Evaluation of Chatbot dialog system

Sourav Dutta

14 February 2019

What to evaluate?

Evaluation Metrics

- Conversational User Experience (CUX)
  - Expectation
  - Behavior
  - Sentiment
  - Trust
- Engagement
- Coherence
- Domain Coverage
- Conversation Depth

[Venkatesh *et al.* , 2018]

How to evaluate?

- Generally evaluated manually (not automatic but needed!)
- **Word-Overlap**
  - **BLEU** [Papineni *et al.* , 2002] – **use this!**
  - Comparing chatbot response to human response
  - Many possible human responses, not feasible enough
- **Automatic Evaluation** (*Neural Evaluation*)
  - **ADEM** [Lowe *et al.* , 2017], **RUBER** [Tao *et al.* , 2017]
  - Trained on labelled human responses
  - Mainly for single-turn response, but maybe extended
- **Adversarial Evaluation** [Kannan & Vinyals, 2017]
  - Train a "Turing-like" evaluator classifier
  - The more it "fools" the evaluator, the better!

References I


Kannan, Anjuli, & Vinyals, Oriol. 2017.
Adversarial Evaluation of Dialogue Models.
CoRR, **abs/1701.08198**.


Lowe, Ryan, Noseworthy, Michael, Serban, Iulian Vlad, Angelard-Gontier, Nicolas, Bengio, Yoshua, & Pineau, Joelle. 2017.
Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses.
CoRR, **abs/1708.07149**.


Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. 2002.
BLEU: A Method for Automatic Evaluation of Machine Translation.
Pages 311–318 of: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

References II

ACL '02.
Stroudsburg, PA, USA: Association for Computational Linguistics.

Tao, Chongyang, Mou, Lili, Zhao, Dongyan, & Yan, Rui. 2017.

RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems.
CoRR, **abs/1701.03079**.

Venkatesh, Anu, Khatri, Chandra, Ram, Ashwin, Guo, Fenfei, Gabriel, Raefer, Nagar, Ashish, Prasad, Rohit, Cheng, Ming, Hedayatnia, Behnam, Metallinou, Angeliki, Goel, Rahul, Yang, Shaohua, & Raju, Anirudh. 2018.
On Evaluating and Comparing Conversational Agents.
CoRR, **abs/1801.03625**.