EANCS 2021

The First Workshop on Evaluations and Assessments of Neural Conversation Systems

Proceedings of the Workshop

Nov 11, 2021
Online

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-00-1

Introduction

Much progress has been made recently to improve conversation systems and chatbots using neural based deep models. While the architectures of conversation models evolve quickly, evaluation techniques over the years remain unchanged. As a matter of fact, evaluating and assessing conversation models has been a decades long challenge. One part of the difficulty comes from the fact that human dialogues exist in a variety of forms. Existing approaches that compare generated conversations from a neural model with ground truth can easily incur large biases due to the lack of diversity. Another part of the difficulty comes from the interactive nature of conversations, which often requires an agent (usually a real person), to conduct the assessments. Human evaluations, on the other hand, can introduce a large amount of variance and are often impractical on a large scale. A third dimension of evaluation has to do with the fairness and reliability of the models, which has become an increasingly important issue for commercial use of neural based systems. Finally, transferability, transparency and ethical issues in evaluations are among some of the other important topics to explore.

Organizing Committee

Wei Wei, Google Cloud AI Research
Bo Dai, Google Brain
Tuo Zhao, Georgia Institute of Technology
Lihong Li, Amazon
Diyi Yang, Georgia Institute of Technology
Yun-Nung Chen, National Taiwan University
Y-Lan Boureau, Facebook AI Research
Asli Celikyilmaz, Microsoft Research
Alborz Geramifard, Facebook AI Research
Aman Ahuja, Virginia Tech University
Haoming Jiang, Amazon Search

Table of Contents

Conference Program

Counterfactual Matters: Intrinsic Probing For Dialogue State Tracking
Yi Huang, Junlan Feng, Xiaoting Wu and Xiaoyu Du

GCDF1: A Goal- and Context- Driven F-Score for Evaluating User Models
Alexandru Coca, Bo-Hsiang Tseng and Bill Byrne

Unsupervised Testing of NLU models with Multiple Views
Radhika Arava, Matthew Trager, Boya Yu and Mohamed Abdelhady

A Comprehensive Assessment of Dialog Evaluation Metrics
Yi-Ting Yeh, Maxine Eskenazi and Shikib Mehri

User Response and Sentiment Prediction for Automatic Dialogue Evaluation
Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu and Dilek Hakkani-Tur

SERI: Generative Chatbot Framework for Cybergrooming Prevention
Pei Wang, Zhen Guo, Lifu Huang and Jin-Hee Cho