

Alexa Prize — State of the Art in Conversational AI

*Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia,
Ashwin Ram, Raefer Gabriel, Rohit Prasad*

■ *To advance the state of the art in conversational AI, Amazon launched the Alexa Prize, a \$2.5 million competition that challenges university teams to build conversational agents, or “socialbots,” that can converse coherently and engagingly with humans on popular topics for 20 minutes. The Alexa Prize offers the academic community a unique opportunity to perform research at scale with real conversational data obtained by interacting with millions of Alexa users, along with user-provided ratings and feedback, over several months. This opportunity enables teams to effectively iterate, improve, and evaluate their socialbots throughout the competition. Eighteen teams were selected for the inaugural competition last year. To build their socialbots, the students combined state-of-the-art techniques with their own novel strategies in the areas of natural language understanding and conversational AI. This article reports on the research conducted over the 2017–2018 year. While the 20-minute grand challenge was not achieved in the first year, the competition produced several conversational agents that advanced the state of the art, that are interesting for everyday users to interact with, and that help form a baseline for the second year of the competition.*

Artificial intelligence is becoming ubiquitous. With advances in technology, algorithms, and sheer compute power, it is now practical to utilize AI techniques in everyday applications in the domains of transportation, healthcare, gaming, productivity, and media. Yet one seemingly intuitive task for humans still eludes computers: natural conversation. While simple for humans, voice communication in everyday language continues to be one of the most difficult challenges in AI. Human conversation requires the ability to understand the meaning of spoken language, relate that meaning to the context of the conversation, create a shared understanding and world view between the parties, model discourse and plan conversational moves, maintain semantic and logical coherence across turns, and generate

natural speech. Conversational agents capable of natural human conversation have applicability in both professional and everyday domains.

Voice-based virtual assistants, an important type of conversational agent, have become very popular in the last several years. The first generation of such assistants — Amazon’s Alexa, Apple’s Siri, Google Assistant, and Microsoft’s Cortana — have been focused on short, task-oriented interactions, such as playing music or answering simple questions, as opposed to the longer free-form conversations that occur naturally in social and professional human interaction. Conversational AI is the study of techniques for creating software agents that can engage in natural conversational interactions with humans. Significant advances in this area are needed to make interactions with virtual assistants and other types of AI agents easier and more natural for everyday use, particularly for open-domain conversations, those that are not bounded to a single task or topic.

Conversational AI is still in its infancy and several leading university research teams are actively pushing research boundaries in this area (Serban et al. 2016; Vinyals and Le 2015). Access to large-scale data and real-world feedback can drive faster progress in research. To address this challenge, Amazon announced the Alexa Prize on September 26, 2016, with the goal of advancing the research in conversational AI. Selected university teams were challenged to build conversational agents, known as “socialbots,” to converse coherently and engagingly with humans on popular topics such as sports, politics, entertainment, fashion, and technology for 20 minutes. The grand challenge is to conduct coherent and engaging conversations for 20 minutes, with an average rating of 4 or higher on a scale of 1 to 5.

Given the complexity of the challenge, Amazon collaborated with the participating teams to provide them with tools, data, and a unique opportunity to perform iterative research with a live system deployed to millions of Alexa users. Through the Alexa Prize competition, participating universities were able to conduct research by building socialbots, training conversational models, and testing hypotheses at scale. Alexa users interacted with socialbots via the “Alexa, let’s chat” experience, engaged in live conversations, and left ratings and feedback for the teams at the end of their conversations. Over 40,000 hours of conversation were logged in the course of the 2017 competition through the finals last November. As users continue to interact with the winning socialbots, this has now become over 130,000 hours.

In this article, we describe the scientific problems related to open-domain conversational systems, the state of the art in addressing these problems, how these approaches were used during the inaugural competition, and the results and scientific advances obtained. We present the technical setup of the Alexa Prize Finals event along with the process of selecting

the winner. We conclude with a summary of the work that we plan to address in the second year of the competition.

The Alexa Prize Experience

The Alexa Prize competition received hundreds of applications from interested universities. After a detailed review of the applications, Amazon announced 12 sponsored and 6 unsponsored teams as the inaugural cohort for the Alexa Prize. The teams that went live for the 2017 competition, listed alphabetically by university, were DeisBot (Brandeis University), Magnus (Carnegie Mellon University), RubyStar (Carnegie Mellon University), Alquis (Czech Technical University in Prague), Emerson (Emory University), What’s Up Bot (Heriot-Watt University), Pixie (Princeton University), Wise Macaw (Rensselaer Polytechnic Institute), Chatty Chat (Seoul National University), Eigen (University of California, Berkeley), SlugBot (University of California, Santa Cruz), Edina (University of Edinburgh), MILA Team (University of Montreal), Roving Mind (University of Trento), and Sounding Board (University of Washington).

The university teams built socialbots using the Alexa Skills Kit (ASK) (Kumar et al. 2017). The Amazon team provided automatic speech recognition (ASR) to convert user utterances to text for the socialbots and text to speech (TTS) to render text responses from the socialbots to speech for the users. All the intermediate steps — natural language understanding (NLU), dialogue modeling, and conversational user experience (CUX) — were handled by the university teams through their socialbots. Teams were allowed to leverage the standard NLU system that is provided with ASK. We also provided live news feeds to enable socialbots to stay current with popular topics and news events that users might want to talk about, and other tools and data as described in this article.

While the Alexa Prize had clear scientific goals and objectives, Alexa users played the key role of providing feedback on the socialbots, helping teams improve their systems and helping us determine which socialbots were the most coherent and engaging. Because users helped drive the direction and result of the competition, it was important for us to ensure an easy and compelling hook into the Alexa Prize socialbots to obtain a statistically significant number of data points for the ratings and feedback needed to improve the socialbots.

Eighteen teams were selected, although only 15 went live. To allow us to randomize traffic to all 15 socialbots without revealing their identity and to set user expectations about the socialbots being early-stage systems, we designed and implemented the Alexa Prize skill with a natural invocation phrase that was easy to remember (“Alexa, let’s chat,” “Alexa,

let's chat about <topic>," and common variants). The user heard a short editorial that educated them about the Alexa Prize and instructed them on how to end the conversation and provide ratings and feedback. We kept it succinct and interesting so we would not lose the user's attention before they had a chance to speak with a socialbot. The editorial and instructions changed as needed to keep the information relevant to the different phrases. For example, at the initial public launch on May 8, 2017, when the 15 socialbots were still in their infancy, the Alexa Prize skill started with the following editorial: "Hi! Welcome to the Alexa Prize Beta. I'll get you one of the socialbots being created by universities around the world. When you're done chatting, say stop."

After listening to the editorial, the user was handed off to one of the competing socialbots selected at random. Socialbots began the conversation with a common introduction phrase ("Hi, this is an Alexa Prize socialbot") without revealing their identity. The user could exit at any time, and thereafter was prompted to provide a verbal rating ("On a scale from one to five stars, how do you feel about speaking with this socialbot again?"). Finally, the user could offer free-form verbal feedback without knowing which socialbot they had interacted with. Ratings and feedback were provided to individual teams to help them improve their socialbots.

Challenges with Conversational AI

Current state-of-the-art systems are still a long way from engaging in truly natural everyday conversations with humans (Levesque 2017). There are a number of major challenges associated with building conversational agents: conversational automatic speech recognition for free-form multiturn speech; conversational natural language understanding for multiturn dialogues; conversational datasets and knowledge ingestion for comprehension; common-sense reasoning for understanding concepts; context modeling for relating past concepts; dialogue planning for driving coherent and engaging conversations; response generation and natural language generation for generating relevant, grammatical, and nongeneric responses; sentiment detection for systematically identifying, extracting, quantifying, and studying affective states and for handling sensitive content (such as profanity, inflammatory opinions, inappropriate jokes, hate speech detection), driving quality conversations; personalization for addressing user preferences; conversation evaluation for evaluating the quality of the conversations and the artificial agent; and conversational experience design for maintaining a great experience for the interactors.

Most voice-based conversational agents follow a similar architecture. First, the agent comprehends speech signals and converts them to text by a process called automatic speech recognition (ASR). After

obtaining the text, the agent tries to understand the meaning and intention of the user using existing knowledge by a process called natural language understanding (NLU). Once the concepts and intents are identified, the agent starts the process of response generation (RG), which involves identifying relevant responses based on context, knowledge, personalization, and some form of planning. Planning may involve optimizing for some reward such as sentiment, serving the goal in a goal-directed dialogue, or increasing user engagement. This process can be managed by a dialogue manager (DM), which acts as an engine for maintaining the state and flow of the conversation. Finally, once the output response is produced in textual form, the agent converts it to speech by a process called text to speech (TTS).

Using these techniques, academia and industries have created virtual assistants to support short, task-oriented dialogues such as playing music or asking for information. Some assistants are capable of longer multiturn dialogues, although most of these are goal directed or designed for specific tasks such as customer support or shopping (for example, eBay's ShopBot). Furthermore, these systems tend to be text based and not capable of natural voice conversation. Long, free-form voice conversations that occur naturally in social and professional human interactions are often open domain. In natural conversations, intents and topics change with time based on the interest of the interactors and the state of the conversation. Furthermore, natural conversations feature many plausible responses at each turn and are highly path dependent: even if two sets of interactors have a similar background and share a similar set of knowledge, they may end up having completely different conversations.

The remainder of this article describes how the challenges listed at the beginning of this section were addressed and what results were obtained.

Addressing Problems in Creating Conversational Agents

The Alexa Prize team developed components to provide conversational speech recognition, conversational intent tracking, conversational topic tracking, inappropriate and sensitive content detection, and conversational quality evaluation. In addition, the team addressed engineering challenges such as traffic allocation, socialbot scalability, socialbot invocation, and a feedback framework. These components and solutions enabled live deployment to a large user base consisting of millions of Alexa customers. To build the socialbots, university teams combined state-of-the-art techniques with their own novel strategies in the areas of natural language understanding, context modeling, dialogue management, response generation, ranking and selection, sentiment analysis, and knowledge acquisition. Subse-

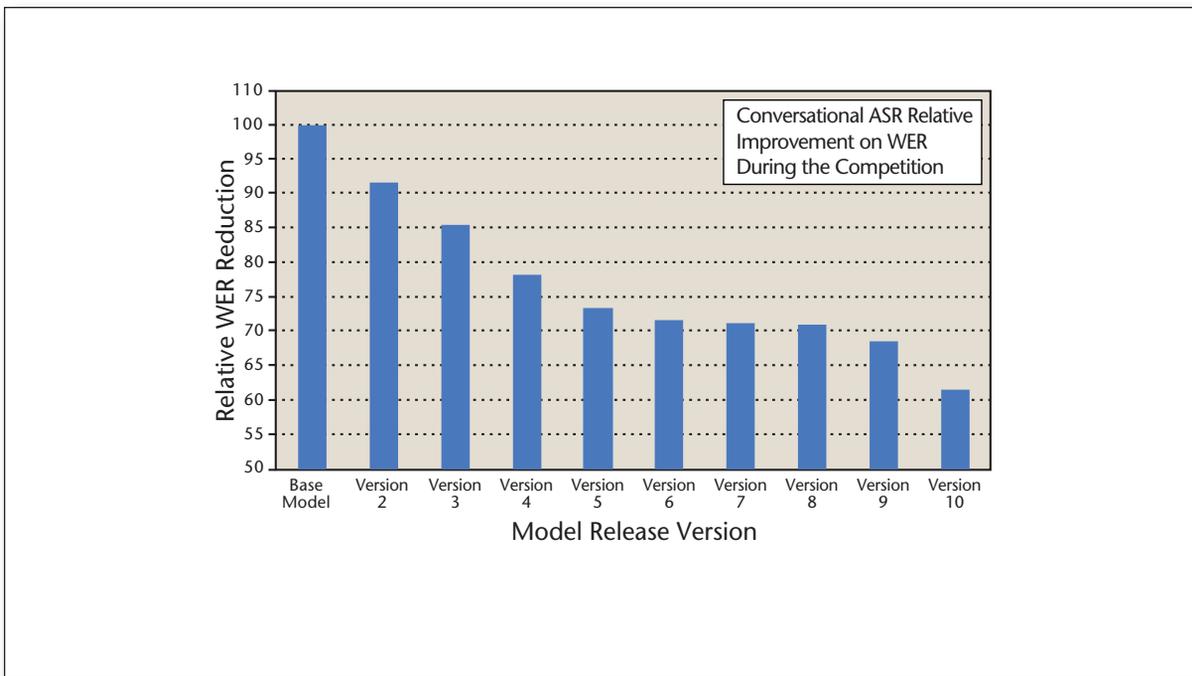


Figure 1. Conversational ASR Performance Improvement.

Relative reduction in WER with respect to base model.

quent sections describe the state of each of these problems and the advancements produced by all these teams.

Conversational Automatic Speech Recognition (ASR)

Speech is the gateway to voice-based agents, and errors in speech recognition can get propagated to later stages such as NLU and the dialogue manager, leading to incorrect or incoherent responses.

ASR is even more difficult in open-domain and non-task-oriented conversational agents. Free-form speech does not necessarily fit a command-like structure. It typically contains longer sentences, and the space of plausible word combinations is much larger. In addition, social conversations are informal, and open ended, they contain many topics, and they have a high out-of-vocabulary rate. Furthermore, production-grade ASR approaches must deal with a much wider array of noise and environmental conditions than the conditions in the normalized research datasets often reported in the literature. All of these make conversational ASR a challenging problem.

We developed a custom language model (LM) targeted specifically at open-ended conversations with socialbots. We initially used publicly available conversational datasets such as Fisher, Switchboard, Reddit comments, Linguistic Data Consortium (LDC)

news, OpenSubtitles, and Yelp reviews, along with the Washington Post news data, for building this conversational language model. As the competition progressed, we incrementally added data from Alexa Prize utterances collected over the course of the competition to our dataset. Performance of the model improved significantly on conversational test sets over the course of the competition. Figure 1 provides the relative improvement in word error rate (WER). We see a reduction by nearly 40 percent relative to the base model.

Conversational Natural Language Understanding (NLU)

Understanding the input of an interactor during a conversation is critical for dialogue systems. If machines cannot comprehend a user's intent or the topics and entities mentioned in an utterance, then the machine will not be able to respond well, which will lead to a poor customer experience. NLU in goal-oriented dialogue systems follows a domain, an intent, and a slot-like approach (Kumar et al. 2017). For example, in the utterance "Play Havana from Camila Cabello," the domain is music, the intent is to play a song, and the entities (modeled as slots) are "song_name: Havana, artist: Camila Cabello." These approaches work well in goal-oriented dialogue systems; however, open-domain dialogue systems are

free-form and are not confined to a predefined set of domains, intents, or entities. The intent or domain may be unclear and the slots may not be well defined. For example, consider the following utterance: “Last night, I went to Justin Bieber’s show. He was great, but the crowd was not. Do you think should I go next time?” The goal of the utterance is not well defined. The utterance consists of multiple intents: information delivery, opinion sharing, and opinion request. Furthermore, there are multiple slots in the same utterance. Traditional NLU systems do not work well for natural conversations. The following NLU components were developed during the Alexa Prize competition to address these challenges.

Conversation Intent

To connect an Alexa user with a socialbot, we first needed to identify whether the user’s intent was to have a conversation with Alexa. We introduced a “conversation intent” within the Alexa NLU model to recognize a range of utterances such as “let’s chat,” “let’s talk,” “let’s chat about <topic>,” and so forth, using a combination of grammars and statistical models. We further expanded the experience to other natural forms of conversational initiators such as “what are we going to talk about,” “can we discuss politics,” “do you want to have a conversation about the Mars Mission,” and so on.

In the production system, if an utterance from an Alexa user is identified as a conversation intent, then one of the Alexa Prize socialbots is invoked and the user interacts with that socialbot until the user says stop. Following the detection of conversational intent, the entire conversation is controlled by the socialbots. Teams used a combination of Alexa Skills Kit NLU along with their own NLU approaches, as will be described.

NLU Techniques Adopted by the Participating Teams

After a user has initiated a conversation, the socialbot requires an NLU system to identify semantic and syntactic elements from the user utterance, including user intent (such as opinion, chit-chat, knowledge, and others), entities and topics (for example, the entity “Mars Mission” and the topic “space travel” from the utterance “what do you think about the Mars Mission”), user sentiment, as well as sentence structure and parse information. A certain level of understanding is needed to generate responses that align well with user intent and expectation and to maximize user satisfaction. Although conversational utterances do not generally follow intent-slot structure, teams brought several workarounds to address this problem.

NLU is difficult because of the inherent complexities within the human language, such as anaphora, elision, ambiguity, and uncertainty, which require contextual inference in order to extract the necessary information to formulate a coherent response. These

problems are magnified in conversational AI since it is an open domain problem where a conversation can be on any topic or entity and the content of the dialogue can also change rapidly. Some specific techniques used by teams are listed in the following paragraphs.

Named Entity Recognition (NER)

Identifying and extracting entities (names, organizations, locations) from user utterances. Teams used various libraries such as StanfordCoreNLP (Manning et al. 2014), *pacy*,¹ and Alexa’s ASK NLU to perform this task. NER is helpful for retrieving relevant information for response generation, as well as for tracking conversational context over multiple turns.

Intent Detection

Intents represent the goal of a user for a given utterance, and the dialogue system needs to detect it to act and respond appropriately to that utterance. Some of the teams built rules for intent detection or trained models in a supervised fashion by collecting the data from Amazon Mechanical Turk or by using open source datasets, such as Reddit comments, with a set of intent classes. Others utilized Alexa’s ASK NLU engine for intent detection.

Anaphora and Coreference Resolution

Finding expressions that refer to the same entity in past or current utterances. Anaphora resolution is important for downstream tasks such as question answering and information extraction in multiturn dialogue systems. Most of the teams used StanfordCoreNLP’s Coreference Resolution System (Manning et al. 2014) to perform this task.

Sentence Completion

Some teams expanded user utterances with contextual information. For example, “Yes” can be transformed to “Yes, I like Michael Jackson” when uttered in the context of a question about the singer, or “I like Michael Jackson” can be extended to “I like Michael Jackson, singer and musician” in a conversation where this entity needs disambiguation. Teams wrote customized wrappers for performing sentence completion, which also involves querying knowledge bases to obtain more information about the entities, as described in the entity-linking section.

Topic and Domain Detection

Classifying the topic (for example, Seattle Seahawks) or domain (such as sports) from a user utterance. Teams used various datasets to train topic detection models, including news datasets, Twitter, and Reddit comments. Some teams also collected data from Amazon Mechanical Turk to train these models.

Entity Linking

Identifying information about an entity. Teams generally used publicly available knowledge bases such as Evi,² FreeBase (Bollacker et al. 2008), and Wikidata.³ Some teams also used these knowledge bases to identify related entities.

Text Summarization

Extracting or generating key information from documents for efficient retrieval and response generation. Some of the teams adopted this technique for summarizing the articles or potential responses for efficient response generation.

Sentiment detection

Identifying user sentiment. Some teams developed sentiment detection modules to help with generating engaging responses. This approach also helped them to better understand a user's intent and generate appropriate responses.

Knowledge Ingestion and Common Sense Reasoning

Currently, available conversational data is limited to datasets that have been produced from online forums (for example, Reddit), social media interactions (for example, Twitter), and movie subtitles (for example, OpenSubtitles, Cornell Movie-Dialogs Corpus). While these datasets are useful at capturing the syntactic and semantic elements of conversational interactions, they also have many issues with data quality, content (profanity, offensive data), query-response pair tracking, context tracking, multiple users interacting without a specific order, and short and ephemeral conversations. In the absence of better alternatives, teams still used these datasets. To address offensive content and profanity, teams built classifiers to detect this content. Furthermore, we shared *Washington Post* (WaPo) live comments, which are conversational in nature and also highly topical. Several teams made use of these comments.

The teams also used various knowledge bases, including Amazon's Evi, Freebase, and Wikidata for retrieving general knowledge, facts, and news, and for general question answering. Some teams also used these sources for entity linking, sentence completion, and topic detection. Ideally, a socialbot should be able to ingest and update its knowledge base automatically; however, this is an unsolved problem and an active area of research. Finally, teams also ingested information from news sources such as the Washington Post and CNN to keep current with news events that users may want to chat about.

For commonsense reasoning, several teams built modules to understand user intent. Some teams pre-processed open source and Alexa Prize datasets and extracted information about trending topics and opinions on popular topics, integrating them within their dialogue manager to make the responses seem as natural as possible. To complement commonsense reasoning, some of the top teams added user satisfaction modules to improve both engagement and conversational coherence.

To make sure that teams were leveraging relevant datasets and knowledge bases, we emphasized early availability of live user interactions to the socialbots,

which helped the teams in identifying relevant data sources before the competition went live.

Dialogue and Context Modeling

A key component of any conversational agent is a robust system to handle dialogues effectively. The system should accomplish two main tasks: help break down the complexity of the open domain problem to a manageable set of interaction modes, and be able to scale as the diversity and breadth of topics expands. A common dialogue strategy used by teams was a hierarchical architecture with a main dialogue manager (DM) and multiple smaller DMs corresponding to specific tasks, topics, or contexts.

Some teams, such as Sounding Board, used a hierarchical architecture and added additional modules such as an error handler to handle cases such as low-confidence ASR output or low-confidence response candidates (Fang et al. 2017).⁴ Other teams, such as Alquist, (Pichl, J. et al. 2017)⁴ used a structured topic-based dialogue manager, where components were broken up by topics, along with intent-based dialogue modules broken up by intents. Generally, teams also incorporated special-purpose modules such as a profanity or offensive content module to filter a range of inappropriate responses and modules to address feedback and acknowledgement and to request clarity or rephrasing from users. Teams experimented with approaches to track context and dialogue states, and corresponding transitions to maintain dialogue flow. For example, Alquist and Slugbot (Bowden et al. 2017)⁴ modeled dialogue flow as a state graph. These and other techniques helped socialbots produce coherent responses in an ongoing multi-turn conversation and guided the direction of the conversation as needed. A few teams, such as Magnus (Prabhumoye et al. 2017),⁴ built finite-state machines (FSMs) (Wright 2005) for addressing specific modules such as movies, sports, and others. One challenge in using this technique for dynamic components is scaling and context switching; however, for small and static modules, FSMs can be useful.

The top teams focused not only on response generation but also on customer experience, and experimented with conversational strategies to increase engagement as discussed in the next section.

Conversational User Experience

Participating teams built several conversational user experience (CUX) modules, which included engagement, personalization, and other user experience-related aspects. CUX modules are relatively easy to build, but such modules may lead to significant gains on ratings and duration. CUX is an essential component, and the teams that focused most of their efforts on NLU and DM, with less emphasis on CUX, were not received as top performers by Alexa users. Following are the five main components built by various teams.

Personalization

Socialbots received an obfuscated (by one-way hash function) user ID to enable personalization for repeat users while maintaining user privacy. Alquist built a personalization module to remember past interactions and users. This module was used to give a more natural and personal touch in initiating conversations. Sounding Board tried multiple strategies to get more information about user preferences on topics. For example, they developed a personality quiz that enabled them to tailor topics based on the user's personality. They found that extroverts, as determined by the personality quiz, correlated with higher ratings and longer turns. Edina added a level of personalization to their DM to track whether a certain topic or type of response is doing particularly well with a user (Krause et al. 2017).⁴

Topic Switching

Alana used a multibot strategy consisting of data-driven bots and rule-based bots, including Eliza/Template, Persona, Quiz Game, NewsBot, Factbot, Evi, and Weatherbot. When presenting the information from a submodule, for example, NewsBot, Alana determined whether to keep or change the topic based on user feedback. Edina identified topical drift to recognize when the customer wants to set a new topic or keep the current one. Emerson focused on machine-driven conversation through a topic-recommendation mechanism for conversation topic transition.

Initiative

Initiative in conversation is another key variable — should the bot guide a user to certain topics of conversation, let the user steer topics, or mix these approaches? SlugBot designed a system-initiative module to direct the conversation through stories, games, and informing the user of various headlines. Edina collected data on various topics through Amazon Mechanical Turk (AMT) and built proactive modules to drive conversation with users. Emerson investigated four levels of initiative: passive, less passive, active, and less active. Based on their experiments, the Emerson team concluded that conversations actively driven by bots give the best performance on rating and duration. ChattyChat used machine-initiated dialogues to drive conversation by asking yes or no questions along with topic suggestion (Yi and Jung 2017).

Sentiment-Based Modules

Sounding Board gauged user reaction through sentiment analysis and user decisions on certain cases, particularly opinion-related utterances. They developed a dissatisfaction detector by training a classifier to detect when a user expresses discontent, which was used to trigger a change of topic. Magnus built a classifier to filter abusive content. RubyStar had a module to avoid explicit topics such as pornography, as well as other sensitive subjects. In such cases, RubyStar responded with predefined templated responses.

Engagement and Greeting Modules

Responses that engage and captivate a user, entertaining them and leaving them wanting to keep talking, are critical to developing great conversation. RubyStar used an approach called engagement re-ranking that was trained on Reddit comments (Liu et al. 2017).⁴ They split comments into engaging (high number of upvotes) and nonengaging (low number of upvotes). Roving Mind considered three semantic dimensions (Cervone et al. 2017)⁴ proposed in Dialogue Act Markup Language (Bunt et al. 2010): social obligation (addressing basic social conventions such as greetings and feedback to user); addressing user feedback to a machine's statement; and task (addressing user actions). WiseMacaw developed a two-layer chatbot framework that can be described as a metagame, where the user could walk to multiple gaming modules (Ji et al. 2017).⁴

Several teams started adding games, quizzes, and related modules in their interactions, which led to significant increase in ratings and duration with certain segments of users. However, such interactions are not necessarily conversational; furthermore, they did not advance the state of conversational AI, which was the main objective of this competition. To address these issues, we guided teams to remove such modules and eliminated their inclusion in the final round of the competition.

Response Generation

There are four main types of approaches for response generation in dialogue systems: template/rule-based, retrieval, generative, and hybrid. A functional system can be an ensemble of these techniques. It can follow a waterfall structure, for example, rules → retrieval → generative. Or it can use a hybrid approach with complementary modules, for example, generative models for retrieval, or generative models to create templates for a retrieval or rule-based module. Some of the teams used AIML, ELIZA (Weizenbaum 1966), or Alicebot⁵ for rule-based and templated responses. Teams also built retrieval-based modules that tried to identify an appropriate response from the dataset of dialogues available. Retrieval was performed using techniques such as n-gram matching and entity matching, or using similarity metrics based on vectors such as TF-IDF, word/sentence embeddings, skip-thought vectors, and dual-encoder systems.

Hybrid approaches leveraging retrieval in combination with generative models are fairly new and have shown promising results in the past couple of years, usually with sequence-to-sequence approaches with some variants. Some of the Alexa Prize teams created novel techniques along these lines and demonstrated scalability and relevance for the open-domain, response-generation models deployed in production systems. MILABot, for example, devised a hierarchical latent variable

encoder-decoder (VHRED) (Serban et al. 2016) model, in addition to other neural network models such as skip thought (Kiros et al. 2015) to produce hybrid retrieval-generative candidate responses. Some teams (such as Pixie)(Adewale et al. 2017)⁴ used a two-level, long short-term memory (LSTM) model (Hochreiter and Schmidhuber 1997) for retrieval. Eigen (Guss et al. 2017)⁴ and RubyStar, on the other hand, used dynamic memory networks (Sukhbaatar et al. 2015) and character-level recursive neural networks (RNN) (Sutskever, Vinyals, and Le 2014) for generating responses. Alquist used a sequence-to-sequence model (Sutskever, Vinyals, and Le 2014) specifically for their chit-chat module. While these teams deployed the generative models in production, other teams also experimented with generative and hybrid approaches offline.

Ranking and Selection Techniques

Open-domain social conversations do not always have a specific goal or target, and the response space can be unbounded. There may be multiple valid responses for a given utterance. As such, identifying the response that will lead to the highest customer satisfaction and help drive the conversation forward is a challenging problem. Socialbots need mechanisms to rank possible responses and select the response that is most likely to achieve the goal of a coherent and engaging conversation in that particular dialogue context. Alexa Prize teams attempted to solve this problem with either rule-based or model-based strategies.

For teams that experimented with rule-based rankers, a ranker module chose a response from the candidate responses obtained from submodules (such as topical modules or intent modules) based on some logic. For model-based strategies, teams utilized either a supervised or reinforcement learning approach, trained on user ratings (Alexa Prize data) or on predefined large-scale dialogue datasets such as Yahoo! Answers, Reddit comments, Washington Post Live comments, and OpenSubtitles. The ranker was trained to provide higher scores to correct responses (for example, follow-up comments on Reddit are considered correct responses) while ignoring incorrect or noncoherent responses obtained by sampling. Alan (Papaioannou et al. 2017),⁴ for example, trained a ranker module on Alexa Prize ratings data and combined that with a separate ranker function that used hand-engineered features. Teams using a reinforcement learning approach developed frameworks where the agent was a ranker, the actions were the candidate responses obtained from submodules, and the agent was trying to maximize the trade-off between selecting a response to satisfy the customer immediately and selecting one that takes into account some long-term reward. MILABot, for example, used this approach and trained a reinforcement learning ranker function on conversation ratings.

The afore-mentioned components form the core of socialbot dialogue systems. In addition, we developed the following components to support the competition.

Conversational Topic Tracker

To understand the intent of a user, it is important to identify the topic of the given utterance and corresponding keywords. Alexa Prize data is highly topical because of the nature of the social conversations. Alexa users interacted with socialbots on hundreds of thousands of topics in various domains such as sports, politics, entertainment, fashion, and technology. This is a unique dataset collected from millions of human-conversational agent interactions. We identified the need for a conversational topic tracker for various purposes such as conversation evaluation (for example, coherence, depth, breadth, diversity), sentiment analysis, entity extraction, profanity detection, and response generation.

To detect conversation topics in an utterance, we adopted deep average networks (DAN) and trained a topic classifier on interaction data categorized into multiple topics. We proposed a novel extension by adding topic-word attention to formulate an attention-based DAN (ADAN) (Guo et al. 2017) that allows the system to jointly capture topic keywords in an utterance and perform topic classification. We fine-tuned the model on the data collected during the course of the competition. The accuracy of the model was obtained to be 82.4 percent on 26 topical classes (sports, politics, movies_TV, and so on). Furthermore, the topic model was also able to extract keywords corresponding to each topic. We used the conversational topic tracker to evaluate the socialbots on various metrics such as conversational breadth, conversational depth, and topical and domain coverage (Venkatesh et al. 2017; Guo et al. 2017). We will explore additional details on evaluation later in this article.

Inappropriate and Sensitive Content Detection

One of the most challenging aspects of delivering a positive experience to end users in socialbot interactions is to obtain high-quality conversational data. The datasets most commonly used to train dialogue models are sourced from internet forums (for example, Reddit, Twitter) or movie subtitle databases (for example, OpenSubtitles, Cornell Movie-Dialogs Corpus). These sources are all conversational in structure in that they can be transformed into utterance-response pairs. However, the tone and content of these datasets are often inappropriate for interactions between users and conversational agents, particularly when individual utterance-response pairs are taken out of context. In order to effectively use dialogue models based on these or other dynamic data sources, an efficient mechanism to identify and filter

different types of inappropriate and offensive speech is required.

We identified several (potentially overlapping) classes of inappropriate responses: (1) profanity, (2) sexual responses, (3) racially offensive responses, (4) hate speech, (5) insulting responses, and (6) violent responses (inducements to violent acts or threatening responses). We explored keyword- and pattern-matching strategies, but these strategies are subject to poor precision (with a broad list) or poor recall (with a carefully curated list), as inappropriate responses may not necessarily contain profane or other blacklisted words. We tested a variety of support vector machines and Bayesian classifiers trained on n-gram features using labeled ground truth data. The best accuracy results were in profanity (>97 percent at 90 percent recall), racially offensive responses (96 percent at 70 percent recall), and insulting responses (93 percent at 40 percent recall). More research is needed to develop effective offensive speech filters. In addition to dataset cleansing, an offensive speech classifier is also needed for online filtering of candidate socialbot responses prior to outputting them to ASK for text-to-speech conversion.

Addressing Problems in Evaluating Conversational Agents

Social conversations are inherently open ended. For example, if a user asks the question “What do you think of Barack Obama?,” there can be thousands of distinct, valid, and reasonable responses. That is, the response space is unbounded for open-domain conversations. This makes training and evaluating social, non-task-oriented, conversational agents extremely challenging. It is easier to evaluate a task-oriented dialogue system because we can measure systems by successful completion of tasks, which is not the case with open-ended systems. As with human-to-human dialogues, an interlocutor’s satisfaction with a socialbot could be related to how engaging, coherent, and enjoyable the conversation was. The subjectivity associated with evaluating conversations is a key element underlying the challenge of building non-goal-oriented dialogue systems.

This problem has been heavily studied but lacks a widely agreed-upon metric. A well-designed evaluation metric for conversational agents that addresses the above concerns would be useful to researchers in this field. There is significant previous work on evaluating goal-oriented dialogue systems. Two of those notable earlier works are TRAINS system and PARADISE (Walker et al. 1997). All of these systems involve some subjective measures that require a human in the loop. Due to the expensive nature of human-based evaluation procedures, researchers have been using automatic machine translation (MT) metrics, such as BLEU, or text summarization metrics, such as ROUGE, to evaluate systems. But as shown by Liu et

al. (2017), these metrics do not correlate well with human expectations.

The Turing Test (Turing 1950) is a well-known test that can potentially be used for dialogue evaluation. However, we do not believe that the Turing Test is a suitable mechanism to evaluate socialbots for the following reasons:

Incomparable elements: Given the amount of knowledge an AI has at its disposal, it is not reasonable to suggest that a human and an AI should generate similar responses. A conversational agent may interact differently from a human, but may still be a good conversationalist.

Incentive to produce plausible but low-information content responses: If the primary metric is just generation of plausible human-readable responses, it is easy to opt out of the more challenging areas of response generation and dialogue management. It is important to be able to source interesting and relevant content while generating plausible responses.

Misaligned objectives: The goal of the judge should be to evaluate the conversational experience, not to attempt to get the AI to reveal itself.

To address these issues, we propose a comprehensive, multimetric evaluation strategy designed to reduce subjectivity by incorporating metrics that correlate well with human judgement. The proposed metrics provide granular analysis of the conversational agents, which is not captured in human ratings. We show that these metrics can be used as a reasonable proxy for human judgment. We provide a mechanism to unify the metrics for selecting the top performing agents, which has also been applied throughout the Alexa Prize competition. The following objective metrics (Guo et al. 2017; Venkatesh et al. 2017) have been used for evaluating conversational agents. The proposed metrics also align with the goals of a socialbot, that is, the ability to converse coherently and engagingly about popular topics and current events.

Conversational user experience (CUX): Different users have different expectations concerning the socialbots, and so their experiences might vary widely since open-domain dialogue systems involve subjectivity. To address these issues, we used average ratings from frequent users as a metric to measure CUX. With multiple interactions, frequent users have their expectations established and they evaluate a socialbot in comparison to others.

Coherence: We annotated hundreds of thousands of randomly selected interactions for incorrect, irrelevant, or inappropriate responses. With the annotations, we calculated the response error rate (RER) for each socialbot, using that figure to measure coherence.

Engagement: Evaluated through performance of conversations identified as being in alignment with socialbot goals. Measured using duration, turns, and ratings obtained from engagement evaluators (a set of Alexa users who were asked to evaluate socialbots based on engagement).

Algorithm	RMSE	Spearman	Pearsonr
Random	2.211	0.052	0.017
HLSTM	1.392	0.232	0.235
GBDT	1.34	0.352	0.351

Table 1. Correlation of the Regression Model with User Ratings.

Domain coverage: Entropy analysis of conversations against the five socialbot domains for Alexa Prize (Sports, Politics, Entertainment, Fashion, Technology). Performance was targeted on high entropy, while minimizing the standard deviation of the entropy across multiple domains. High entropy ensures that the socialbot is talking about a variety of topics, while a low standard deviation gives us confidence that the metric is applied equally across domains.

Topical diversity: Obtained using the size of topical vocabulary for each socialbot. A higher topical vocabulary within each domain implies more topical affinity.

Conversational depth: We used the topical model to identify the domain for each individual utterance. Conversational depth for a socialbot was calculated as the average of the number of consecutive turns on the same topical domain, where single turn corresponds to user utterance and corresponding bot response pair within a conversation. Conversational depth evaluates the socialbot's ability to have multiturn conversations on specific topics within the five domains.

Selecting Alexa Prize Finalists

The Alexa Prize competition was structured to allow users to participate in the selection of finalists. Two finalists were selected purely on the basis of user ratings averaged over all the conversations with those socialbots. At the end of the conversation, users were asked to rate how coherent and engaging the conversation was.

In addition, one finalist was selected by Amazon based on internal evaluation of coherence and engagement of conversations by over one thousand Amazon employees who volunteered as Alexa Prize judges, on analysis of conversational metrics computed over the semifinals period, and on scientific review of the team's technical papers by senior Alexa scientists. The quality of all the socialbots was also analyzed based on the metrics mentioned above. We observed that a majority of those metrics correlate well with user ratings, frequent ratings, and ratings from Alexa Prize judges, with a correlation coefficient

greater than 0.75. A simple combination of the metrics correlated strongly with Alexa user ratings (0.66), suggesting that the "wisdom of crowds" (Surowiecki 2004) is a reasonable approach to evaluation of conversational agents when conducted at scale in a natural setting. The average rating across all socialbots was lower by 20 percent for the judge's pool as compared with the general public.

Teams also evaluated the quality of their socialbots and made necessary improvements during the competition by leveraging the ratings and feedback from users. Alexa users had millions of interactions and over 100,000 hours of conversations with socialbots throughout the duration of the competition.

Automatic Evaluation of Open-Domain Conversations

If we are able to build a model that can predict the rating of an Alexa Prize conversation with reasonable accuracy, then it is possible to remove humans from the loop for evaluating non-task-oriented dialogues.

To automate the evaluation process, we did a preliminary analysis on 60,000 conversations and ratings, and we trained a model to predict user ratings. We observed the Spearman and Pearson correlations of 0.352 and 0.351 respectively (table 1) with significantly low p-value with a model trained using a gradient-boosted decision tree (GBDT). Although the results for GBDT are significantly better than random selection for five classes and the model trained using hierarchical LSTM, there is a need to extend this study to millions of Alexa Prize interactions. Furthermore, some of the evaluation metrics (coherence, topical depth, topical breadth, domain coverage) obtained at conversation level can also be used as features. With a significantly higher number of conversations combined with topical features, we hypothesize that the model would perform much better than the results obtained in the preliminary analysis shown in table 1. Given subjectivity in ratings, we appropriately found interuser agreement to be quite low for ratings analysis. Users may have their own

criteria to evaluate the socialbots. Therefore, as a part of the future work, we will train the model with user-level features as well.

The Alexa Prize Finals

Following the conclusion of the semifinals, three finalists were identified: Alquist (Czech Tech University), Alana (Heriot-Watt University), and Sounding Board (University of Washington). These finalists remained online for the remainder of the competition. They entered the last phase of the competition with an average rating of 2.77. After two months of additional interaction with Alexa users, they went to the finals with an average rating of 3.48, an improvement of 26 percent from their earlier ratings. The Alexa Prize finals were conducted in November as a closed-door event with three rounds of conversations with each socialbot. We invited professional judges well versed in conversation: talk show hosts, authors, speakers, and AI experts. Three judges were “conversationalists” and asked to engage in conversations with the socialbots; five judges were “evaluators” and asked to listen to and rate the conversations. The winner was Sounding Board from the University of Washington with an average score of 3.17 on a 5-point scale and an average conversation duration of 10 minutes and 22 seconds.

The objective of the finals was to evaluate socialbots on their ability to conduct coherent and engaging voice conversations with humans and to chain relevant entities into a dialogue. Highlights of these conversations are available in a short film.⁶ The best interactions combined many of the techniques discussed in this article.

For example, a conversation that started off with a request to discuss the book *The Dark Tower* evolved in the following manner: The Dark Tower (book) → The Dark Tower (movie) → Matthew McConaughey’s performance in the movie → Other movies that he’s acted in and that the user may find interesting → a topical humorous suggestion, “Someone needs to make a mashup of *Interstellar* and *Contact* where Matthew McConaughey tries to prevent Matthew McConaughey from going into space.” Such relevant transitions and natural injection of humor give us a glimpse into the future of conversational AI.

Another example that combines entity resolution and reference disambiguation in the context of an introductory ice-breaker experience:

Socialbot: Hi. This is an Alexa Prize socialbot. How’s it going?

User: It’s going great, how about you?

Socialbot: Awesome. I’m glad you’re doing well today. I’m having a good day, thanks for asking. Ready to enjoy our chat. If you like, you can tell me your name?

User: My friends call me John, but you can call me Mr. Pickles.

Socialbot: Excellent, I’m happy to meet you Mr. Pickles.

There were some examples of failure points as well (such as ASR failures). The field has a long way to go yet, but there is definite promise and improvement in the accomplishments of this first year, and we look forward to seeing more gains made through Alexa Prize 2018.

The finalist socialbots will continue to be a part of the “let’s chat” experience until the Alexa Prize 2018 socialbots become available to the public in May 2018. To speak with the socialbots yourself, simply say “Alexa, let’s chat” to any Alexa-enabled device.

Results

The Alexa Prize was designed as a framework to support research on conversational AI at scale in a real-world setting. The scientific advances described above (and detailed in individual team papers) resulted in significant improvements in socialbot quality and a significant amount of user engagement.

User Engagement

Customer engagement remained high throughout the competition. Alexa Prize ranked in the top 10 Alexa skills by usage, with over 40,000 hours of conversations spanning millions of utterances by the end of the finals. Customers chatted on a wide range of popular and current topics with movies/TV, music, politics, celebs, business, and scitech being the highest frequency (most popular) topics. The most popular topics from the post-semifinals feedback phase were movies/TV (with an average rating of 3.48), scitech (3.60), travel/Geo (3.51), and business (3.48). Based on user ratings, the three lowest rated topics were arts (with an average rating of 2.14), shopping (2.63), and education (3.03).

It is still early in the Alexa Prize journey towards natural human conversation, but the high level of engagement and feedback (over 130,000 hours of conversation to date) demonstrates that users are interested in chatting with socialbots and supporting their development.

Socialbot Quality

Over the course of the competition, socialbots showed a significant improvement in customer experience. The three finalists improved their ratings by 29.6 percent (from 2.77 to 3.59) over the duration of competition. All 15 socialbots had an average customer rating of 2.87, with a median conversation duration of 1:35 minutes and a 90th percentile of 5:43 minutes by the end of the semifinal phase. The conversation duration of finalists across the entire competition was 1:41 minutes (median) and 8:02 minutes (90th percentile), improving 19.4 percent and 58.26 percent respectively from the start of the competition, with 10 turns (median) per conversation.

We measured response error rate (RER) through the

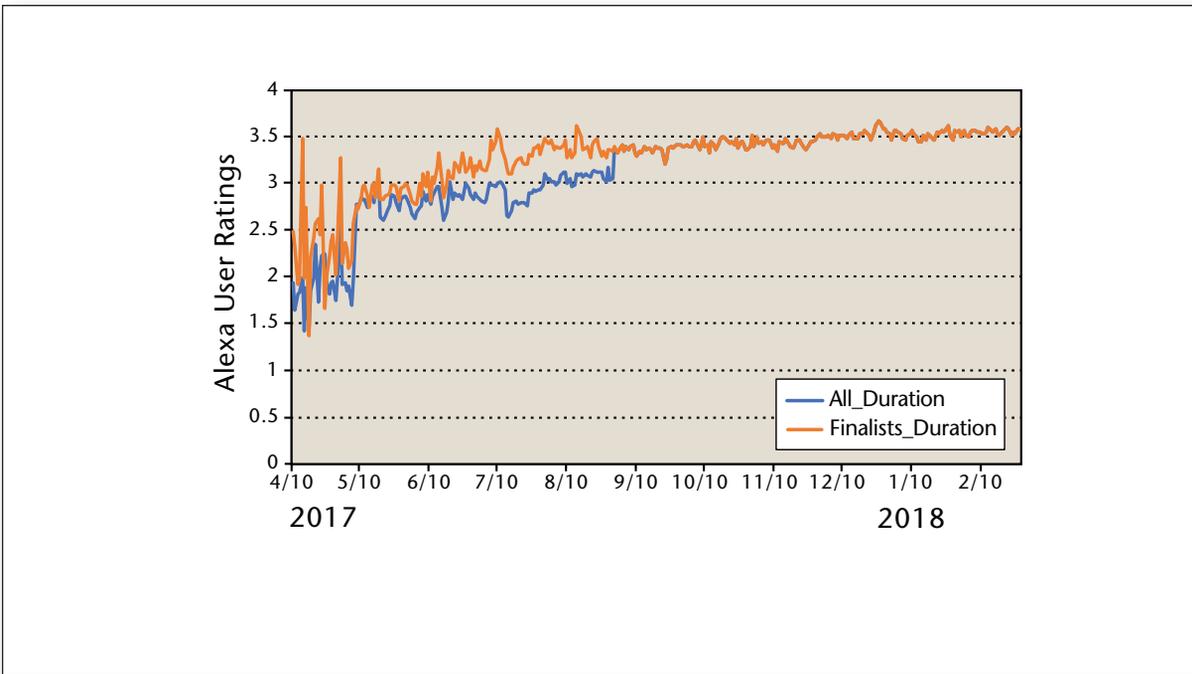


Figure 2. Daily Ratings for Socialbots During the Competition.

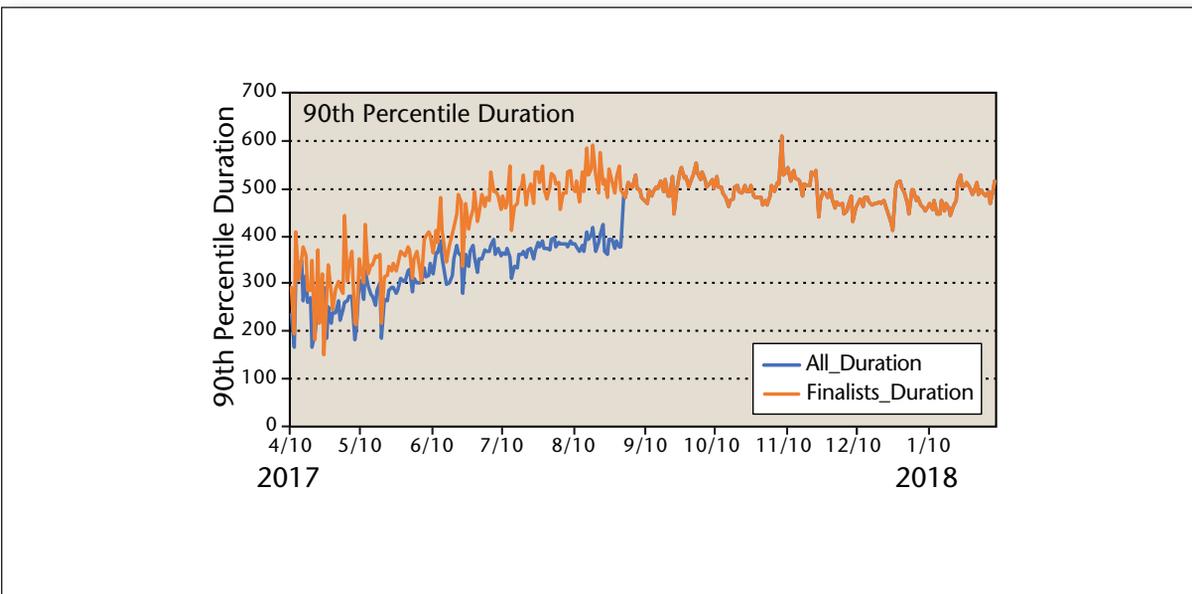


Figure 3. Conversation Duration Median and 90th Percentile for Socialbots During the Competition.

manual annotation of a large fraction of user utterance-social response pairs by trained data analysts to identify incorrect, irrelevant, or inappropriate responses. RER was quite irregular during the first 30 days of the launch, from May 8 to June 8, 2017, fluctuating between 8.5 percent and 36.7 percent. This fluctuation was likely due to rapid experimentation

by teams in response to initial user data. During the semifinals, from July 1 to August 15, 2017, RER was in the range of 20.8 percent to 28.6 percent. The three finalists improved further over the post-semifinals feedback phase, and their average RER was at 11.21 percent (L7D: last 7 days) as they went into finals.

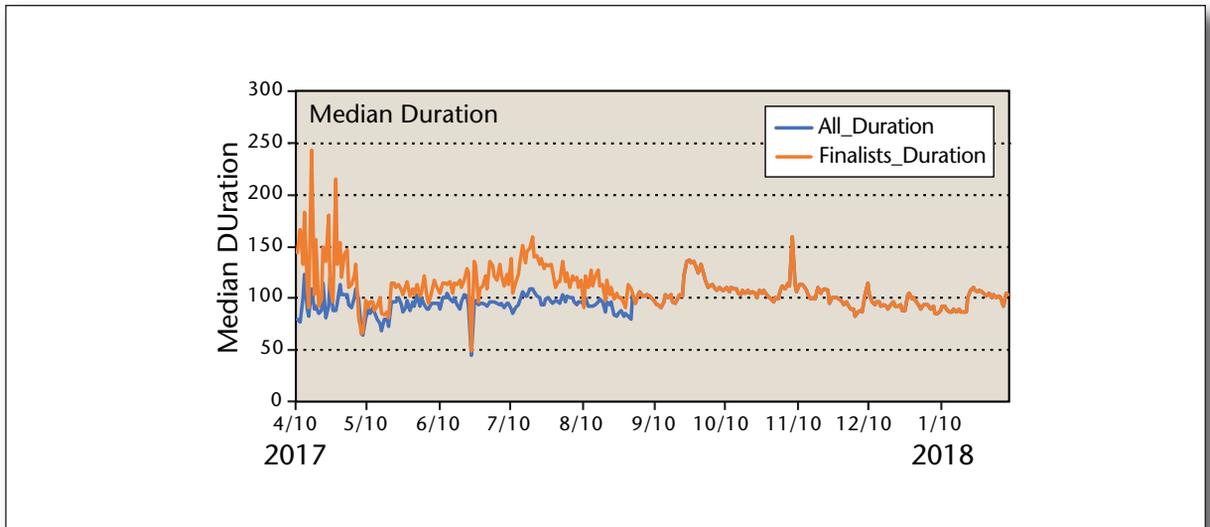


Figure 4. Conversation Duration 90th Percentile for Socialbots During the Competition.

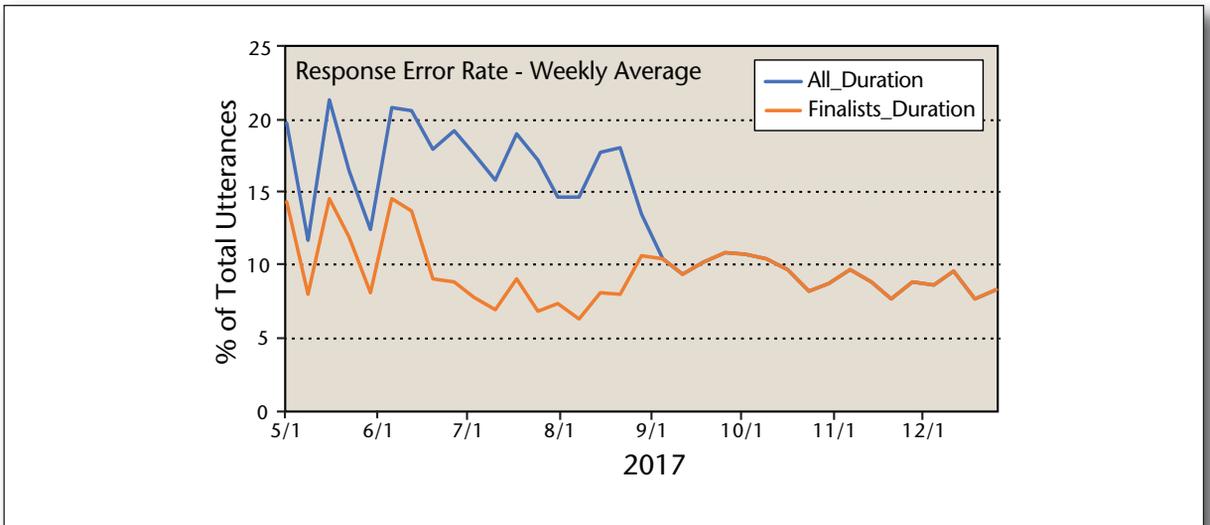


Figure 5. Response Error Rate for Individual Utterance-Response Pairs for All Socialbots and Finalists During the Competition.

Conclusion and Future Work

Fifteen teams went live in the inaugural Alexa Prize, and customer ratings improved by about 24 percent over the duration of the competition (May 8 to Nov 13). An analysis of the technical and scientific detail for each team in relation to their performance in the competition led to the following findings:

The following components are key for building an effective socialbot: (1) dialogue manager (DM), (2) natural language understanding (NLU) module, (3)

knowledge module, (4) response generation, (5) conversational user experience (CUX) handler, (6) ranking and model selection policy module.

Teams that focused on building CUX modules saw significant gains on ratings and duration. CUX is an essential component, and the teams who focused most of their efforts on NLU and DM, with less emphasis on CUX, were not received as top performers by Alexa users.

A robust NLU system supported by strong domain coverage leads to high coherence. Teams who invest-

ed in building a strong NLU and knowledge components had the lowest response error rates leading to higher user ratings.

Different conversational goals call for different response-generation techniques, suggesting that retrieval, generative, and hybrid mechanisms may all be required within the same system. When the performance of a socialbot has converged, generative and hybrid modules combined with a robust ranking and selection module can lead to a better conversational agent.

A response ranking and selection model greatly impacts socialbot quality. The teams who built a strong model-selection policy had significant improvements in ratings and average number of dialogue turns.

Even if a socialbot has strong response-generation and ranker modules, lack of good NLU and DM components adversely affect user ratings.

We expected that the grand challenge of 20-minute conversations would take many years to achieve — the Alexa Prize was set up as a multiyear competition to enable sustained research on this problem. Despite the difficulty of the challenge, it is extremely encouraging to see the work that the inaugural cohort of the Alexa Prize has achieved in year one of the competition. We have seen significant advancements in research, and in the quality of socialbots as observed through the customer ratings, but much remains to be achieved. With the help of Alexa users and the science community, Alexa Prize 2018 will continue to work towards the goal of 20-minute-long coherent and engaging social conversations, and continue to advance the state of conversational AI.

Acknowledgments

We would like to thank all the university students and their advisors (Alexa Prize Teams 2017) who participated in the competition. We would also like to thank the entire Alexa Prize team (Eric King, Kate Bland, Qing Liu, Jeff Nunn, Ming Cheng, Ashish Nagar, Yi Pan, Han Song, SK Jayadevan, Amanda Wartick, Anna Gottardi, Gene Hwang, Art Pettigrew, and Nate Michel) for their contribution in making the Alexa Prize competition a success. We would also like to thank Amazon leadership and Alexa principals for their vision and support through this entire program; the marketing, public relations, and legal departments for helping drive the right messaging and a high volume of traffic to the Alexa Prize skill, ensuring that the participating teams received real-world feedback for their research; Alexa engineering for all the support and work on enabling the Alexa Prize skill and supporting a custom Alexa Prize ASR model, while always maintaining operational excellence; and Alexa machine learning for continued support with NLU and data services, which allowed us to capture user requests to initiate conversations and also provide high-quality annotated feedback to

the teams. We also want to thank ASK leadership and the countless teams in ASK who helped us with the custom APIs for Alexa Prize teams, enabling skill beta testing for the Alexa Prize skills before it went general availability, and who further supported us with skill management, QA, certification, marketing, operations, and solutions. We would also like to thank the Alexa experiences organization for exemplifying customer obsession by providing us with critical input to share with the teams on building the best customer experiences and driving us to track our progress against customer feedback.

Finally, thank you to the Alexa customers who engaged in tens of thousands of hours of conversations spanning millions of interactions with the Alexa Prize socialbots and who provided the feedback that helped teams improve over the course of the year.

Notes

1. pacy.io.
2. www.evi.com.
3. www.wikidata.org/wiki/Wikidata:Main_Page.
4. See the Alexa Prize Proceedings, developer.amazon.com/alexaprize/proceedings.
5. alicebot.sourceforge.net/alice_page.htm.
6. www.youtube.com/watch?v=WTGuOg7GXU&feature=youtu.be.
7. developer.amazon.com/alexaprize/2017/teams.

References

- Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1247–50. New York: Association for Computing Machinery.
- Adewale, O.; Beatson, A.; Buniatyan, D.; Ge, J.; Khodak, M.; Lee, H.; Prasad, N.; Saunshi, N.; Seff, A.; Singh, K.; Suo, D.; Zhang, C.; Arora, S. 2017. Pixie: A Social Chatbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Bowden, K. K.; Wu, J.; Oraby, S.; Misra, A.; Walker, M. 2017. Slugbot: An Application of a Novel and Scalable Open Domain Socialbot Framework. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; Soria, C.; and Traum, D. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the International Conference on Language Resources and Evaluation*. Luxembourg: European Language Resources Association.
- Cervone, A.; Tortoreto, G.; Mezza, S.; Gambi, E.; Riccardi, G. 2017. Roving Mind: A Balancing Act between Open-Domain and Engaging Dialogue Systems. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Fang, H.; Cheng, H.; Clark, E.; Holtzman, A.; Sap, M.; Ostendorf, M.; Choi, Y.; Smith, N. 2017. Sounding Board – University of Washington’s Alexa Prize Submission. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

- Guo, G.; Metallinou, A.; Khatri, C.; Raju, A.; Venkatesh, A.; and Ram, R. 2017. Topic-Based Evaluation for Conversational Bots. Paper presented at the NIPS 2017 Workshop on Conversational AI — Today's Practice and Tomorrow's Potential. Long Beach, CA, December 8.
- Guss, W. H.; Bartlett, J.; Kuznetsov, P.; Patil, P. 2017. Eigen: A Step Towards Conversational AI. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–80.
- Ji, J.; Wang, Q.; Battad, Z.; Gou, J.; Zhou, J.; Divekar, R.I.; Carlson, C.; Si, M.. 2017. A Two-Layer Dialogue Framework for Authoring Social Bots. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Skip-Thought Vectors. In *Annual Conference on Neural Information Processing Systems*. Advances in Neural Information Processing Systems 28. Red Hook, NY: Curran Associates, Inc.
- Krause, B.; Damonte, M.; DobreM.; Duma, D.; Fancellu, F.; Kahembwe, E.; Cheng, J.; Fainberg, J.; Webber, B. 2017. Edina: Building an Open Domain Socialbot with Self-Dialogues. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Kumar, A.; Gupta, A.; Chan, J.; Tucker, S.; Hoffmeister, B.; Dreyer, M.; Peshterliev, S.; Gandhe, A.; Filiminov, D.; Rastrow, A.; Monson, C.; and Kuymar, A. 2017. Just ASK: Building an Architecture for Extensible Self-Service Spoken Language Understanding. arXiv preprint arXiv:1711.00549 [cs.CL]. Ithaca, NY: Cornell University Library.
- Levesque, H. J. 2017. *Common Sense, the Turing Test, and the Quest for Real AI: Reflections on Natural and Artificial Intelligence*. Cambridge, MA: The MIT Press.
- Liu, H.; Lin, T.; Sun, H.; Lin, W.; Chang, C.-W.; Zhong, T.; Rudnicky, A. 2017. RubyStar: A Non-Task-Oriented Mixture Model Dialog System. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations*, 55–60. Stroudsburg, PA: Association for Computational Linguistics.
- Papaioannou, I.; Curry, A.; Part, J.; Shalyminov, I.; Xu, X.; Yu, Y.; Dusek, O.; Rieser, V.; Lemon, O. 2017. Alana: Social Dialogue Using an Ensemble Model and a Ranker Trained on User Feedback. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Pichl, J.; Marek, P.; Konrád, J.; Matulík, P.; Nguyen, H. L.; Sedivy, J. 2017. Alquist: The Alexa Prize Socialbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Prabhumoye, S.; Botros, F.; Chandu, K.; Choudhary, S.; Keni, E.; Malaviya, C.; Manzini, T.; Pasumarthi, R.; Poddar, S.; Ravichander, A.; Yu, Z.; Black, A. 2017. Building CMU Magnus from User Feedback. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Serban, I.; Sankar, C.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Chandar, S.; Ke, N. R.; Rajeswar, S.; de Brebisson, A.; Sotelo, J. M. R.; Suhubdy, D.; Michalski, V.; Nguyen, A.; and Bengio, Y. 2017. The Octopus Approach to the Alexa Competition: A Deep Ensemble-Based Socialbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.
- Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. arXiv preprint arXiv:1605.06069[cs.CL]. Ithaca, NY: Cornell University Library.
- Sukhbaatar, S.; Szymanski, A.; Weston, J.; and Fergus, R. 2015. Weakly Supervised Memory Networks. arXiv preprint arXiv:1503.08895[cs.NE]. Ithaca, NY: Cornell University Library.
- Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business*. New York: Doubleday.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Annual Conference on Neural Information Processing Systems*, Advances in Neural Information Processing Systems 27, 3104–12. Red Hook, NY: Curran Associates, Inc.
- Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460.
- Venkatesh, A.; Khatri, C.; Ram, A.; Guo, F.; Gabriel, R.; Nagar, A.; Prasad, R.; Cheng, M.; Hedayatnia, B.; Metallinou, A.; Goel, R.; Yang, S.; and Raju, A. 2017. On Evaluating and Comparing Conversational Agents. Paper presented at the NIPS 2017 Workshop on Conversational AI — Today's Practice and Tomorrow's Potential. Long Beach, CA, December 8.
- Vinyals, O., and Le, Q. 2015. A Neural Conversational Model. Paper presented at the ICML Deep Learning Workshop. Lille, France, July 10.
- Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. arXiv preprint arXiv:cmp-lg/9704004 [cs.CL]. Ithaca, NY: Cornell University Library.
- Weizenbaum, J. 1966. Eliza — A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9(1): 36–45.
- Wright, D. R. 2005. Finite State Machines: CSC215 Class Notes. Unpublished Class Notes. Raleigh, NC: North Carolina State University.
- Yi, S.; Jung, K.. 2017. A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy. A Two-Layer Dialogue Framework for Authoring Social Bots. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Chandra Khatri is an AI scientist at Amazon Lab126 with a research and development team responsible for making Alexa conversational. Currently, he is the leading scientist for Alexa Prize competition. Some of his recent work involves open-domain dialogue planning and evaluation, conversational speech recognition, conversational natural language understanding, and topic modeling. Prior to Alexa, Khatri was a research scientist at eBay in an applied science group. At eBay, he led various deep learning and NLP initiatives, such as automatic text summarization and automatic content generation within the ecommerce domain. He holds degrees in machine learning and computational science and engineering from the Georgia Institute of Technology and the Birla Institute of Technology and Science.

Anu Venkatesh is a technical program manager on the Alexa AI team and is responsible for execution of the Alexa Prize, a university competition to advance conversational

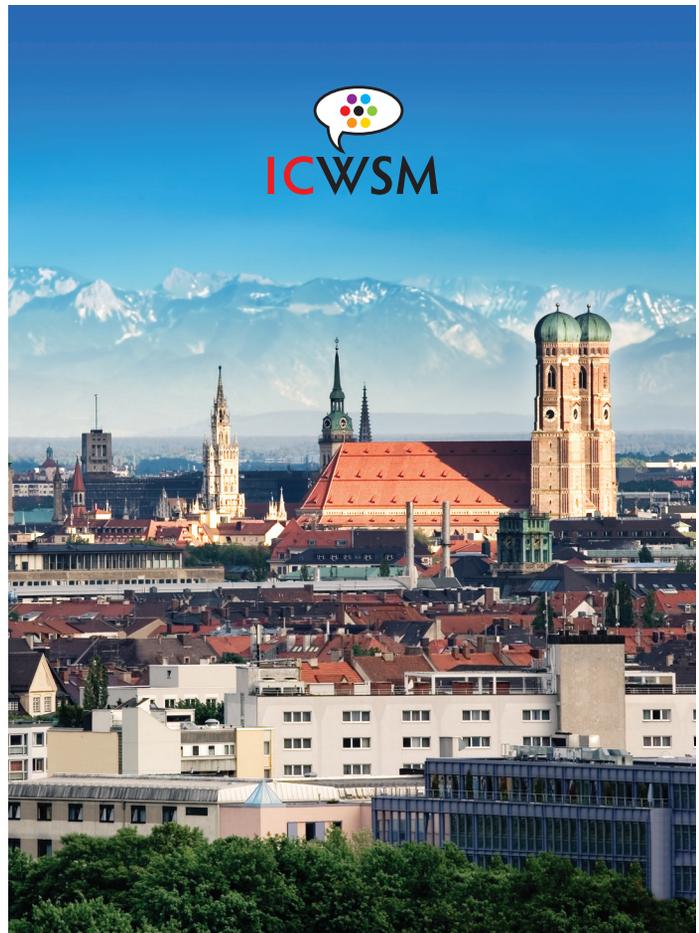
AI, and additional technical initiatives targeted at making Alexa more conversational. Her background is in speech, dialogue, and natural language understanding. Prior to Alexa, Venkatesh was a research scientist at Nuance Communications working on their speech recognition and virtual assistant technologies. Venkatesh received her master's degree in AI from the Georgia Institute of Technology in 2009 and a bachelor's degree in computer science from the PES Institute of Technology, India in 2007.

Behnam Hedayatnia is an applied scientist on the Alexa AI team. He has been working on conversational AI since 2017. His main focus is on driving conversational ASR and dialogue for the Alexa Prize. He received his MS in electrical engineering at the University of California, San Diego, in 2017, while working as a research scientist at the San Diego Super Computer Center. He received his BS in electrical engineering from the University of California, San Diego, in 2015.

Ashwin Ram is technical director of AI in the office of the CTO for Google Cloud. He focuses on bringing Google AI to the world through deep personalized engagement with the leadership of top companies to reimagine their businesses by leveraging the power of AI. He also works with Google's AI teams to drive new technologies and capabilities that address customer needs. Prior to Google, Ram was senior manager of AI science for Amazon Alexa. He led cross-functional R&D initiatives to create advanced conversational AI technologies for intelligent agents, including the university-facing Alexa Prize competition. Ram received his PhD from Yale University in 1989, his MS from University of Illinois in 1984, and his BTech from IIT Delhi in 1982.

Raefer Gabriel is the lead for the Alexa Prize, and manager of AI software development at Alexa AI. He focuses on driving the next generation of natural and social interaction on Alexa, and on leading the teams building models, infrastructure, and development toolkits to support open and multidomain conversation. Prior to Amazon, Gabriel was the CEO of machine intelligence studio Delvv, Inc., and chief scientist and cofounder of Reputation.com. He has founded other companies, including TruExchange, a futures trading software company, and worked in the hedge fund industry as a quantitative analyst. Gabriel received his MBA from Columbia Business School in 2007. He graduated from Harvard University with an AB in physics in 2000.

Rohit Prasad is vice president and head scientist of Alexa Artificial Intelligence. Prasad leads Alexa research and development in speech recognition, natural language understanding, and machine learning technologies. Prior to Amazon, Prasad was deputy manager and senior director of the speech, language, and multimedia business unit at Raytheon BBN Technologies. In that role, he directed US Government-sponsored research and development initiatives in speech-to-speech translation, psychological health analytics, document image translation, and STEM learning. Prasad earned his master's degree in electrical engineering at the Illinois Institute of Technology, Chicago, and a bachelor's degree in electronics and communications engineering from Birla Institute of Technology, India.



Save the Date!

ICWSM 2019

Munich, Germany

11–14 June 2019

General Chair
Jürgen Pfeffer

Program Committee Cochairs
Ceren Budak, Yu-Ru Lin, Fred Morstatter

Local Chair
Mirco Schönfeld

www.icwsm.org/2019