NLPCC2019 Shared Task Guideline:
Open Domain Conversation Evaluation

RSVP.ai

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.

In NLPCC2019, 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 will be evaluated from five aspects. We will also provide human-annotated real data for researchers, to contribute to the community.

1. Task Description

We consider to scenarios:

Scenario A. Single-turn Conversation.

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

Scenario B. Multi-turn Conversation.

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

2. Evaluation

Both scenario A and B tasks will be evaluated by human assessors.

Scenario A. Single-turn Conversation

We define five aspects used in the evaluation of scenario A participating systems: Syntax, content expression, emotional expression, topic divergence and contextual association, as shown in Table 1.

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

Aspects	Explanation	Good Cases	Bad Cases
Syntax	Correctness and smoothness of syntax.	今天天气不错	今天天气错不
		挺好的	好的挺
Content Expression	Clear content without ambiguity. Appropriate amount of information. Esp. no inappropriate (violence, sexual, sensitive) content is allowed.	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	Subjective attitude or obvious moods. Causes mood changes (becoming glad or sad).	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～	问：你喜欢我吗？答：嗯。
Topic Divergence	Mentioning new topics or entities, causing successive turns.	问：今天好冷啊！答：咱们去吃火锅吧。	问：今天好冷啊！答：是啊，好冷。
		问：你叫什么啊？答：我叫张三，那你呢？	问：你叫什么啊？答：我叫张三。
Contextual Association	Following the same topic from context, content or entities.	问：你喜欢什么颜色？答：红色。	问：你喜欢什么颜色？答：苹果（不关联）
		问：你连上网了吗？答：着啥急？	问：你连上网了吗？答：然后呢？（不自然）

Note: More examples and explanations will be provided at the time of release.

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 * (1 + 1) * 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.

Scenario B. Multi-turn Conversation

The evaluation of multi-turn conversations consists of two categories, as shown in Table 2.

Table 2 Multi-turn Conversation Evaluation Aspects

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

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*4*5 = 40 points.

3. Dataset

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

We classified them into 16 topical domains and 2 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 1700 sentences.

Before the evaluation, a sample conversation set (200 sentences and replies) will be provided. 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 will be used as our testing dataset. For the multi-turn evaluation, we will only test with 20 (initial) sentences. The remaining sentences will be posted for research purpose at the end of this evaluation.

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

The sample dataset consists of:

- Column A: Input question (sentence).
- Column B: Sample reply (by rsvp.ai)
- 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.

Please note that:

1. The data is owned by rsvp.ai.
2. The data is used for research purpose only, within the scope of this NLPCC evaluation.
3. We reserve the rights of legal and copyrights of the data.

4. Contact

For more information, please contact yshan@rsvp.ai.

NLPCC2019 开放领域对话质量评测 邀请函

薄言 RSVP.ai

近年来，自然语言对话作为一种新的交互方式，得到广泛应用。研究者们探索了基于检索、端到端生成、问答、推荐等多种算法的对话框架，并在智能家居、情感闲聊、客服、手机助手等场景以及教育、医疗、汽车等行业中投入使用。然而，对话质量的评价仍然没有统一标准，对话的好坏随场景、目标的不同，有较大的差别和主观性。

在本届 NLPCC2019 会议中，薄言信息技术有限公司（RSVP.ai）设立了开放领域对话质量评测任务。针对开放领域对话和真实的应用需求，本任务提出了多维度的评价体系，涉及句法分析、命名实体识别、主题/意图分类、情感分析、消歧等多项基础自然语言处理技术。评价体系在兼容对话主观性的前提下，尽量采用量化或明确的评分标准，减少标注员的主观分歧，使对话质量具有可比性。此外，本任务将提供真实对话数据集（中文），并通过人工标注和测试，为研究者提供真实评价数据，推动学术交流和对话应用的发展。

一、任务描述

子任务一：单轮对话，即输入一个自然语句（提问或闲聊），要求参赛系统输出对应的对话回复。

子任务二：多轮对话。从一个初始句开始，由参赛系统与测试员对话，围绕同一主题、或发散至其他主题进行多轮对话。

二、评价标准

单轮和多轮的对话质量评价均由人工标注员完成。

1. 单轮对话

单轮对话质量从 5 个维度进行评价，如表 1 所示：

表 1 单轮对话质量评价维度及示例

维度	说明	好的示例	不好的示例
语言形式	是否有语法错误、是否自然规范等。	今天天气不错	今天天气错不
		挺好的	好的挺
内容表达	内容是否明确、信息量适当；特别地，没有色情、暴力等敏感或消极内容。	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）

情感表达	是否有主观态度或明显的情绪，是否会引发开心或难过等心情变化。	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～	问：你喜欢我吗？答：嗯。
话题发散	是否发散到新的主题或实体，引发新一轮对话。	问：今天好冷啊！答：咱们去吃火锅吧。	问：今天好冷啊！答：是啊，好冷。
		问：你叫什么啊？答：我叫张三，那你呢？	问：你叫什么啊？答：我叫张三。
上下文关联	是否延续上文主题，是否有相关的内容或实体。	问：你喜欢什么颜色？答：红色。	问:你喜欢什么颜色？答：苹果（不关联）
		问：你连上网了吗？答：着啥急？	问：你连上网了吗？答:然后呢？（不自然）

注：更多示例和说明将在任务正式开放时发布。

对于每个维度，评测员将对参赛队伍返回的全部结果进行“是/否”的打分，并根据满足“是”的百分比给出得分。例如：

情感表达维度共有 2 个指标。对于一个回复，有主观态度或明显情绪表达，得 1 分；能够引发开心或难过等心情变化，得 1 分。一共有 200 个测试例，由 3 位测试员标注，满分 200×（1+1）×3=1200 分。对于参赛队伍的实际结果（0～1200 分），按线性折合为百分制。

综合评价将 5 个维度的总分相加（满分 500 分）进行排名。此外，考虑到参赛队伍对话系统的应用场景不尽相同，我们还将提供各个维度得分（满分 100 分）的排名供参考。

2. 多轮对话

多轮对话的质量评价从两个角度出发：（1）单轮对话的质量；（2）持续多轮的数量。如表 2 所示。

表 2 多轮对话质量评价维度

类别	维度	说明	得分
单轮对话质量	逻辑关联	每一轮内回复与提问的关联程度，参考单轮对话的“上下文关联”。	每轮最高 2 分。
	对话牵引	每一轮内回复是否能引发下一轮对话，参考单轮对话的“话题发散”。	每轮最高 2 分。

持续多轮数量	对话总轮数	对话进行的总轮数（一问一答为 1 轮）。	每轮得 2 分。
	同主题对话轮数	与初始话题相同的对话轮数。	每轮得 2 分。

在测试时，先由测试员人工与参赛系统进行对话。对话终止（例如系统回复：“哦。”）或第 5 轮对话完成后，测试员停止测试，由标注员进行标注。

综合评分取 4 项指标相加。最高为 2×4×5=40 分。

三、数据集

评测所用问题集（即输入，不一定是疑问句）来自真实用户对话日志和互联网社交媒体公开数据（亿级）。我们从互联网对话数据提取了 16 个主题和 2 个无主题类别。每个主题随机选取 50 个问题，另 2 个无主题类别一共选取 100 个问题，总计 17 类共 1700 个问题，作为评测的完整数据集。

在比赛开始时，我们将从中随机选取 200 个样本问题，并提供示例回复（由 rsvp.ai 对话系统生成）。这个回复可视做基线系统。剩余 1500 个问题，其中 500 个问题用于测试，由参赛队伍提交其机器系统生成的回复。对于多轮对话，考虑到人工测试的工作量，我们只选取 20 个问题用作测试，这些问题作为初始第一句话，供人机交互。其余问题将在比赛最后放出供研究使用。

考虑到对话任务的开放性，参赛队伍可以使用其他资源用于机器自动生成回复。

示例的样本数据，我们也提供了标注，供参赛队伍理解相关指标。数据字段包括：

- A 列：测试问题；
- B 列：系统返回答案（供参考）；
- D 列：该问题标注人数；
- E～O 列：每一项评价指标的认可人数。例如，某问题的回答共 3 个人标注，在某指标上只有 1 个人认可（，说明不是很符合该指标）；
- 数据最后两行是该（样本）数据集的整体统计（200 题*3 人标注），供参考。最终评测结果与之一致。

数据使用说明：

（1）评测相关数据为 rsvp.ai 所有；

（2）该数据集只能在 NLPCC 会议相关评测中使用。未经许可，请勿传播和分享数据；

（3）如果未经允许，数据被传播或公开，我司有权追究法律责任。

四、联系方式

更多信息请联系 yshan@rsvp.ai。