An Evaluation of Chinese Human-Computer Dialogue Technology (Postprint)

Authors: Zixian, Feng, Caihai Zhu, Zhang, Weinan, Zhigang, Chen, Wanxiang, Che, Minlie, Huang, Linlin, Li, Weinan, Zhang

Date: 2022-11-27T00:00:00+00:00

Abstract

There is a growing interest in developing human-computer dialogue systems which is an important branch in the field of artificial intelligence (AI). However, the evaluation of large-scale Chinese human-computer dialogues is still a challenging task. To attract more attention to dialogue evaluation work, we held the fourth Evaluation of Chinese Human-Computer Dialogue Technology (ECDT). It consists of few-shot learning in spoken language understanding (SLU) (Task 1) and knowledge-driven multi-turn dialogue competition (Task 2), the data sets of which are provided by Harbin Institute of Technology and Tsinghua University. In this paper, we will introduce the evaluation tasks and data sets in detail. Meanwhile, we will also analyze the evaluation results and the existing problems in the evaluation.

Full Text

Preamble

An Evaluation of Chinese Human-Computer Dialogue Technology

Zixian Feng[1], Caihai Zhu[1], Weinan Zhang[1]†, Zhigang Chen[2], Wanxiang Che[1], Minlie Huang[3] & Linlin Li[4]
[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]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[4]Consumer BG, Huawei Technologies CO., LTD., Nanjing 210012, China

Keywords: Chinese human-computer dialogue evaluation; Evaluation data; Few-shot learning; Knowledge-driven multi-turn dialogue

Citation: Feng, Z.X., et al.: An evaluation of Chinese human-computer dialogue technology. Data Intelligence 3(2), 274-286 (2021). doi: 10.1162/dint_a_{00090}
Received: November 18, 2020; **Revised:** January 27, 2021; **Accepted:** February 3, 2021

ABSTRACT

There is growing interest in developing human-computer dialogue systems, an important branch of artificial intelligence (AI) research. However, evaluating large-scale Chinese human-computer dialogues remains a challenging task. To draw greater attention to dialogue evaluation efforts, we organized the fourth Evaluation of Chinese Human-Computer Dialogue Technology (ECDT), which comprises few-shot learning in spoken language understanding (SLU) (Task 1) and a knowledge-driven multi-turn dialogue competition (Task 2). The datasets for these tasks were provided by Harbin Institute of Technology and Tsinghua University. This paper introduces the evaluation tasks and datasets in detail, analyzes the evaluation results, and discusses existing problems in the evaluation process.

1. INTRODUCTION

At the end of the 20th century, rapid advances in computer technology gave rise to human-computer interaction research [1]. In the 21st century, this field has attracted increasing attention [2, 3]. Since the Turing test [4], human-computer dialogue systems have become a major research focus for many scholars. Traditional dialogue systems can be divided into two categories [5, 6]: task-oriented dialogue systems [7, 8] that help users accomplish complex tasks through multi-turn conversations, and open-domain dialogue systems [9, 10] designed purely for casual conversation. However, evaluating large-scale Chinese human-computer dialogues remains challenging.

Two critical tasks in dialogue systems are few-shot learning in spoken language understanding (SLU) and knowledge-driven multi-turn dialogue generation. Few-shot learning aims to train a model that can leverage prior experience from existing (source) domains and quickly adapt to new (target) domains with minimal labeled samples—typically one or two examples per class. In recent years, AI has achieved remarkable success through deep learning methods, yet these approaches require large amounts of labeled training data, which is often difficult and expensive to obtain [11]. In task-oriented dialogue development, for instance, it is frequently challenging to acquire real user corpora for functions under development. Even when raw corpora are available, manual annotation costs are prohibitively high. Furthermore, AI applications like dialogue systems often face rapidly changing requirements, forcing data labeling efforts to be re-

peated continuously. In stark contrast, humans can learn new tasks from just a few examples. This striking difference has inspired researchers to explore AI systems capable of learning from previous experience and small amounts of data, much like humans do.

The second task is knowledge-driven multi-turn dialogue generation, which requires producing responses that align with both knowledge graph information and contextual logic when the dialogue context and all relevant knowledge graph information are provided [12].

To advance evaluation technologies for human-computer dialogue systems and provide a valuable communication platform for academic researchers and industry practitioners, we organized the Evaluation of Chinese Human-Computer Dialogue Technology during the Ninth China National Conference on Social Media Processing (SMP2020-ECDT)[1]. The evaluation consists of two tasks:

Few-shot Learning in SLU. This task focuses on few-shot learning scenarios where only a handful of labeled examples are available for each test category. Models are first trained on domains with sufficient data and then tested on a new domain.

Knowledge-driven Multi-turn Dialogue Competition. Submitted models must generate dialogue responses that conform to knowledge graph information and contextual logic when both the context and all knowledge graph information are known. The knowledge graph is represented as a series of triples (e.g., <head entity, relationship, tail entity>). Generated responses must be fluent, semantically relevant to the dialogue context, and consistent with the relevant knowledge graph information.

Compared to SMP2019[2]-ECDT, we provided new datasets [13] for both tasks this year, incorporating natural language understanding for few-shot scenarios in Task 1 and adding knowledge integration to the dialogue competition. The remainder of this paper is organized as follows: Section 2 details the two tasks, Section 3 describes the datasets, Section 4 presents partial evaluation results, and Section 5 concludes the paper.

---

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

This section provides a brief overview of the evaluation tasks.

2.1 Task 1: Few-shot Learning in SLU

This evaluation focuses on few-shot learning where only a few labeled examples are available for each test category. Models are first trained on domains with sufficient data and then tested in a new domain. Participants are given a labeled support set as reference and must annotate any unseen query set with user

intentions and slots. As illustrated in Figure 1 [Figure 1: see original paper], when provided with a support set and the query sentence "Play Avatar," the model must predict the intent as "Play movie" and the slot as [movie: Avatar].

Many text categorization tasks use F1-score as an evaluation metric [14]. For the few-shot slot filling task, we use F1-score as the evaluation index: $F = 2PR/(P + R)$, where P represents average precision and R represents average recall. A predicted slot is considered correct when its key-value combination exactly matches a key-value combination in the ground truth. For intent recognition, we use intent accuracy (Intent acc) as the evaluation metric. To comprehensively assess model capabilities, we employ sentence accuracy (Sentence acc) to measure the joint performance of intent recognition and slot filling. We provide three separate rankings for reference, with the final competition ranking determined by Sentence acc.

2.2 Task 2: Knowledge-Driven Multi-Turn Dialogue Competition

Task 2 is defined as follows: Given the dialogue context and all knowledge graph information, models must generate responses that conform to the knowledge graph information and contextual logic.

In the preliminary stage, we evaluate submitted systems using automatic metrics. For Task 2, we selected the following metrics:

BLEU-4 [15]: Evaluates n-gram overlap between generated responses and ground truth.
Distinct-2 [16]: Assesses response diversity.

We calculate separate rankings for each model on these two metrics and use the average of the two rankings as the basis for overall ranking. In the final stage, the top 10 dialogue systems are selected for manual evaluation. During manual evaluation, 100 dialogue samples are drawn from test sets across three domains, and responses generated by each team are assessed via crowdsourcing on two aspects:

Informativeness: The amount of relevant knowledge graph information contained in generated responses, scored on an integer scale from 0 to 2.
Appropriateness: Whether generated responses conform to everyday human communication habits, scored on an integer scale from 0 to 2.

The final ranking is based on manual evaluation results.

---

3. EVALUATION DATASET

The Task 1 dataset is FewJoint, provided by Harbin Institute of Technology. It contains 59 real domains, making it one of the most domain-rich datasets available. It reflects the true difficulty of real-world natural language processing

tasks, overcoming current limitations of few-shot NLP research that has focused primarily on simple artificial tasks like text classification. The user corpus comes from two main sources: (1) real user data from the iFLYTEK AIUI[3] platform, and (2) artificially constructed corpora created by domain experts. The ratio between these two sources is approximately 3:7.

After annotating each data record with user intent and semantic slots, we divided all 59 domains into three parts: 45 training domains, 5 development domains, and 9 test domains. We reconstructed the test and development domain data into a few-shot learning format: each domain contains an artificially constructed K-shot support set and a query set composed of the remaining data. Table 1 shows the dataset statistics for Task 1. The Task 1 dataset is available for reference[4].

The Task 2 dataset is KdConv, a Chinese multi-domain dataset for multi-turn knowledge-driven conversation provided by Tsinghua University. KdConv contains 86K utterances and 4.5K dialogues across three domains: film, music, and travel. Each utterance is annotated with relevant knowledge facts from the knowledge graph, which can serve as supervision for knowledge interaction modeling. Table 2 shows the dataset statistics for Task 2. The Task 2 dataset is available for reference[5].

---

4. EVALUATION RESULTS

This section presents partial evaluation results for Task 1 and Task 2, along with qualitative analysis. Complete leaderboards are provided in Appendix A.

4.1 Task 1

For Task 1, we received eight submitted systems on the test dataset, with partial results shown in Table 3 . All teams performed well on intent recognition, likely because it is a relatively simple classification task, while slot filling is more complex. Interestingly, the second team outperformed the first in Intent acc and Slot F1, but achieved a lower final ranking. This suggests that the first-place model possesses stronger joint training capabilities.

4.2 Task 2

Five groups submitted systems for Task 2. We list the Informativeness and Appropriateness scores for the three domains separately, with the final score representing the overall results. Partial results are shown in Table 4 . All teams performed well on Appropriateness, indicating that researchers have gradually learned how to make machines communicate more like humans. However, most models struggled to effectively utilize knowledge, with only Model 1 achieving better Informativeness scores above 1 across all three domains. Additionally, all teams obtained higher scores in the travel domain compared to the others.

4.3 Analysis

The human-computer dialogue evaluation has been successfully concluded. All participating teams objectively evaluated their models on our provided datasets and can now optimize their approaches based on the evaluation results.

4.3.1 Task 1 To address few-shot data scarcity, participating teams employed pre-training models such as BERT [17] and ERNIE [18]. Since pre-training models can learn generalized language information from large amounts of unlabeled text, they are often used as basic encoders to transform natural language sentences into hidden states. Teams focused on leveraging dependencies between labels [11] or using rules to complete the mapping from support set to test set.

To explore participants' methods, we detail the approaches of the top three teams and compare their differences.

China Merchants Bank AILab-CC. One approach to solving data scarcity in NLP is data augmentation. For slot tagging augmentation, they explored sentence generation methods to create additional labeled samples. First, they used synonym expansion to augment slot recognition data and balanced the dataset to help the model learn information from different slots. Second, following Hou' s paper [11], they used Roberta-wwm-ext [19] as a benchmark model and fine-tuned it on the support set. Finally, for intent recognition, they incorporated intent information into slot recognition (e.g., intent: cov_{length}, slot: srcLengthUni, result: srcLengthUni-cov_{length}). However, experiments showed that introducing intent information actually reduced model effectiveness. To achieve better competition results, they trained a BERT+BI-LSTM+CRF [17] model for sequence labeling and Joint-BERT [20] for intent recognition, using a voting method for model fusion. They first merged all models, then removed them one by one, keeping any model whose removal decreased performance.

Shanghai Jiao Tong University-SpeechLab. They also used BERT as the encoder. Their model employed ProtoNet [21] based on Hou's paper [11] to map the support set to the test set, achieving desirable results. They encoded the support set into hidden states using BERT, converted it into sentence vectors by averaging word vectors, and then merged it with input x via vector dot product. Finally, they completed intent recognition and sequence labeling through softmax or CDT-CRF.

Peking University. Their method was relatively simple. They built a few-shot language understanding model using pre-training models and rules. Specifically, they used ERNIE as the pre-trained language model, fine-tuned it on the support set, and applied rule-based corrections. For data processing, they constructed a slot dictionary to improve accuracy.

4.3.2 Task 2 Task 2 presents three main challenges: (1) how to model knowledge, (2) how to incorporate knowledge information into the model, and (3) how

to ensure the model selects correct knowledge from candidate knowledge. Most teams used encoders to represent knowledge and then fed it into pre-training models to integrate knowledge with context.

Suzhou KidX.AI Education Technology Co., Ltd. They trained a topic extraction model to identify all topics related to knowledge in the context and established connections with knowledge entities. They then used an inverted index model to index all knowledge entities. During generation, for each topic word appearing in the context, they added corresponding knowledge to the input. They experimented with three methods to integrate knowledge and context.

NetEase Fuxi Lab. They stored all knowledge in a knowledge base and applied heuristic rule-based knowledge intervals. The heuristic rules included:

- **Relation screening:** Calculating commonly used relations based on statistics of triples in the training set.

- **Head entity screening:** Matching head entities from the test set to entities easily confused with common words (e.g., "dao," "yes"), using word frequency statistics of head entity matches in the training set across three knowledge base types (within dialogue units). Additionally, for confusing entity information such as numbers, years, and dates (e.g., "1998"), they applied regular expression filtering.

- **Confusing entity screening:** Some header entities include bracketed explanations in the knowledge base, but bracket content does not appear in dialogues. Other entities appear both with and without parenthetical annotations, referring to different knowledge (e.g., "Recognize it" vs. "Recognize it (Eason Chan Album)"). They created a de-parenthesis dictionary, first matching without parentheses. If no match was found in the knowledge base, they looked up the parenthetically annotated entity from the dictionary.

They encoded knowledge and context using an encoder and applied different attention mechanisms in the decoder to attend to context and knowledge separately, finally combining the outputs.

Soochow University. They used separate knowledge and context encoders, applying KL loss in KG Fusion to help the model learn to select correct knowledge. The knowledge and context selected by KG Fusion were fed to the decoder to learn how to generate knowledge-informed responses. To ensure semantic relevance, they also added reconstruction loss.

Through this evaluation, researchers have begun paying greater attention to few-shot learning and knowledge-driven technologies. Based on participation rates, we find that human-computer dialogue evaluation has attracted extensive attention from both academia and industry.

---

5. CONCLUSION

We successfully organized the fourth Evaluation of Chinese Human-Computer Dialogue Technology. This paper introduced the two evaluation tasks and their corresponding evaluation metrics, detailed the two datasets, and analyzed the evaluation results. We hope our work provides valuable insights for future human-machine dialogue evaluation research.

ACKNOWLEDGEMENTS

We thank the Social Media Processing Committee of the Chinese Information Processing Society of China (CIPS-SMP) for their strong support of this evaluation. We also thank Huawei Technologies Co., Ltd. for financial support, iFLYTEK Co., Ltd. for data and evaluation support, and Kaiyan Zhang and Jiale Zhang for their indispensable assistance during the evaluation. This work is supported by the National Natural Science Foundation of China (No. 62076081, No. 61772153, and No. 61936010).

AUTHOR CONTRIBUTIONS

This work represents a collaboration among all authors. C.H. Zhu (chzhu@ir.hit.edu.cn) designed the overall evaluation framework. W.N. Zhang (wnzhang@ir.hit.edu.cn) served as the leader of 2020-ECDT. W.X. Che (car@ir.hit.edu.cn), Z.G. Chen (zgchen@iflytek.com), M.L. Huang (aihuang@tsinghua.edu.cn), and L.L. Li (lilinlin@huawei.com) guided the evaluation process and summarized the conclusion. Z.X. Feng (zxfeng@ir.hit.edu.cn) compiled the datasets and results of SMP2020-ECDT and drafted the manuscript. All authors contributed meaningfully to revising and proofreading the final manuscript.

DATA AVAILABILITY STATEMENT

All data are available in the Science Data Bank repository, https://doi.org/10.11922/sciencedb.j00104.00091,

REFERENCES

[1] Zhang, W.N., et al.: A Chinese intelligent conversational robot. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pp. 13–18 (2017)
[2] Serban, I.V., et al.: A deep reinforcement learning chatbot. arXiv preprint arXiv: 1709.02349v2 (2017)
[3] Zhang, W.N., et al.: The first evaluation of Chinese human-computer dialogue technology. arXiv preprint arXiv:1709.10217v2 (2017)
[4] Turing, A.M. Computing machinery and intelligence. Mind 59(236), 433–460 (1950)
[5] Wang, X.J., Yuan, C.X.: Recent advances on human-computer dialogue. CAAI Transactions on Intelligence Technology 1(4), 303–312 (2016)

[6] Chen, H.S., et al.: A survey on dialogue systems: Recent advances and new frontiers. arXiv preprint arXiv: 1711.01731 (2017)
[7] Zhang, Y.Z., Zhang, W.N., Liu, T.: Survey of evaluation methods for dialogue systems (in Chinese). SCIENTIA SINICA Informationis 47(8), 953-966 (2017)
[8] Mesnil, G., et al.: Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio Speech Language Processing 23(3), 530–539 (2015)
[9] Yan, R., Zhao, D.Y.: 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 (KDD ' 18), pp. 2574–2583 (2018)
[10] Zhang, W.N., et al.: Neural personalized response generation as domain adaptation. World Wide Web 22, 1427–1446 (2019)
[11] Hou, Y.: Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. arXiv preprint arXiv:2006.05702 (2020)
[12] Zhou, H., et al.: KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7098-7108 (2020)
[13] Feng, Z.X., et al.: Chinese human-computer dialogue technology dataset. Available at: https://doi.org/10.11922/sciencedb.j00104.00091. Accessed 5 February 2021
[14] Tang, B., Kay, S., He, H.B.: Toward optimal feature selection in Naive Bayes for text categorization. arXiv preprint arXiv: 1602.02850 (2016)
[15] Li, J., et al.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
[16] Papineni, K., et al. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
[17] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805v2 (2018)
[18] Sun, Y., et al.: ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223 (2019)
[19] Cui, Y., et al.: Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv:1906.08101 (2019)
[20] Chen, Q., Zhuo, Z., Wang, W.: BERT for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019)
[21] Zhu, S., et al.: Vector projection network for few-shot slot tagging in natural language understanding. arXiv preprint arXiv:2009.09568 (2020)

APPENDIX A: COMPLETE LEADERBOARD

Table A1. The complete leaderboard of Task 1.

Ranking	Participant	Intent acc	Slot F1	Sentence acc
1	AILab-CC, China Merchants Bank			
2	SpeechLab, Shanghai Jiao Tong University			
3	Peking University			
4	MOE Key Laboratory of High Confidence Software Technologies, The Chinese University of Hong Kong			
5	ICRC, Harbin Institute of Technology (Shenzhen)			
6	Laiye Network Technology Co., Ltd.			
7	1STEP.AI			
8	Spoken Dialogue System Lab, South China Agricultural University			

Table A2. The complete leaderboard of Task 2.

Ranking	Participant	Appropriateness	Informativeness	Final
		Film	Music	Travel
1	Suzhou KidX.AI Education Technology Co., Ltd.			

Ranking	Participant	Appropriateness	Informativeness	Final
2	NetEase Fuxi Lab			
3	Soochow University			
4	Laiye Network Technology Co., Ltd.			
5	Ping An Life Insurance Company of China			
6	TMG, Harbin Institute of Technology (Shenzhen)			

AUTHOR BIOGRAPHY

Zixian Feng is a postgraduate student at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. Her current research interests focus on human-computer dialogue system evaluation. ORCID: 0000-0002-6337-7126

Caihai Zhu is a postgraduate student at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His current research interests focus on conversational recommendation. ORCID: 0000-0003-3714-1512

Weinan Zhang is an associate professor at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interests include human-computer dialogue, natural language processing, and information retrieval. ORCID: 0000-0001-5981-4752

Zhigang Chen joined iFLYTEK Corporation in 2003 and is currently Vice President of the AI Research Institute of iFLYTEK Corporation, primarily responsible for cognitive intelligence research and productization.

Wanxiang Che is a professor at the School of Computer Science and Technology, Harbin Institute of Technology (HIT). His main research area is natural language processing (NLP), and he currently leads research projects sponsored by the National Natural Science Foundation of China and the National 973 Project. ORCID: 0000-0002-3907-0335

Minlie Huang is an associate professor at the Department of Computer Science and Technology, Tsinghua University. His research interests include artificial intelligence, deep learning, reinforcement learning, and natural language processing. ORCID: 0000-0001-7111-1849

Linlin Li is the leader of Intelligent Voice Assistant at Huawei Consumer BG, in charge of Huawei Xiaoyi' s voice assistant business. Her work involves collaboration across pan-terminal equipment for voice services, multi-language internationalization, multi-modal semantic understanding, and end-to-end intelligence.

[1] http://smp2020.cips-smp.org/
[2] http://smp2019.cips-smp.org/
[3] https://aiui.xfyun.cn/index-aiui
[4] https://atmahou.github.io/attachments/FewJoint.zip
[5] https://github.com/thu-coai/KdConv

Note: Figure translations are in progress. See original paper for figures.