Instructions for ACL 2023 Proceedings

Anonymous ACL submission

Abstract

TBD

1 Introduction

In many customer-facing dialogue applications, customer service interactions must follow a set of guidelines for safety, which have a natural sequential order. If a customer is locked out of their account and requests a password reset, the agent must first verify that the customer is indeed the owner of the account. This if-then structure is common to flows in guidelines.

Both human and robot agent must follow safety guidelines. As a result, safety guidelines are often written in natural language.

Our goal is to train dialogue agents that not only follow a set of guidelines, but justify their actions by pointing to the guidelines. This allows others to verify their actions, and whether the guidelines have been followed.

We propose a generative model of dialogue, that justifies decisions by aligning to a guidelines, utilizes the sequential structure of guidelines, and does not require supervision.

Experiments show that our model is accurate, intepretable, and works at a range of supervision levels.

Datasets include a variety of guidelines. In ABCD, the guidelines are given to us Chen et al. (2021). In SGD, we write the guidelines ourselves, using the generative model to aid development. In doc2dial, we show that our method works for alignment to general document-guided dialogue as well.

2 Related work

The adaptation of large languge models to task-oriented dialogue has allowed for impressive results in zero-shot generalization, where models are tested in scenarios that they have not previously seen (). The key idea behind this success is the use of a natural language interface: specify scenario-specific details using natural language, and take advantage of the generalization abilities of large language models.

3 Problem setup

We are interested in a generative model of dialogue that justifies its actions by aligning to a document. The model first chooses its alignments $z \sim p(z)$,

$$p(x, z) = p(x \mid z)p(z)$$

References

Derek Chen, Howard Chen, Yi Yang, Alex Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. *CoRR*, abs/2104.00783.

A Example Appendix

This is a section in the appendix.