End-to-End Conversation Modeling: Moving beyond Chitchat

DSTC7 Track Proposal

Michel Galley, **Chris Brockett**, **Bill Dolan**, and **Jianfeng Gao**
Microsoft AI&R, One Microsoft Way, Redmond, WA 98052, USA
{mgalley,chrisbkt,billdol,jfgao}@microsoft.com

1 Motivation

Recent work [1, 2, 3, 4, 5, etc.] has shown that conversational models can be trained in a completely end-to-end and data-driven fashion, without any hand-coding. However, such prior work has been mostly applied to chitchat, as this is the salient trait of the social media data (e.g., Twitter [1]) utilized to train these systems. To effectively move beyond chitchat, fully data-driven models would need grounding in the real world and access to external knowledge (textual or structured), in order to produce system responses that are both substantive and "useful". Figure 1 illustrates this desideratum: while an ideal response would directly include entities relevant to the current conversation, most state-of-the-art neural conversation models produce responses that are

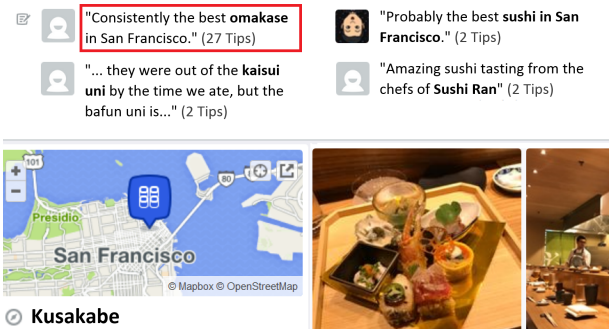

Figure 1: Contrasting chatty and bland response of a fully data-driven conversation model (Neural model) with an ideal response (Human).

conversationally appropriate, but often bland [6], purely chatty, and lacking entities and factual content. Such models contrast with conventional dialog systems, which can readily inject entities and facts into responses, but often at the cost of significant hand-coding (e.g., slot filling).

In DSTC6 [7], the "End-to-End Conversation Modeling" track (Track 2) offered to build fully data-driven systems trained on Twitter customer support data, therefore providing a valuable opportunity to investigate the possibility of fully data-driven conversation in a more goal-oriented scenario than casual chitchat. However, the generation of goal-oriented responses from social media faces several challenges as explained in our DSTC6 system description [8]. First, social media is constituted almost entirely of chitchat, which requires very aggressive filtering (i.e., by customer support user IDs) of Twitter data to produce a goal-oriented dataset. This kind of filtering limits the ability of the model to learn the backbone of general conversation, and hinders its ability to mix goal-oriented responses with chatty replies. Second, task-oriented respondents (e.g., customer support) generally avoid responding publicly—often because of customer privacy concerns—and therefore tend to take the exchange offline (e.g., by email or Twitter direct message) early in the conversation, well before the task has been completed or abandoned. Such a behavior makes it difficult to devise reward functions and to measure success in the task.

By contrast, we argue that dialog shouldn't necessarily be either completely goal-oriented or completely chitchat. This is often reflected in real human-human data, which often combines the two genres. There is also a wide continuum on which a dialog system can have a practical purpose: On one end, task-oriented dialog systems are designed for concrete goals, but to this day still require hand-crafting a fair amount of information specific to the domain or task. On the other end of the continuum, there are chitchat dialog systems that are sometimes seen as less useful, even though they do fulfill a social role and promote user engagement. Following our work "A Knowledge-Grounded Neural Conversation Model" [9], we take a middle-ground approach, combining the benefits of fully data-driven conversation with a purpose that goes beyond purely social.

2 Response Generation Task

We propose an end-to-end conversational modeling track that extends the experimental setting of [9], a work made public in February 2017 and recently accepted at AAAI-18. In that paper, the goal is to generate conversational responses that go beyond chitchat, by injecting informational responses that are grounded in external knowledge (e.g., Foursquare as in [9], or possibly also Wikipedia, Goodreads, or TripAdvisor). Note that in this proposed track, there is no specific or predefined goal (e.g., booking a flight, or reserving a table at a restaurant), so this task does not constitute what is commonly called either goal-oriented, task-oriented, or task-completion dialog. We do not see this as a limitation—even in human-human dialogs—the underlying goal is often ill-defined or not known in advance, even in work and other productive environments (e.g., brainstorming meetings).

The task follows the data-driven framework established in 2011 by Ritter et al. [1],

which prevents systems from hand-coding any linguistic, domain, or task-specific information. In the knowledge-grounded setting of [9], we extend that framework, with each system input consisting of two parts:

- **Conversational input**: Similarly to DSTC6 Track 2 [7], all preceding turns of the conversation are available to the system. For practical purposes, we might truncate the context to the K most recent turns, at least in the case of the training data.
- **Contextually-relevant "facts"**: The system is given up to N factual snippets of text ($N = 20$ in [9]) that are relevant to the context of the conversation. Note that these facts may vary turn-by-turn as the discussion progresses. These snippets of text are not drawn from any conversational data, and are instead extracted from external knowledge sources such as Wikipedia or Foursquare.

From that input, the task it to produce a single response that is both conversationally appropriate and informative. The proposed evaluation setup is presented in Section 4.

Note that the selection of contextually-relevant facts from a much larger pool of facts (e.g., the entire Wikipedia or Foursquare, or a large subset thereof) is not part of the proposed response generation task, as we want to evaluate systems on their ability to produce good responses, and not on their retrieval performance. Therefore, the set of contextually-relevant facts is given in advance and fixed. Optionally, we might allow participants to evaluate their systems on a supplemental and optional task, in which they would select these contextually-relevant facts themselves from the full collection of facts.

3 Data

As in DSTC6 Track 2 [7] and our knowledge-grounded work [9], the conversational dataset would be drawn from Twitter data. We would use and extend the download scripts provided for DSTC6 Track 2, as Twitter doesn't allow sharing of any Twitter data. Similarly to DSTC6, we would define a timeline for extracting data and building models.[1] We propose to use a similar timeline, except that we suggest extending the data collection period from 2 weeks to 1-2 months, as we think models such as [9] need to be trained on more data to be effective. Indeed, DSTC6 Track 2 data contained only about 1M Twitter responses, while [9] was trained on 23M responses. We think systems should be trained on at least 4M responses and corresponding contexts to be conversationally appropriate in free-form dialog.

For a subset of the tweets, we would identify up to $N = 20$ contextually-relevant facts extracted from Foursquare. As we would probably not be able to redistribute this data, we would also provide a way for participants to extract data on their own (e.g., through the Foursquare API[2]). Systems in [9] were grounded in Foursquare data (sample inputs and output shown in Figure 2), but we will investigate the feasibility of using

[1] `http://workshop.colips.org/dstc6/proposals/Easy3stepDataCollection.pdf`

[2] `https://developer.foursquare.com/docs/api/venues/tips`

Facts: (5 out of 50)
Get **Rocco's salad** and the Ontario prosciutto pizza.
Probably the best pizza this side of the Atlantic Ocean.
Sit at the bar of you can. Staff are great! Try the vanilla Affogato for dessert!
Pizza, **calamari** always amazing... the prix fixe is a great deal as well.
Mama's Hot Pepper Pizza: A recipe chef **Rocco** Agostino got straight from his mother.
Conversation history:
T_1: Looking forward to trying @pizzalibretto tonight! My expectations are high.

Response:
T_2: Get the **Rocco salad**. Can you eat **calamari**?

Facts: (5 out of 50)
Interesting. A perfect place to bring kids.
Planetarium tickets. **Shows** are very **popular** and therefore very crowded.
The aquariums are the best part!
Definitely don't miss the **planetarium**, the best **show** there.
Always something new, and so much fun for kids.
Conversation history:
T_1: I'm at California Academy of Science.

Response:
T_2: Make sure you catch the **show** at the **Planetarium**. **Tickets** are usually **limited**.

Figure 2: Sample inputs and gold responses. The highlighting (boldface) is our own, to show similarity between facts and responses.

other data sources such as Wikipedia, and may propose them as optional conditions for this data-driven track.

4 Evaluation

We will evaluate response quality using both automatic and human evaluation. Since we are not aiming task-oriented dialog, there is no pre-specified task and therefore no extrinsic way of measuring task success. Instead, we will perform a per-response evaluation judging (via Mechanical Turk) each system response's:

- **Appropriateness**: this evaluation criterion asks whether the system response is conversationally appropriate and relevant given the K immediately preceding turns.[3] Note that this judgment has nothing to do with grounding in external sources, and is similar to quality human judgments for prior data-driven conversation models (e.g., [2]). Nevertheless, this judgment is important, as we do not want systems to produce much less appropriate responses for the sake of being more informational.

[3]We need to limit the number of preceding turns as to not overload MTurk workers with too much information.

- **Informativeness & Utility**: this evaluation criterion measures the degree to which the produced response includes information that is relevant to the user input (or more generally to the preceding K previous turns) and whether that information has potential utility with respect to some possible decision or action by the user.[4]

We will score both evaluation criteria on a 5-point Likert scale, and finally combine the two judgments by weighting them equally. We might use either absolute scores, or assess systems relative to [9] that could be available as a baseline. The system with best average Appropriateness and Informativeness+Utility will be determined the winner.

In order to provide participants with preliminary results to include in their system descriptions, we will also perform automatic evaluation using standard machine translation metrics such as BLEU [10] and METEOR [11].[5]

References

[1] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," *EMNLP*, 2011.

[2] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," *NAACL-HLT*, 2015.

[3] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *ACL-IJCNLP*, 2015.

[4] O. Vinyals and Q. Le, "A neural conversational model," *ICML*, 2015.

[5] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *AAAI*, 2016.

[6] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," 2016.

[7] C. Hori and T. Hori, "End-to-end conversation modeling track in DSTC6," *arXiv:1706.07440*, 2017.

[8] M. Galley, C. Brockett, B. Dolan, and J. Gao, "The MSR-NLP system at dialog system technology challenges 6," in *Dialog System Technology Challenges 6*, 2017.

[9] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," *AAAI (to appear)*, 2018.

[4] As in [9], we will not measure the accuracy of that information against the facts provided as input to the systems, as it would be impractical to show 20 or 50 such facts to the Turkers. We simply ask them to exercise their best judgment as to whether the information in the response is plausible – i.e., for a query *going to pizzeria X tonight*, a response *try their sushi, the best in town* should get a low score.

[5] Those metrics would work as a proxy to the appropriateness judgment, not informativeness. We would use single-reference BLEU and METEOR, as Twitter provides one reference for each response.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," *ACL*, 2002.

[11] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proc. of the Second Workshop on Statistical Machine Translation*, StatMT '07, (Stroudsburg, PA, USA), pp. 228–231, Association for Computational Linguistics, 2007.