ICEB 2021 Proceedings (Nanjing, China)

International Conference on Electronic Business (ICEB)

Winter 12-3-2021

A Survey of Customer Service System Based on Learning

Sihan Yin

Xudong Luo

Yin, S. & Luo, X. (2021). A survey of customer service system based on learning. In *Proceedings of The 21st International Conference on Electronic Business* (pp. 34-41). ICEB'21, Nanjing, China, December 3-7, 2021.

A Survey of Customer Service System Based on Learning

Sihan Yin [1,2]
Xudong Luo [1,3,*]

*Corresponding author
[1] Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin, China,
[3] Student, 1469214793@qq.com
[2] Professor, luoxd@mailbox.gxnu.edu.cn

ABSTRACT

With the rapid development of artificial intelligence, people have moved from manual customer service to handling affairs, and now they are more inclined to use intelligent customer service systems. The intelligent customer service system is generally a chat robot based on natural language processing, and it is a dialogue system. Therefore, it plays a vital role in many fields, especially in the field of e-commerce. In this article, to help researchers further study the customer service system for e-commerce, we survey the learning-based methods in dialogue understanding, dialogue management and dialogue response generation in the customer service system. In particular, we compare the advantages and disadvantages of these methods and pointed out further research directions.

Keywords: E-commerce, customer service system, learning, dialogue understanding, dialogue management, dialogue response generation.

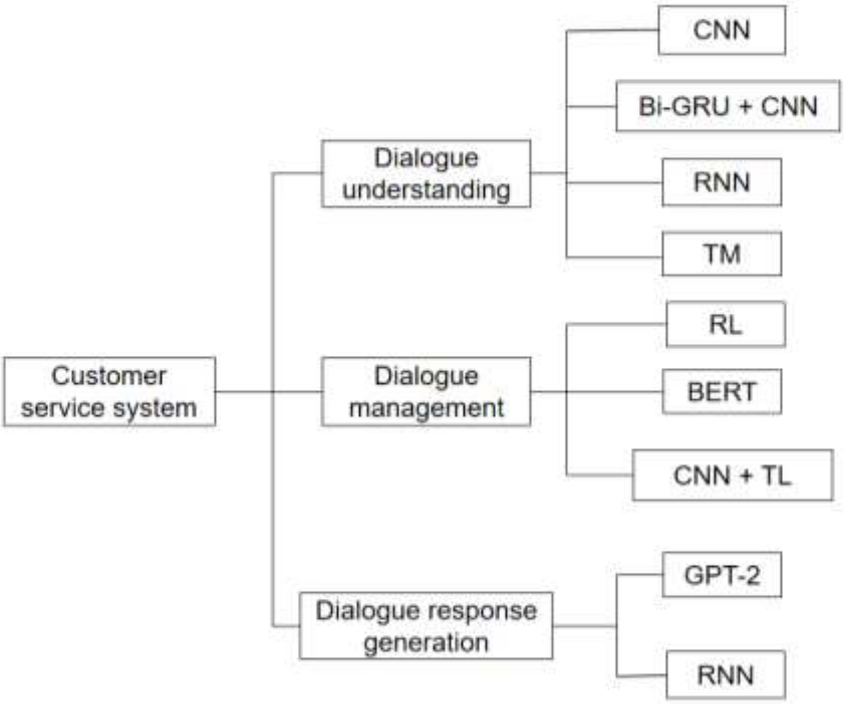
INTRODUCTION

With the rapid development of artificial intelligence and e-commerce, customer service is essential in real life. Customer service refers to manually or non-manually providing customers with pre-sales and after-sales support to help them get a relaxing and pleasant experience. Nowadays, customer service is performed through the phone and e-mail, the Internet, Short Message Service (SMS), social media, and even artificial intelligence technology. If customer service is represented by a computer system or realized by Natural Language Processing (NLP) technology, we call it a customer service system. In the past ten years, many companies have developed self-service intelligent customer service systems, such as AliMe from Alibaba (Li *et al.*, 2017), Google Now from Google (Ehrenbrink, Osman & Möller, 2017), and CarPlay from Apple (Holstein *et al.*, 2015), etc. Customers can talk to them at any time to find the answers they want.

In the field of e-commerce, traditional customer service is primarily manual. In a period of the rapid increase in traffic, such as Taobao's Double Eleven shopping carnival, it will overload the call centre, and manual customer service cannot handle customers quickly and online 24 hours a day. It is also impossible to fully understand the customer's information due to the concurrent needs of the customer, which will reduce the efficiency of customer service to solve the problem, resulting in a long waiting time for customers. Finally, customers are disappointed and dissatisfied with the interactive experience. Research shows that nearly 75% of customers have experienced poor customer service (Siddiqui & Sharma, 2010). Therefore, driven by big data and artificial intelligence technology, merchants are increasingly inclined to use intelligent customer service systems instead of manual customer service to improve their competitiveness in the e-commerce market. The essence of customer service systems is a dialogue system.

A dialogue system mainly consists of three comments: dialogue understanding, dialogue management, and dialogue response generation (Fan, Luo & Lin, 2020a; Fan, Luo & Lin, 2020b; Fan & Luo, 2020). Dialogue understanding aims to convert textual information into a semantic representation that a computer can process. For computers, it is not essential to understand the exact meaning of each word in a sentence. Instead, it is necessary to understand the importance of the sentence. Dialogue management is the core part of the dialogue system. Its primary function is to maintain and update the dialogue's state and give the next dialogue strategy based on the current dialogue state and the knowledge of the knowledge base. Dialogue response management is an integral part of interaction with users. It uses a generative model to generate responses based on the dialogue strategy given by dialogue management to facilitate user interaction. A good generative model can generate sentences suitable for new scenarios by using a small amount of training corpus to provide users with accurate information (Mi *et al.*, 2019).

The existing surveys on the customer service system (Nuruzzaman & Hussain, 2018; Arora, Batra & Singh, 2013; Wang & Yuan, 2016; Chen *et al.*, 2017) are not comprehensive enough, do not involve the latest work in recent years, or are outdated. Therefore, to facilitate people to further study the customer service system, in this paper, we focus on its three essential components: dialogue understanding, dialogue management, and dialogue response generation. Specifically, we will survey the most advanced learning-based methods concentrate on these three aspects. As shown in Figure 1, these methods are Convolutional Neural Network (CNN) (LeCun *et al.*, 1989), Bi- Gate Recurrent Unit (Bi- GRU) (Cho *et al.*, 2014) + CNN, Recurrent Neural Network (Mikolov *et al.*, 2010), Topic Model (Papadimitriou *et al.*, 2000), Reinforcement Learning (RL) (Kaelbling, Littman & Moore, 1996), BERT (Devlin *et al.*, 2018), CNN + Transfer Learning (TL) (Weiss, Khoshgoftaar & Wang) and GPT-2 (Radford *et al.*, 2019). In particular, we compare their pros and cons and point out possible future jobs. We hope that the survey in this paper can provide some clues for customer service developers to choose the proper method for their specific task.



Source: This study

Figure 1: A taxonomy of learning methods used for the customer service system

We organise the rest of this paper as follows. The second section briefs some learning-based methods for dialogue understanding. The third section recaps some learning-based methods for dialogue management. The fourth section reviews some learning-based methods for dialogue response generation. The fifth section discusses this paper. Finally, the sixth section summarises this paper.

DIALOGUE UNDERSTANDING

Intent recogisation and slot filling

CNN-based method

To reduce the working load of manual customer service, Li *et al.* (2017) develops an intelligent customer service system, named AliMe, based on CNN for Alibaba the e-commerce platform. It can answer customers' simple, repetitive, and common questions, thus it saves labour costs of an online shopping platform and improves the customers' online shopping experience. So, AliMe is a Question Answering (Q & A) system. To answer the customer's question well, it must recognise the customer's intents through context. AliMe uses CNN for intent recognition of a piece of customer's dialogue (Kim, 2014), and uses fastText (Bojanowski *et al.*, 2017) in the embedding layer of CNN to pre-train the data, and then further fine-tune the training results in the CNN model.

The overall accuracy of AliMe has reached 89.91%, outperforming an SVM based model and an ME based model. It is 7.21% higher than the combination of the SVM model and the ME model.

In the future, it is worth using RL to guide shopping, improve people's online shopping experience, or give AliME the ability to recognize images, and improve the multi-round interaction ability of AliMe to improve human-computer dialogue.

Bi-GRU + CNN-based method

An essential task of customer service is to perform semantic analysis, including two subtasks: intention recognition and slot filling. To reduce the accumulation of errors caused by the independent modelling of the two subtasks, Wu, Mao, and Feng (2021) propose a model, which integrates deep learning with cloud computing and jointly model CNN, Bi-GRU and CRF to improve the performance of semantic analysis and deploy it in the intelligent dialogue system. First, using two Bi-GRU models to obtain intent classification and slot filling's features, respectively. Then their work uses a CNN model for intent classification, and finally, a CRF model is used to model the slot filling's sequence information.

The experiment uses the ATIS dataset and the PCSF dataset. The first experiment set a different structure for the model on the ATIS data set. After performing ablation experiments, the experimental results show that the accuracy of the proposed model is the best, reaching 89.6%. Then, they carry out the second experiment on the PCSF data set, and other conditions remain unchanged. Again, the experimental results show that the accuracy of the proposed model is the best, reaching 85.6%. Finally, they compare the proposed model with other advanced baseline models on two data sets, and the experimental results show that the proposed model is very effective.

In the future, they will improve the proposed network to process complex sentences to obtain a better user experience. They will also develop a lighter version for the low resource platform so that the model can better serve the common resource platform.

Sentiment detection

RNN-based method

To track the speaker's status throughout the dialogue process and use this information for sentiment detection, Majumder *et al.* (2019) propose a novel RNN-based model called DialogueRNN. Their model assumes three main aspects related to emotion in the dialogue: the speaker, the context of the previous sentence, and the feeling of the last few sentences. For these three states, DialogueRNN first uses Global GRU to capture the context of a given utterance through dialogue and speaker status, and the speaker's status depends on this context. Then, since the dialogue process is mutual and dynamic, DialogueRNN uses Speaker GRU to update the speaker's emotional state and uses Party GRU to ensure that the model knows the state of the speaker of each utterance. Finally, the updated speaker state is input into the emotion GRU, and the word's emotion representation is decoded, thereby completing sentiment detection.

During the experiment, they use two sentiment detection data sets: IEMOCAP (Busso *et al.*, 2008) and AVEC (Schuller *et al.*, 2012), and divide them into a training set and test set according to the ratio of 8:2. The experiment also compares DialogueRNN with the latest baseline model, using accuracy, F1 score and mean absolute error as evaluation indicators to evaluate the model's performance. Experimental results show that the proposed model is superior to the current state-of-the-art baseline model on both data sets. In addition, the proposed model achieves the best average accuracy values, F1 score and mean absolute error, reaching 63.4%, 62.75%, and 2.102, respectively.

Although their work has achieved the best performance, the proposed model can only perform sentiment detection for a single speaker. In future work, we can extend the current work to perform sentiment detection for multiple speakers.

TM-based method

The main challenge of sentiment detection in a dialogue system is to classify sentiment types. To obtain comprehensive information from certain types of conversations to improve the performance of sentiment classification, Wang *et al.* (2020) propose a novel Topic-aware Multi-task Learning (TML) based on TM. TML uses three topic models to capture the whole, customer, and agent's topic information to learn the rich discourse representations in customer service conversations. After combining with the context-aware utterance representation obtained after BERT pre-training and context iterative modelling, the sentiment label is finally obtained after gated fusion.

The conversation data set used in the experiment comes from an online customer service system of a top e-commerce company in China, and the experiment uses the Macro-F1 score to evaluate the model's overall performance. The experiment also compared the proposed model with other state-of-the-art baselines. The experimental results show that the proposed TML method is significantly better than the baseline model in terms of performance, with a Macro-F1 score of 75.9%.

It is worth studying to deal with the imbalance problem of emotion classification in customer service dialogue in future work. There are much fewer negative samples in the dialogue dataset than other samples in their work, but in reality, it is sometimes more important to identify negative samples.

DIALOGUE MANAGEMENT

Dialogue strategy

RL-based method

To support negotiation between a seller and a buyer, Bagga *et al.* (2020) develop such a system, called ANEGMA, based on deep reinforcement learning. This model can enable agents to conduct bilateral negotiations in an unknown and dynamic electronic market and use multi-thread synchronization to reduce time costs and increase the speed of model operation by using reinforcement learning further to express the strategy of the deep neural network and pre-training the strategy to reduce the learning time in the negotiation process.

Their experiments show that ANEGMA outperforms those with two well-known negotiation strategies in one-to-many contemporaneous bilateral negotiations in different market environments. For example, for the overall performance of the model, in the first experiment of training and testing on conceder time-dependent seller strategy, ANEGMA integrates RL with Supervised Learning (SL) achieves the best performance on U_{avg} and S%, reaching 0.29 and 87.12, respectively, but on T_{avg}, the performance of ANEGMA+RL+SL is not so good as that of ANEGMA+RL (67.75s/0.77s). In the second experiment of training and testing on relative tit for tat behaviour dependent seller strategy, ANEGMA+RL+SL achieves the best performance on U_{avg} and S%, reaching 0.29 and 85.15, respectively, but on T_{avg}, the performance of ANEGMA+RL+SL is not so good as that of ANEGMA+RL (36.33s/0.76s). For the adaptive performance of the model, in the first experiment of training on relative tit for tat behaviour dependent seller strategy and testing on conceder time-dependent seller strategy, ANEGMA+RL+SL achieves the best performance on U_{avg} and S%, reaching 0.26 and 86.13, respectively. Still, on T_{avg}, The performance of ANEGMA+RL+SL is not so good as that of ANEGMA+RL (38.40s/0.74s). In the second experiment of training on conceder time-dependent seller strategy and testing on relative tit for tat behaviour dependent seller strategy, ANEGMA+RL+SL achieves the best performance on U_{avg} and S%, reaching 0.28 and 84.16, respectively, but on T_{avg}, the performance of ANEGMA+RL+SL is not so good as that of ANEGMA+RL (19.30s/0.81s).

In the future, it is worth improving ANEGMA in some aspects because the actual market environment is even more complicated than those where ANEGMA works well. So, sellers and buyers are more inclined to do multi-issue negotiation using dynamic strategy in a real market.

BERT-based method

Due to the limitations of the shallow connection of the pre-trained language model, for example, it cannot adapt to dynamic speaking scenarios, weakens the relationship between contexts, and cannot handle sequences composed of a large number of tags. So, to apply BERT to the task of multi-round response selection in e-commerce, Gu *et al.* (2020) propose a novel model called Speaker-Aware BERT (SA-BERT):

1) They add speaker embeddings to specific tags to enhance the model so that Speaker-Aware BERT knows the speaker changes during the conversation.
2) Since BERT is not good at processing sequences composed of tokens exceeding the limit, they propose a heuristic speaker-perceived disentanglement strategy to solve complex entangled conversations, which helps select a few but the most critical utterances based on speaker information.
3) They design domain adaptations to integrate specific domain knowledge into pre-trained language models.

The experiment thoroughly tested the performance of SA-BERT on Ubuntu Dialogue Corpus V1 (Lowe *et al.*, 2015), Ubuntu Dialogue Corpus V2 (Lowe *et al.*, 2017), Douban Conversation Corpus (Wu *et al.*, 2017), E-commerce Dialogue Corpus (Zhang *et al.*, 2018) and DSTC 8-Track 2-Subtask 2 Corpus (Kim *et al.*, 2019), and used recall, mean average precision, mean reciprocal rank, and precision as evaluation indicators. The experiment compares the performance of SA-BERT and the baseline model on the first four data sets. The experimental results show that SA-BERT has the best performance on the four data sets, with average recall reaching 93.28%, 92.43%, 54.70%, and 85.60%, respectively. Furthermore, SA-BERT has the best mean average precision, mean reciprocal rank and precision in Douban Conversation Corpus, reaching 61.90%, 0.659 and 49.60%, respectively. Finally, SA-BERT is used to evaluate the performance of the proposed strategy in DSTC 8-Track 2-Subtask 2 Corpus. The experimental results show that SA-BERT has the best mean reciprocal rank and recall, reaching 0.594 and 47.70%, respectively.

Although their work has achieved the best results on the five data sets, in a horizontal comparison, the proposed model results on Douban Conversation Corpus are much worse than that of Ubuntu Corpus V1 and Ubuntu Corpus V2. Therefore, we can continue to adjust the pre-training language model to adapt to multiple rounds of response selection and design new strategies to achieve better results in future work.

Dialogue behavior recognition

CNN + TL-based method

To match the best answer in the question-answering knowledge base in the Q & A system, and solve the lack of training data and achieve high Queries Per Second (QPS), Yu *et al.* (2018) propose a CNN + TL model. Researchers roughly divide the CNN-based models that solve Paraphrase Identification (PI) and Natural Language Inference (NLI) problems into two categories, sentence-encoding-based method and sentence interaction-based method. In Yu *et al.* work, they combine the two approaches to capture sentence representation and interaction structure. Make the model reach high QPS to meet industrial needs. At the same time, the TL model is used to obtain labelled data from a resource-rich source domain to a resource-poor target domain and reveals the inter-domain and intra-domain relations. Furthermore, to improve the anti-noise ability of the proposed TL framework, they introduce the positive semi-defined covariance matrix and the adaptive loss function to model the information weight between and within domains.

Their experiments show that compared with other baseline models, the CNN model achieves the best performance in terms of both area under curve and accuracy (*i.e.*, 82.2% and 81.2%, respectively). Moreover, the DRSS model proposed based on the TL method achieves the best performance in both the PI and NLI tasks. So, the performance of the overall model is more efficient than several baseline models, and its performance on six source-target pairs is also better than all existing frameworks.
In addition, the framework has been deployed to the online chat system of AliExpress, a cross-border e-commerce platform, significantly improving its performance.

In the future, it is worth improving the model's performance and the transferring ability or learning cross-language chatbots to achieve multilingual services.

DIALOGUE RESPONSE GENERATION

GPT-2-based method

The pre-trained model can generate smooth dialogue responses, but there is still the problem of generating blanks. Existing models (Lian *et al.*, 2019; Kim, Ahn, and Kim, 2020; Roller *et al.*, 2021; Chen *et al.*, 2020; Zhao *et al.*, 2020) generate richer responses by retrieving relevant knowledge from a large corpus and adding dialogue history, but these two methods are time-consuming and labour-intensive. To reduce the retrieval steps and only rely on the generative model to generate dialogue responses, Xu *et al.* (2021) propose a new end-to-end retrieval-free framework based on GPT-2, called KnowExpert. Their work is to integrate the prior knowledge obtained from unstructured data into the pre-training model with lightweight adapters (Bapna & Firat, 2019) for the first time to get a dialogue response generation model. This method can avoid retrieving the knowledge base, speed up the reasoning time, improve the memory utilization efficiency and significantly improve the generation efficiency of the model.

The experiment uses Wizzard of Wikipedia (WoW) (Dinan *et al.*, 2018) and CMU Document Grounded Conversations (CMU_DoG) (Zhou, Prabhumoye & Black, 2018) as data sets, compares the proposed model with the baseline model and uses perplexity and F1 score as evaluation indicators to reflect the performance of the model. Experimental results show that on the visible WoW data set, the proposed model has the best perplexity and F1 score, reaching 15.3 and 18.8%. On the other hand, on the invisible WoW data set, the performance of the proposed model is worse than that of the visible WoW data set, reaching 21 and 16.6% in PPL and F1 score. Finally, on CMU_DoG, the model achieved 17.9 and 12.0% in perplexity and F1 score. In general, the performance of KnowExpert is comparable to the retrieval-based model, indicating that the non-retrieval dialogue generation model has a promising future.

It can be seen from the experimental results that the performance of the proposed model in the visible and invisible domains is quite different. Therefore, studying topic modelling to narrow the performance gap between the visible and invisible domains is worthwhile in future work.

RNN-based method

Nowadays, many researchers have widely used neural network methods in dialogue response generation, among which the more advanced model is Hierarchical Recurrent Encoder-Decoder (HRED) based on RNN. However, existing work rarely generates informative dialogue responses, so to solve this limitation, Wang *et al.* (2020) extend RNN-based HRED to capture more information content so that the model develops more informative dialogue responses. Their work has followed the three RNNs of HRED. Encoder RNN is mainly used to encode input sentences, and context RNN is used to encode information at the dialogue level, such as the state and intention of the entire dialogue. Finally, the decoder RNN is used for decoding. Moreover, they also expanded the scope of their attention to capturing more information content.

The experiment used two data sets, a preprocessed version of Ubuntu Dialogue Corpus (Serban *et al.*, 2017) and a multi-round dialogue data set collected from Sina Weibo, which from http://tcci.ccf.org.cn/conference/2018/dldoc/taskgline05.pdf. The experiment uses BLEU and embedding average as indicators to measure the dialogue response generated by the model and compares the proposed model with three baseline models. The experimental results show that the performance of the proposed model is the best on both datasets. The proposed model achieves the best BLEU and embedding average on the Ubuntu dataset, reaching 1.0276 and 0.5716, also in the Weibo dataset, reaching 1.2198 and 0.8260, indicating that the proposed model is effective.

Future work is worth exploring how to better incorporate external knowledge into technical-domain dialogue generation, such as the Ubuntu Troubleshoot scenario.

DISCCUSION

Table 1: Various methods based on customer service system.

Reference	citation	Method	Issue addressed	Advantage	Disadvantage
(Li *et al.*, 2017)	69	CNN	Intent recogisation and slot filling	Efficient	No explanation
(Wu, Mao & Feng, 2021)	0	Bi-GRU + CNN	Intent recogisation and slot filling	Reduce the accumulation of errors	Complicated and large amount of calculation
(Majumder *et al.*, 2019)	193	RNN	Sentiment detection	Provide better context	Single speaker
(Wang *et al.*, 2020)	6	TM	Sentiment detection	Get overall information to improve classification performance	Class imbalance
(Bagga *et al.*, 2020)	6	RL	Dialogue strategy	Efficient	Long exploration time
(Gu *et al.*, 2020)	36	BERT	Dialogue strategy	Domain adaptation, resolve entangled dialogue, adapt to dynamic environment	Few strategies to resolve entanglement dialogue
(Yu *et al.*, 2018)	72	CNN + TL	Dialogue behavior recognition	Focus on inter-domain and intra-domain relation	Not easy to converge
(Xu *et al.*, 2021)	3	GPT-2	Dialogue response generation	Bypass the retrieval process	The performance gap between the domains is large

(Wang *et al.*, 2020)	0	RNN	Dialogue response generation	Selectively focus on the more semantically important parts of the conversation	Not making better use of external knowledge

Source:This study.

Table 1 shows several commonly used methods for customer service systems. CNN is often used for text classification to help dialogue systems better perform intent recognition. The advantage of CNN is that it does not need to extract features, which saves feature extraction manually. However, the interpretability of CNN in text classification is poor, so it is not easy to adjust the elements according to the training results. The joint modelling of Bi-GRU and CNN can prevent the accumulation of errors caused by the two subtasks of independent modelling semantic understanding. However, joint modelling will also cause the model to be more complex and generate many calculations. Using RNN for sentiment detection can process each speaker's utterance and consider the speaker's characteristics, which can provide better context to improve the performance of the sentiment detection model. However, this model is currently only for one speaker and cannot be used in the dialogue of multiple speakers. Using TM for sentiment detection can get comprehensive information to improve sentiment classification performance, but this method can still not deal with the class imbalance in sentiment classification. RL is also efficient enough for dialogue systems because it does not require labelling samples, which reduces the costs for manual labelling. However, because RL takes too long to explore the best strategy, it is necessary to use deep learning to accelerate the RL model's convergence speed. The BERT-based method can understand the change information of the speaker, select a small amount of the actual utterances as the filtered context, and perform domain adaptation. However, the model's multi-turn response selection and entanglement strategy are too simple to handle more complex conversations. CNN + TL enables the model to focus on the relationship between source and target domains, rather than just learning the shared feature space between fields. However, the model parameters of TL are not easy to converge. The GPT-2 can bypass the retrieval step in the knowledge base and only rely on the generation model to generate the response, which significantly improves the model's efficiency. However, this method has a significant performance gap between the visible domain and the invisible domain. Finally, the RNN-based dialogue response generation can selectively focus on the more semantically essential parts of the dialogue to generate more reasonable responses. However, this method has not used external knowledge well into dialogue generation.

SUMMARY

More and more researchers are interested in customer service systems. One of the important reasons is that people's demand for customer service systems in e-commerce and daily life is increasing, and intelligent customer service systems also bring people's lives. Great convenience. In this paper, we survey the learning-based dialogue understanding, dialogue management and dialogue response generation methods. They are the convolutional neural network, Bi-gate recurrent unit + convolutional neural network, recurrent neural network, topic model, reinforcement learning, BERT, convolutional neural network + transfer learning and GPT-2.

In particular, many problems need to be solved in the future. For example, how to use RL to guide shopping, improve people's online shopping experience, give the model the ability to recognise images, and improve the multi-round interaction ability of the model to enhance human-computer dialogue and process complex sentences to obtain a better user experience. Or extend the current work to perform sentiment detection for multiple speakers, adapt to various rounds of response selection, design new strategies to achieve better results, etc.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61762016) and Graduate Student Innovation Project of School of Computer Science and Engineering, Guangxi Normal University (JXXYYJSCXXM-2021-001).

REFERENCES

Arora, S., Batra, K., & Singh, S. (2013). Dialogue system: A brief review. *arXiv Preprint arXiv:1306.4134.*

Baegga, P., Paoletti, N., Alrayes, B. & Stathis, K. (2020). A deep reinforcement learning approach to concurrent bilateral negotiation. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and 17th Pacific Rim International Conference on Artificial Intelligence* (pp. 297-303). IJCAI-PRICAI'20, Yokohama, Japan, January 7-15. https://doi.org/10.24963/ijcai.2020/42

Bapna, A., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1538-1548). Hong Kong, China. https://doi.org/10.18653/v1/D19-1165

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics,* 5, 135-146. https://doi.org/10.1162/tacl_a_00051

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation,* 42(4), 335-359. https://doi.org/10.1007/s10579-008-9076-6

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter,* 19(2), 25-35. doi: 10.1145/3166054.3166058

Chen, X., Meng, F., Li, P., Chen, F., Xu, S., Xu, B., & Zhou, J. (2020). Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3426-3437). https://doi.org/10.18653/v1/2020.emnlp-main.275

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv Preprint ArXiv:1406.1078.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805.*

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv Preprint ArXiv:1811.01241.*

Ehrenbrink, P., Osman, S., & Möller, S. (2017). Google now is for the extraverted, Cortana for the introverted: investigating the influence of personality on IPA preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction* (pp. 257-265). doi: 10.1145/3152771.3152799

Fan, Y., & Luo, X. (2020). A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 1202-1209). IEEE. https://doi.org/10.1109/ICTAI50040.2020.00182

Fan, Y., Luo, X., & Lin, P. (2020a). A Survey of Response Generation of Dialogue Systems. *International Journal of Computer and Information Engineering,* 14(12), 461-472.

Fan, Y., Luo, X., & Lin, P. (2020b). On Dialogue Systems Based on Deep Learning. *International Journal of Computer and Information Engineering,* 14(12), 508-516.

Gu, J. C., Li, T., Liu, Q., Ling, Z. H., Su, Z., Wei, S., & Zhu, X. (2020). Speaker-Aware Bert for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 2041-2044). CIKM’20, Ireland, October 19-23. https://doi.org/10.1145/3340531.3412330

Holstein, T., Wallmyr, M., Wietzke, J., & Land, R. (2015). Current challenges in compositing heterogeneous user interfaces for automotive purposes. In *International Conference on Human-Computer Interaction* (pp. 531-542). Springer, Cham. https://doi.org/10.1007/978-3-319-20916-6_49

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence research,* 4, 237-285. https://doi.org/10.1613/jair.301

Kim, B., Ahn, J., & Kim, G. (2020). Sequential latent knowledge selection for knowledge-grounded dialogue. *ArXiv Preprint ArXiv:2002.07510.*

Kim, S., Galley, M., Gunasekara, C., Lee, S., Atkinson, A., Peng, B., Schulz, H., Gao, J., Li, J., Adada, M., Huang, M., Lastras, L., Kummerfeld, J. K., Lasecki, W. S., Hori, C., Cherian, A., Marks, T. K., Rastogi, A., Zang, X., Sunkara, S. & Gupta, R. (2019). The eighth dialog system technology challenge. *ArXiv Preprint ArXiv:1911.06394*.

Kim,Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746-1751). EMNLP, Doha, Qatar, October 25-29. https://doi.org/10.3115/v1/D14-1181

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation,* 1(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541

Li, F. L., Qiu, M., Chen, H., Wang, X., Gao, X., Huang, J., Ren, J., Zhao, Z., Wang, L., Jin, G. & Chu, W. (2017). AliMe Assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2495-2498). CIKM’17, Singapore, November 6-10. 9 https://doi.org/10.1145/3132847.3133169

Lian, R., Xie, M., Wang, F., Peng, J., & Wu, H. (2019). Learning to Select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 5081-5087). IJCAI 2019, Macao, China, August 10-16.

Lowe, R., Pow, N., Serban, I. V., & Pineau, J. (2015). The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 285-294). SIGDIAL 2015, Prague, Czech Republic, September 2-4. https://doi.org/10.18653/v1/W15-4640

Lowe, R., Pow, N., Serban, I. V., Charlin, L., Liu, C. W., & Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse,* 8(1), 31-65.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 33(1), 6818-6825. https://doi.org/10.1609/aaai.v33i01.33016818

Mi, F., Huang, M., Zhang, J., & Faltings, B. (2019). Meta-learning for low-resource natural language generation in task-oriented dialogue systems. *ArXiv Preprint ArXiv:1905.05644.*

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech,* 2(3), 1045-1048.

Nuruzzaman, M., & Hussain, O. K. (2018). A survey on chatbot implementation in customer service industry through deep neural networks. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 54-61). IEEE. https://doi.org/10.1109/ICEBE.2018.00019

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences,* 61(2), 217-235. https://doi.org/10.1006/jcss.2000.1711

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog,* 1(8), 9.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y. L & Weston, J. (2021, April). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 300-325). https://doi.org/10.18653/v1/2021.eacl-main.24

Schuller, B., Valster, M., Eyben, F., Cowie, R., & Pantic, M. (2012). Avec 2012: The Continuous Audio/Visual Emotion Challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 449-456). ICMI'12, Santa Monica, California, USA, October 22-26. http://dx.doi.org/10.1145/2388676.2388776

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., & Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 31(1), 3295-3301.

Siddiqui, M. H., & Sharma, T. G. (2010). Analyzing customer satisfaction with service quality in life insurance services. *Journal of Targeting, Measurement And Analysis For Marketing,* 18(3), 221-238. https://doi.org/10.1057/jt.2010.17

Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., Zhang, M. & Zhou, G. (2020). Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34(5), 9177-9184. https://doi.org/10.1609/aaai.v34i05.6454

Wang, X., & Yuan, C. (2016). Recent advances on human-computer dialogue. *CAAI Transactions on Intelligence Technology,* 1(4), 303-312. https://doi.org/10.1016/j.trit.2016.12.004

Wang, Y., Rong, W., Zhou, S., Ouyang, Y., & Xiong, Z. (2020). Dynamic multi-level attention models for dialogue response generation. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 62-71). Springer, Cham.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data,* 3(1), 1-40. https://doi.org/10.1186/s40537-016-0043-6

Wu, Y., Mao, W., & Feng, J. (2021). AI for online customer service: intent recognition and slot filling based on deep learning technology. *Mobile Networks and Applications,* 1-13.

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2017). Sequential matching 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*, 1, 496-505. Vancouver, Canada, July 30-August 4. https://doi.org/10.18653/v1/P17-1046

Xu, Y., Ishii, E., Liu, Z., Winata, G. I., Su, D., Madotto, A., & Fung, P. (2021). Retrieval-free knowledge-grounded dialogue response generation with adapters. *ArXiv Preprint ArXiv:2105.06232.*

Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., & Chen, H. (2018). Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 682-690). WSDM 2018, Marina Del Rey, CA, USA, February 5-9. https://doi.org/10.1145/3159652.3159685

Zhang, Z., Li, J., Zhu, P., Zhao, H., & Liu, G. (2018). Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3740-3752). Santa Fe, New Mexico, USA, August 20-26.

Zhao, X., Wu, W., Xu, C., Tao, C., Zhao, D., & Yan, R. (2020). Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3377-3390).

Zhou, K., Prabhumoye, S., & Black, A. W. (2018). A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 708-713). Brussels, Belgium, October 31-November 4.