SigDial 2020 BreakOut Session
Friday 3rd July 2020

Best practices in human evaluation of dialogue and NLG - More art than science?

Session Chair: Verena Rieser

Session Brief:
Human judgement is often considered the ultimate evaluation metric for response generation in dialogue systems. For example, we measure correlation with human judgement to determine whether automatic metrics are 'meaningful', e.g. (Novikova et al., EMNLP 2017). But what if the human judgements themselves are not reliable? For example, in our previous work we show that inter-rater agreement can be heavily influenced by the experimental setup (Novikova et al., NAALC 2018)

In this session, participants are invited to review current practices in human evaluation for dialogue systems and Natural Language Generation (NLG) -- where the running hypothesis is that many of the parameters are decided ad hoc rather than based on scientific research. Participants will also be invited to share their experience in conducting experiments themselves (what worked, what didn't?). The desired outcome of this session is to get a better understanding of evaluation consensus and identify open research questions.

For example, we will discuss:

- How do you determine how many users you need? What's a good sample size? How do you make sure your participants are representative of the target population? (inclusion matters!)
- What concepts do you measure? Fluency, adequacy, etc? Which questions to ask and how to frame them?
- Should you ask people to rate or rank? Should you use absolute or relative ratings?
- Are you using ordinal Likert Scales? Or continuous sliders? How do you define the numeric ranges on Likert/sliders?
- Expert vs. naive user rating?
- Overhearer vs. engaged users? at the end? right way? online dialogue cues/ interactive feedback? Online vs. offline?

Video Recording:
https://drive.google.com/drive/folders/1TakdBHpzFsZLUZcwqm3EeTAGmUdcdw78?usp=sharing

Notes from the session:

How do you determine sample size?

- Current practice of “p-hacking”
- Power analysis answers questions like “how much statistical power does my study have?” and “how big a sample size do I need?”. See e.g. https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/

How do you design a questionnaire?

- Need for standardised questionnaires!
- How to determine which questions to ask: device a list of questions and measure which dimensions they are using, see previous work such as SASSI or work by Sebastian Möller.

How to get more informative feedback from users?

- Online questionnaire vs. reflective feedback after the interaction (via structured interview)
- Analyse reflective feedback: use techniques from conversation analysis, e.g. use tags to identify common topics, report interesting quotes.

User preference vs. usability of system?

- Neighbouring fields pay more attention to long-term usability of systems, such as IVA or CHI.
- They use longitudinal studies and repeated measures of a system used in a realistic environment.
- NLP often relies on binary ratings (forced choice) of single outputs, e.g. for NLG. see ACUTE-EVAL (2019).
- Measure intra-annotator agreement (agree with yourself)

User population and crowdsourcing

- How do you determine whether your user group represents your target population?
- Crowdsourcing is good to get “an average user from the street” (?)
- For a system aimed at a special group (experts), it’s better to have a pool of users from that group

How do you get good crowdworkers:

- Build you profile on Turkopticon: Pay decent fees (USD15/h), communicate clearly to your workers
- Pay everyone (to keep reputation) but block bots from further tasks & Report bots to AMT
- Design a task which cannot be easily hacked by bots
- Use CMU DialCrowd Tool https://dialcrowd.herokuapp.com/

Some very recent overview papers (both published end of June 2020)

Deriu et al (2020) Survey on Evaluation Methods for Dialogue Systems
https://arxiv.org/pdf/1905.04071.pdf

Celikyilmaz et al. (2020) Evaluation of Text Generation: A Survey
https://arxiv.org/pdf/2006.14799.pdf

Snapshots of attendees: