DATA PAPER

An Evaluation of Chinese Human-Computer Dialogue Technology

Zhengyu Zhao[1†], Weinan Zhang[1], Wanxiang Che[1], Zhigang Chen[2] & Yibo Zhang[3]

[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China

[2]AI Research Institute, IFLYTEK Co., Ltd., Hefei 230088, China

[3]Huawei Technologies Co., Ltd., Shenzhen 518129, China

Keywords: Chinese human-computer dialogue evaluation; Evaluation data; User intent classification; Task-oriented dialogue

Citation: Z. Zhao, W. Zhang, W. Che, Z. Chen, & Y. Zhang. An evaluation of Chinese human-computer dialogue technology. Data Intelligence 1(2019), 187-200. doi: 10.1162/dint_a_00007

Received: November 20, 2018; Revised: February 12, 2019; Accepted: February 19, 2019

ABSTRACT

The human-computer dialogue has recently attracted extensive attention from both academia and industry as an important branch in the field of artificial intelligence (AI). However, there are few studies on the evaluation of large-scale Chinese human-computer dialogue systems. In this paper, we introduce the Second Evaluation of Chinese Human-Computer Dialogue Technology, which focuses on the identification of a user's intents and intelligent processing of intent words. The Evaluation consists of user intent classification (Task 1) and online testing of task-oriented dialogues (Task 2), the data sets of which are provided by iFLYTEK Corporation. The evaluation tasks and data sets are introduced in detail, and meanwhile, the evaluation results and the existing problems in the evaluation are discussed.

1. INTRODUCTION

With the development of artificial intelligence, human-computer dialogue technology has become increasingly popular and has attracted growing attention [1]. Human-computer dialogue systems are conversation agents, which are normally divided into two classes [2, 3]: task-oriented dialogue systems [4, 5, 6] and none-task-oriented systems [7, 8]. In this paper, we mainly focus on task-oriented dialogue systems.

[†] Corresponding author: Zhengyu Zhao (Email: zyzhao@ir.hit.edu.cn; ORCID: 0000-0003-1678-9694).

There are two important tasks in a task-oriented dialogue system. One is concerned with classification of a user's intents, which is a text categorization task. Its purpose is to recognize the user's chat intentions, such as task-based interaction, a knowledge quiz or chit-chat. It is the foundation for building a large and complex human-machine dialogue system [9] and it is a clear but difficult task because of a limited number of corpora available for training the algorithms and difficulties in understanding semantic meanings. Recently, there have been some evaluations with user intent classification tasks. For example, Task 1 in the 17th China National Conference on Computational Linguistics (CCL2018)[①], which is based on Chinese corpora, is a user intent classification task in the customer service field. They provide some open data to allow participants to build systems and then test them on hidden data sets. However, the range of data sets they provide is limited to Q&A data from China Mobile Communications Group Co., Ltd.[②], including the query categories, data processing categories and business consulting categories.

The other is to accomplish tasks in a specific domain in a human-computer dialogue. A complete human-computer dialogue system should be capable of understanding the tasks that users want to accomplish and assist them in completing a specific domain task, such as inquiring for train information or booking a ticket. This is a fairly complex task, which can fully reflect the intelligence of a human-machine dialogue system. Another challenge is how to evaluate and compare these systems, and what influencing factors we need to pay attention to. A similar evaluation based on English corpora is the 6th Dialog System Technology Challenges (DSTC6) held in 2017 [10]. In DSTC6, participants need to build a system that responds to a user's utterances based on the context of the conversation, where they can use external data. Both objective and subjective indicators are used to evaluate the submitted systems [11]. However, the focus of the task for participants in DSTC6 is on text generation instead of the complete process of accomplishing the given task. As far as we know, the last manual evaluation of the end-to-end task-based dialogue system was the Spoken Dialog Challenge 2010 [12], which was held eight years ago.

In short, in order to promote the development of the evaluation technology for human-computer dialogue systems, and to attract more people to pay attention to the above two key issues in human-computer dialogue systems, the Second Evaluation of Chinese Human-Computer Dialogue Technology was held during the 7th China National Conference on Social Media Processing[③] (SMP2018-ECDT), which consists of two tasks:

1) **User intent classification.** There are 31 categories in total, which include one chit-chat category and 30 vertical categories of 30 specific tasks such as accessing apps and inquiring about the weather. The submitted systems need to determine which category the user's input belongs to among all of the 31 categories.
2) **Online testing of task-oriented dialogues.** The submitted systems should complete the corresponding tasks about tickets inquiring or reservation through online real-time dialogues with testers.

[①] http://www.cips-cl.org/static/CCL2018/call-evaluation.html

[②] http://www.10086.cn

[③] http://smp2018.cips-smp.org/

This Evaluation has not only automatic evaluation (for user intent classification tasks) but also online manual testing (for online testing of task-oriented dialogues). Compared with CCL2018 Task 1 and DSTC6, this Evaluation bears the following features:

- Compared to CCL2018 Task1, as organizers of the competition our data set contains a more comprehensive and more general set of tags, not just in one area. Specifically, we provide a data set which contains 31 user intents that appear frequently in general-purpose chatbots.
- Compared to DSTC6, we select several reviewers to evaluate the complete process of accomplishing a given task, and the reviewers will give their scores for a submitted system during each process.
- In order to avoid revealing the hidden test set and thereby reducing the possibility of manual intervention, we have modified the traditional evaluation method to allow the participating teams to set up services to respond to our requests so that the participants do not have to submit the code. At the same time, in order to avoid participants obtain the complete test set, we add a lot of noise to the test set.

In addition, compared to SMP2017-ECDT [13], this year we add new data sets for each of the two tasks. Our data sets provided by iFLYTEK Corporation[4] are all labeled manually. Different from Task 1 last year, we cancel the evaluation of the closed domain and only remain the open domain evaluation. The difference between the closed domain and the open domain is that users can not only use the provided training data but also collect data by themselves in the open domain. However, there is no guarantee that the participating teams will just use the evaluation data provided by us for training and developing their systems if we do not ask them to provide the code.

The rest of the paper is organized as follows. We introduce two tasks in detail in Section 2 and describe the data sets of two tasks in Section 3. Parts of the evaluation results are given in Section 4 and finally the conclusion is drawn in Section 5.

2. THE SECOND EVALUATION OF CHINESE HUMAN-COMPUTER DIALOGUE TECHNOLOGY

In this section, we give a brief introduction to evaluation tasks.

2.1 Task 1: User Intent Classification

The specific descriptions of Task 1 are as follows: build a system that can classify a user's input into the most relevant category, including chit-chat or task subcategories, e.g.,

[4] http://www.iflytek.com/

- What have you done recently? -chat
  你最近干嘛呢？
- What's the big news? -news
  有什么重大新闻？
- I want to read free novels. -novel
  我要看免费的小说

In Task 1, participating teams do not need to consider the overall intention of multiple rounds of a task-based dialogue, but to pay attention to a single round of dialogue. In addition, they are provided with a template of an example system[5] to facilitate the unification of the interface.

There are many text categorization tasks that use F1-measure as evaluation indicators, such as [14, 15, 16]. In order to avoid the imbalance of category distribution and meanwhile take into account each category, we also evaluate submitted systems based on the F1-measure obtained from precision and recall. Specifically, we first construct a confusion matrix for calculating the Precision P_i and Recall R_i value of each category, and then take the average precision as $\overline{P}=\frac{1}{N}\sum_{i=1}^{N}P_i$ and take the average recall as $\overline{R}=\frac{1}{N}\sum_{i=1}^{N}R_i$, and F1-measure is calculated by Equation (1):

$$F_1=\frac{2\overline{PR}}{\overline{P}+\overline{R}}, \tag{1}$$

where N denotes the total number of categories.

2.2 Task 2: Online Testing of Task-Oriented Dialogues

Task 2 of the Evaluation is described as follows: For a complex task on booking a flight, a train ticket, or a hotel room, build a system to guide the user to complete the corresponding task based on the given relevant database. In this evaluation, we evaluate submitted systems online manually. Research in [17] suggests that the use of crowdsourcing technology is feasible and it can provide reliable results, and our reviewers are professional testers from iFLYTEK Corporation, which will be more likely to produce accurate results. A complete intent of a flight reservation task is described as:

"帮我订一张从北京到上海的飞机票，早上或者中午都行"
"Booking a flight from Beijing to Shanghai in the morning or at noon".

The whole dialogue process of this flight reservation task is shown in Table 1, where U denotes the utterance of the user and R denotes the response of the agent.

[5] https://github.com/WindInWillows/SMP2018-ECDT-TASK1

Table 1. An example of air ticket booking.

Role	Questions and answers
U	查询明天从北京去上海的机票。 Check out the ticket from Beijing to Shanghai tomorrow.
R	请问您只要机票吗？ Do you only need an air ticket?
U	是的！ Yes!
R	请问您要明天什么时候出发呢？ When are you leaving tomorrow?
U	上午或中午吧。 Morning or noon.
R	以下是帮您查询到的机票信息，是否需要预定？ The following is the ticket information for you to check, would you like to book a ticket?
U	也行，就订这个吧。 OK, I'll take it.
R	已经帮您预订该航班机票，将跳转至付款页面！ The flight ticket has already been booked for you. Now we go to pay for the ticket!

Considering a variety of important factors on evaluation of a task-oriented dialogue system, we use the following indicators to evaluate the submitted systems in Task 2:

- **Task completion ratio**: The number of tasks completed during the test divided by the total number of tasks.
- **Average number of dialogue turns**: The number of utterances during the process of completing a task.
- **Satisfaction score**: The subjective score of the system marked by the tester, including 5 integers from -2 to 2.
- **Fluency degree of response**: Subjective scoring, including 3 integers from -1 to 1.
- **Uncovered data guidance capability**: Subjective scoring, including 0 and 1.

The core purpose of a task-oriented dialogue system is to help users complete a specific task. Then, the two most direct indicators for evaluating a task-oriented dialogue system is the *task completion ratio* and the *average number of dialogue turns* [18, 19]. The task completion ratio indicates the completion of the task and is the most important indicator that can reflect the system's capabilities. In Task 2, a complete intent may contain multiple subtasks, such as booking a flight first, and then booking a train ticket, and at last booking a hotel room. In order to demonstrate the ability of the system to complete composite tasks, when all the subtasks of a composite task are completed, we mark the completion of the task. For the average number of dialogue turns, it is counted by the evaluation system. When the task completion ratio is the same, the smaller the number of dialogue rounds, the better the system performs. In order to ensure that the number of dialogue turns of unfinished subtasks must be greater than or equal the number of rounds of completed subtasks, we take the number of dialogue turns of unfinished subtasks as the theoretical maximum number. If the maximum number of rounds is exceeded during the test, the current round of testing will be terminated. The remaining indicators are the subjective scores of the three reviewers, all of which are average scores. They reflect the performance of the dialogue system in the three aspects, respectively.

Actually, the test method of this evaluation is not only applicable to the Chinese Human-Computer Dialogue Technology Evaluation but also can be applied to the same evaluation tasks in other languages without too much modification except for the corpus.

3. EVALUATION OF DATA SETS

The evaluation data set[⑥] in Task 1 is provided by iFLYTEK Corporation, all of which are labeled manually. Some specific examples of this data set are shown in Table 2. There are 31 categories of intent data and Table 3 shows how the data set is divided.

Table 2. Some examples in training set of Task 1.

Input message	Intent category
给我讲个长篇小说。 Tell me a novel.	Novel
你最近干嘛呢？ What have you done recently?	Chat
打开 Chrome 浏览器。 Open Chrome browser.	App
帮我写一封邮件。 Write an email for me.	Email
打电话给我哥。 Call my brother.	Telephone
中国银行股票怎么样？ How about the stock of Bank of China?	Stock

Table 3. Statistics of the intent data set in Task 1.

	Train	Dev	Test
Count	2,299	770	1,550

The data set of Task 2 contains information on flights, train tickets and hotels. It mainly includes the origin and destination of the flight or the train, the departure time and arrival time, the price, the type of tickets of the flight or the train, and the price and location of the hotel. Participants need to build a task-based dialogue system based on this information. In addition, we provide testers with some test cases and corresponding starting sentences that contain individual intentions and mixed intentions for tasks on booking air tickets, train tickets and hotels.

[⑥] The data set in Task 1 is available at https://worksheets.codalab.org/worksheets/0x27203f932f8341b79841d50ce0fd684f/.

4. EVALUATION RESULTS

In this section, we show partial results of Task 1 and Task 2. Meanwhile, we analyze the results and summarize some frequently occurring problems of the two tasks. The complete leaderboards are shown in Appendix A.

4.1 Task 1

For Task 1, we have received 21 submitted systems in total, and part of the evaluation results are shown in Table 4.

Table 4. The top 8 teams of Task 1 ranked by F1 score.

Ranking	Participant	F1 score
1	达闼科技（北京）有限公司 CloudMinds (Beijing)	0.8339
2	深思考人工智能机器人科技（北京）有限公司 iDeepWise Artificial Intelligence (Beijing)	0.8276
3	北京智能一点科技有限公司 ABitAI Technology Co., Ltd.	0.8008
4	华南农业大学口语对话系统研究室 Spoken Dialogue System Lab, South China Agricultural University	0.7923
5	北京来也网络科技有限公司 Laiye Networktechnology Co., Ltd.	0.7846
6	山西大学计算机与信息技术学院 School of Computer & Information Technology, Shanxi University	0.7735
7	同济大学 Tongji University	0.7722
8	山西大学 Shanxi University	0.7648

After evaluating and ranking the submitted systems, we find that the average F1 score (0.8079) of the top five entries in this year's competition is much lower than that of last year (0.9268). The main reason is perhaps that the test set of this year is completely new and it is created later than the training set and the development set, which makes the test set in the different distribution with the training set and the development set. Therefore, the model trained in the training set performs worse on this year's test set than on last year's test set. This also indicates that many of the current models for text classification tasks have considerable losses after migration.

4.2 Task 2

Since Task 2 is much more difficult and complex than Task 1, the number of submitted systems is also relatively small. A total of 10 systems are submitted in Task 2 (Table 5). The main reference indicators are **C** (task completion rate) and **T** (the average number of dialogue turns: the smaller the score **T**, the better the system). In this task, 34.29 is the theoretical maximum number of **T**, and the maximum penalty is made when **C** is zero.

Table 5. The top 5 teams of Task 2.

Ranking	Participant	C	T	Sa	F	G
1	深思考人工智能机器人科技（北京）有限公司 iDeepWise Artificial Intelligence	0.397	26.13	0.667	0.333	0.762
2	深圳市人马互动科技有限公司 Centaurs Technologies Co., Ltd.	0.349	21.86	0.429	0.286	0.714
3	华南理工大学-CIKE 实验室 CIKE Lab, South China University of Technology	0.270	28.73	0.2381	-0.064	0.524
4	北京桔子互动科技有限公司 BatOrange Interactive Technology Co., Ltd.	0.222	30.32	0.556	0.349	0.698
5	北京来也网络科技有限公司 Laiye Networktechnology Co., Ltd.	0.127	31.84	-0.222	-0.238	0.191

Note: **C** denotes task completion ratio, **T** denotes average dialogue turns, **Sa** denotes user satisfaction score, **F** denotes fluency degree of response, and **G** denotes uncovered data guidance capability. All these indicators are average scores of all test cases.

The results shown in Table 5 are ranked by **C** firstly, then ranked by **T**, **Sa**, **F** and **G** in order. Among these indicators, **C**, **Sa**, **F** and **G** are manually labeled and **T** is calculated by the evaluation system. There are three reviewers to score each test case for each participating system. The final score for each indicator is the average of its scores on all test cases, given by reviewers or the evaluation system.

4.3 Analysis

According to the results, this evaluation has been completed smoothly. Each participating team has verified their system on the provided data set and has achieved results that are consistent with their expectations. Through this evaluation, some key problems in the human-computer dialogue have attracted more people's attention. In addition, this evaluation mainly focuses on the application of human-computer dialogue systems, so it provides some references for the industry to solve the problem of constructing a human-computer dialogue system. In the meanwhile we found an interesting phenomenon from the evaluation results that the top three teams in the two tasks are almost all from the industry, which demonstrates the importance of experience in natural language processing evaluation tasks.

5. CONCLUSION

We introduce the Second Evaluation of Chinese Human-Computer Dialogue Technology, which has made some adjustments and improvements to solve the problems of the first session of the competition in 2017. In this paper, we introduce Task 1 and Task 2 of this Evaluation, respectively, and explain the updated indicators of the two tasks and the calculation methods of them. In addition, we illustrate the data sets of the two tasks. Finally, we show the evaluation results and analyze the problems in the evaluation.

ACKNOWLEDGMENTS

We would like to thank Social Media Processing Committee of Chinese Information Processing Society of China (CIPS-SMP) for its strong support for this evaluation. Then we would especially thank Wenxia Feng, Xinyi Chen, Shu Fang, etc. They are the very serious and responsible testers from iFLYTEK Corporation, and it is them who complete the online evaluation of Task 2 patiently and impartially. Thanks to Huawei Technologies Co. Ltd. for providing financial support which sums up to RMB 50,000 as a bonus for this evaluation. Thanks to Lingzhi Li, Caihai Zhu, Yiming Cui, Haoyu Song and Yuanxing Liu for their indispensable support during the evaluation.

AUTHOR CONTRIBUTIONS

This work was a collaboration between all of the authors. W. Zhang (wnzhang@ir.hit.edu.cn) is the leader of SMP 2018-ECDT, who drew the whole picture of the evaluation. W. Che (car@ir.hit.edu.cn), Z. Chen (zgchen@iflytek.com) and Y. Zhang (yibo.cheung@huawei.com) supervised the evaluation process. They summarized the conclusion part of this paper. Z. Zhao (zyzhao@ir.hit.edu.cn, corresponding author) summarized the data sets and results of SMP2018-ECDT and drafted the paper. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

REFERENCES

[1] I.V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N.R. Ke, S. Mudumba, A. de Brébisson, J. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, & Y. Bengio. A deep reinforcement learning chatbot. arXiv preprint. arXiv:1709.02349, 2017.

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

[3] H. Chen, X. Liu, D. Yin, & J. Tang. A survey on dialogue systems: Recent advances and new frontiers. arXiv preprint. arXiv:1711.01731, 2018.

[4] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, & M. Zhou. Superagent: A customer service chatbot for ecommerce websites. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, 2017, pp. 97–102. doi: 10.18653/v1/P17-4017.

[5] B. Liu, G. Tur, D. HakkaniTur, P. Shah, & L. Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 2060–6069. Available at: http://www.aclweb.org/anthology/N18-1187.

[6] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, & D. Yu. Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio Speech Language Processing 23(3)(2015), 530–539. doi: 10.1109/TASLP.2014.2383614.

[7] R. Yan, & D. Zhao. Coupled context modeling for deep chit-chat: Towards conversations between human and computer. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, 2018, pp. 2574–2583. doi: 10.1145/3219819.3220045.

[8] I.V. Serban, A. Sordoni, Y. Bengio, A.C. Courville, & J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016, pp. 3776–3784. Available at: https://dl.acm.org/citation.cfm?id=3016435.

[9] A. Bhardwaj, & A. Rudnicky. User intent classification using memory networks: A comparative analysis for a limited data scenario. arXiv preprint. arXiv: 1706.06160, 2017.

[10] DSTC6: Dialog System Technology Challenges. Available at: http://workshop.colips.org/dstc6/.

[11] C. Hori, & T. Hori. End-to-end conversation modeling track in DSTC6. arXiv preprint. arXiv: 1706.07440, 2017.

[12] A.W. Black, S. Burger, B. Langner, G. Parent, & M. Eskenazi. Spoken Dialog Challenge 2010. In: 2010 IEEE Spoken Language Technology Workshop, 2010, pp. 448–453. doi: 10.1109/SLT.2010.5700894.

[13] W. Zhang, Z. Chen, W. Che, G. Hu, & T. Liu. 2017. The first Evaluation of Chinese Human-Computer Dialogue Technology. arXiv preprint. arXiv: 1709.10217, 2017.

[14] G. Chen, D. Ye, Z. Xing, J. Chen, & E. Cambria. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In: International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377–2383. doi: 10.1109/IJCNN.2017.7966144.

[15] B. Tang, S. Kay, & H. He. Toward optimal feature selection in NaiveBayes for text categorization. arXiv preprint. doi: 10.1109/TKDE.2016.2563436.

[16] F. Rousseau, E. Kiagias, & M. Vazirgiannis. Text categorization as a graph classification problem. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 1702–1712. doi: 10.3115/v1/P15-1164.

[17] F. Jurcicek, S. Keizer, M. Gasic, F. Mairesse, B. Thomson, K. Yu, & S. Young. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011, pp. 3061–3064. Available at: http://mi.eng.cam.ac.uk/~sjy/papers/jkgm11.pdf.

[18] P.-H. Su, M. Gasic, N. Mrk-sic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, & S. Young. On-line active reward learning for policy optimisation in spoken dialogue systems. arXiv preprint. arXiv: 1605.07669, 2016.

[19] A.W. Black, S. Burger, A.-I Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, J.D. Williams, K. Yu, S. Young, & M. Eskenazi. Spoken Dialog Challenge 2010: Comparison of live and control test results. In: Proceedings of the SIGDIAL2011 Conference, 2011, pp. 2–7. Available at: https://dl.acm.org/citation.cfm?id=2132892.

APPENDIX A: COMPETE LEADERBOARD

Table A1. The complete leaderboard of Task 1 ranking by F1 score.

Ranking	Participant	F1 score
1	达闼科技（北京）有限公司 CloudMinds (Beijing)	0.833949
2	深思考人工智能机器人科技（北京）有限公司 iDeepWise Artificial Intelligence (Beijing)	0.827594
3	北京智能一点科技有限公司 ABitAI Technology Co., Ltd.	0.800823
4	华南农业大学口语对话系统研究室 Spoken Dialogue System Lab, South China Agricultural University	0.792296
5	北京来也网络科技有限公司 Laiye Networktechnology Co., Ltd.	0.784645
6	山西大学计算机与信息技术学院 School of Computer & Information Technology, Shanxi University	0.773488
7	同济大学 Tongji University	0.772231
8	山西大学 Shanxi University	0.764794
9	华南理工大学-CIKE 实验室 CIKE Lab, South China University of Technology	0.762546
10	哈尔滨工业大学 Wang, Harbin Institute of Technology	0.759060
11	广东外语外贸大学 NLP 实验室 NLPLab, Guangdong University of Foreign Studies	0.748618
12	北京桔子互动科技有限公司 BatOrange Interactive Technology Co., Ltd.	0.742506
13	北京大学网络所 NC&IS, Peking University	0.742133
14	广东外语外贸大学 NLP 实验室 GDUFS_NLP, South China University of Technology	0.729600
15	众安信息技术服务有限公司 ZhongAn Techology	0.725358
16	西北师范大学自然语言处理研究组 NLP Group, Northwest Normal University	0.720373
17	义语智能科技（上海）有限公司 DeepBrain	0.714655
18	复旦大学 KELAB KELAB, Fudan University	0.692646
19	哈工大深圳 Harbin Institute of Technology, Shenzhen	0.682747
20	郑州大学自然语言处理实验室 NLP lab, Zhengzhou University	0.496503
21	山西大学小虎队 Little Tiger, Shanxi University	0.187605

Table A2. The complete leaderboard of Task 2.

Ranking	Participant	C	T	Sa	F	G
1	深思考人工智能机器人科技（北京）有限公司 iDeepWise Artificial Intelligence	0.3970	26.13	0.667	0.333	0.762
2	深圳市人马互动科技有限公司 Centaurs Technologies Co., Ltd.	0.3490	21.86	0.429	0.286	0.714
3	华南理工大学-CIKE 实验室 CIKE Lab, South China University of Technology	0.2700	28.73	0.238	-0.064	0.524
4	北京桔子互动科技有限公司 BatOrange Interactive Technology Co., Ltd.	0.2220	30.32	0.556	0.349	0.698
5	北京来也网络科技有限公司 Laiye Networktechnology Co., Ltd.	0.1270	31.84	-0.222	-0.238	0.191
6	山西大学 Shanxi University	0.0159	33.11	-0.825	-0.492	0.286
7	复旦大学 KELAB KELAB, Fudan University	0.0159	34.29	-0.921	-0.619	0.064
8	山西大学小虎队 Little Tiger, Shanxi University	0.0000	34.29	-0.984	-0.508	0.444
9	西北师范大学自然语言处理研究组 NLP Group, Northwest Normal University	0.0000	34.29	-1.825	-0.952	0.032
10	北京大学网络所 NC&IS, Peking University	0.0000	34.29	-1.968	-1.000	0.000

Note: **C** denotes task completion ratio, **T** denotes average dialogue turns, **Sa** denotes user satisfaction score, **F** denotes fluency degree of response, and **G** denotes uncovered data guidance capability. All these indicators are average scores of all test cases.

AUTHOR BIOGRAPHY

Zhengyu Zhao is a postgraduate in Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His current research interests are mainly in conversational robot and user profiling.

Dr. Weinan Zhang is a Lecturer in Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interest includes human-computer dialogue, natural language processing and information retrieval.

Dr. Wanxiang Che is a professor of School of Computer Science and Technology at Harbin Institute of Technology. His main research area lies in Natural Language Processing (NLP). His projects are funded by National Natural Science Foundation of China and National Key Basic Research Program of China (973 Program).

Dr. Zhigang Chen joined iFLYTEK Corporation in 2003 and is currently the vice president of the AI Research Institute of iFLYTEK Corporation. He is mainly responsible for cognitive intelligence research and productization.

Yibo Zhang received his PhD degree from Beijing University of Posts and Telecommunications in 2004. He is currently the chief scientist of the Intelligence Engineering Department, the Huawei Consumer Business Group. Before joining Huawei, he worked at IBM China Research Lab between 2004 and 2011, Noah's Ark Lab between 2011 and 2015, and Microsoft Search Technology Center Asia between 2015 and 2018. His recent focus areas include intent understanding, task completion and dialogue management.