models are more likely to generate meaningless universal replies, such as 'I don't know'. The amount of information, appropriateness and logic of the generated content is still inadequate, and there is a long way to go before semantic understanding in the true sense of the word. In terms of consistency, it is easy to create semantic identity and personality conflicts in the

interaction. In terms of interactivity, current intelligent QA systems have significant shortcomings in terms of emotional interaction and strategic response. It is not able to adaptively adjust its own strategies to the user's topic status, such as topic strategies, active-passive strategies and emotional expression strategies, and it also makes it impossible to achieve smooth and natural human-computer interaction. With the advancement of technology and media publicity, people are more inclined to think of an intelligent QA system as their life companion rather than a machine that can only be used to perform a task. To meet this expectation of users, the future intelligent QA systems will need to have some emotional intelligence. They can carry out emotional response and interaction, and reflect personality, language style and personality in dialogue and interaction.

References

1. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Gai, M.: MultiwOZ - a large-scale multi-domain wizard-of-OZ dataset for task-oriented dialogue modelling (2018)
2. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: recent advances and new frontiers. ACM SIGKDD Explor. Newslett. **19**(2), 25–35 (2017)
3. Chen, W., Chen, J., Qin, P., Yan, X., Wang, W.Y.: Semantically conditioned dialog response generation via hierarchical disentangled self-attention (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018)
5. Eric, M., Manning, C.D.: Key-value retrieval networks for task-oriented dialogue (2017)
6. Goo, C.W., Gao, G., Hsu, Y.K., Huo, C.L., Chen, Y.N.: Slot-gated modeling for joint slot filling and intent prediction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (2018)
7. Kim, S., Kang, I., Kwak, N.: Semantic sentence matching with densely-connected recurrent and co-attentive information (2018)
8. Kim, Y.B., Lee, S., Stratos, K.: OneNet: joint domain, intent, slot prediction for spoken language understanding. Amazon Alexa Brain, Seattle, WA; Microsoft Research, Redmond, WA; Toyota Technological Institute, Chicago, IL
9. Kondrak, G.: *N*-Gram similarity and distance. In: Consens, M., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 115–126. Springer, Heidelberg (2005). https://doi.org/10.1007/11575832_13
10. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
11. Lee, S., et al.: ConvLab: multi-domain end-to-end dialog system platform (2019)
12. Li, C., Li, L., Qi, J.: A self-attentive model with gate mechanism for spoken language understanding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3824–3833 (2018)
13. Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D.: Adversarial learning for neural dialogue generation (2017)
14. Liang, P., Lan, Y., Guo, J., Xu, J., Cheng, X.: Text matching as image recognition (2016)
15. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling (2016)

16. Luo, L., Huang, W., Qi, Z., Nie, Z., Xu, S.: Learning personalized end-to-end goal-oriented dialog (2018)
17. Madotto, A., Wu, C.S., Fung, P.: Mem2Seq: effectively incorporating knowledge bases into end-to-end task-oriented dialog systems (2018)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Comput. Sci. (2013)
19. Qiu, M., et al.: AliMe chat: a sequence to sequence and Rerank based chatbot engine. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 498–503. Association for Computational Linguistics, Vancouver, July 2017. https://doi.org/10.18653/v1/P17-2079. https://aclanthology.org/P17-2079
20. Ritter, A., Cherry, C., Dolan, B.: Data-driven response generation in social media. In: Empirical Methods in Natural Language Processing (EMNLP), January 2011. https://www.microsoft.com/en-us/research/publication/data-driven-response-generation-in-social-media/
21. Shao, L., Gouws, S., Britz, D., Goldie, A., Strope, B., Kurzweil, R.: Generating high-quality and informative conversation responses with sequence-to-sequence models (2017)
22. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural network for web search. In: Proceedings WWW, pp. 373–374 (2014)
23. Sordoni, A., et al.: A neural network approach to context-sensitive generation of conversational responses. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May-Jun 2015
24. Vaswani, A., et al.: Attention is all you need. arXiv (2017)
25. Wen, T.H., Gasic, M., Mrksic, N., Su, P.H., Vandyke, D., Young, S.: Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. Comput. Sci. (2015)
26. Williams, J., Raux, A., Ramachandran, D., Black, A.: The dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 Conference (2013)
27. Wu, C.S., Socher, R., Xiong, C.: Global-to-local memory pointer networks for task-oriented dialogue (2019)
28. Wu, Y., Wu, W., Yang, D., Xu, C., Li, Z., Zhou, M.: Neural response generation with dynamic vocabularies (2017)
29. Xu, Z., et al.: Neural response generation via GAN with an approximate embedding layer. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 617–626 (2017)
30. Yang, L., et al.: A hybrid retrieval-generation neural conversation model. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (2019)
31. Zhang, X., Wang, H.: A joint model of intent determination and slot filling for spoken language understanding (2016)