COMPETITION ON KNOWLEDGE DRIVEN DIALOGUE

A PREPRINT

Baidu NLP
Baidu Inc.,
Beijing, China

March 10, 2019

ABSTRACT

Human-machine conversation is one of the most important topics in artificial intelligence (AI) and has received much attention across academia and industry in recent years. Currently dialogue system is still in its infancy, which usually converses passively and utters their words more as a matter of response rather than on their own initiatives, which is different from human-human conversation. Therefore, we set up this competition on a new conversation task, named **knowledge driven dialogue**, where machines converse with humans based on a built knowledge graph. It aims at testing machines ability to conduct human-like conversations.

Keywords Dialogue · Competition · Knowledge

1 Motivation

Building a human-like conversational agent is one long-cherished goal in Artificial Intelligence (AI) [1]. Various kinds of conversational agents have been proposed during the past years, from the handcrafted rule based systems [2] to the neural dialogue systems that purely driven by dialogue corpus [3, 4]. Although great progress has been made, currently the dialogue system is still in its infancy: it usually converses passively and utters their words more as a matter of response rather than on their own initiatives, which is different from human-human conversation.

Our preliminary investigation suggests that the major challenge, for existing dialogue systems, to build human-like dialogue systems is to endow the machine the ability of proactively leading the conversation, such as introducing a new dialogue topic or maintaining the current topic on purpose. Motivated by this investigation, in this paper, we set up a competition on the new conversation task: the **knowledge driven dialogue**. Our assumption is that knowledge is the key to build human-like conversational agents, a bot without knowledge can never truly control the dialogue as human beings.

Practically, our knowledge driven dialogue task focuses on building fully data-driven systems, which is able to naturally shift the conversation topic to arbitrary one by exploiting knowledge. Unlike existing knowledge grounded dialogue tasks [5, 6], our knowledge driven dialogue task gives each competing system a sequence of topics as the dialogue goal and asks those competing systems to conduct human-like conversations following the given topic sequence. A set of goal-related background knowledge is also provided for each competitor for naturally and coherently control the shifting of dialogue topics.

Moreover, we create a large-scale human-human corpus, which is grounded with related knowledge , to facilitate training and evaluation. Our dataset includes information of Dialogue Goal, Background Knowledge and Conversation. Each conversation is generated by two annotators, one of which plays the agent role and the other plays the user role. The agent was asked to lead the conversation with the given knowledge to achieve the setting goal, and the user just needs to talk without any given information.

To test the usability of our dataset, we propose a knowledge-aware neural dialogue generator and train it with our dataset as benchmarks. Experimental results demonstrate that while the task of leading the conversation to given topics is very challenging, the dialogue generation quality and controllability can be significantly improved by leveraging our crowdsourced dialogue corpus as well as its related background knowledge.

2 Task

Given a dialogue goal g and a set of topic-related background knowledge $K = k_1, k_2, ..., k_n$, a participating system is expected to output an utterance "x_t for the current conversation $X = x_1, x_2, ..., x_{t-1}$, which keeps the conversation coherent and informative under the guidance of the given goal. During the dialogue, a participating system is required to proactively lead the conversation from one topic to another. The dialog goal g is given like this: "**START - > TOPIC_A - > TOPIC_B**", which means the machine should lead the conversation from any start state to topic A and then to topic B. The given background knowledge includes knowledge related to topic A and topic B, and the relations between these two topics.

3 Dataset

The background knowledge provided in the dataset is collected from the domain of movies and stars, including information such as box offices, directors, reviews and etc., organized as triples {entity, property, value}. The topics given in the dialogue goal are entities, i.e., movies or stars. The data set includes 30k sessions, about 120k dialogue turns, of which 100k are training set, 10k are development set and 10k are test set. Each conversation is generated by two annotators, one of which plays the agent role and the other plays the user role. The agent was asked to lead the conversation with the given knowledge to achieve the setting goal, and the user just needs to talk without any given information. The agent starts the conversion and talks with the user. The data set includes:

- Part of the Training Data: 400 dialogue turns.
- All Training Data: 100k dialogue turns.
- Development Data: 10k dialogue turns.
- Testing Data 1: 5k sample. After submitting your results, you can see the rankings on the leaderboard.
- Testing Data 2: 10k sample. After submitting your results, you can see rankings on the leaderboard, which is the basis for human evaluation.

{"goal": [["START", "阳光灿烂的日子", "王朔"],
["王朔", "代表作", "阳光灿烂的日子"]],
"knowledge": [["阳光灿烂的日子", "时光网 短评", "70 年代 少年 人 的 成长 经历 ， 太 过 真实 ， 再回首 至于 刺眼 的 日光 灼 目"],
["阳光灿烂的日子", "主演", "宁静"],
["阳光灿烂的日子", "上映 时间", "1994年9月9日"],
["阳光灿烂的日子", "类型", "剧情"],
["阳光灿烂的日子", "领域", "电影"],
["王朔", "评论", "才华横溢 ！ "],
["王朔", "毕业 院校", "北京四十四中学"],
["王朔", "主要 成就", "第53届洛迦诺国际电影节主 竞赛 单元 - 金豹奖"],
["王朔", "性别", "男"],
["王朔", "职业", "编剧"],
["王朔", "领域", "明星"],
["阳光灿烂的日子", "是否 上映", "已 上映"],
["阳光灿烂的日子", "时光网 短评", "有点 西西里 的 感觉 。"],
["阳光灿烂的日子", "时光网 评分", "8.5"],
["阳光灿烂的日子", "导演", "姜文"]],
"conversation": ["我 发现 姜文 的 电影 产量 不 高 ， 但是 质量 都 挺 高 的 。",
"同感 ， 那 你 觉得 你 印象 最深 的 一部 姜文 的 作品 是 什么 ？ ",
"阳光灿烂的日子 吧 ， 有点 西西里 的 感觉 。",
"我 也 觉得 这部 电影 不错 ！ ",
"嗯 呀 ， 它 是 一个 年代 的 缩影 吧 。",
"对 呀 ， 可能 姜文 只是 把 他 自己 经历 的 给 拍 了 出来 吧 。",
"但是 里面 那位 主演 真 的 是 才华横溢 。",
"你 说 的 是 哪 一位 ？ "
"王朔 啊 ， 是 北京四十四中学 毕业 的 那位 。"]}

(a). One training example

{"goal": [["START", "卓别林", "小罗伯特 · 唐尼"],
["卓别林", "主演", "小罗伯特 · 唐尼"]],
"knowledge": [["卓别林", "描述 标签", "英式 口音"],
["卓别林", "时光网 短评", "太 像 啦 ！ ！ ！ "],
["卓别林", "口碑", "口碑 不错 的 喜剧 电影"],
["卓别林", "类型", "传记"],
["卓别林", "领域", "电影"],
["卓别林", "主演", "小罗伯特 · 唐尼"],
["小罗伯特 · 唐尼", "评论", "可 帅可萌可骚"],
["小罗伯特 · 唐尼", "搭档", "乔恩 · 费儒"],
["小罗伯特 · 唐尼", "评分", "9"],
["小罗伯特 · 唐尼", "性别", "男"],
["小罗伯特 · 唐尼", "职业", "制作人"],
["小罗伯特 · 唐尼", "领域", "明星"],
["小罗伯特 · 唐尼", "家人", "罗伯特 · 唐尼"],
["卓别林", "时光网 短评", "骚 大叔 最好 的 一次 表演"], ["卓别林", "获奖", "奥斯卡金像奖 _（ 1993 ； 第65届 ） _ 提名 _ 奥斯卡奖 - 最佳配乐 _ 约翰 · 巴里 John Barry"],
["卓别林", "时光网 评分", "8.4"],
["卓别林", "类型", "喜剧"]],
"history": ["你 觉得 把 喜剧 演 的 最 出神 的 人 是 谁 。",
"喜剧 么 ？ 有 好多 喜剧演员 啊 ， 这 怎么 说 得 清 。",
"那 你 认为 卓别林 怎么样 呢 ？ ",
"哇塞 ， 大佬 啊 ， 我 不敢 评价 。",
"那 你 觉得 如果 有 一部 电影 完美 的 呈现 了 卓别林 的 一生 ， 你 会 想 看 么 ？ ",
"当然 啦 ， 这 才是 一个 很好 的 了解 大佬 的 一生 的 机会 啊 。",
"这部 剧 的 主演 我很喜欢 ， 他 的 搭档 是 乔恩 · 费儒 ， 你 知道 是 谁 吗 ？ ",
“这 还 真 不知道 呢 ， 我 猜 他 肯定 很 有名 。"],
“response": "他 叫 小罗伯特 · 唐尼 ， 评论 说 他 可 帅可萌可骚 。"}

(b). One test example

Figure 1: Examples in our dataset.

The training set and the development set are organized in the form of session. Each session includes Dialogue Goal, Background Knowledge and Conversation. The test set is organized in samples. Each sample includes Dialogue Goal, Background Knowledge and History, the participating model is required to lead the conversation according to the current dialogue history, that is, it only needs to simulate the actions of the agent. The various parts of the data are described below:

- Dialogue Goal (goal): It contains two lines: the first contains the given dialogue path i.e., ["Start", TOPIC_A, TOPIC_B]. The second line contains the relationship of TOPIC_A and TOPIC_B.
- Knowledge: Background knowledge related to TOPIC_A and TOPIC_B.
- Conversation: 4 to 8 turns of conversation.
- Dialogue History: Conversation sequences before the current utterance, empty if the current utterance is in the start of the conversion

Figure 1 presents an example from training/test sets respectively.

4 Method

To test the usability of our dataset, we propose a generation-based neural dialogue generator, akin to the work of Lian et al., [7]. Figure 2 demonstrates its structure, which is comprised of three major parts: the **Encoder**, the **Knowledge Manager** and the **Decoder**.

Let X and Y represent the dialogue context (multi-turn) and the true replies generated by our annotators respectively. The **Encoder** encodes the context X into a vector **x**, and feeds it into the

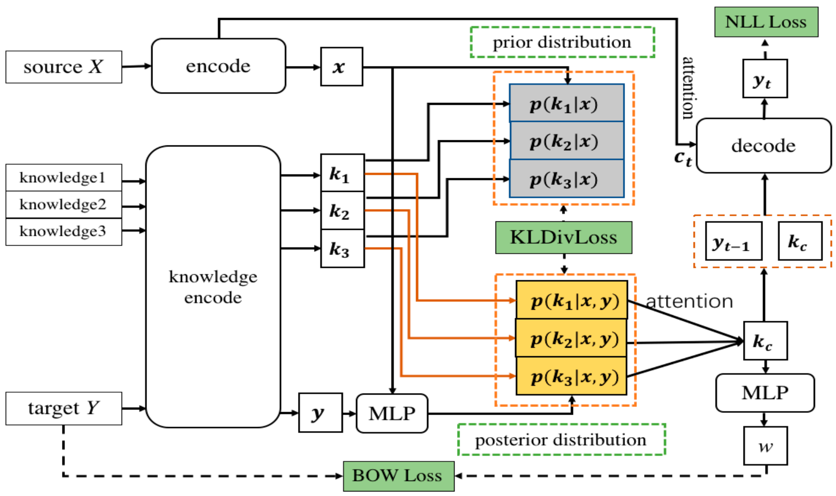

Figure 2: Our benchmark model.

knowledge manager. Practically, it leverages a bi-directional GRU [] to read the context from left to right (right to left) and produce the vector representation for X using the concatenation of the two last hidden states from those two directions. Each knowledge k_i is also encoded in a similar way as the context X. It is worthy noticing that the true response Y is also feed into the encoder during training, namely $\mathbf{y}$.

The representations of context X, knowledge K and true response Y (only in training) are then fed to the **Knowledge Manager**. The responsibility of Knowledge Manager is to select related background knowledge in order to naturally lead the conversation. Specifically, during training phase, two probabilities are estimated in the Knowledge Manager, i.e., (1) the true knowledge reasoning distribution $P(k_j|X,Y)$ and (2) the learned knowledge reasoning distribution $P(k_j|X)$, those two distributions are defined as:

$$P(k_j|X,Y) = \frac{exp(\mathbf{k}_j \cdot MLP([\mathbf{x};\mathbf{y}]))}{\sum_{j=1}^{N} exp(\mathbf{k}_i \cdot MLP([\mathbf{x};\mathbf{y}]))} \tag{1}$$

$$P(k_j|X) = \frac{exp(\mathbf{k}_j \cdot \mathbf{x})}{\sum_{i=1}^{N} exp(\mathbf{k}_i \cdot \mathbf{x})} \tag{2}$$

the goal of knowledge manager is to learn to utilize knowledge in the way that human does. To this end, we introduce an auxiliary loss, namely the KullbackLeibler divergence loss (KLDivLoss), to measure the proximity between $P(k_j|X,Y)$ and $P(k_j|X)$, formulated as:

$$L_{KL}(\theta) = -\frac{1}{N}\sum_{j=1}^{N} P(k_j|X,Y) log \frac{P(k_j|X,Y)}{P(k_j|X)} \tag{3}$$

The decoder is implemented with the **Hierarchical Gated Fusion Unit** described in the work of Lian et al., which is a standard GRU based decoder enhanced with external knowledge gates, we strongly recommend readers refer to their work for more technological information [7].

Our generative baseline has three training objects. Besides the KLDivLoss defined in Equation (3), it also has:

NLL Loss. The objective of NLL loss is to quantify the difference between the true response and the response generated by our baseline. It minimize the Negative Log-Likelihood (NLL) :

$$L_{NLL}(\theta) = -\frac{1}{m}\sum_{t=1}^{m} P_\theta(Y_t|Y_{<t}, X, K) \tag{4}$$

where Y_t denotes the tth word in response Y and $Y_{<t}$ denotes the sub-text from the sentence beginning to the $y - t$ th word.

BOW Loss The BOW loss [8] is designed to ensure the accuracy of the fused knowledge k by enforcing the relevancy between the knowledge and the target response. Specifically, let $w = MLP(k) \in R|V|$ where $|V|$ is the vocabulary size, and we define:

$$P(r_t|k) = \frac{exp(w_{r_t})}{\sum exp(w_v)} \tag{5}$$

Then, the BOW loss is defined to minimize:

$$L_{BOW}(\theta) = -\frac{1}{m}\sum_{t=0}^{m} logP(r_t|k) \tag{6}$$

In summary, the final loss of our generative model is:

$$L(\theta) = L_{KL}(\theta) + L_{NLL}(\theta) + L_{BOW}(\theta) \tag{7}$$

5 Evaluation

The evaluation of our dialogue competition has two phases, i.e., the **automatic evaluation** and the **human evaluation**.

5.1 Automatic Evaluation

Similar to most existing works, we apply several most widely used metrics for automatic evaluation, including:

- F1: char-based F-score of output responses against golden responses, the main metric for dialogue systems.
- BLEU: word-based precision of output responses against golden responses, the auxiliary metric for dialogue systems.
- DISTINCT: diversity of the output responses, the auxiliary metric for dialogue systems.

Based on the evaluation results, we will rank all systems on the leaderboard.

5.2 Human Evaluation

The top 10 models on the leaderboard will be evaluated by humans. Each model talks with a volunteer and leads the conversation given conversation goals and the related knowledge. The generated dialogues will be evaluated by humans on criteria of coherence and goal completion.

Coherence: measures the overall fluency of the whole dialogue, it has four levels:

- score 0 (bad): over 2 responses irrelevant or logically contradictory to the previous context.

- score 1 (fair): only 2 bot responses irrelevant or logically contradictory to the previous context.
- score 2 (good): only 1 response irrelevant or logically contradictory to the previous context.
- score 3 (perfect): no response irrelevant or logically contradictory to the previous context.

Goal Completion: measures how good the given conversation goal is finished, it has three levels:

- score 0 (bad): do not mention “topic_a" or “topic_b" following the given sequence.
- score 1 (fair): mention “topic_a" or “topic_b", but the whole dialogue is very boring and using less than 2 different knowledge triplets during topic exchanging.
- score 2 (good): mention “topic_a" or “topic_b" and use equal to or more than 2 different knowledge triplets during topic exchanging.

The final rankings and winners will be determined based on the human evaluation results

References

[1] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[2] Joseph Weizenbaum. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[3] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[4] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[5] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205*, 2018.

[6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

[7] Fan Wang Jinhua Peng Hua Wu Rongzhong Lian, Min Xie. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*, 2019.

[8] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.