ServiceGroup: A Human-Machine Cooperation Solution for Group Chat Customer Service

Minghui Yang, Hengbin Cui, Shaosheng Cao, Yafang Wang, Xiaolong Li
{minghui.ymh,alexcui.chb,shaosheng.css,yafang.wyf,xl.li}@antfin.com
Ant Financial Services Group, Hangzhou, China

ABSTRACT

With the rapid growth of B2B (Business-to-Business), how to efficiently respond to various customer questions is becoming an important issue. In this scenario, customer questions always involve many aspects of the products, so there are usually multiple customer service agents to response respectively. To improve efficiency, we propose a human-machine cooperation solution called ServiceGroup, where relevant agents and customers are invited into the same group, and the system can provide a series of intelligent functions, including question notification, question recommendation and knowledge extraction. With the assistance of our developed ServiceGroup, the response rate within 15 minutes is improved twice. Until now, our ServiceGroup has already supported thousands of enterprises by means of millions of groups in instant messaging softwares.

CCS CONCEPTS

• **Information systems → Question answering**; *Recommender systems.*

KEYWORDS

Customer Service, Group Chat, Question Answering

ACM Reference Format:
Minghui Yang, Hengbin Cui, Shaosheng Cao, Yafang Wang, Xiaolong Li. 2020. ServiceGroup: A Human-Machine Cooperation Solution for Group Chat Customer Service. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401414

1 INTRODUCTION

Customer service has always been an important part of business behavior, both in B2C (business to consumer) and B2B (business to business) scenarios. In the B2C scenario, the agent is an employee of the company, and the service target is a customer who uses the company’s products. Although the total number of customers is often large, the consulting behavior is low frequency and questions are often simple and common. While in the B2B scenario, the agents and customers come from two enterprises or organizations with business relationships. The number of customers is relatively small, but the communication between customers and agents is much more frequent.

Traditional customer service solutions use one-to-one communication in B2C, such as work tickets, hotlines, and online chat. These solutions usually only support one agent to communicate with one customer. If more agents or customers are required to cooperate to solve the problem, it is difficult to collaborate. For example, a cloud service company provides cloud server, Database and DNS services. When a customer encounters a network exception, which might involve multiple aspects of services. Through the traditional one-to-one customer service, work tickets and conversations will be transferred between different responsible agents. If relevant agents and customers can cooperate in a group, the problem will be solved much more efficiently.

Nowadays, with the increasing popularity of instant messaging (IM) software, such as DingTalk, Slack, WeChat, and Snapchat, group chats are becoming more common. When a message is sent, all group members can receive it. Based on group chat, many-to-many customer service solutions can be implemented. Compared with one-to-one solutions, group chat can support multiple agents and customers to communicate and solve problems together. In addition, daily communication between agents and customers in a group chat can further improve the efficiency of solving problems. However, group chat based customer service is facing huge challenges:

- Customers also send messages with each other in a group chat, which reduces the efficiency of the agent;
- Similar questions are usually proposed by different customers, but the answers are not well reused;
- Customer’s questions can not be answered when no agent is working.

To overcome the challenges, we present a new customer service solution called ServiceGroup. Our main contributions include:

- Our developed ServiceGroup can assist human-machine cooperation, which has already supported thousands of enterprises by means of millions of chat groups in real world;
- Based on our designed functions, the response rate within 15 minutes is improved twice, and more than 30% answers replies on our functions.

2 SYSTEM OVERVIEW

In our case, there are two kinds of roles,*i.e.*, customer and agent. In the B2B scenarios, our ServiceGroup provides services to multiple customers and agents in a group chat.

As illustrated in Figure 1, the system consists of three main components. The front end interface is a group chat embedded

SIGIR ’20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401414

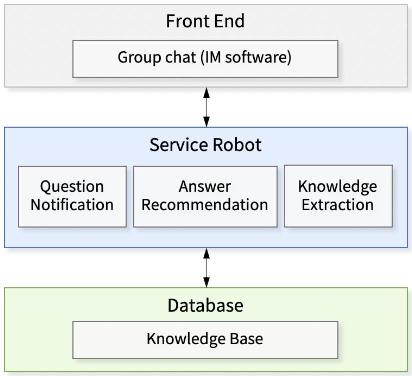


Figure 1: The Overview of ServiceGroup

in an IM software. And service robot provides three important functions, including question notification, answer recommendation and knowledge extraction. Besides, our knowledge base stores question-answering pairs to support the service robot.

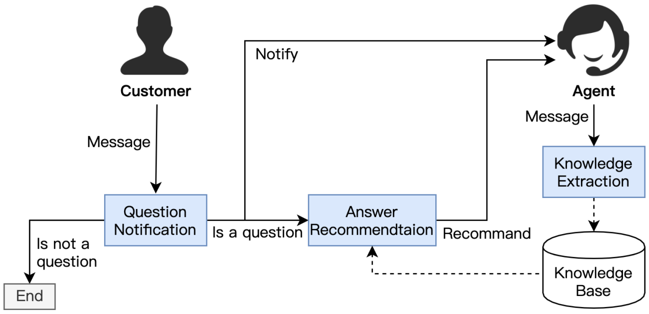


Figure 2: The Workflow of ServiceGroup

As demonstrated in Figure 2, when a customer sends a message in the group chat, question notification function will first distinguish whether it is a question or not. If it is a question, question notification will send a prompt message to agents immediately. At the same time, the question is sent to answer recommendation, which is a response to push recommended answers to a suitable agent. After the agent reply to the question, our knowledge extraction function will collect potential question-answering pairs and write into knowledge base, which supports answer recommendation.

3 SYSTEM IMPLEMENTATION

In this section, we will describe more details about our system.

3.1 Question Notification

When a customer sends a question in a group chat, ServiceGroup reminds the suitable agent to respond. The first key step of question notification is to identify whether a message is a question or not. In this work, we define a question as a message that should be answered or solved by agents, including a query, request, and command. Finally, we consider utilizing text classification models to achieve question notification function.

We first directly utilize an LSTM [5] model, one of the most popular text classification models, to implement question identification. In this case, the input is a customer's message in a group chat, and the output is a label indicating whether the message is a question or not. We also periodically collect labels from agents' service records.

However, based on our analysis, the judgment of a question is highly relevant to the business scenario in B2B customer service. For example, "How is the weather today?" is a chit-chat for the most cases, but it is indeed a question in weather situations. Therefore, it is necessary to train different identification models for different business scenarios. However, the total amount of data is not large enough to simply train multiple classifiers based on each scenario. To alleviate this issue, we adopt an adversarial multi-task training method [8]. Finally, the average F1 score is improved from 80% to 92%.

3.2 Answer Recommendation

When a question is identified, answer recommendation function will recommend the most relevant answers to agents. Following [6, 15], we also employ an information retrieve based framework to implement recommendation. We collect question-answering pairs from manual annotation and our implemented knowledge extraction function. These pairs are stored in the inverted index constructed by Elasticsearch[1] after our normalization preprocessing. In the online real-time recall process, we retrieve the most likely related questions and answers based on the proposed questions.

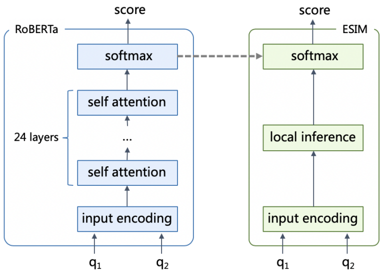
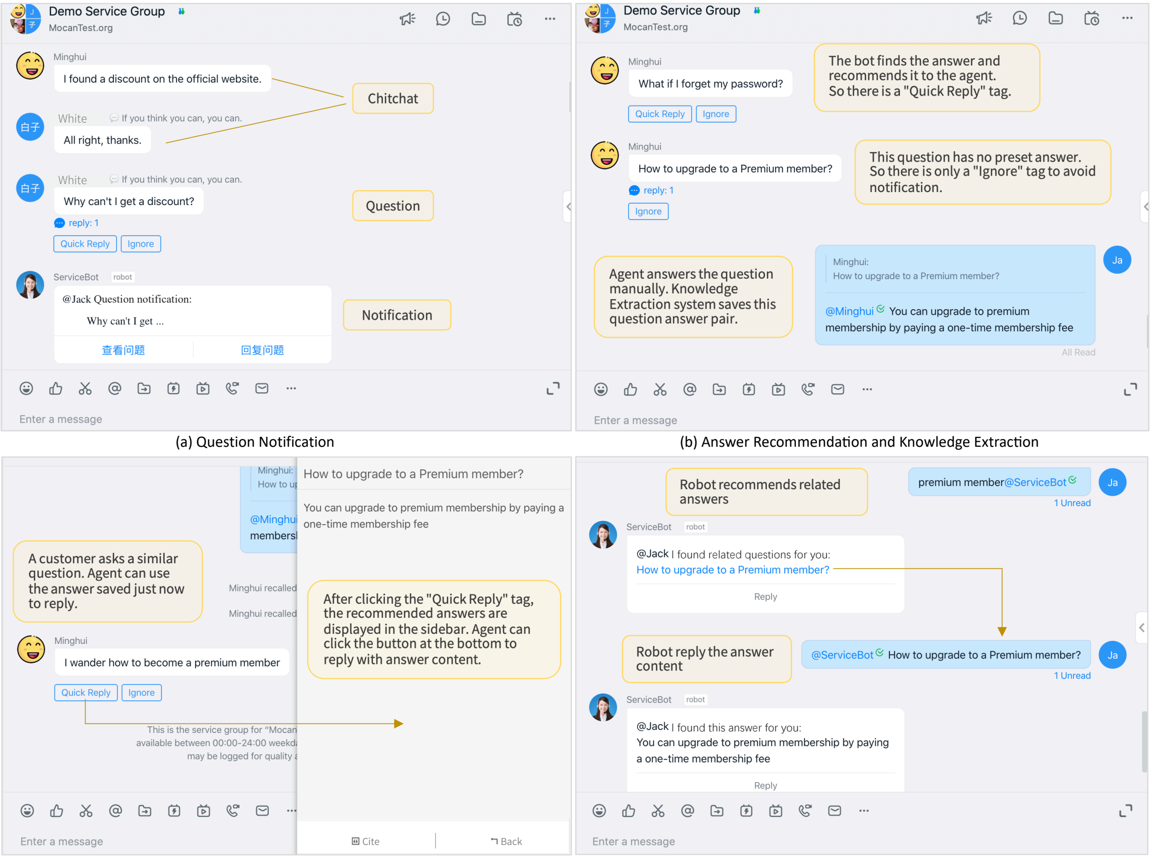


Figure 3: Our Distillation Model

In the second phase, we calculate text similarity to sort the recalled question-answering pairs to find the most relevant answer. In the report of [9], RoBERTa model achieved state-of-the-art results on text similarity tasks. However, the model has a large number of parameters, which deteriorates online real-time prediction performance. In order to address this problem, we introduce a new knowledge distillation model. As shown in Figure 3, we leverage a 24-layer RoBERTa model as a teacher, and an ESIM [1] model as a student. From the experimental results, based on our manually labeled 10,000 test sentences, our distillation model obtains 20 times faster than the original RoBERTa for online inference, with slight accuracy loss from 84.9% to 84.2%.

Besides, a customer also can ask the robot a question via "@" symbol in ServiceGroup, and the robot will directly reply to the customer with our answer recommendation function. In this way, the question will be answered even though no agent is working.

[1] https://www.elastic.co/



Minghui: I found a discount on the official website.

White (If you think you can, you can.): All right, thanks.

White (If you think you can, you can.): Why can't I get a discount?

ServiceBot: @Jack Question notification: Why can't I get ...

(a) Question Notification

Minghui: What if I forget my password?

The bot finds the answer and recommends it to the agent. So there is a "Quick Reply" tag.

Minghui: How to upgrade to a Premium member?

This question has no preset answer. So there is only a "Ignore" tag to avoid notification.

Agent answers the question manually. Knowledge Extraction system saves this question answer pair.

Minghui: How to upgrade to a Premium member?

@Minghui You can upgrade to premium membership by paying a one-time membership fee

(b) Answer Recommendation and Knowledge Extraction

A customer asks a similar question. Agent can use the answer saved just now to reply.

Minghui recalled

Minghui recalled

Minghui: I wander how to become a premium member

This is the service group for "Mocan
available between 00:00-24:00 weekd
may be logged for quality

How to upgrade to a Premium member?

You can upgrade to premium membership by paying a one-time membership fee

After clicking the "Quick Reply" tag, the recommended answers are displayed in the sidebar. Agent can click the button at the bottom to reply with answer content.

(c) Answer Recommendation and Knowledge Extraction

Robot recommends related answers

premium member@ServiceBot

ServiceBot: @Jack I found related questions for you: How to upgrade to a Premium member?

Robot reply the answer content

@ServiceBot How to upgrade to a Premium member?

ServiceBot: @Jack I found this answer for you: You can upgrade to premium membership by paying a one-time membership fee

(d) Question Answering

Figure 4: ServiceGroup Demonstration

3.3 Knowledge Extraction

Mining high quality question-answering pairs is critical to answer recommendation. We will introduce our designed knowledge extraction function to collect pairs automatically.

As mentioned before, we provide function “@”, which cites the question to be answered. When an agent gives an answer in this way, it is suggested to be a potential question-answering pair. For the other cases without symbol “@”, we first focus on the agents’ reply message, which is regarded as our candidate answer, and next we collect context sentences of the answer, aiming to obtain the potential questions.

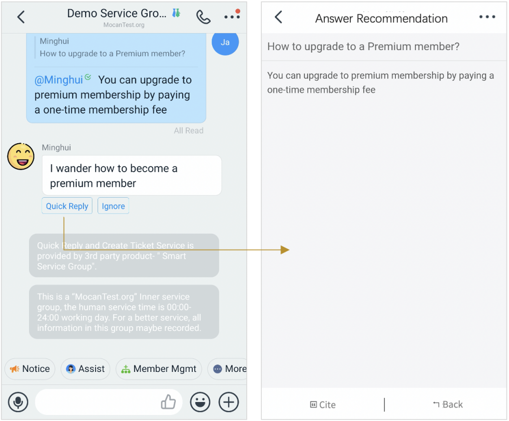
After that, our relevance model will yield probability scores for each possible pair, and the pairs whose scores are higher than the predefined threshold will be written into the knowledge base warehouse. For the relevance model, we simply train a BERT [2] to predict the relevance probability score. Based on our offline evaluation, the accuracy of the model achieves 87.9% on 5,000 annotated test instances.

4 DEMONSTRATION

We show an illustrated example in Figure 4. There are two customers and one agent in the group.

In Figure 4(a), at first, two customers were talking about discount without asking questions. Then White sent a question about the discount in the group. The system recognized the question, but the agent did not respond immediately. One minute later (a preset time interval), the service robot reminded the agent to reply in time. In Figure 4(b), the customer asked two questions continuously. The system recognized them and displayed the “Ignore” button. The agent also can click this button to avoid notification. The question answering function found the answer to the first question, so the “Quick Reply” tag was displayed. Nevertheless, the second question had no relevant answer. The agent answered the second question manually. Based on the replied answer and the corresponding question, the knowledge extraction function updated knowledge base.

In Figure 4(c), when another customer asked a similar question, the question-answering function recommended the answer just saved to the agent. After the agent clicked the “Quick Reply” tag, the sidebar displayed the recommended answer. The agent could send this answer to the customer by clicking the button at the bottom of the sidebar. In Figure 4(d), when the customer directly asked the robot through “@” function, the robot recommended the most relevant answers to the questioner. ServiceGroup can also be used on mobile phones, as shown in Figure 5. More about our product features are shown in video[2].



Minghui
How to upgrade to a Premium member?

@Minghui You can upgrade to premium membership by paying a one-time membership fee

All Read

Minghui
I wander how to become a premium member

Quick Reply and Create Ticket Service is provided by 3rd party product- " Smart Service Group".

This is a "MocanTest.org" Inner service group, the human service time is 00:00-24:00 working day. For a better service, all information in this group maybe recorded.

How to upgrade to a Premium member?

You can upgrade to premium membership by paying a one-time membership fee

Figure 5: ServiceGroup on Mobile Phone

With the assistance of our ServiceGroup, the response rate within 15 minutes is improved twice, and more than 30% replied answers are shown with our functions. Besides, 9% of questions are directly solved by service robots via "@" function.

5 RELATED WORK

In this section, we investigate related literature, including dialogue systems, studies on group chat and text matching.

5.1 Dialogue Systems

Dialogue system has been studied in many years. Early robots, such as Eliza [13], were mainly based on hand-crafted rules. In recent years, assistant robots, such as Microsoft's Cortana[3], Amazon's Alexa[4], Apple Siri[5] and Alibaba's TmallGenie[6], have been widely used. On the other hand, dialogue systems have also been successfully applied in the field of customer service, such as IBM Watson [3], Alibaba's AliMe [10] and jd.com's JIMI [16], which achieved good results in one-to-one customer service scenarios.

5.2 Group Chat

Li and Rosson [7] introduced instant annotations to record important information in group chat. Tepper et al. [12] proposed Collabot for personalized group chat summarization. Most of the previous studies focus on message annotation and summarization in a group chat, but there is a lack of research in customer service scenarios, especially on the efficiency of agent's problem-solving.

5.3 Text Matching

Statistics-based methods, such as BM25 [11], are first proposed to model text matching, and word embedding based methods like Charagram [14] have a clear improvement. Recently, BERT [2] and RoBERTa [9] models achieved new state-of-the-art accuracy. However, BERT-based methods have a large number of parameters, which reduce online prediction efficiency in real-world applications. To address this problem, knowledge distillation [4] technology is getting more and more attention.

6 CONCLUSION

In this paper, we propose a human-machine cooperation solution called ServiceGroup, which provides a number of functions. Our deployed system has supported thousands of enterprises in real world. In the future, we will further explore more methods and better system implementations to achieve better performance.

7 ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive and valuable advice, as well as Ant Financial's Customer Communication Management team and Alibaba's DingTalk team for their assistance on system implementation and deployment.

REFERENCES

[1] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *ACL*. Association for Computational Linguistics, 1657–1668.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. Association for Computational Linguistics, 4171–4186.
[3] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59–79.
[4] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).
[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[6] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988* (2014).
[7] Na Li and Mary Beth Rosson. 2014. Using annotations in online group chats. In *CHI*. ACM, 863–866.
[8] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* (2017).
[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019).
[10] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *ACL*. Association for Computational Linguistics, 498–503.
[11] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.
[12] Naama Tepper, Anat Hashavit, Maya Barnea, Inbal Ronen, and Lior Leiba. 2018. Collabot: Personalized Group Chat Summarization. In *WSDM*. ACM, 771–774.
[13] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
[14] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. *arXiv preprint arXiv:1607.02789* (2016).
[15] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*. 55–64.
[16] Xiaoming Zhu. 2019. Case II (Part A): JIMI's Growth Path: Artificial Intelligence Has Redefined the Customer Service of JD. Com. In *Emerging Champions in the Digital Economy*. Springer, 91–103.

[2] https://www.youtube.com/playlist?list=PL9l1h3zzCJD3BA5v0VmaGrY4cBamFE2Jw
[3] https://www.microsoft.com/en-us/cortana
[4] https://developer.amazon.com/en-US/alexa
[5] https://www.apple.com/siri
[6] https://en.wikipedia.org/wiki/Tmall_Genie