Overview of the NLPCC 2018 Shared Task: Multi-turn Human-Computer Conversations

Juntao Li[2] and Rui Yan[1,2(✉)]

[1] Institute of Computer Science and Technology (ICST), Peking University, Beijing 100871, China
ruiyan@pku.edu.cn

[2] Institute of Big Data Research, Peking University, Beijing 100871, China
lijuntao@pku.edu.cn

Abstract. In this paper, we give an overview of multi-turn human-computer conversations at NLPCC 2018 shared task. This task consists of two sub-tasks: conversation generation and retrieval with given context. Data-sets for both training and testing are collected from Weibo, where there are 5 million conversation sessions for training and 40,000 non-overlapping conversation sessions for evaluating. Details of the shared task, evaluation metric, and submitted models will be given successively.

Keywords: Multi-turn conversation · Conversation generation
Conversation retrieval · Sequence matching

1 Task Background

Building multi-turn conversation system in open-domain is a research hotpot in academia and draws mounting attention in the industry. In earlier years, researchers mainly design rule-based or template-based systems for task-oriented conversation. With the access to myriad conversation data, it is prompting to develop conversation systems for open-domain. Existing open-domain conversation systems in recent few years can be roughly summarized as two categories: retrieval-based models and generation-based approaches.

Most conversation generation approaches are based upon the sequence to sequence framework [1]. For example, Vinyals and Le propose to utilize such a sequence to sequence model for addressing the issue of conversation generation [2]. However, the simple sequence to sequence model can not adequately model context information in conversations. With limited context information, improper or even unrelated output will be yielded by conversation systems. To address the obstacle of "out of context", context-sensitive conversation generation approach was proposed [3,4]. Serban et al. [5,6] further propose to model conversation context as a hierarchical structure to avoid data sparsity.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 446–451, 2018.
https://doi.org/10.1007/978-3-319-99501-4_42

Besides, specific information in context are extracted such as topic [7], diversity [8,9], and persona [10], for generating conversations.

For retrieval-based conversation systems, it can be constructed as a sequence matching problem by computing the matching degree of candidate responses and user-issued query. Yan et al. [11] treat responses retrieval as a ranking problem through incorporating multi-dimension ranking evidences, where conversation context in a continuous session in multi-turns is also captured. To better extract matching information from conversation context, the sequential matching network is designed and achieves the state-of-art performance [12].

Although substantial progress of multi-turn human-computer conversation has been achieved by either retrieval-based or generation-based systems, there are still room for improvement. Herein the NLPCC 2018 shared task 5 is designed for putting Chinese multi-turn conversation forward. This task consists of two sub-tasks: conversation generation and response retrieval. There are 10 teams targeting at conversation generation and 5 teams for response retrieval.

2 Task Description

2.1 Task Formulation

The multi-turn conversation generation task is formulated as follows: given the previous $n-2$ sentences in an continuous conversation session $X = (x_1, ..., x_{n-2})$ as *context* and the $(n-1)$-th sentence x_{n-1} as *query*, the goal is to generate the *response* x_n. Table 1 gives two conversation sessions used for generation.

Table 1. Two conversation sessions used for generation.

Conversation Sessions	
Context	谢谢你所做的一切
Context	你开心就好
Context	开心
Context	嗯因为你的心里只有学习
Query	某某某，还有你
Response	这个某某某用的好
Context	你们宿舍都是这么厉害的人吗
Query	眼睛特别搞笑，这土也不好捏但是就是觉得挺可爱
Response	特别可爱啊

For the sub-task of response retrieval in multi-turn conversations, it can be defined as follows: given the *context* $X = (x_1, ..., x_{n-2})$ and *query* x_{n-1}, the proper *response* x_n will be retrieved from 10 response candidates $\{c_1, c_2, ..., c_{10}\}$. As illustrated in Table 2, c_4 is the right answer for the given query and will be selected.

Table 2. An example of response retrieval.

Conversation Session	
Context	你们宿舍都是这么厉害的人吗
Query	眼睛特别搞笑，这土也不好捏但是就是觉得挺可爱
Candidate1	佛山有这个地方吗
Candidate2	佛山都是以吃的为准，怎么会有海呢，珠海，潮汕惠州就有
Candidate3	男朋友不是赵阳玩你手机的话那你简直可怕
Candidate4	特别可爱啊
Candidate5	那么什么时候一起游西湖
Candidate6	自己宿舍不给进那是大四认识的学妹宿舍
Candidate7	我今年参加高考，想报你们学校，云南的学生希望大吗?
Candidate8	王小懒小瘦子你票都没买号反正要聚餐啊
Candidate9	我只能说这厕能长到是要积了多少德
Candidate10	你是怎么赔偿的阿

2.2 Datasets

The datasets employed in this task are collected from Sina Weibo, which contain training set and testing set. We clean the datasets by removing emojis and repeated utterances. For conversation generation, there are 5,000,000 conversation sessions in the training set and extra 40,000 conversation sessions in the testing set. Each session contains at least 3 sentences. As for response retrieval, there are 5,000,000 conversation sessions for training and non-overlap 10,000 sessions for testing. Participants are asked to submit at least one system for either one of the sub-tasks or both.

2.3 Evaluation Metric

It is still challenging for evaluating the results of conversation generation. Although automatic evaluation is not aligned with user experience, we still use BLEU score [13] as the evaluation metric for conversation generation insomuch as human evaluation is not affordable for this task. For the sub-task of response retrieval, we use precision of selected candidates as the evaluation metric.

3 Results Statistics

Ultimately, 10 teams submitted their results for the sub-task of conversation generation and 5 teams participated in the response retrieval. As shown in Table 3, the best result for conversation generation is yielded by the system Yiwise-DS, i.e. 16.58. It can be seen that other five systems also achieve presentable results, e.g. G930, Lmll, BD-chatbot. Table 4 shows the evaluation results of response retrieval. The best performance is generated by ECNU and Wyl-buaa achieves a comparable results while other systems have a large space for improvement.

Table 3. Results for conversation generation.

Submitted systems	Score
DialogMind	5.24
BLCU-NLP	0.19
Yiwise-DS	16.58
Jiaoyanqiaowuwang	1.54
Laiye-rocket	12.05
G930	12.85
BD-chatbot	15.51
Phantomgrapes	11.51
Lmll	12.98
ECNU	0.91

Table 4. Results for response retrieval.

Submitted systems	Precision
BLCU-NLP	10.54
Yiwise-DS	26.68
Laiye-rocket	18.13
Wyl-buaa	59.03
ECNU	62.61

4 Representative Systems

In this part, we give a brief analysis of 4 representative systems in NLPCC 2018 shared task 5, where there are 2 systems for conversation generation and response retrieval respectively.

For conversation generation, G930 re-implements the VHRED model [6] or part of KgCVAE [9] on the Weibo Dataset released by NLPCC 2018. The VHRED model utilizes the latent variable for addressing the wording novelty issue of RNNs. The hierarchical encoder of VHRED can effectively take conversation context into account and thus yields relatively good responses. The Lmll model augments HERD [5] with keywords and an attention mechanism for addressing the issue of topic irrelevance in generated responses.

For response retrieval, Wyl-buaa presents a novel RCMN model for addressing the relationship between utterances in context through utilizing the self-matching information in context. To obtain the relevance between response and each utterances in both word-level and sentence-level, the Sequential Matching Network (SMN) [12] is used. The SMN and the RCMN are further combined as an ensemble model and achieve rank second in the shared task of NLPCC 2018. ECNU presents a framework that combines NLP features, SMN, and memory-based matching network (MBMN) for addressing the issue of response selection.

Specifically, the MBMN is designed for learning global context information and important long-distance dependence on the query, while SMN is utilized for modeling sequential relationships of contexts. Inspired by the performance improvement yielded by NLP features, three features are designed in this system. The overall combination of the three parts achieves rank 1st in all teams.

5 Conclusion

In this paper, we present the details of NLPCC 2018 shared task 5, including task formulation, datasets, evaluation metrics, results of submitted systems, and representative systems. This task investigates the performance of various systems on Chinese multi-turn conversations. We release a large corpus which contains more than 5 million multi-turn conversation sessions. There are 10 teams targeting at conversation generation and 5 teams at retrieval. As presented, there is still a long way for both multi-turn conversation generation and response retrieval models. It is expected that there are more and more researchers focusing on multi-turn conversation and moving the state-of-art forward.

References

1. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
2. Vinyals, O., Le, Q.V.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
3. Sordoni, A., et al.: A neural network approach to context-sensitive generation of conversational responses. In: HLT-NAACL, pp. 196–205 (2015)
4. Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D.: How to make context more useful? an empirical study on context-aware neural conversational models. In: ACL, vol. 2, pp. 231–236 (2017)
5. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI, pp. 3776–3784 (2016)
6. Serban, I.V. et al.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: AAAI, pp. 3295–3301 (2017)
7. Xing, C., et al.: Topic aware neural response generation. In: AAAI, pp. 3351–3357 (2017)
8. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: HLT-NAACL, pp. 110–119 (2016)
9. Zhao, T., Zhao, R., Eskenazi, M.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: ACL, vol. 1, pp. 654–664 (2017)
10. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, W.B.: A persona-based neural conversation model. In: ACL, vol. 1, pp. 994–1023 (2016)
11. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: SIGIR, pp. 55–64 (2016)

12. Wu, Y., et al.: Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots. In: ACL, vol. 1, pp. 496–505 (2017)
13. Papineni, K., Roukos, S., Ward, T., et al.: IBM research report bleu: a method for automatic evaluation of machine translation. In: ACL, vol. 2, pp. 311–318 (2002)