Overview of the NLPCC 2019 Shared Task: Open Domain Conversation Evaluation*

Ying Shan, Anqi Cui, Luchen Tan, and Kun Xiong

RSVP.ai, Waterloo, ON, Canada
{yshan, caq, lctan, kun}@rsvp.ai

Abstract. This paper presents an overview of the Open Domain Conversation Evaluation task in NLPCC 2019. The evaluation consists of two sub-tasks: Single-turn conversation and Multi-turn conversation. Each of the reply is judged from four to five dimensions, from syntax, contents to deep semantics. We illustrate the detailed problem definition, evaluation metrics, scoring strategy as well as datasets. We have built our dataset from commercial chatbot logs and public Internet. It covers a variety of 16 topical domains and two non-topical domains. We prepared to annotate all the data by human annotators, however, no teams submit their systems. This may due to the complexity of such conversation systems. Our baseline system achieves a single-round score of 55 out of 100 and a multi-round score of 292 out of 400. This indicates the system is more of an answering system rather than a chatting system. We would expect more participation in the succeeding years.

Keywords: Chatbot · Conversation Systems · Conversation Evaluation

1 Introduction

Natural language conversation as an advanced user interface has created a wide range of applications. Researchers have been working on different approaches to generate natural replies, including retrieval-based, end-to-end generation, question-answering and recommendation systems. We have already seen chatbots all around us, from smart home devices to smart phone assistants, from customer service to chatting. However, there is no standard to evaluate conversations. The quality of conversations varies from different applications and goals, and is sometimes very subjective.

This open problem has addressed much attention among researchers. Typically, conversation evaluation is treated from two sides: automatic scoring and human evaluations. While many are still exploring metrics to reflect conversation quality comprehensively, automatic models and algorithms are also studied in recent years [10].

Inspired from machine translation and summarization, metrics such as BLEU [9], METEOR [1] and ROUGE [5] are usually considered as baselines for conversations. However, they are still less informative and precise to reflect conversation

* Supported by China's National Key R&D Program of China 2018YFB1003202.

quality both qualitatively and quantitatively [6, 7]. Therefore, researchers begin to learn end-to-end scores to evaluate conversations' validity [2], topic coherency and diversity [3].

On the other hand, human annotations are still important and suitable for conversation evaluation. To reduce human subjectivity, the problem is usually narrowed down to either task-oriented or open domain [10]. Crowd-sourcing is also suitable to label a large amount of corpus [4], however, it is still difficult to collect standard annotations across different datasets (movie subtitles [2, 8], switchboard corpus, tweets [7] or chatbot logs [3]).

In NLPCC 2019, we setup a task to evaluate human-computer conversations. All participating systems will be talking with human annotators, live user-in-the-loop. In the task, understanding natural language inputs (which can be questions or statements) is crucial, as well as providing smooth responses. The responses are evaluated from five aspects. We also provide human-annotated real data for researchers, to contribute to the community.

2 Task Description

We consider two scenarios:

2.1 Single-turn Conversation

In this scenario, a set of natural language sentences is given to the participating systems. The systems should provide corresponding relies for each sentence just as human conversation.

2.2 Multi-turn Conversation

In this scenario, we begin with an initial sentence. Human testers will interact with participating systems manually.

3 Evaluation

Both the scenario tasks are designed to be evaluated by human assessors.

3.1 Single-turn Conversation

We define five aspects used in the evaluation of participating systems:

1. Syntax: Correctness and smoothness of syntax.
2. Content Expression: Clear content without ambiguity. Appropriate amount of information. Esp. no inappropriate (violence, sexual, sensitive) content is allowed.
3. Emotional Expression: Subjective attitude or obvious moods. Causes mood changes (becoming glad or sad).

4. Topic Divergence: Mentioning new topics or entities, causing successive turns.
5. Contextual Association: Following the same topic from context, content or entities.

Some examples are illustrated in Table 1.

Table 1. Single-turn Conversation Evaluation Aspects and Examples.

Aspects	Good Cases	Bad Cases
Syntax	今天天气不错	今天天气错不
	挺好的	好的挺
Content Expression	老虎有四条腿	嗯嗯（言之无物）
		跳楼吧，像你这样我早就跳楼了（内容消极）
Emotional Expression	我好开心。	今天天气不错。
	天哪，好疼啊！	好疼。
	问：你喜欢我吗?	
	答：我最喜欢你啦，么么哒～	答：嗯。
Topic Divergence	问：今天好冷啊！	
	答：咱们去吃火锅吧。	答：是啊，好冷。
	问：你叫什么啊?	
	答：我叫张三，那你呢?	答：我叫张三。
Contextual Association	问：你喜欢什么颜色?	
	答：红色。	答：苹果（不关联）
	问：你连上网了吗?	
	答：着啥急?	答：然后呢？（不自然）

Each aspect is judged by asking human assessors yes/no questions, scoring 1/0 respectively. Each reply will be judged by three human annotators separately.

For example, the Emotional Expression aspect has two evaluation metrics: (1) If the response has subjective attitude or obvious moods, earns one point. (2) If it causes changing of moods, earns one point. For a total of 200 test cases, with three annotators, the full score of Emotional Expression is $200 \times (1+1) \times 3 = 1200$ points. The participant's actual score (ranged between 0 and 1200) is then linearly converted to a max score of 100.

The overall score is the sum of scores from five aspects, a max of 500. We will rank the participants according to this score. In addition, we will also rank individual aspects, since different applications may focus on only a part of these aspects.

3.2 Multi-turn Conversation

The evaluation of multi-turn conversations consists of two categories. Each category contains two factors (with their scores shown below):

1. Single Turn Evaluation:
   (a) Logical Association (max 2 per turn): The association between question and response. Please refer to "Contextual Association" in Table 1.
   (b) Conversation Trigger (max 2 per turn): Whether or not the response could trigger another turn. Please refer to "Topic Divergence" in Table 1.
2. Multi-turn Evaluation:
   (a) Total Turns (2 per turn): Number of turns of this conversation (a question-answer pair is defined as one turn).
   (b) Total Topical Turns (2 per turn): Number of turns that have the same topic with the initial sentence.

During the testing, human testers will interact with participating systems. When the conversation ends (e.g. responding "OK.") or after the fifth turn has finished, the testers will stop. Annotators will label the whole conversations.

The overall score is the sum of all four aspects, at most $2 \times 4 \times 5 = 40$ points per topic.

4 Dataset

The dataset is adopted from commercial chatbot logs and public Internet social media conversations.

We classified them into 16 topical domains and two non-topical domains. For each topical domain, we selected 100 sentences and for the two non-topical domains, we selected 100 sentences altogether. In total, there are $1,700$ sentences.

Before the evaluation, a sample conversation set (200 sentences and replies) is provided. The dataset contains the following columns:

- Column A: Input question (sentence).
- Column B: Sample reply.
- Column D: Number of annotators.
- Columns E – O: How many annotators agree on that metric for this reply.
- The last two lines (rows) of the file is an overall statistics on this dataset (200×3). Similarly, we will also evaluate participants' systems with this method.

The replies are provided by a baseline conversation system by *rsvp.ai*. Along with the sentence/reply pairs, human annotations of the replies are provided as well. When the evaluation begins, 500 sentences are used as our testing dataset. For the multi-turn evaluation, we only test with 20 (initial) sentences. The remaining sentences are posted for research purpose at the end of this evaluation, downloadable at https://github.com/RSVP-Technologies/nlpcc2019-conversation.

Considering the difficulty of open domain conversations, participants can use external resources to train or build their own conversation systems.

5 Evaluation Results

At the beginning of this task, a total of 18 teams registered for subtask Single-turn conversation and 19 teams for subtask Multi-turn conversation. About 15% teams are from companies and the rest are from colleges or institutions.

Unfortunately, none of the teams submit their system (API) at the end of the evaluation. Instead of showing the results of participants, here we list the scores from our baseline system:

5.1 Single-turn Conversation

1. Syntax: To achieve a high Syntax score, the reply should not contain any offensive words. It also needs to be clear without ambiguity. The baseline system earns a point of $997/1200 = 83\%$.
2. Content Expression: We examine the reply with natural (not too formal) and appropriate amount of information (not too much nor too little). Our system has a point of $690/1200 = 58\%$.
3. Emotional Expression: Reply with obvious emotional expression, and leads reader to be happy / sad, for our system, $288/1200 = 24\%$.
4. Topic Divergence: The reply could motivate readers with more entities or more rounds of conversations: $538/1200 = 45\%$.
5. Contextual Association: The topic or entities continue in the reply. Our system has a point of $761/1200 = 63\%$.

Total: $83 + 58 + 24 + 45 + 63 = 273$ out of 500: 55%.

5.2 Multi-turn Conversation

We ask three experts to annotate our baseline system. Out of 20 initial sentences, our system generally interacts with people about 2.6 turns per seed, mostly under the same topic. The highest score of the interaction is 34/40. In total, the baseline system has an average score of 292 out of $40 \times 20 = 800$.

6 Conclusion

This paper briefly presents an overview of the Open Domain Conversation Evaluation task in NLPCC 2019. Detailed problem definition and evaluation design are introduced with samples. Although no participants submit their final results, we see this a pivot organization in conversation evaluations.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)

2. Bruni, E., Fernandez, R.: Adversarial evaluation for open-domain dialogue generation. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 284–288 (2017)
3. Guo, F., Metallinou, A., Khatri, C., Raju, A., Venkatesh, A., Ram, A.: Topic-based evaluation for conversational bots. arXiv preprint arXiv:1801.03622 (2018)
4. Jurčíček, F., Keizer, S., Gašić, M., Mairesse, F., Thomson, B., Yu, K., Young, S.: Real user evaluation of spoken dialogue systems using amazon mechanical turk. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
5. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
6. Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2122–2132 (2016)
7. Lowe, R., Noseworthy, M., Serban, I.V., Angelard-Gontier, N., Bengio, Y., Pineau, J.: Towards an automatic turing test: Learning to evaluate dialogue responses. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1116–1126 (2017)
8. Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: On the evaluation of dialogue systems with next utterance classification. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 264–269 (2016)
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
10. 张伟男, 张杨子, 刘挺: 对话系统评价方法综述. Chinese Science Bulletin **57**, 3409 (2012)