Introduction to Alexa Prize 2018 Proceedings

Dilek Hakkani-Tür

Amazon Alexa AI
hakkanit@amazon.com

Building conversational systems that enable natural language interactions with machines has been attractive to mankind since the early days of computing, as exemplified by earlier text-based systems such as ELIZA [Weizenbaum, 1966]. Previous work on conversational systems generally falls into two categories, task-oriented and socialbots. Task-oriented systems aim to help users accomplish a specific task through multi-turn interactions, whereas socialbots focus on engaging and natural open-domain conversations. In natural interactions, even when conversation participants have a task or goal in mind, they can easily say things that are out of the boundaries of that task domain. The most common solution to handling such utterances in the current applications is through indicating to the user that the system cannot handle these yet, which is not the ideal behavior from the user's viewpoint. Hence, the ability to engage in knowledgeable social interactions and gracefully transition back to the task is also important for task-oriented systems.

For the past two years, Amazon has been organizing the Alexa Prize to advance human-computer interaction through conversations. University teams are supported to create socialbots that can converse coherently and engagingly with humans on a range of current events and popular topics such as entertainment, sports, politics, technology, and fashion.

One of the main obstacles in conversational systems research is the scarcity of conversational datasets that include real interactions with users. Alexa Prize have been providing a unique opportunity for university teams to connect their systems with millions of real users for spoken interactions. Furthermore, automatically evaluating quality of social conversations is critical for advancing the quality of conversational systems and still remains an open question [Liu et al., 2016]. Real user ratings coupled with these conversations provide university teams a large-scale experimentation framework, accelerating the advancement of open domain conversational response generation systems and socialbots.

1. Overview of Alexa Prize 2018

In 2018, eight university teams were selected to participate in the Alexa Prize.[1] The participants were announced in February and received a research grant, Alexa-enabled devices, free Amazon Web Services (AWS) to support their development efforts, and access to other tools (such as the Cobot tools as described later in the proceedings), data sources, and Alexa team support.

The semifinals period took place between July 2 and August 15, 2018. During the semifinals, users interacting with Alexa were connected with the teams' socialbots by saying "*Alexa, let's chat*" to

[1] The announcement can be found at: https://developer.amazon.com/blogs/alexa/post/9f406f35-c997-4d17-a6ac-89a35f69b661/announcing-the-2018-alexa-prize-participants

2nd Proceedings of Alexa Prize (Alexa Prize 2018).

any Alexa-powered device. Competing socialbots were randomly invoked in response to this utterance. At each turn, user's spoken utterances were transcribed using automatic speech recognition and these utterances along with several other high-level metadata (such as confidence scores) were provided to the socialbots, which in return produced textual responses (possibly with Amazon speech synthesis markup language markings to format the prosody of these responses) that were converted to spoken utterances by text-to-speech synthesis. At the end of the interaction, users were prompted to provide a rating from 1 to 5 stars (5 being the highest) on how they felt about speaking with that socialbot. After the rating, the users were also asked if they would like to leave a verbal feedback to the university team that built the socialbot they just interacted with. The user ratings determined two of the teams moving on to the finals, and Amazon selected the wildcard finalist based on a combination of the following criteria:

- Ratings from Alexa customers
- Ratings from Amazon judges
- Depth and breadth of topics covered
- Appropriateness and accuracy of responses
- Scientific merit as determined by the content of their technical paper

The semifinalist socialbots continued being available to Alexa users until the finals and after. During this period the semifinalists were able to improve their systems according to the feedback given by the users as well as by performing A/B tests.

2. Related Work on Conversational Systems

Building conversational agents has been an attractive research area for a long time, however task-oriented conversational systems that millions of users can speak to became common only in the past few decades starting with customer care center applications [Gorin et al, 1997]. In this approach, users can speak naturally, and the aim of the machine is to detect the intent of the user from a handful of pre-defined intents and transferring them to human customer care agents who are experienced for handling such intents. Language understanding in the form of intent detection has later been extended with detection of related arguments and entities [Gupta et al, 2006; among others]. In such systems, dialogue management is usually handled by manually crafted enterprise policies. With the rise of virtual personal assistants after 2010, conversational systems have started to become ubiquitous in mobile phones (e.g., Siri) and later in stand-alone smart speakers (e.g., Alexa). In such applications, while the main backbone is similar to call center solutions, the breadth of domains and associated intents has exploded, powered by both 1st and 3rd party experiences and skills. As mentioned earlier, since the notion of task domain and its boundaries are not necessarily clear to the users, they may easily switch from the tasks to social interactions and back during their conversations, motivating innovations for socialbots.

Work on socialbots mainly relied on rule-based, pattern matching approaches [Wallace, 2009]. Recently, data-driven, end-to-end approaches to conversational response generation treated the task as statistical machine translation, where the goal is to generate a response given the previous dialogue turn [Ritter et al., 2010; Vinyals & Le, 2015]. While these studies resulted in a paradigm change compared to earlier work based on rule- and retrieval-based approaches, responses generated by these sequence-to-sequence models are not always coherent and contextually appropriate. This results from considering conversations as request-response pairs and not including mechanisms to represent longer term conversation context.

In order to have a better representation of conversation context as input to response generation, hierarchical recurrent encoder-decoder (HRED) networks have been proposed [Serban et al., 2016]. HRED combines two RNNs, one at the token level, modeling individual system and user turns, and one at the dialogue level, inputting turn representations from the token-level RNNs. However, utterances generated by such neural response generation systems are noted to be often generic and

lack interesting content [Vinyals and Le, 2015]. To improve the diversity and content of generated responses, HRED was later extended with a latent variable that aims to model the higher-level aspects (such as topic) of the generated responses, resulting in the variational HRED (VHRED) approach [Serban et al., 2017]. Another issue that remains with these previous approaches is that the generated dialogues focus on naturalness and discourse coherence but do not ground responses on related or relevant knowledge.

Previous work also investigated ways of including world knowledge and common sense into the conversational response generation, in a way shifting the research towards more task-oriented conversations, which already have language understanding [Hakkani-Tür et al, 2016; among others] and dialogue state tracking [Rastogi et al., 2017; among others] to be able to look up relevant information for task completion before formulating system turns [Liu et al, 2017; Fazel-Zarandi et al, 2017; among others]. [Ghazvininejad et al., 2017] used end-to-end memory networks to base the generated responses on knowledge, where an attention over the knowledge relevant to the conversation context is estimated, and multiple knowledge representations are included as input during the decoding of responses. In this work, we use end-to-end memory networks as one of the baselines. Along similar lines, [Liu et al., 2018] used pattern matching, named entity recognition and linking to find facts relevant to the current dialogue and other related entities from a knowledge base. Their work combined the information as an additional input feature representing context during the decoding phase, generating responses. [Zhou et al., 2018] instead retrieves relevant knowledge graphs given the previous conversation and then encodes the graphs with a static graph attention mechanism. Then, during word generation, the decoding model attentively reads the retrieved knowledge graphs and the knowledge triples within each graph to facilitate better generation through a dynamic graph attention mechanism. All three studies demonstrated higher quality response generation due to the integration of knowledge.

Grounding conversational response generation on related knowledge has also been the focus for two tracks in this year's dialogue system technology challenges (DSTCs), where sentence selection approaches from Ubuntu and Flex Data are targeted in one track [Polymenakos et al., 2018] and sentence generation using the Reddit conversation is targeted in the other [Galley et al., 2018].

In all of these previous studies, with the exception of the earlier DSTCs [Henderson et al., 2014], the mode of interaction is mainly limited to textual interactions, off-line interactions or interactions between crowd workers. Alexa Prize is unique enabling conversational bots built by research teams to be accessed by real users in real-time and enriched with real user ratings.

3. Brief Overview of Scientific Contributions

Understanding natural language in open-domain spoken conversations is a daunting task, first of all, one would need to recognize speech accurately. Furthermore, as also explored in the previous studies described above, entities and topic of the current conversation could be useful in finding relevant knowledge for generating the following system responses, increasing the importance of entity recognition, and topic detection. In 2018, Amazon Alexa Prize team has provided a suite of tools and models to the participants including automatic speech recognition with improved word and entity accuracies, the Conversational Bot (CoBot) toolkit, dialogue act and topic detection, sensitive content detection, and conversation evaluator models. The first article in the proceedings from the Alexa Prize team presents these advancements. These are in addition to the scientific contributions from the 2017 challenge that have been summarized in [Ram et al., 2018].

Natural language and speech processing tools commonly used (built in-house or off-the-shelf) by the teams include sentence segmentation, syntactic parsing and part-of-speech tagging, entity extraction and linking, topic, intent and dialogue act detection. The participating teams also built novel approaches including but not limited to dialogue management for social dialogues, user and system personality management and tracking and discourse relation dialogue modeling. In the

following chapters, the Alexa Prize proceedings is organized to include overviews and contributions of the participating university systems. Many thanks to the Amazon Alexa Prize team and participating university teams for making the Amazon Alexa Prize established and influential.

4. References

Fazel-Zarandi M, Li SW, Cao J, Casale J, Henderson P, Whitney D, Geramifard A. Learning Robust Dialog Policies in Noisy Environments. arXiv preprint arXiv:1712.04034. 2017 Dec 11.

Galley M, Brockett C, Gao X, Dolan W, Gao J. End-to-End Conversation Modeling: Moving beyond Chitchat. Dialog System Technology Challenge 7, Track 1, 2018. http://workshop.colips.org/dstc7/proposals/DSTC7-MSR_end2end.pdf

Ghazvininejad M, Brockett C, Chang MW, Dolan B, Gao J, Yih WT, Galley M. A knowledge-grounded neural conversation model. arXiv preprint arXiv:1702.01932. 2017 Feb 7.

Gorin AL, Riccardi G, Wright JH How may I help you? Speech communication. 1997 Oct 1;23(1-2):113-27.

Gupta N, Tur G, Hakkani-Tür D, Bangalore S, Riccardi G, Gilbert M. The AT&T spoken language understanding system. IEEE Transactions on Audio, Speech, and Language Processing. 2006 Jan;14(1):213-22.

Hakkani-Tür D, Tür G, Celikyilmaz A, Chen YN, Gao J, Deng L, Wang YY. Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In Interspeech 2016 Jun (pp. 715-719).

Henderson M, Thomson B, and Williams J. The second dialog state tracking challenge. In 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2014, vol. 263.

Liu CW, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023. 2016 Mar 25.

Liu B, Tur G, Hakkani-Tur D, Shah P, and Heck L. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. In NIPS Conversational AI Workshop.

Liu S, Chen H, Ren Z, Feng Y, Liu Q, Yin D. Knowledge diffusion for neural dialogue generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2018 (Vol. 1, pp. 1489-1498).

Polymenakos L, Gunasekara C, Lasecki W, Kummerfeld J. NOESIS: Noetic End-to-End Response Selection Challenge. Dialog System Technology Challenge 7, Track 1, 2018. http://workshop.colips.org/dstc7/proposals/Track%201%20Merged%20Challenge%20Extended%20Desscription_v2.pdf

Ram A, Prasad R, Khatri C, Venkatesh A, Gabriel R, Liu Q, Nunn J, Hedayatnia B, Cheng M, Nagar A, King E, Bland K, Wartick A, Pan Y, Song H, Jayadevan S, Hwang G, Pettigrue A. Conversational AI: The science behind the Alexa prize. arXiv preprint arXiv:1801.03604.

Rastogi A, Hakkani-Tür D, Heck L. Scalable multi-domain dialogue state tracking. In Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE 2017 Dec 16 (pp. 561-568). IEEE.

Ritter A, Cherry C, Dolan B. Unsupervised modeling of twitter conversations. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics 2010 Jun 2 (pp. 172-180).

Serban IV, Sordoni A, Bengio Y, Courville AC, Pineau J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In AAAI 2016 Feb 12 (Vol. 16, pp. 3776-3784).

Serban IV, Sordoni A, Lowe R, Charlin L, Pineau J, Courville AC, Bengio Y. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In AAAI 2017 Feb 4 (pp. 3295-3301).

Vinyals O, Le Q. A neural conversational model. arXiv preprint arXiv:1506.05869. 2015 Jun 19.

Wallace RS. The anatomy of ALICE. In Parsing the Turing Test 2009 (pp. 181-210). Springer, Dordrecht.

Weizenbaum, J. Eliza – A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1):36–45, Jan. 1966. ISSN 0001-0782. https://web.stanford.edu/class/linguist238/p36-weizenabaum.pdf

Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In IJCAI 2018 (pp. 4623-4629).