How should we call it? - Introducing the PhotoBook Conversation Task and Dataset for Training Natural Referring Expression Generation in Artificial Dialogue Agents

J. Haber

Abstract

The past few years have seen an immense interest in developing and training computational agents for visually-grounded dialogue, the task of using natural language to communicate about visual input. While the resulting dialogue agents often already achieve reasonable performance on their respective task, none of the models can produce consistent and efficient outputs during a multi-turn conversation. We argue that this is primarily due to the fact that they cannot properly utilise the dialogue history. Human interlocutors on the other hand are shown to collaboratively establish a shared repository of mutual information during a conversation. This common ground then is used to optimise understanding and communication efficiency. We therefore propose that implementing a similar representation of dialogue context for computational dialogue agents is a pivotal next step in improving the quality of their dialogue output.

One of the main reasons why current research seems to eschew modelling common ground is that common ground is a conversation model concept that cannot be assessed directly in actual conversations. In order to address this problem, we propose to first investigate the generation of referring expressions: Being an indirect representation of a referent object, they too are not absolute, but conventions established with a specific conversation partner - based on the common ground established so far. By tracking the development of referring expressions during a conversation we therefore obtain a proxy of the underlying processes in the emerging common ground.

In order to develop an artificial dialogue agent that can utilise the conversation’s common ground, we propose to implement a data-driven, modular agent architecture in an end-to-end training framework. With this setup, the dialogue agent is expected to learn the correct usage of referring expressions from recorded dialogue data directly and can be evaluated on downstream task performance. Opting for this approach however requires a large amount of dedicated dialogue training data that has never been collected before. To initiate this new track in dialogue modelling research, we therefore introduce a novel conversation task called the PhotoBook task that can be used to collect rich, human-human dialogue data for extended, goal-oriented conversations. We use the PhotoBook task to record more than 2,500 dialogues stemming from over 1,500 unique participants on crowd-sourcing platform Amazon Mechanical Turk (AMT). The resulting data contains a total of over 160k utterances, 130k actions and spans a vocabulary of close to 12k unique tokens. An extensive analysis of the data validates that the recorded conversations closely resemble the dialogue characteristics observed in natural human-human conversations. We therefore argue that this data provides a pivotal new repository to be used in further research which has the potential to significantly improve the dialogue output consistency, efficiency and naturalness of artificial dialogue agents.