The 2018 Alexa Prize: Several Steps Forward in Conversational AI

Mari Ostendorf

Department of Electrical & Computer Engineering, University of Washington, Seattle, WA
ostendor@uw.edu

In September 2016, Amazon introduced a new problem in conversational AI, the Alexa Prize, which challenged student teams to build a socialbot that could converse with Alexa on a wide range of topics. The socialbot differs from task-oriented dialog systems that address explicit user goals associated with a constrained domain, and also differs from chatbot systems that only handle social chitchat. The socialbot must handle chitchat, and needs to inform and exchange opinions with a user about recent news and other topics of interest, serve evolving user interests and implicit information sharing goals. Because of these differences, many of the conventional approaches that had been developed for dialog systems were difficult to apply to the socialbot problem. Further, because this was essentially a new problem, there was no existing conversational data that was well-matched to the types of interactions that Alexa users had with the socialbots. Initially, even applying simple machine learning was challenging, let alone end-to-end system design. Moreover, the implementation had to be scalable to a high volume of user interactions. However, access to millions of real users and the creativity of student teams led to a good start on this challenging problem.

A nice summary of the 2017 efforts is provided in [1], but it is worth highlighting a few areas to put this year's contributions in context. Much effort was devoted to natural language understanding (NLU), with teams exploring several different directions, including named entity recognition, intent recognition, anaphora and co-reference resolution, sentence completion, topic detection, entity linking, text summarization and sentiment detection. A common strategy for dialog modeling was to use a hierarchical architecture with a main dialog manager controlling smaller components that focused on specific tasks or t opics. To handle user questions, systems used Amazon's Evi and several different knowledge bases. For response generation, there was a mix of template-based, retrieval and generative approaches, but virtually all systems included some modules with template-based strategies. A challenge with using content scraped from the internet is that it can be offensive or controversial. The same is true about interactions with a broad range of real users: some of the interactions are adversarial and contain vulgar expressions. Systems were required to be family friendly and to deflect potentially problematic topics (e.g., a user asking for advice), so substantial effort was devoted to detecting inappropriate user input and informational content.

All of these broad trends continued with the 2018 systems, although most returning teams had major (and sometimes complete) reimplementations of their systems. In addition, some features of the top systems were adopted more broadly. For example, the 2017 Alquist system [2] used specialized, structured modules for some popular topics (movies, video games, etc.), an approach that was used by most 2018 teams. Many more systems incorporated named entity recognition, entity linking, co-reference and anaphora resolution, used last year by Alana [3]. Multiple teams added modules associated with uplifting news and interesting facts or thoughts, scraping content from the subreddits found to be most useful in Sounding Board [4]. Similarly, more efforts were aimed at user engagement, with most teams incorporating sentiment detection, several leveraging conversational grounding strategies, and more teams controlling for speech prosody with the Speech Synthesis Markup Language (SSML) to obtain appropriately emotive speech.

While the 2017 student teams were busy exploring different socialbot architectures, Amazon was busy improving the Alexa infrastructure. Automatic speech recognition (ASR) word error rate reduced by

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

nearly 33%, plus a new topic tracker and an offensive language detector were developed [1]. These advances were made available to the 2018 teams, together with modules for named entity detection, sentiment classification, and dialog act modeling. In response to request from the teams, Amazon started providing word timing/pause information and some punctuation. Amazon also made available a new graph database service (Neptune) and general tools to make it easier for new teams to get started in buidling a conversational system (CoBot). In addition, they developed a conversation evaluation service for the teams and provided weekly metrics on team performance in different topical domains. These advances from Amazon clearly had an impact on the 2018 teams. In particular, the performance of systems in discussing named entities was noticibly improved. There has been a nice synergy in the advances coming out of this industry-university collaboration.

The 2017 experience – as well as further experiments by the 2018 teams – demonstrated that existing datasets (movie subtitles, Twitter and Reddit interactions) were not very useful for training sequence-to-sequence response generation modules or even retrieval-based response generation. The 2017 Edina team developed a mechanism for crowdsourcing socialbot-like dialogues for training the dialogue system [5], and at least one of the 2018 teams took advantage of the data they made available. Returning teams could benefit from data collected from their previous system, though of course the performance of that system limits what can be learned. A couple of the 2018 teams found it useful to collect crowdsourced conversational data. In particular, Fantom [6] introduced an interesting new approach that automatically detects where more content is needed for system training based on ongoing interactions with users. Some out-of-domain data was used successfully for specialized NLU components, such as the Switchboard data for training dialog act taggers. With the availability of more appropriate data, teams were able to make more effective use of machine learning throughout the systems.

In addition to building on past advances, improved services and better data, the 2018 teams have many new contributions that will move the field forward. To improve the ASR output for NLU, a few of the teams explore new uses of word confidence information, but Gunrock went further with an automatic correction interface leveraging dialog context and sentence segmentation [7]. Several teams used truecasing to improve entity detection [8, 9, 6]. Tartan developed a semantic grammar for understanding a broader range of intents, and they introduced a special module for handling use statements or questions that serve more as asides to avoid unintended interruptions or topic changes [9]. Iris improved understanding of user intent with contextualized topic detection [10]. In discussions about named entities, Alana incorporated clarification questions to resolve ambiguities in entity linking [8]. While many teams leveraged knowledge graphs for content knowledge, Fantom used a dialog graph to capture social interaction knowledge for retrieval-based response generation [6]. Eve introduce an approach to characterizing conversational flow with utterance embeddings to improve response retrieval [11]. Slugbot [12] outlined a new dialog theory appropriate for socialbots that impacts content curation and structuring. Alquist [13] adapted hybrid code networks [14] to the socialbot scenario to facilitate dialog manager training. They also implemented a new structured dialogue authoring work-flow, with a web-based editor for creating a dialogue structure from which a module could be automatically trained and Java code generated, separating the creative design process from much of the detailed implementation. In contrast, Alana [8] introduced an ontology bot that provided a more general mechanism for handling topic-dependent structure in subdialogues. A range of new ideas were explored in the area of user modeling to personalize topic suggestions, including conditioning on initial user mood [12], topic interests [10, 8], usesr opinions about entities [11, 12], and user utterances [9]. There were also efforts to influence bot opinion generation (sentiment) about topics based on reflection of user interests [11] or opinions expressed in social media [7, 13, 10] and debate opinions [7].

For brevity, I have listed just a few examples that I think are likely to be leveraged in future systems, but there are many more ideas and analyses presented in these proceedings that will be of interest researchers in conversational AI. I congratulate all the teams for their tremendous achievements.

References

[1] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue. Conversational AI: The science behind the Alexa Prize. In *Proc. Alexa Prize 2017*, 2017.

[2] J. Pichi, P. Marek, J. Konrád, M. Matulík, H. L. Nguyen, and J. Šedivý. Alquist: The Alexa Prize socialbot. In *Proc. Alexa Prize*, 2017.

[3] I. Papaioannou, A. Curry, J. Part, I. Shalyminov, X. Xu, Y. Yu, O. Dušek, V. Rieser, and O. Lemon. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *Proc. Alexa Prize*, 2017.

[4] H. Fang, H. Cheng, E. Clark, A. Holtzman, M. Sap, M. Ostendorf, Y. Choi, and N. Smith. Sounding board – university of washington's alexa prize submission. In *Proc. Alexa Prize 2017*, 2017.

[5] B. Krause, M. Damonte, M. Dobre, D. Duma, J. Fainberg, F. Fancellu, E. Kahembwe, J. Cheng, and B. Webber. Edina: Building an open domain socialbot with self-dialogues. In *Proc. Alexa Prize*, 2017.

[6] P. Jonell, M. Bystedt, F. Dogan, P. Fallgren, J. Ivarsson, M. Slukova, U. Wennberg, J. Lopes, J. Boyce, and G. Skantze. Fantom: A crowdsourced social chatbot using an evolving dialog graph. In *Proc. Alexa Prize 2018*, 2018.

[7] C.-Y. Chen, D. Yu, W. Wen, Y. M. Yang, J. Zhang, M. Zhou, K. Jesse, A. Chau, A. Bhowmick, S. Iyeer, G. Sreenivasulu, R. Cheng, A. Bhandare, and Z. Yu. Gunrock: Building a human-like social bot by leveraging large scale real user data. In *Proc. Alexa Prize 2018*, 2018.

[8] A. Curry, I Papaioannou aand A. Suglia, S. Agarwal, I. Shalyminov, X. Xu, O. Dusek, A. Eshghi, I. Konstas, V. Rieser, and O. Lemon. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. In *Proc. Alexa Prize 2018*, 2018.

[9] G. Larionov, Z. Kaden, H. Dureddy, G. Kalejaiye, M. Kale, S. Potharaju, A. Shah, and A. Rudnicky. Tartan: A retrieval-based socialbot powered by a dynamic finite-state machine architecture. In *Proc. Alexa Prize 2018*, 2018.

[10] A. Ahmadvand, I. Choi, H. Sahijwani, J. Schmidt, M. Sun, S. Volokhin, Z. Wang, and E. Agichtein. Emory IrisBot: an open-domain conversational bot for personalized information access. In *Proc. Alexa Prize 2018*, 2018.

[11] N. Fulda, T. Etchart, W. Myers, D. Ricks, Z. Brown, J. Szendre, B. Mudoch, A. Carr, and D. Wingate. BYU-EVE: Mixed initiative dialog via structured knowledge graph traversal and conversational scaffolding. In *Proc. Alexa Prize 2018*, 2018.

[12] K. Bowden, J. Wu, W. Cui, J. Juraska, V. Harrison, B. Schwarzmann, N. Santer, and M. Walker. SlugBot: Developing a computational model and framework of a novel dialogue genre. In *Proc. Alexa Prize 2018*, 2018.

[13] J. Pichi, P. Marek, J. Konrád, M. Matulík, and J. Šedivý. Alquist 2.0: Alexa prize socialbot based on sub-dialogue models. In *Proc. Alexa Prize 2018*, 2018.

[14] J. Williams, K Asadi, and G. Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv:1702.03274*, 2017.