Evaluating Response Generation Systems

Amanda Cercas Curry[◇], **Igor Shalyminov**[◇], **Helen Hastie**[◇], **Oliver Lemon**[◇], **Rafael E. Banchs**[⋆], **Verena Rieser**[◇]
[◇]The Interaction Lab, Heriot-Watt University, Edinburgh, UK
[⋆]Human Language Technology Institute for Infocomm Research A*Star, Singapore
Contact: v.t.rieser@hw.ac.uk

Abstract

We propose a shared task to evaluate non-task oriented, end-to-end response generation systems. We aim to address the following needs created by this newly emerging field: (1) a comprehensive comparison of diverse approaches in a controlled setting; (2) a set of common metrics to evaluate non-task based systems; (3) a sufficiently large data set to build advanced quality estimation models for training the next generation of systems.

1 Overview

This shared task will follow on from the successful WOCHAT shared tasks series[1] and the current Alexa Prize[2] in evaluating open-domain, non-task oriented conversational agents (aka. 'chatbots'), with a special focus on end-to-end response generation systems. 'End-to-end' systems are ones which are developed directly from "raw" dialogue data, for example transcripts of movie conversations, and do not rely on semantic or pragmatic annotations, such as Dialogue Acts. The aims of this shared task are to:

1. Benchmark current state-of-the-art response generation techniques using a common set of metrics and release a detailed report.
2. Evaluate state-of-the-art metrics and propose new ones (as outlined in Section 5).
3. Gather, annotate and release data of humans interacting with these systems and user ratings of their interactions.

[1] http://workshop.colips.org/wochat/shared.html

[2] https://developer.amazon.com/alexaprize

2 Relevance

End-to-end response generation for non-task-based interaction has recently received a lot of attention from research and industry. In contrast to Spoken Dialogue Systems (which were the focus of previous editions of DSTC), these systems do not attempt to solve a specific task, nor are they restricted to a specific domain. To address this challenging task, a variety of methods were proposed, including neural approaches, e.g. (Vinyals and Le, 2015; Sordoni et al., 2015), as well as Machine Translation (MT) e.g. (Ritter et al., 2011), Information Retrieval e.g. (Banchs and Li, 2012), and deep Reinforcement Learning approaches e.g. (Li et al., 2016a). Also, some hybrid models, incorporating handwritten rules have been proposed (Yu et al., 2016). Up to this point, there has not been a comprehensive cross-system evaluation of how these different methods compare: there is no standard baseline model nor a standard evaluation framework. The majority of works published in this area rely on automatic metrics, such as BLEU or METEOR. However, a recent study suggests that these do not sufficiently reflect human preferences (Liu et al., 2016). Several other metrics have been proposed, e.g. (Lowe et al., 2016; Li et al., 2017), however it is unclear how they relate to human ratings.

3 Proposed Task

In this challenge, we invite the research community to submit their systems in order to compare them in a unified experimental setup. To qualify for the challenge the systems must be non task-based and not restricted to a specific domain. We will explore a

wide range of existing and new metrics, and evaluate their correlation with human ratings. We will also analyse the effects of the implemented algorithms and datasets used. In contrast to previous challenges in this area, such as WOCHAT and Alexa, we aim to gather, annotate and release a sufficiently large dataset of system outputs and human ratings, with the aim to inspire future work on (a) automatic quality estimation for response generation, following research in MT, e.g. (Specia et al., 2010); (b) reranking of system utterances based on manual annotations of e.g. appropriateness.

4 Setup

Target: We invite submissions of existing and new systems capable of open domain, non-task based interaction, also see target conversation in Appendix A. Note that interactions will be typed, rather than spoken. We propose two sub-tracks:

1. *Restricted*: Systems trained only on OpenSubtitles data (Tiedemann, 2009) for a rigorous experimental setup.
2. *Unrestricted*: All types of non-task based systems can enter independently of dataset, (which includes rule-based chatbots) to encourage wider participation.

Baseline: We will provide a trained sequence-to-sequence neural network based on the approach in (Vinyals and Le, 2015), for system comparisons.

Human evaluation: We will collect data of human experts and crowd-workers interacting with the systems, and rating the interactions with final scores.

Annotations: We will also add some detailed annotations at the utterance level, see Section 5.2. We will be using a mixture of (controlled) crowdsourcing and expert annotations, measuring inter-annotator agreement to produce reliable results.

Funding: We will encourage participants to serve as expert annotators/users. We will also apply for industrial sponsorship to cover crowd-sourcing costs. Finally, we have access to a limited amount of funding through UK research grants held by HWU.

5 Metrics

In contrast to previous DTSC challenges, our task has no clear measure of final task success. Consequently, evaluation is notoriously hard. With this in mind, we will gather and automatically annotate a range of potential metrics and investigate how they relate to human ratings. These metrics will include, but are not limited to:

5.1 Automatic metrics

At the turn-level:

- Baseline metrics, such as BLEU, METEOR.
- Coherence: semantic similarity between consecutive turns, e.g. (Li et al., 2016b).
- Flow: semantic dis-similarity between utterances of the same speaker (Li et al., 2016b).
- Next utterance classification (Lowe et al., 2016)
- Perplexity and information retrieval metrics such as Recall and Precision
- Automatic sentiment analysis including engagement using an off-the-shelf tool.

At the dialogue-level:

- Conversational depth: the number of consecutive turns belonging to the same topic (Yu et al., 2016).
- Information gain/ lexical diversity: the number of unique words that are introduced into the conversation from both the system and the user (Yu et al., 2016).
- Approximations of user engagement across the whole dialogue: number of turns, user reactivity, sentiment, etc. (Yu et al., 2016).
- System ease of answering: e.g. negative log-likelihood of responding to that utterance with a dull response (as defined by a blacklist) (Li et al., 2016b).
- Entrainment to the linguistic style of the user (Niederhoffer and Pennebaker, 2002).

5.2 Human Ratings

Annotations at the turn-level:

- Appropriateness, i.e. is a given response valid in a given context, with annotations similar to the ones in WOCHAT, e.g. `</VALID>`, `</ACCEPTABLE>`, `</INVALID>`.
- Engagement, i.e. does a given response incite the user to continue the conversation.
- Pairwise response choice.

User ratings at the dialogue-level:

- Likert-scale ratings on Future use etc.

Acknowledgments

This research is supported by the EPSRC projects GUI (EP/L026775/1), DILiGENt (EP/M005429/1), MaDrIgAL (EP/N017536/1), and BABBLE (EP/M01553X/1).

References

Rafael E. Banchs and Haizhou Li. 2012. IRIS: A chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 37–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016a. Deep reinforcement learning for dialogue generation. *CoRR*, abs/1606.01541.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of EMNLP*, pages 1192–1202.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Learning to decode for future success. *CoRR*, abs/1701.06549.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, TX, USA. arXiv:1603.08023.

Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 264–269.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Alan Ritter, Colin Cherry, and Bill Dolan. 2011. Data-driven response generation in social media. Empirical Methods in Natural Language Processing (EMNLP), January.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Zhou Yu, Ziyu Xu, Alan W. Black, and Alexander I. Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 404–412.

Appendix A: Annotations and Target Conversations

This challenge plans to follow a similar format to the WOCHAT series. The following example is taken from the WOCHAT evaluation of the IRIS system (Banchs and Li, 2012) with human annotations highlighted.

```
<dialogue id="iris_00050" lang="en">
  <system_name>IRIS</system_name>
  <user_name>305838235865</user_name>
  <timestamp>10-Feb-2016 at 8:44</timestamp>
  <turn id="1">
    <speaker>SYSTEM</speaker>
    <utterance>Hello... USER_NAME</utterance>
    <annotator id="wochat-ijrnotpijgepw">
      </VALID> </NON_ENGAGING>
    </annotator>
    <annotator id="wochat-wifwfkfjporpp">
      </VALID> </NEUTRAL>
    </annotator>
  </turn>
  <turn id="2">
    <speaker>USER</speaker>
    <utterance>hi SYSTEM_NAME ! It's so sunny today !</utterance>
    <annotator id="wochat-ijrnotpijgepw">
      </VALID> </ENGAGING>
    </annotator>
    <annotator id="wochat-wifwfkfjporpp">
      </VALID> </ENGAGING>
    </annotator>
  </turn>
  <turn id="3">
    <speaker>SYSTEM</speaker>
    <utterance>Hi USER_NAME, what's wrong today?</utterance>
    <annotator id="wochat-ijrnotpijgepw">
      </VALID> </ENGAGING>
    </annotator>
    <annotator id="wochat-wifwfkfjporpp">
      </ACCEPTABLE> </ENGAGING>
    </annotator>
  </turn>
  <turn id="4">
    <speaker>USER</speaker>
    <utterance>nothing is wrong !</utterance>
    <annotator id="wochat-ijrnotpijgepw">
      </VALID> </NEUTRAL>
    </annotator>
    <annotator id="wochat-wifwfkfjporpp">
      </VALID> </NON_ENGAGING>
    </annotator>
  </turn>
```

```
<turn id="5">
  <speaker>SYSTEM</speaker>
  <utterance>I thought we could walk into town this morning.</utterance>
  <annotator id="wochat-ijrnotpijgepw">
    </VALID> </NEUTRAL>
  </annotator>
  <annotator id="wochat-wifwfkfjporpp">
    </ACCEPTABLE> </ENGAGING>
  </annotator>
</turn>
<turn id="6">
  <speaker>USER</speaker>
  <utterance>that’s a good idea, we could enjoy the blue sky</utterance>
  <annotator id="wochat-ijrnotpijgepw">
    </VALID> </ENGAGING>
  </annotator>
  <annotator id="wochat-wifwfkfjporpp">
    </VALID> </ENGAGING>
  </annotator>
</turn>
<turn id="7">
  <speaker>SYSTEM</speaker>
  <utterance>One day, out of the blue clear sky, I got a letter from USER_NAME..
  <annotator id="wochat-ijrnotpijgepw">
    </INVALID> </NON_ENGAGING>
  </annotator>
  <annotator id="wochat-wifwfkfjporpp">
    </ACCEPTABLE> </NEUTRAL>
  </annotator>
</turn>
```