

The First Evaluation of Chinese Human-Computer Dialogue Technology

Wei-Nan Zhang[†], Zhigang Chen[‡], Wanxiang Che[†], Guoping Hu[‡], Ting Liu[†]

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

[‡]Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Hefei, China

{wnzhang, car, tliu}@ir.hit.edu.cn, {zgchen, gphu}@iflytek.com

Abstract

In this paper, we introduce the first evaluation of Chinese human-computer dialogue technology. We detail the evaluation scheme, tasks, metrics and how to collect and annotate the data for training, developing and test. The evaluation includes two tasks, namely **user intent classification** and **online testing of task-oriented dialogue**. To consider the different sources of the data for training and developing, the first task can also be divided into two sub tasks. Both the two tasks are coming from the real problems when using the applications developed by industry. The evaluation data is provided by the iFLYTEK Corporation. Meanwhile, in this paper, we publish the evaluation results to present the current performance of the participants in the two tasks of Chinese human-computer dialogue technology. Moreover, we analyze the existing problems of human-computer dialogue as well as the evaluation scheme itself.

Keywords: Chinese dialogue evaluation, user intent classification, online task-oriented dialogue test, evaluation data

1. Introduction

Recently, human-computer dialogue has been emerged as a hot topic, which has attracted the attention of both academia and industry. In research, the natural language understanding (NLU), dialogue management (DM) and natural language generation (NLG) have been promoted by the technologies of big data and deep learning (Shang et al., 2015; Serban et al., 2016a; Serban et al., 2016b; Li et al., 2015; Xing et al., 2016; Li et al., 2016; Vinyals and Le, 2015; Wen et al., 2016; Wen et al., 2015b). Following the development of machine reading comprehension (Cui et al., 2016; Hermann et al., 2015; Kadlec et al., 2016; Liu et al., 2017; Wang et al., 2017; Cui et al., 2017), the NLU technology has made great progress. The development of DM technology is from rule-based approach and supervised learning based approach to reinforcement learning based approach (Young et al., 2013). The NLG technology is through pattern-based approach, sentence planning approach and end-to-end deep learning approach (Mairesse and Young, 2014; Mairesse et al., 2010; Wen et al., 2015a). In application, there are massive products that are based on the technology of human-computer dialogue, such as Apple Siri¹, Amazon Echo², Microsoft Cortana³, Facebook Messenger⁴ and Google Allo⁵ etc.

Although the blooming of human-computer dialogue technology in both academia and industry, how to evaluate a dialogue system, especially an open domain chit-chat system, is still an open question. Figure 1 presents a brief comparison of the open domain chit-chat system and the task-oriented dialogue system.

From Figure 1, we can see that it is quite different between the open domain chit-chat system and the task-oriented dialogue system. For the open domain chit-chat system, as it has no exact goal in a conversation, given an input mes-

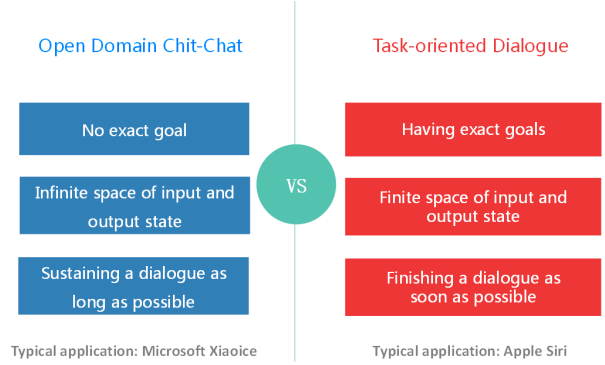


Figure 1: A brief comparison of the open domain chit-chat system and the task-oriented dialogue system.

sage, the responses can be various. For example, for the input message “How is it going today?”, the responses can be “I’m fine!”, “Not bad.”, “I feel so depressed!”, “What a bad day!”, etc. There may be infinite number of responses for an open domain messages. Hence, it is difficult to construct a gold standard (usually a reference set) to evaluate a response which is generated by an open domain chit-chat system. For the task-oriented system, although there are some objective evaluation metrics, such as the number of turns in a dialogue⁶, the ratio of task completion⁷, etc., there is no gold standard for automatically evaluating two (or more) dialogue systems when considering the satisfaction of the human and the fluency of the generated dialogue. To promote the development of the evaluation technology for dialogue systems, especially considering the language characteristics of Chinese, we organize the first evaluation

¹<https://www.apple.com/ios/siri/>

²https://en.wikipedia.org/wiki/Amazon_Echo

³<https://www.microsoft.com/en-us/windows/cortana>

⁴https://en.wikipedia.org/wiki/Facebook_Messenger

⁵<https://allo.google.com/>

⁶The number of utterances in a task-completed dialogue. Given a same task, the less number of turns, the better of the dialogue system.

⁷The number of completed tasks divided by the total number of tasks.

Input message	Intent category
你好啊，很高兴见到你！ Hello, nice to meet you!	闲聊类 Chit-chat
我想订一张去北京的机票 I want to book an air ticket to Beijing.	任务型（订机票） Task-oriented dialogue (Booking air tickets)
我想找一家五道口附近便宜干净的快捷酒店 I want to book a neat and low-priced inn near Wudaokou.	任务型（订酒店） Task-oriented dialogue (Booking hotels)

Table 1: An example of user intent with category information.

of Chinese human-computer dialogue technology. In this paper, we will present the evaluation scheme and the released corpus in detail.

The rest of this paper is as follows. In Section 2, we will briefly introduce the first evaluation of Chinese human-computer dialogue technology, which includes the descriptions and the evaluation metrics of the two tasks. We then present the evaluation data and final results in Section 3 and 4 respectively, following the conclusion and acknowledgements in the last two sections.

2. The First Evaluation of Chinese Human-Computer Dialogue Technology

The First Evaluation of Chinese Human-Computer Dialogue Technology includes two tasks, namely **user intent classification** and **online testing of task-oriented dialogue**.

2.1. Task 1: User Intent Classification

In using of human-computer dialogue based applications, human may have various intent, for example, chit-chatting, asking questions, booking air tickets, inquiring weather, etc. Therefore, after receiving an input message (text or ASR result) from a user, the first step is to classify the user intent into a specific domain for further processing. Table 1 shows an example of user intent with category information. In task 1, there are two top categories, namely, chit-chat and task-oriented dialogue. The task-oriented dialogue also includes 30 sub categories. In this evaluation, we only consider to classify the user intent in single utterance.

It is worth noting that besides the released data for training and developing, we also allow to collect external data for training and developing. To considering that, the task 1 is indeed includes two sub tasks. One is a **closed evaluation**, in which only the released data can be used for training and developing. The other is an **open evaluation** that allow to collect external data for training and developing. For task 1, we use F1-score as evaluation metric.

2.2. Task 2: Online Testing of Task-oriented Dialogue

For the task-oriented dialogue systems, the best way for evaluation is to use the online human-computer dialogue. After finishing an online human-computer dialogue with a dialogue system, the human then manually evaluate the system by using the metrics of user satisfaction degree, dialogue fluency, etc. Therefore, in the task 2, we use an online testing of task-oriented dialogue for dialogue

U	有到北京的软卧火车票吗？ Are there any soft berth tickets to Beijing?
R	请问您要从哪里出发？ Where do you leave from?
U	哈尔滨 Harbin.
R	您要订哪天的呢？ When will you leave?
U	明天吧 Tomorrow.
R	发车时间呢？ What about the departure time?
U	晚上吧，正好睡觉。 Evening is okay, just sleep.
R	上铺还是下铺？ Do you want an upper or lower berth?
U	都行。 Either is okay.
R	以下是为您查询到的车票信息。 The tickets information is as following.
U	谢谢，再见。 Thanks, see you.
R	不客气，再见。 You're welcome. Goodbye.

Table 2: An example of the task-oriented human-computer dialogue.

systems. For a human tester, we will give a complete intent with an initial sentence, which is used to start the online human-computer dialogue. Table 2 shows an example of the task-oriented human-computer dialogue. Here “U” and “R” denote user and robot respectively. The complete intent is as following:

“查询明天从哈尔滨到北京的晚间软卧火车票，上下铺均可。

Inquire the soft berth ticket at tomorrow evening, from Harbin to Beijing, either upper or lower berth is okay.”

In task 2, there are three categories. They are “air tickets”, “train tickets” and “hotel”. Correspondingly, there are three type of tasks. All the tasks are in the scope of the three categories. However, a complete user intent may include more than one task. For example, a user may first inquiring the air tickets. However, due to the high price, the user decide to buy a train tickets. Furthermore, the user may

also need to book a hotel room at the destination.

We use manual evaluation for task 2. For each system and each complete user intent, the initial sentence, which is used to start the dialogue, is the same. The tester then begin to converse to each system. A dialogue is finished if the system successfully returns the information which the user inquires or the number of dialogue turns is larger than 30 for a single task. For building the dialogue systems of participants, we release an example set of complete user intent and three data files of flight, train and hotel in JSON format. There are five evaluation metrics for task 2 as following.

- **Task completion ratio:** The number of completed tasks divided by the number of total tasks.
- **User satisfaction degree:** There are five scores -2, -1, 0, 1, 2, which denote very dissatisfied, dissatisfied, neutral, satisfied and very satisfied, respectively.
- **Response fluency:** There are three scores -1, 0, 1, which indicate nonfluency, neutral, fluency.
- **Number of dialogue turns:** The number of utterances in a task-completed dialogue.
- **Guidance ability for out of scope input**⁸: There are two scores 0, 1, which represent able to guide or unable to guide.

For the number of dialogue turns, we have a penalty rule that for a dialogue task, if the system cannot return the result (or accomplish the task) in 30 turns, the dialogue task is end by force. Meanwhile, if a system cannot accomplish a task in less than 30 dialogue turns, the number of dialogue turns is set to 30.

3. Evaluation Data

In the evaluation, all the data for training, developing and test is provided by the iFLYTEK Corporation⁹.

For task 1, as the descriptions in Section 2.1., the two top categories are chit-chat (chat in Table 3) and task-oriented dialogue. Meanwhile, the task-oriented dialogue also includes 30 sub categories. Actually, the task 1 is a 31 categories classification task. In task 1, besides the data we released for training and developing, we also allow the participants to extend the training and developing corpus. Hence, there are two sub tasks for the task 1. One is closed test, which means the participants can only use the released data for training and developing. The other is open test, which allows the participants to explore external corpus for training and developing. Note that there is a same test set for both the closed test and the open test.

For task 2, we release 11 examples of the complete user intent and 3 data file, which includes about one month of flight, hotel and train information, for participants to build their dialogue systems. The current date for online test is set to April 18, 2017. If the tester says “today”, the systems developed by the participants should understand that he/she indicates the date of April 18, 2017.

⁸During the dialogue, the testers may inquire the information that is not existing in the data files. These inquiry inputs are called out of scope inputs.

⁹<http://www.iflytek.com/en/index.html>

4. Evaluation Results

There are 74 participants who are signing up the evaluation. The final number of participants is 28 and the number of submitted systems is 43. Table 4 and 5 show the evaluation results of the closed test and open test of the task 1 respectively. Due to the space limitation, we only present the top 5 results of task 1. We will add the complete lists of the evaluation results in the version of full paper.

Note that for task 2, there are 7 submitted systems. However, only 4 systems can provide correct results or be connected in a right way at the test phase. Therefore, Table 6 shows the complete results of the task 2.

5. Conclusion

In this paper, we introduce the first evaluation of Chinese human-computer dialogue technology. In detail, we first present the two tasks of the evaluation as well as the evaluation metrics. We then describe the released data for evaluation. Finally, we also show the evaluation results of the two tasks. As the evaluation data is provided by the iFLYTEK Corporation from their real online applications, we believe that the released data will further promote the research of human-computer dialogue and fill the blank of the data on the two tasks.

Acknowledgements

We would like to thank the Social Media Processing (SMP) committee of Chinese Information Processing Society of China. We thank all the participants of the first evaluation of Chinese human-computer dialogue technology. We also thank the testers from the voice resource department of the iFLYTEK Corporation for their effort to the online real-time human-computer dialogue test and offline dialogue evaluation. We thank Lingzhi Li, Yangzi Zhang, Jiaqi Zhu and Xiaoming Shi from the research center for social computing and information retrieval for their support on the data annotation, establishing the system testing environment and the communication to the participants and help connect their systems to the testing environment.

This paper is supported by National Key Basic Research Program of China (No. 2014CB340503) and NSFC (No. 61502120), the Fundamental Research Funds for the Central Universities(30620170037).

Reference

- Cui, Y., Liu, T., Chen, Z., Wang, S., and Hu, G. (2016). Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan.
- Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602. Association for Computational Linguistics.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in*

	app	bus	calc	chat	cinemas	contacts	cookbook	datetime	email	epg	flight
Train	36	24	24	456	24	30	269	18	24	107	62
Dev	18	8	8	114	10	10	88	6	8	36	21
Test	18	8	8	50	8	10	90	6	8	36	21
Sum	72	40	40	662	42	50	447	30	40	179	104
	health	lottery	map	match	message	music	news	novel	poetry	radio	riddle
Train	55	24	68	24	63	66	58	24	402	24	34
Dev	19	8	23	8	21	22	19	8	34	8	11
Test	18	8	23	8	21	22	19	8	34	8	11
Sum	92	40	114	40	105	110	96	40	470	40	56
	schedule	stock	telephone	train	translation	tvchannel	video	weather	website	Total #	
Train	29	71	63	70	61	71	182	66	54	2,583	
Dev	9	24	21	23	21	23	60	22	18	729	
Test	10	24	21	23	20	24	61	22	18	688	
Sum	48	119	105	116	102	118	303	110	90	4,000	

Table 3: The statistics of the released data for task 1.

Ranking	Participant	F1 score
1	华南农业大学口语对话系统研究室 Spoken Dialogue System Lab, South China Agricultural University	0.9391
2	义语智能科技（上海）有限公司 DeepBrain Corporation	0.9288
3	山西大学计算机与信息技术学院 School of Computer & Information Technology, Shanxi University	0.9089
4	北京邮电大学智能科学与技术中心 Intelligent science and technology center, Beijing University of Posts and Telecommunications	0.9082
5	哈尔滨工业大学(深圳) Harbin Institute of Technology (Shenzhen)	0.9028

Table 4: Top 5 results of the closed test of the task 1.

Ranking	Participant	F1 score
1	华南农业大学口语对话系统研究室 Spoken Dialogue System Lab, South China Agricultural University	0.9414
2	义语智能科技（上海）有限公司 DeepBrain Corporation	0.9288
3	中国科学院自动化研究所-出门问问语言智能与人机交互联合实验室 Institute of automation, Chinese Academy of Sciences	0.9258
4	广东外语外贸大学 Guangdong University of Foreign Studies	0.9255
5	山西大学计算机与信息技术学院 School of Computer & Information Technology, Shanxi University	0.9123

Table 5: Top 5 results of the open test of the task 1.

- Neural Information Processing Systems*, pages 1693–1701.
- Kadlec, R., Schmid, M., Bajgar, O., and Kleindienst, J. (2016). Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918, Berlin, Germany, August. Association for Computational Linguistics.
- Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. *NAACL*.
- Li, J., Monroe, W., Ritter, A., and Dan, J. (2016). Deep reinforcement learning for dialogue generation. *EMNLP*.
- Liu, T., Cui, Y., Yin, Q., Zhang, W.-N., Wang, S., and Hu, G. (2017). Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111, Vancouver, Canada, July. Association for Computational Linguistics.
- Mairesse, F. and Young, S. (2014). Stochastic language generation in dialogue using factored language models.

Ranking	Participant	Ratio	Satisfaction	Fluency	Turns	Guide
1	深思考人工智能机器人科技（北京）有限公司 iDeepMind Artificial Intelligence	0.3175	64.53	0	-1	2
2	上海葡萄纬度科技有限公司 Shanghai Putao Technology Co., Ltd.	0.1905	72.28	-1	1	3
3	北京邮电大学信息与通信工程学院 School of Information and Communication Engineering Beijing University of Posts and Telecommunications	0.1905	78.72	0	1	3
4	中国科学院自动化研究所 出门问问语言智能与人机交互联合实验室 Institute of automation, Chinese Academy of Sciences	0.1111	71.39	-2	-1	3

Table 6: The results of the task 2. Ratio, Satisfaction, Fluency, Turns and Guide indicate the task completion ratio, user satisfaction degree, response fluency, number of dialogue turns and guidance ability for out of scope input respectively.

Computational Linguistics, 40(4):763–799.

- Mairesse, F., , M., Jurcicek, Ek, F., Keizer, S., Thomson, B., Yu, K., and Young, S. (2010). Phrase-based statistical language generation using graphical models and active learning. In *ACL*, pages 1552–1561.
- Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016a). Building end-to-end dialogue systems using generative hierarchical neural network models. *Computer Science*.
- Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2016b). A hierarchical latent variable encoder-decoder model for generating dialogues.
- Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586.
- Vinyals, O. and Le, Q. (2015). A neural conversational model. *Computer Science*.
- Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.
- Wen, T. H., Gasic, M., Kim, D., Mrksic, N., Su, P. H., Vandyke, D., and Young, S. (2015a). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *SIGDial*.
- Wen, T. H., Gasic, M., Mrksic, N., Su, P. H., Vandyke, D., and Young, S. (2015b). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *EMNLP*.
- Wen, T.-H., Gašić, M., Mrkšić, N., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., and Young, S. (2016). Multi-domain neural network language generation for spoken dialogue systems. In *NAACL*, pages 120–129.
- Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W. Y. (2016). Topic augmented neural response generation with a joint attention mechanism.
- Young, S., Gasic, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, pages 1160–1179.