Further Advances in Open Domain Dialog Systems in the Third Alexa Prize Socialbot Grand Challenge

Raefer Gabriel[1] **Yang Liu**[1] **Anna Gottardi**[1] **Mihail Eric**[1]
Anju Khatri[1] **Anjali Chadha**[1] **Qinlang Chen**[1] **Behnam Hedayatnia**[1]
Pankaj Rajan[1] **Ali Binici**[1] **Shui Hu**[1] **Karthik Gopalakrishnan**[1]
Seokhwan Kim[1] **Lauren Stubel**[1] **Kate Bland**[1] **Arindam Mandal**[1]
Dilek Hakkani-Tür[1]

[1]Amazon Alexa AI
{raeferg, yangliud, gottardi, mihaeric}@amazon.com
{anjkh, anjch, qinlangc, behnam}@amazon.com
{pnrajan, binicm, shuihu, karthgop}@amazon.com
{seokhwk, stubells, kateblan, arindamm}@amazon.com
hakkanit@amazon.com

Abstract

Building open domain conversational systems that allow users to have engaging conversations on topics of their choice is a challenging task. The Alexa Prize Socialbot Grand Challenge was launched in 2016 to tackle the problem of achieving natural, sustained, coherent and engaging open-domain dialogs. In the third iteration of the competition, university teams have moved the needle on the state of the art, bringing together common sense knowledge representations, neural response generation models, NLU systems enhanced by large-scale transformer models and improved dialog policies to switch between graph-based representations or retrieval-based or templated dialog fragments, along with generated responses. The Third Socialbot Grand Challenge included an improved version of the CoBot (conversational bot) toolkit from the prior competition, along with topic and dialog act detection models, conversation evaluators, and a sensitive content detection model so that the competing teams could focus on building knowledge-rich, coherent and engaging multi-turn dialog systems. This paper outlines the advances developed by the university teams as well as the Alexa Prize team to move closer to the Grand Challenge objective. We address several key open-ended problems such as conversational speech recognition, open domain natural language understanding, commonsense reasoning, statistical dialog management and dialog evaluation. These collaborative efforts have driven improved ratings in the Semifinals of the competition from 3.19 in the prior competition cycle to 3.47 across all teams, an increase of 8.8%. As of the end of the final feedback phase, the top 7-day average rating achieved by a socialbot was 3.71 (out of 5), with the top 90th percentile conversation duration at 14 minutes 19 seconds.

1. Introduction

Conversational AI is one of the most challenging problem domains in artificial intelligence, due to the subjectivity involved in interpreting human language. Broadly speaking, we consider the Conversational AI domain to include a range of tasks in natural language understanding, knowledge representation, common-sense reasoning, dialog evaluation and natural language generation. Complete solutions to these problems will likely require a system at parity with human intelligence (Hassan et al., 2018; Xiong et al., 2016). With recent advances in deep learning, we have made significant progress toward solutions for problems within other very challenging domains, such as speech recognition and image recognition in computer vision. However, many of these advances have occurred due to the objective nature of evaluating

solutions to these problems and the resulting availability of high quality labeled data with objective ground truth. Such is not the case with many tasks in the Conversational AI domain. The language and response generation task in particular has a potentially unbounded response space and a resulting lack of objective success metrics, making it a highly challenging problem to model effectively.

Voice assistants such as Alexa and Google Assistant have significantly advanced the state of the art for goal-directed conversations and successfully deployed these systems in production. However, building agents that can carry on multi-turn, open-domain conversations is still far from a solved problem. To address these challenges and further advance the state of Conversational AI, Amazon launched the Alexa Prize Socialbot Grand Challenge in 2016. The grand challenge objective is to build agents that can converse coherently and engagingly with humans for 20 minutes, and obtain a 4 out of 5, or higher rating, from humans interacting with them. Apart from the Alexa Prize, there have been challenges like Dialog System Technology Challenge (DSTC) (Williams et al., 2016) (task based and closed domain) and Conversational AI Challenge (ConvAI) (Burtsev et al., 2018) (persona based, chit-chat and challenges with evaluation). Both of these challenges are text based as opposed to speech-based. Achieving natural, sustained, coherent and engaging open-domain dialogs in spoken form, and in real time, which can be evaluated and measured for success, is the primary goal of the Alexa Prize Socialbot Grand Challenge. Through this competition, participating universities have been able to conduct research and test hypotheses by buildings socialbots that interact with real Alexa customers.

As in the last two cycles of the Alexa Prize Socialbot Grand Challenge, upon receiving a request to engage in a conversation with Alexa, e.g. “Alexa, Let's Chat”, Alexa customers were read a brief message, then connected to one of the 10 participating socialbots. Customers were provided instructions on how to end the conversation and provide ratings and feedback. The introductory message and instructions changed through the competition to keep the information relevant to the different phrases. After exiting the conversation with the socialbot, which the user could do at any time, the user was prompted for a verbal rating: “How do you feel about speaking with this socialbot again?”, followed by an option to provide additional freeform feedback. Ratings and feedback were both subsequently shared with the teams to help them improve their socialbots.

The Third Socialbot Grand Challenge was launched to a cohort of Amazon employees on October 28, 2019, followed by a public launch on December 2, 2019 at which time all US Alexa customers could interact with the participating socialbots. Unlike in prior years, we ran an initial feedback phase, followed by a competitive quarterfinals from February 3 through March 17, 2020. One team was eliminated from competition after the quarterfinals, and the remaining teams participated in the Semifinals from March 20 through April 29, 2020. Five teams qualified to participate as Finalists, and participated in an additional feedback phase through May 22. The closed door Finals were held on July 14-16, 2020. Throughout the competition, the teams were required to maintain anonymity in their interactions to ensure fairness in the competition. To drive maximum feedback to the teams and improve user engagement, the Alexa Prize experience was promoted through Echo customer emails, social media, and blogs.

Over the course of the Third Socialbot Grand Challenge competition, we drove over 240,000 hours of conversations spanning tens of millions of interactions, 255% higher than we saw in the 2018 competition. University teams have improved the state of the art for open domain dialog systems in various aspects, including better NLU systems, neural response generation models, common sense knowledge modeling, and dialog policies for more smooth and engaging conversations. The Third Socialbot Grand Challenge also included an improved version of the CoBot (conversational bot) toolkit from the prior competition, along with improved speech recognition, NLU components and neural response generation module. These collaborative efforts have resulted in improved ratings and longer conversations.

2. Capabilities Provided to Teams

To help teams focus more on scientific advances and reduce time invested into infrastructure, hosting and scaling complex model-driven bots, we created CoBot, a conversational bot toolkit in Python for natural language understanding and dialog management, as described in detail in Khatri et al. (2018) and shown in Figure 1. The toolkit provides a set of tools, libraries and base models designed to help develop, train and deploy open-domain or multi-domain conversational experiences through the Alexa Skills Kit (ASK). For the Third Socialbot Grand Challenge, we released an updated version of this software, CoBot v2.

The primary goal of CoBot is to drive improved quality of conversational agents by providing a toolkit that is modular, extensible, and scalable, and provide abstractions for infrastructure and low-level tasks. CoBot provides a continuous integration pipeline where experimentation in language understanding, dialog management techniques, and response generation strategies can be integrated and tested by a single command. This enables seamless scaling from development and test to production workloads on AWS. CoBot uses many of the same principles found in the Node.js, Python, and Java SDKs of the Alexa Skills Kit or ASK (Kumar et al., 2017), as well as general dialog toolkits like Rasa (Bocklisch et al., 2017) and DeepPavlov (Burtsev et al., 2017). CoBot exposes generalized dialog management, state tracking, and Natural Language Understanding (NLU) capabilities that are modular in nature, as illustrated in Figure 2. Unlike other toolkits, CoBot places an emphasis on infrastructure to host models and handle massive scale at runtime.

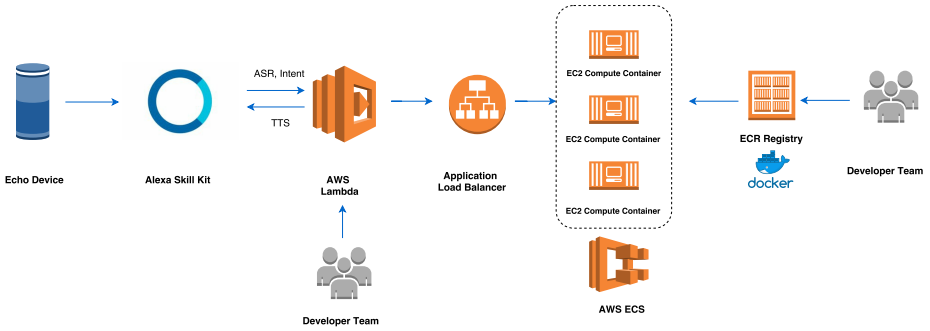


Figure 1 CoBot System Diagram and Workflow

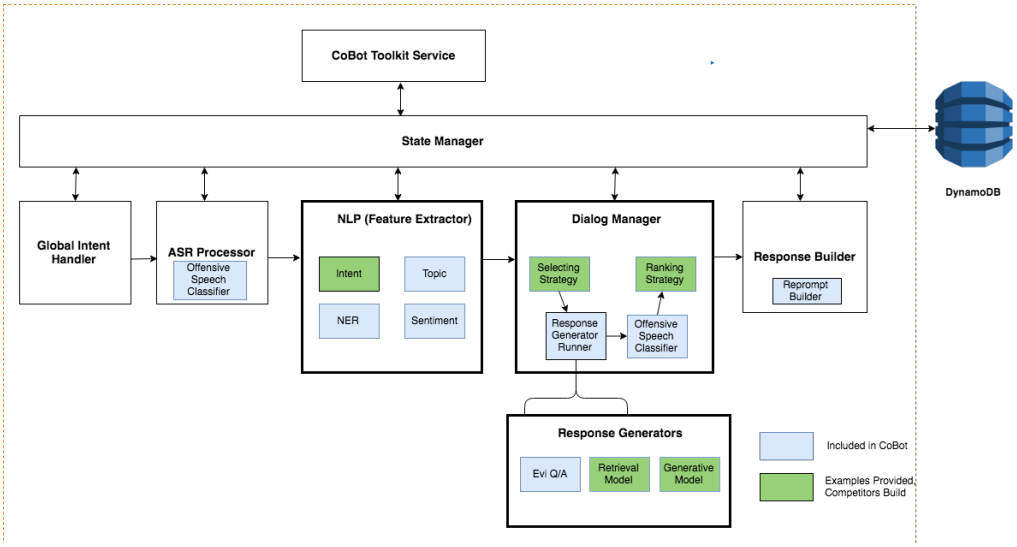


Figure 2 CoBot Architecture

3. Scientific Advancements

3.1 From the Alexa Prize Participants

Overall, the socialbot framework and structure used by most teams is similar to the previous years of the competition, consisting of NLU, dialog manager, and response generation. Teams have improved NLU performance to better understand user's utterances. For response generation, significant effort has been spent on increasing the topical coverage of the socialbot, using external knowledge sources, as well as exploratory work in common sense reasoning to improve response quality, and neural response generation models. In dialog management and policy, different approaches have been adopted to manage multiple skills or topics. It is also worth noting that teams also investigated methods to build more personalized and empathetic socialbots. In the following we briefly describe a few advancements from the university teams. Please refer to the papers from the teams for more details.

3.1.1 NLU

NLU is still a critical component in the socialbot systems since most current systems heavily rely on intermediate NLU results to select and produce system responses, rather than an end-to-end approach that takes just the user's utterance. The general NLU pipeline is mostly similar to previous years, however teams have continued to improve different components in the pipeline, mostly notably using improved neural network models, adapting methods to the Alexa Prize domain, and introducing new components. For example, property detection (identifying relationship between two entities or in triples) and tense classification have been introduced to Alquist (Pichl et al. 2020).

3.1.1.1 Sentence Segmentation

ASR results for user utterances do not have punctuation marks or the notion of sentences. Therefore, the first step of many systems is to perform sentence segmentation in order to obtain the proper units for subsequent understanding components. Neural models, particularly using BERT or Transformer-based techniques (Devlin et al. 2019), have been adopted by several teams for this task including DREAM (Kuratov et al. 2020), Alquist, Athena (Harrison et al. 2020), and Gunrock (Liang et al. 2020). These models are often trained using out-of-domain data, such as Switchboard or Movie corpus, and then fine tuned with in-domain data.

3.1.1.2 Dialog Act, Intent, and Topic Classification

Recognizing the user's intent and topic is necessary to generate a system response and carry on a conversation. In the Cobot toolkit, a Dialog Act (DA) and topic classification service is provided for teams. In addition to using that service, some teams also developed their own models. For example, for topic classification, Tartan (Chen et al. 2020) explored different CNN-based modeling approaches; Emora (Finch et al. 2020) further classified utterances when their topic labels are 'other'. For DA and intent classification, Gunrock used active learning to improve model performance; Athena expanded the DA ontology used by Gunrock, with more types for user questions, and used an ensemble model for DA classification; Alquist also introduced hierarchical DA classes; Chirpy Cardinal (Paranjape et al. 2020) built on the DA model based on Gunrock and optimized for their own socialbot; Tartan developed an LSTM model for question detection.

3.1.1.3 Entity Linking

Important information is conveyed by the entities in user's utterances, therefore most teams have tried to leverage entity information for user utterance understanding and response generation. This year some teams used the Evi knowledge graph provided by Amazon. Some explored other external services. For example, Gunrock used Google knowledge graph, and used similarity measures based on the BERT representation to further improve performance. Other teams developed their own entity linking systems, such as Athena and Chirpy Cardinal. Another effort related to entity linking is the use of a personal graph, in addition to the entities for general domains such as Wikipedia, as in Alquist.

3.1.1.4 Sentiment and Emotion Recognition

Recognizing the sentiment and emotion in user's utterances is also key to generating high quality responses. Sentiment analysis was performed in some teams in previous years, and this work continued this year. In particular, many teams used available sentiment classifiers such as Vader (used by Audrey (Hong et al. 2020), Bernard (Majumder et al. 2020), Emora). Some developed their own models, for example, Gunrock built a BERT-based model. In addition to sentiment, there was more exploration of emotion recognition. Another BERT-based model was used in DREAM, while Audrey used the TL-ERC emotion detection model, and Chirpy Cardinal used the RoBERTa model (Liu et al. 2019). This work was important in taking steps toward developing bots that can show empathy and are socially intelligent.

3.1.2 Response Generation

Generating responses to users' utterances followed several approaches, similar to previous years, including scripted dialogs for some domains or topics, template-based methods that rely on NLU results, and retrieval-based methods. To deliver high quality responses, teams have heavily used external knowledge sources, such as Wikipedia, Reddit, or other dialog data resources. Even for template-based results, some teams investigated approaches to generate responses with variations that fit the context. For example, Alquist performed operations including lexicalization and changing expressions based on quality (singular vs. plural), pronouns, and temporal or tense information. Many teams also included responses to discuss and express opinions. It is worth noting that the covered topics in this year's socialbot systems have significantly increased.

In the following sections, we briefly describe areas of significant new explorations in this year's competition, neural response generation, and common sense reasoning.

3.1.2.1 Neural Response Generation

Neural network-based language generation has seen great progress in the past several years. Chitchat systems that are purely based on end-to-end neural models have demonstrated strong performance (Adiwardana et al. 2020, Roller et al. 2020). To improve the coverage of the system responses or to address the long tail problem, teams have evaluated or included neural response generation in their systems.

Amazon provided a GPT-based (Radford et al. 2018) neural response generation service during the competition and migrated this to GPT-2 (Radford et al. 2019) during the competition, however, as it was not available at the competition start, many teams had already planned their own work in this area or did not adopt these releases. Some teams conducted preliminary studies using this service (Zotbot and Athena), and evaluated NRG with respect to different parameters, for example, with or without some selected knowledge, and different dialog context (Schallock et al. 2020).

Several other teams developed their own NRG models. Audrey, Bernard and Chirpy Cardinal trained GPT or GPT-2 models using the Topical Chat dataset (Gopalakrishnan et al. 2019). In Gunrock's system, a method was proposed to explicitly adjust the token prediction probability depending on the knowledge, rather than just using the knowledge as conditions in the GPT-2 model. Audrey used GPT models for empathic response generation.

There are other efforts related to neural generation models for response generation. Zotbot explored using reading comprehension to generate accurate answers, as well as question generation (based on GPT-2 models) to converse about an article in a more natural manner. Chirpy Cardinal also investigated neural generation approaches to paraphrase responses from formal written text to a conversational style using Topical Chat data.

While most of these studies show promise and can indeed address some of the issues in other response generation methods, significant limitations in their training techniques and inference performance still exist. Additional research is required before the methods in these studies can be fully adopted in real-time socialbot settings.

3.1.2.2 Commonsense Response Generation

In socialbot conversations, users often talk about personal experiences. For example, Tartan's analysis showed about 20% of the utterances are related to personal experience. However, current NLP models and many conversational AI techniques lack common sense reasoning ability. Some teams have tackled this problem in order to generate more sensible and natural responses in their socialbots to any general topics such as hobbies and daily activities.

DREAM used the COMeT (Bosselut et al. 2019) and Atomic (Sap et al. 2019) models to generate predictions for different aspects related to daily activities (e.g., what a person feels during or after the activity, what a person needed for the activity), ask clarification questions about daily activities, and also generate information needed for templates for a conversation about personal events. Zotbot also used COMeT to generate an object output given a subject and a relation from the knowledge base. These efforts have demonstrated promise in use of this technology for improving coverage and response quality, while reducing incorrect or low quality responses.

3.1.3 Dialog Management and Dialog Policy

The response generation modules above may be developed for a particular domain, topic, or skill. However, an overall dialog manager is responsible for implementing a general dialog policy, which includes selecting the right responses for the given dialog context, choosing the next topic, and smoothly transitioning to topics that users may find engaging. Specific implementations of the dialog graph in dialog managers vary greatly among different teams, ranging from state machines, dialog flows, to treelets (Chirpy Cardinal). In several bots, there is a global dialog manager that decides the overall flow, selects a skill or responder (or multiple ones), and finally selects the responses to use. Each topic or skill may have its own dialog manager.

The dialog policy used in this dialog manager is rule-based in many cases. Teams recognize the importance of acknowledging the user's utterances, and have this component embedded into the policy (e.g., DREAM, Audrey, Gunrock). Alquist used adjacency pairs in their dialog policy to capture predefined relations between utterances, and the action creator in their dialog manager took actions based on which responses are generated. DREAM used a learning-based response selector, with features including those derived from the most recent neural models, such as textual entailment based on RoBERTa models. In Audrey, under certain conditions, the topic selection problem is formulated as a Markov decision process, and

reinforcement learning is leveraged to select the next topic based on customers' explicit preference for certain topics and their ordering.

Personalization is another important aspect of the socialbot. Many teams designed the system such that it can talk with the user not just about factual knowledge, or even having opinion-oriented conversation, but rather carry personal conversations (e.g, Emora, Tartan). Several teams proposed new solutions for personalization. Gunrock built user profiles, and used open questions to determine the best topic based on user's responses. In Audrey, a personality understanding module was introduced to infer the user's interests, and a reinforcement learning based approach (as mentioned above) was used to select the most personalized topics.

3.2 From the Alexa Prize Team

3.2.1 Dialog Act and Topic Classification

We developed a new model for both dialog act (or intent) and topic classification tasks. The model is based on hierarchical recurrent neural network (HRNN) that includes the following two RNNs (as shown in Figure 3). The first RNN takes the word sequence of each utterance and learns the utterance representation. The other RNN takes a sequence of the utterance representations in a given dialog context and learns the dialog context representation. The model was trained by multi-task learning with two objectives for dialog topic and intent classification tasks. Specifically, we used unidirectional gated recurrent units (GRUs) for both RNN layers, each of which outputs its final hidden state as the representation of the utterance and its context, respectively. The context representation at the end of the second RNN is fed into two fully-connected layers with softmax, which generate the probability distributions over the topic and intent labels.

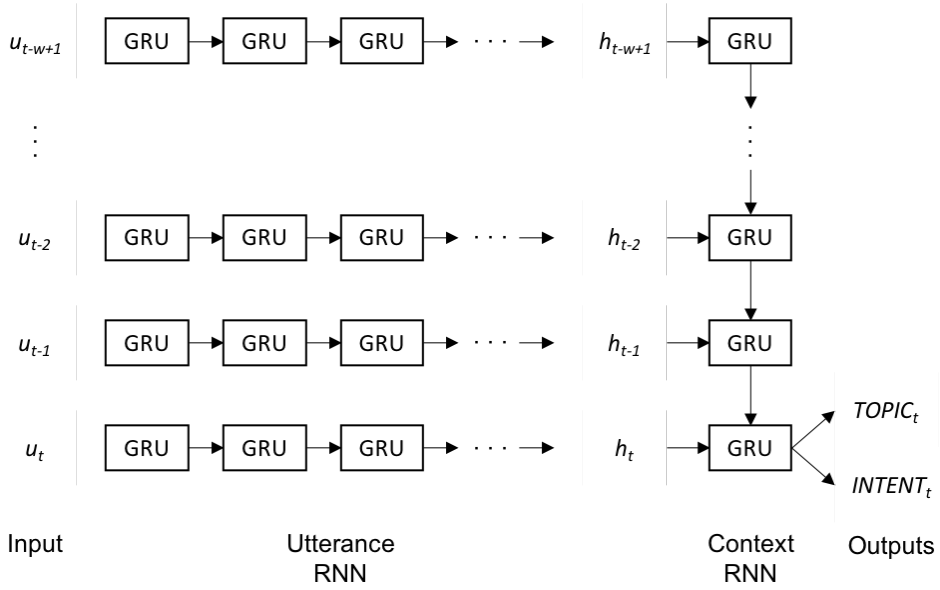


Figure 3. HRNN Topic and Intent Classification

To demonstrate the effectiveness of our new model architecture, we conducted experiments on the same portion of Alexa Prize conversations as in [1], which describes our previous models. Figure 4 compares the performances of topic and intent classification with alternative approaches. HRNN models outperformed the previous models with Deep Average Network (DAN) and BiLSTMs on both topic and intent

classification tasks. The joint HRNN model achieved further performance improvement over the single task models for both tasks.

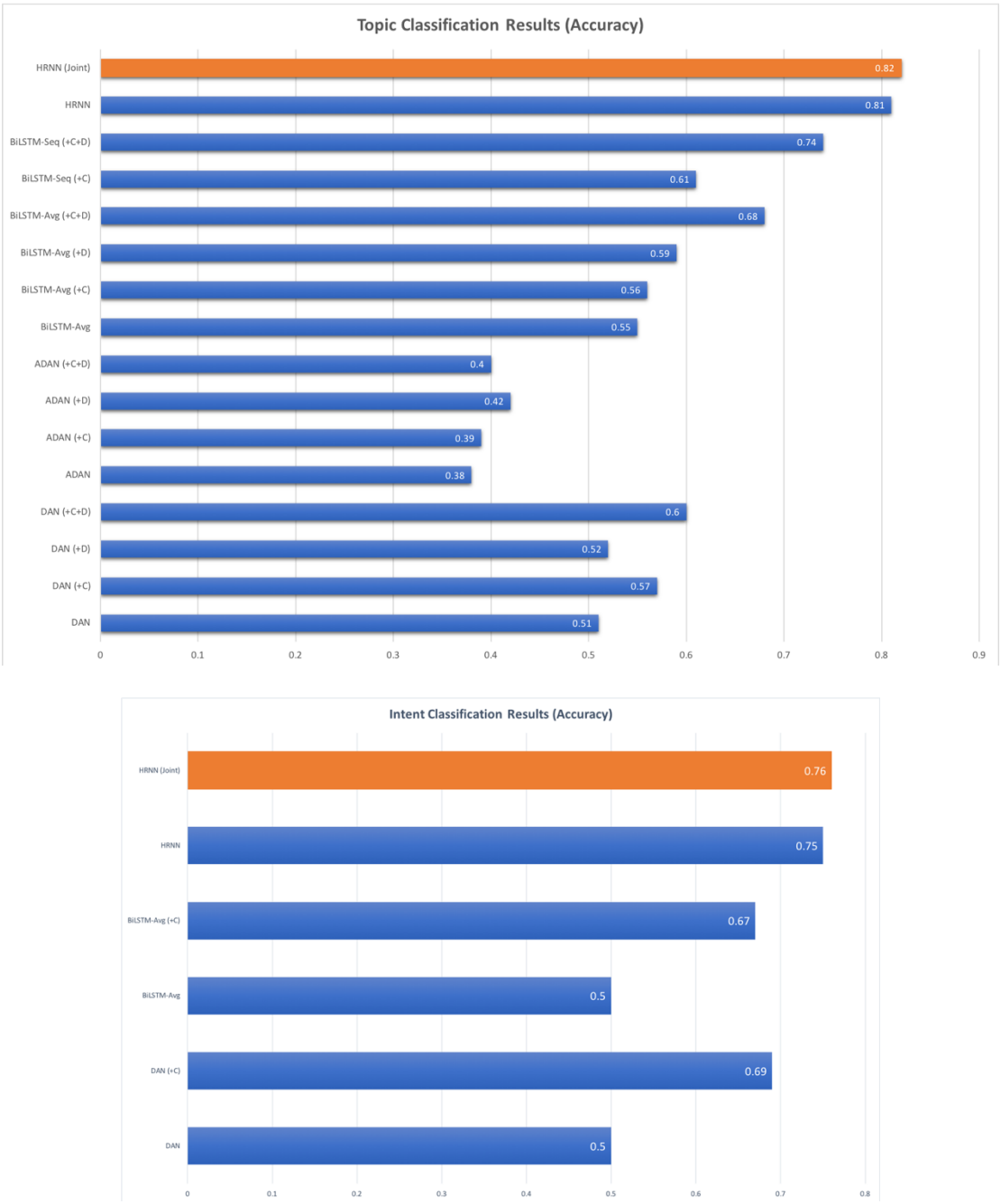


Figure 4. Topic and Intent Classification Results

3.2.2 Topical Chat Dataset

To enable Alexa Prize teams to build high-quality open-domain socialbots in a data-driven fashion, we collected and open-sourced a unique dataset, Topical-Chat, containing more than 10k human-human

dialogues that are grounded in knowledge spanning 8 broad topics: Fashion, Books, Movies, Music, Sports, Politics, Science & Technology and General Entertainment (Gopalakrishnan et al. 2019). Topical-Chat is unique among other knowledge-grounded dialogue datasets in several ways: no explicit Wizard-Apprentice style roles were assigned for partners, the topical breadth, depth and transitions in each dialogue, long utterances and extended dialogue lengths. The entities underlying the Topical-Chat dataset are popular entities that users discussed with Alexa Prize socialbots in prior years of the competition.

3.2.3 Neural Response Generation

Building on the latest advances in large pre-trained language models, we fine-tuned GPT and then GPT-2 on Topical-Chat and exposed a Neural Response Generation service utilizing these models to Alexa Prize teams. The API arguments were defined for simplicity: users pass a dialogue history and some relevant knowledge, and the API returns a knowledge-grounded response. The API also allowed for batch input, to enable users to create and exploit variation in both online and offline settings. Several Alexa Prize teams experimented with the API and leveraged it in their production socialbots. This module was also used by some teams for offline experiments.

3.2.4 ASR Robustness of Neural Response Generation Models

Speech-based voice assistants like Alexa also have imperfect speech recognition, making it important for underlying NLU and NLG models to be built to be robust to speech recognition errors. Given the focus on textual data in the community, we created an augmentation of Topical-Chat, called Topical-Chat ASR, with simulated and actual ASR errors introduced to the dialogues. We empirically studied and found that our GPT-based generative dialogue models were very sensitive to speech recognition errors introduced to the dialogue history during inference time. We then trained our models with simulated ASR errors to be more robust to such errors. We also open-sourced Topical-Chat ASR to enable the community to benchmark their models for speech robustness.

3.2.5 Policy-Driven Neural Response Generation

For response generation in open-domain conversations, it is common to use sequence to sequence models where the model takes in as input dialogue history along with some structured or unstructured knowledge. The use of knowledge helps to generate informative and engaging responses for social dialogue. However, there is still a lack of controllability when generating responses, which can result in abrupt transitions as well as other response defects. To ensure a better user experience, many teams first acknowledge what the user said, and ensure that responses end with a question, in order to better continue the conversation.

To address lack of controllability in neural response generation models, we proposed a policy-driven neural response generation or PD-NRG approach that consists of a mechanism for open domain conversations, i.e., a dialogue policy, which can enable such higher-level control of generated responses (Hedayatnia et al. 2020). The dialogue policy provides a sequential organization plan or action plan. The action plan specifies the order and relationship of sentences within a turn targeting engaging responses to users throughout the interaction.

Each turn in the dialogue contains an action plan that consists of one frame for each sentence. The frames, formed of attributes and values, may include:

- Dialogue acts at a sentence-level to help control the style of the generated response.
- Topics at a turn-level to generate topically coherent responses.
- Knowledge at a turn or sentence-level to generate interesting and informative responses. The knowledge is represented as a sentence drawn from an unstructured knowledge corpus.

- A use-knowledge flag that signals whether or not to use the knowledge attribute at the turn or sentence-level

As illustrated in Figure 5, the proposed PD-NRG approach has two parts: a dialogue policy that determines the action plan based on the dialogue context, and a response generation model that takes the action plan and the dialogue context as input to generate a response. The dialogue policy has components that predict the individual elements of the action plan: knowledge selection and dialogue act planning. Knowledge selection determines the knowledge to be integrated in the response by finding sentences from a knowledge document corpus that are relevant to the dialogue context. Dialogue act (DA) planning determines the style of the response in the form of dialogue acts to be realized. We design a simple hand-crafted dialogue policy that consists of our knowledge selection component and a knowledge-dependent DA planning that looks at the output of the knowledge selection model along with previous dialogue acts to predict the dialogue acts for the next response.

For our PD-NRG model we use the GPT model (Radford et al., 2018) and finetune in a TransferTransfo fashion (Wolf et al., 2019) on the Topical-Chat dataset (Gopalakrishnan et al., 2019). The PD-NRG model is given the action plan (AP) and the dialogue context as input to perform sentence-level prediction.

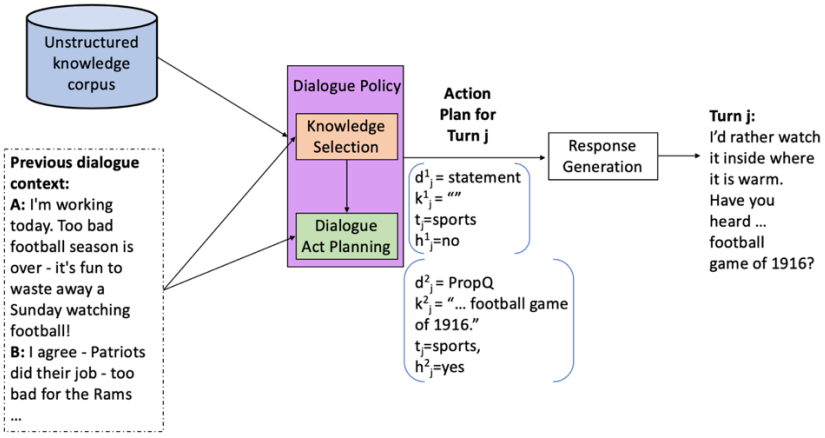


Figure 5. PD-NRG Architecture

Table 1 presents our automatic metrics for variants of our PD-NRG model. As can be seen from the table, adding dialogue acts increases diversity for all the proposed models. The F1-scores of the PD-NRG w/ DA model are lower than the Baseline-Turn model because the PD-NRG model generates shorter sentences resulting in lower recall. The PD-NRG w/ DA model with the addition of previous dialogue acts as input achieves the lowest perplexity for both frequent and rare test sets.

To evaluate controllability of our proposed approach, we manually annotated whether the model's responses realize the dialogue acts and their respective knowledge sentence in their input. As seen in Table 2, the PD-NRG w/ DA + knowledge flag model has the highest accuracy in realizing the input action plan, achieving 80.6% accuracy on the dialogue acts of the generated responses, and 52.1% accuracy in correctly integrating the provided knowledge sentences.

For a more realistic comparison of our dialogue policy model to our baseline, we ran human evaluation. We provided a set of crowd workers outputs from two models along with the dialogue context, and asked them "Which final response is more appropriate for the given conversation?". Crowd-workers were provided with three options: first response, second response and not sure (limited to those cases when the two responses are equally good/bad). Table 3 presents results from the manual evaluations, and shows that

the Knowledge-dependent DA planning (KD-DA-P) responses were chosen over the Baseline-Turn (B-Turn) model by a large margin.

Models	Past DA	PPL	BLEU-1	ROUGE-L	Avg # words	Avg # sentences
Human		- / -	- / -	- / -	24.3 / 25.0	2.10 / 2.15
Baseline-Turn (Wolf et al., 2019)		12.92 / 13.53	0.024 / 0.028	0.134 / 0.130	20.7 / 21.7	1.87 / 1.93
Baseline-Sent		13.85 / 14.36	0.016 / 0.021	0.107 / 0.103	13.7 / 13.9	2.09 / 2.15
PD-NRG w/ DA		12.72 / 13.01	0.024 / 0.027	0.121 / 0.118	18.5 / 19.3	2.05 / 2.10
PD-NRG w/ DA	✓	12.39 / 12.80	0.021 / 0.021	0.115 / 0.111	16.0 / 15.8	1.77 / 1.77
+knowledge flag		12.66 / 12.99	0.025 / 0.027	0.122 / 0.118	17.3 / 18.1	2.03 / 2.08
+knowledge flag	✓	**12.25** / **12.62**	0.019 / 0.020	0.113 / 0.108	15.2 / 15.3	1.68 / 1.76
+knowledge flag +topic		12.76 / 13.07	0.023 / 0.026	0.123 / 0.117	16.8 / 18.2	2.10 / 2.14
+knowledge flag +topic	✓	12.28 / 12.65	0.019 / 0.020	0.115 / 0.109	16.3 / 16.7	1.82 / 1.85
		F1	Precision	Recall	Corpus Diversity *n*=1	Corpus Diversity *n*=2
Human		- / -	- / -	- / -	0.037 / 0.050	0.266 / 0.326
Baseline-Turn (Wolf et al., 2019)		**0.249** / **0.253**	0.275 / 0.272	0.229 / 0.236	0.018 / 0.027	0.118 / 0.165
Baseline-Sent		0.220 / 0.220	0.258 / 0.252	0.191 / 0.195	0.018 / 0.026	0.115 / 0.156
PD-NRG w/ DA		0.241 / 0.240	0.281 / 0.279	0.210 / 0.212	0.018 / 0.027	0.123 / 0.165
PD-NRG w/ DA	✓	0.230 / 0.227	0.291 / 0.287	0.185 / 0.185	0.021 / **0.032**	0.133 / 0.180
+knowledge flag		0.240 / 0.242	0.280 / 0.280	0.209 / 0.213	0.018 / 0.027	0.122 / 0.164
+knowledge flag	✓	0.222 / 0.223	0.287 / 0.281	0.180 / 0.180	**0.032** / **0.022**	**0.137** / **0.181**
+knowledge flag + topic		0.244 / 0.245	0.276 / 0.274	0.210 / 0.213	0.018 / 0.027	0.118 / 0.159
+knowledge flag + topic	✓	0.224 / 0.221	0.272 / 0.271	0.187 / 0.186	0.020 / 0.029	0.136 / 0.177

Table 1. Automated Metrics with Ground Truth Action Plan on Test Freq/Rare

Models	Past DA	% DA	%K
Baseline-Turn (Wolf et al., 2019)		26.7	-
Baseline-Sent		59.4	30.8
PD-NRG w/ DA		69.1	47
PD-NRG w/ DA	✓	69.7	39.3
+knowledge flag		**80.6**	**52.1**
+knowledge flag	✓	68.1	47.8
+knowledge flag +topic		77.8	47.4
+knowledge flag +topic	✓	69.0	45.3

Table 2. % of Dialogue Acts (DA) and Knowledge (K) Realized for PD-NRG Models

Dialogue policy	%W	%T	%L	IAA
KD-DA-P vs. B-Turn*	40.8	30.3	28.9	0.43
KI-DA-P(Seq2Seq) vs. B-Turn*	25.1	35.7	39.2	0.47
KD-DA-P vs. KI-DA-P(PropQ)**	54.2	5.5	40.2	0.46
KD-DA-P vs. KI-DA-P(AllQ)**	54.1	7.4	38.3	0.48
KD-DA-P vs. Human response**	16.7	35.3	48.0	0.53

Table 3. % of Wins (W), Ties (T) and Losses (L) for the baseline models vs. PD-NRG model on appropriateness. Krippendorff's alpha is computed for Inter-annotator Agreement (IAA).

4. Socialbot Quality

Over the course of the competition and since the end of the prior year's competition, socialbots demonstrated significant improvement in customer experience. In this section we provide various metrics to evaluate the socialbot quality in the Third Socialbot Grand Challenge and compare with that observed in the 2018 competition.

4.1 Ratings

After each conversation, Alexa users were asked to rate their interaction on a scale of 1-5. Socialbot ratings have improved when compared to last year - overall ratings were 8.8% higher for the full Semifinals period (3.47 vs. 3.19 last year). We also examined several cohorts among this, looking at the longest duration interactions (up 6.8%) and repeat customer ratings (up 8.4%), which reinforced the overall point. The top-rated socialbot improved by 8.5% across the full Semifinals (3.70 vs. 3.41 last year), and the cohort of Finalists ended the Semifinals period with a 3.61 average rating over a 7 day period, as seen in Figure 6.

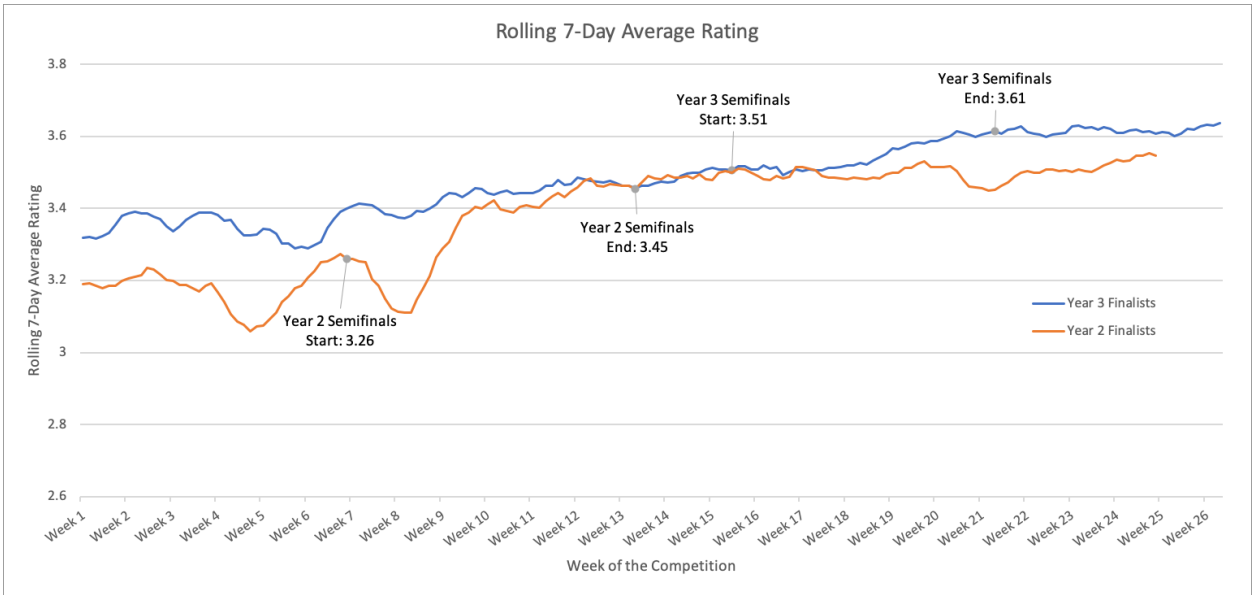


Figure 6. Finalist ratings over time, Year 2 vs Year 3

To better calibrate measurements of year-over-year performance, we surfaced a small portion of traffic to the winning socialbot from the 2018 competition (with refreshed data sources), to maintain as a reference point, correcting for confounding factors in measuring performance – we refer to this below as the "reference socialbot". Over the full Semifinals, 3 of 5 Finalists had higher ratings than this reference socialbot. After the close of the competition (June 4th), we observed a full 7 days of interactions for the Finalist socialbots, and all 5 of them achieved 7-day average ratings higher than last year's reference socialbot (between 3.56 and 3.73, vs. 3.56 for the reference socialbot).

4.2 Conversation Duration

Our primary success metric for conversation duration is the p90 duration, or the 90th percentile duration of conversations (i.e. 10% of conversations have a duration longer than this number). We consider this to be

a strong proxy for the maximum duration of an interesting and engaging conversation between the socialbot and a dedicated interactor. We also track the median (p50) conversation duration, as in Figure 7.

During the Third Alexa Prize Socialbot Grand Challenge, we saw the p90 conversation duration track significantly higher than in the prior year's competition, as seen in Figure 8. At the end of Semifinals, p90 duration was an average of 739 seconds, or over 12 minutes, vs. 598 seconds, or just under 10 minutes, at the end of the prior year Semifinals.

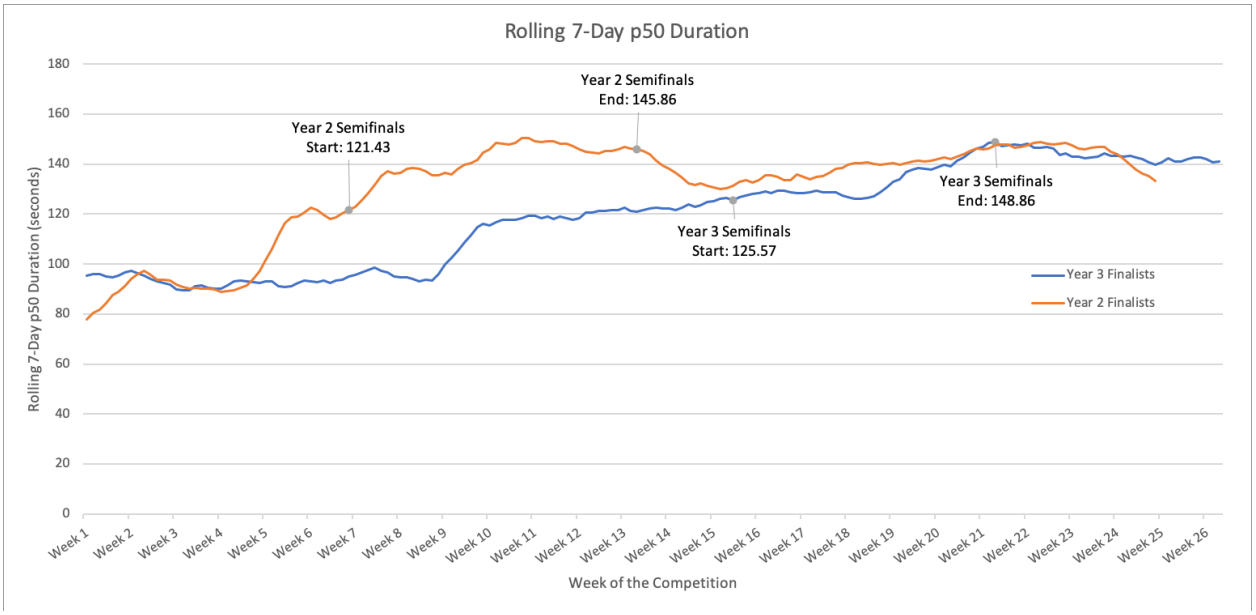


Figure 7. Finalist p50 (median) durations over time, Year 2 vs Year 3

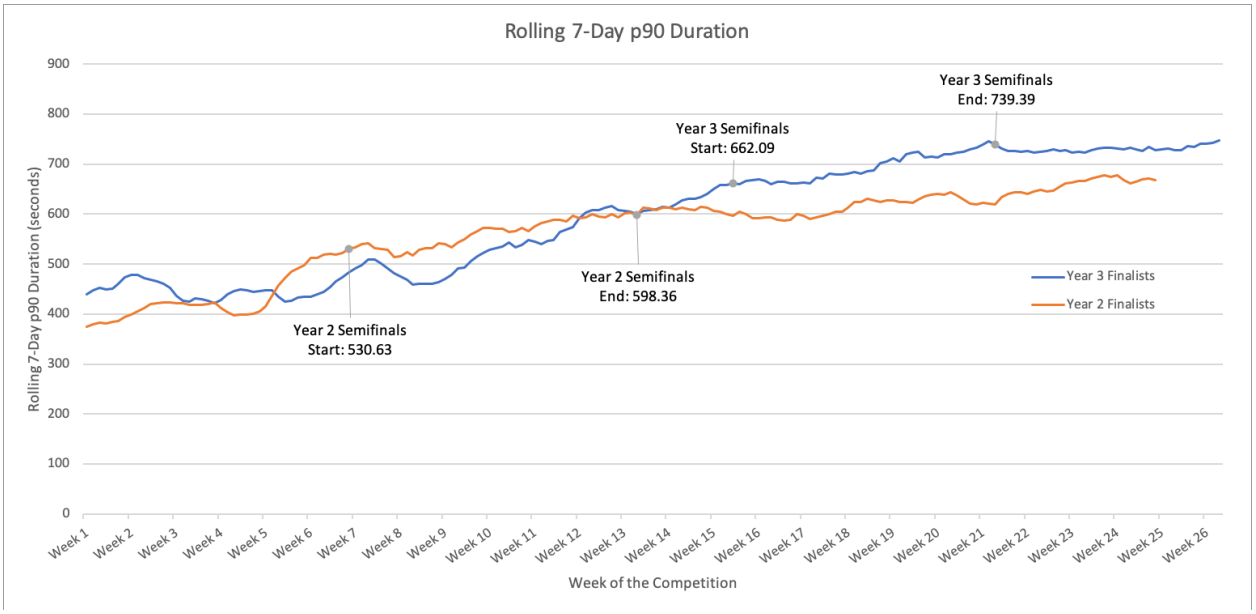


Figure 8. Finalist p90 durations over time, Year 2 vs Year 3

4.3 Response Quality

We obtained annotations for response quality from the Alexa Prize conversations across all the Finalist socialbots. We defined five classes to describe the quality of a response: (1) Poor (response is not comprehensible or bot didn't understand user or response is sensitive) (2) Not Good (the socialbot had some understanding of what user said but contains incorrect or inappropriate information) (3) Passable (response is on topic but is generic or contains too little or too much of information) (4) Good (response is logical, on topic, contains accurate information but lacks personality or is not presented well and obviously coming from a bot), and (5) Excellent (contains accurate information, is complete and it is hard to identify if the response was coming from a bot). By treating these classes as numeric values on a 5-point scale, we can compute an average turn rating.

In the 2018 competition, the Finalist teams averaged a 2.77 turn rating after the close of the Semifinals. This year, we evaluated the reference socialbot interactions, which achieved an annotated turn rating of 2.88. During the Semifinals, three of five Finalist teams sustained an average turn rating higher than the reference socialbot (2.96, 2.94 and 2.90) while two were somewhat lower (2.78 and 2.74). The best performing of this year's socialbots had 17.1% of turns marked as Poor or Not Good, vs. 22.5% for the reference socialbot.

4.4 Response Error Rate

We define Response Error Rate (RER) as the ratio of irrelevant or inappropriate responses to total responses, as obtained through human annotations. RER is an important metric for evaluating socialbots. Given that socialbots are non-goal-oriented open-domain systems, it is often hard to define a correct response, however it is easier to define and identify an incorrect, inappropriate or incoherent response. From Figure 9, we can observe that the best socialbots in the Third Grand Challenge have significantly improved their RER over the best socialbots in the 2018 competition. The reference socialbot obtained a 10.9% RER overall during the Semifinals period, while the 5 finalist socialbots averaged a 9.5% RER during the Semifinals period, a 13% relative improvement. The highest rated socialbot during this same period averaged an 8.6% RER, a 23% relative improvement.

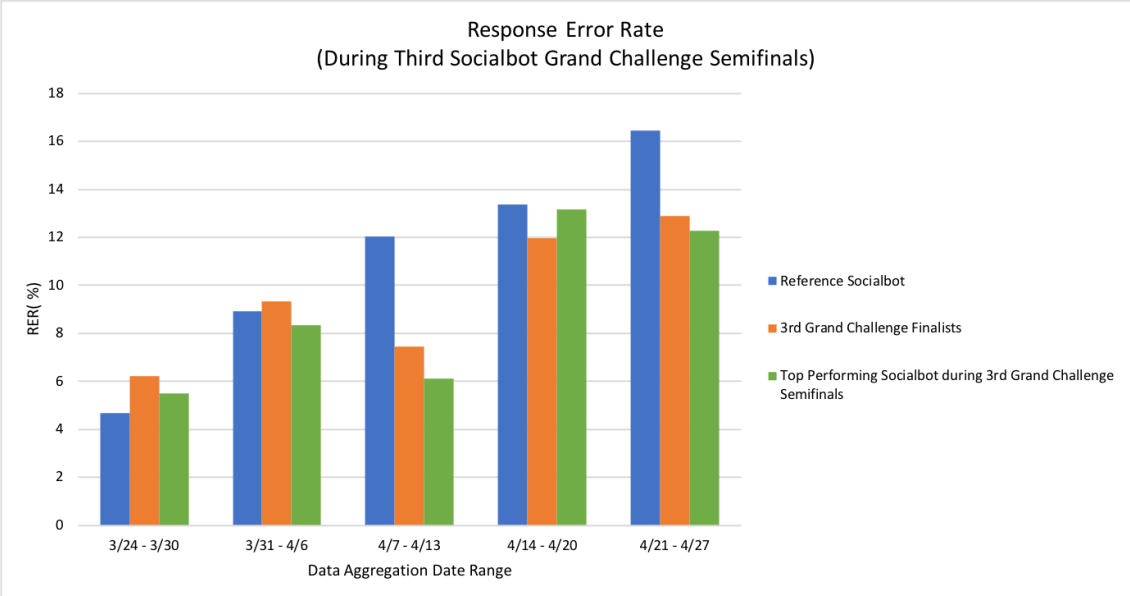


Figure 9. Response Error Rate during 3rd Socialbot Grand Challenge Semifinals

4.5 Coherent and Engaging Responses

Similar to response quality, we performed turn-level annotations for coherent and engaging responses as depicted in Figure 10. Coherent responses have been measured using two criteria: 1) response is comprehensible and 2) response is on topic. Engaging responses are measured using the following criteria: 1) response is interesting, and 2) the user would want to continue the conversation.

During the Semifinals of the Third Socialbot Grand Challenge, the best performing socialbot was annotated as coherent 70% of the time and engaging 67% of the time, while across all the Finalist socialbots, 65% of turns were annotated as coherent and 62% were annotated as engaging. By comparison, the reference socialbot was annotated as coherent 64% of the time, and engaging 62% of the time.

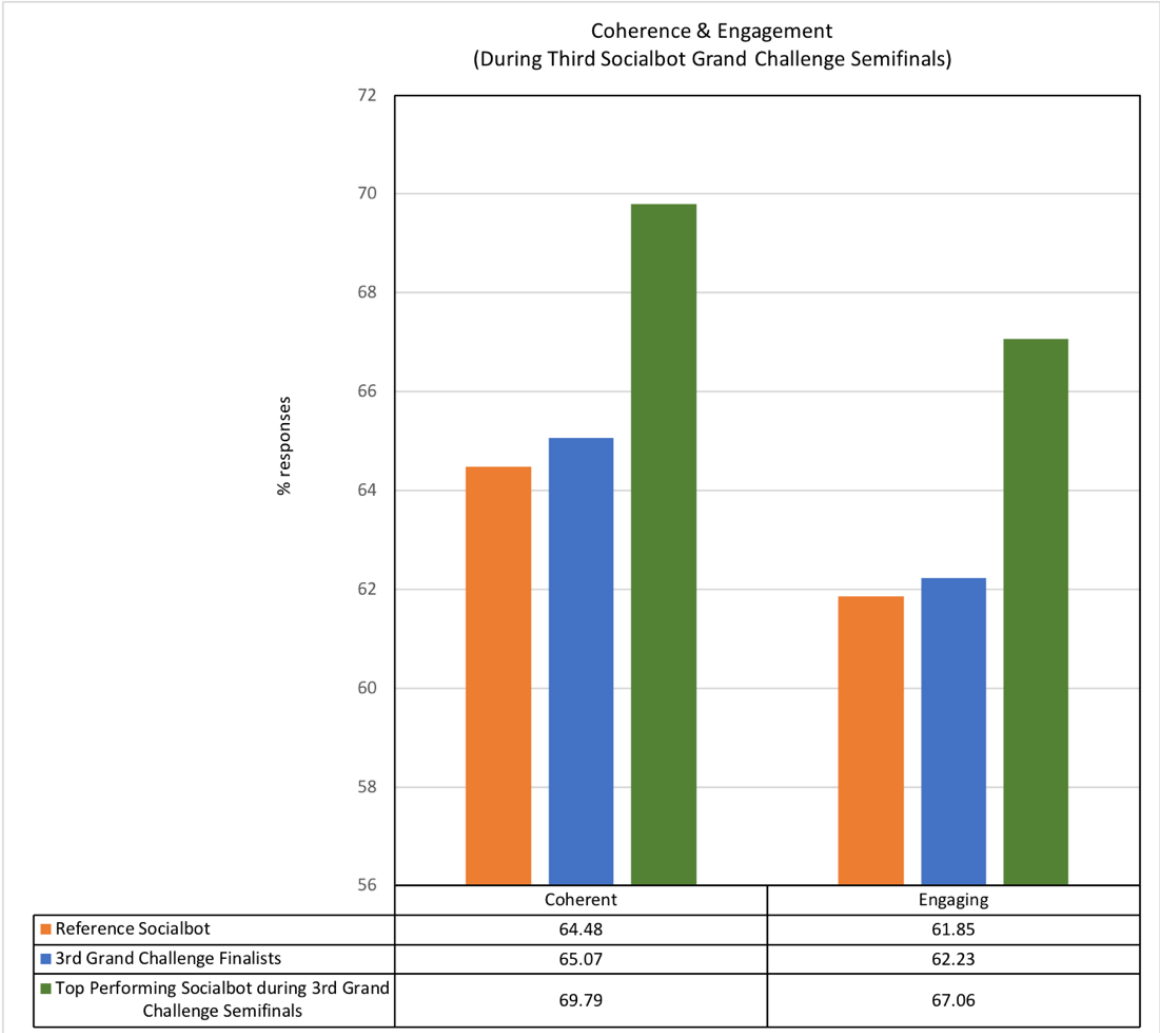
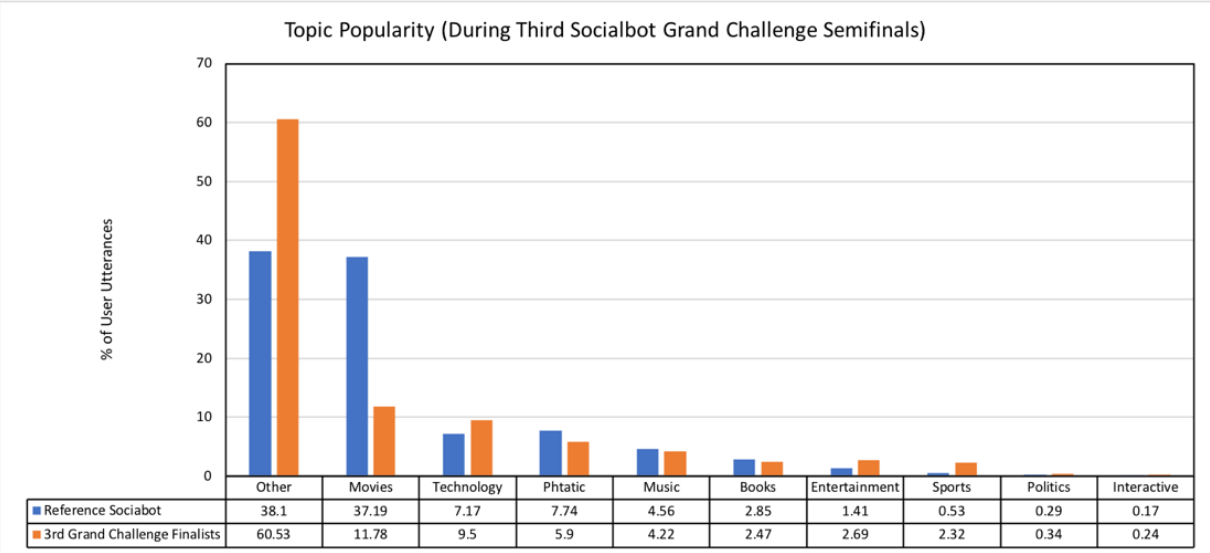


Figure 10. Annotations for Coherent and Engaging Conversations

4.6 Topical Analysis

To improve the quality of socialbots, we shared a variety of data with the teams including topical analysis of conversations. Overall, "Movies", "Music" and "Technology" are the most popular topics apart from "Other". "Other" corresponds to any topic beyond the provided list of topics, as well as utterances containing multiple topics. It is important to note the semifinals of the Third Grand Challenge occurred during March and April 2020, and conversations related to changes in lifestyle, work and social circumstances due to the COVID-19 pandemic may have contributed significantly to the Other classification.

Based on topical annotation during semifinals, coverage of Sports, Technology and Other have increased year-over-year, with significantly fewer interaction in the Movies topic (17% in last year's Finalists, vs. 12% this year). These numbers were confirmed by annotation of interactions with the reference socialbot, of which 37% were related to Movies, compared to a range of 5% to 20% in this year's Finalists, as seen in Figure 11.



	Other	Movies	Technology	Phtatic	Music	Books	Entertainment	Sports	Politics	Interactive
Reference Sociabot	38.1	37.19	7.17	7.74	4.56	2.85	1.41	0.53	0.29	0.17
3rd Grand Challenge Finalists	60.53	11.78	9.5	5.9	4.22	2.47	2.69	2.32	0.34	0.24

Figure 11. Distribution of Topics during Semifinals

5. Discussions and Recommendations

From the results in Section 4, we observe that teams have achieved improvements in response error rate and driven longer conversations through Semifinals. Despite challenges in predicting ratings in individual conversations, longer conversation durations at the P90 level, breadth of topical coverage and percentage of engaging responses in socialbots have all been associated with higher average ratings for a socialbot across multiple years of the Alexa Prize Socialbot Grand Challenge.

This year, teams consistently found advantage in refactoring their existing template-based dialog mechanisms into smaller constituent components, and federating response generation across multiple responses, allowing for more flexible dialog policies that could recombine these fragments into coherent dialogs. Alquist achieved this using an adjacency pair approach, with hierarchical dialog structure information and a robust templated response generation mechanism. Athena relied on an approach with a

federated set of response generators, each acting under a set of discourse constraints, that could then be interleaved together in the overall dialog.

A continuing theme in the Socialbot Grand Challenge is that emotionally captivating, engaging responses and a sense of persona are important to success. Tartan, Bernard and Emora all focused on well-defined persona elements to drive consistent, interesting conversations. Chirpy Cardinal also integrated fine-grained emotional classification, alongside opinion generation and requesting.

Significant advances in this year's competition came from integrating neural response generation techniques together with templated dialog responses for greater diversity and recall, while maintaining sufficient control of response structure to simultaneously optimize for duration and overall dialog rating. However, teams have had mixed results integrating these techniques into existing socialbots that already have high precision templated dialog mechanisms, particularly if they already invested in those dialogs and attempted to integrate neural response generation later in their development process. Athena and DREAM reported no benefit from the use of neural response generation techniques, while Gunrock and Chirpy Cardinal deeply integrated neural response models into their templated responders throughout. Chirpy Cardinal relied on their neural response generation mechanisms to rapidly scale and adjust tone and emotion of their responses.

6. Conclusion and Future Work

The problem of engaging in coherent, engaging conversations is one of the most challenging problems in the artificial intelligence field. Several tasks associated with Conversational AI such as language understanding, knowledge representation, commonsense reasoning and dialog evaluation are believed to be "AI Complete", or truly human-intelligence equivalent problems. To address these challenges, Amazon launched the Alexa Prize Socialbot Grand Challenge, wherein some of the best research groups across the world work towards a common goal of advancing the state of the art in Conversational AI.

The Third Socialbot Grand Challenge ran from September 2019 through May 2020, and the participating university teams have successfully built socialbots that can converse with Alexa users coherently and engagingly on a wide variety of topics. Teams built on the success of the first two years of the competition with sophisticated statistical dialog management, improved personalization, complex utterance handling, increased use of large-scale Transformer-based models for a wide variety of tasks, and incremental steps toward runtime generation of natural language in dialog.

There have been significant further scientific advancements brought by the Third Grand Challenge and have seen improvements in ratings, conversation duration and human-annotated measures of socialbot quality, when compared to the 2018 Grand Challenge. It is still Day One for Conversational AI, as the capabilities of natural language generation, large-scale transformer-based models, and integration of traditional AI approaches into deep learned systems continues. We expect to see even more advancements in coming years, with deeper follow-ups on topics and much better fallback modes of interaction, utilizing natural language generation at runtime with both structured and unstructured knowledge sources. As these higher recall approaches are accelerated sufficiently for runtime execution and we develop better mechanisms to subsume the high precision approaches of templated and partially templated responses, we look forward to today's Grand Challenge objective becoming tomorrow's reality.

References

Williams, J., Raux, A., & Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, *7*(3), 4-33.

Burtsev (2018). *ConvAI2*. Retrieved from http://convai.io/

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M. & Liu, S. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Botvinovsky, E., Bushkov, N., Gureenkova, O., Kamenev, A., Konovalov, V. & Kuratov, Y. (2018). DeepPavlov: An Open Source Library for Conversational AI.

Bocklisch, T., Faulker, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A. & King, E. (2018). Conversational AI: The science behind the Alexa Prize. *arXiv preprint arXiv:1801.03604*.

Khatri, C. Hedayatnia, B., Venkatesh A., Nunn J., Pan Y., Liu Q., Song H., Gottardi A., Kwatra S., Pancholi S., Cheng M., Chen Q., Stubell S., Gopalakrishnan K., Bland K., Gabriel R., Mandal A., Hakkani-Tür D., Hwang G., Michel N., King E., & Prasad R. (2018). Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize. 2nd Proceedings of the Alexa Prize.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J. & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. *In Proc. NAACL , pages 196–205, 2015*

Vinyals, O., & Le, Q. (2015). A neural conversational model. *In Proc. ICML, 2015.*

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*

Smith E., Williamson M., Shuster K., Weston J., and Boureau Y.. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

Kumar, A., Gupta, A., Chan, J., Tucker, S., Hoffmeister, B., Dreyer, M., ... & Monson, C. (2017). Just ASK: building an architecture for extensible self-service spoken language understanding. *arXiv preprint arXiv:1711.00549*.

Honnibal, M. (2016). SpaCy (Version 2.0). Retrieved from: https://spacy.io/

Gopalakrishnan K., Hedayatnia B., Chen Q., Gottardi A., Kwatra S., Venkatesh A., Gabriel R., and Hakkani-Tür D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. Proc. Interspeech 2019,1891–1895.

Hedayatnia, B., Gopalakrishnan K., Kim S., Liu Y., Eric M., Hakkani-Tür D. (2020). Policy-Driven Neural Response Generation for Knowledge-Grounded Dialogue Systems. *arXiv preprint arXiv:2005.12529*

Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf*.

Devlin J., Chang M., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. arXiv preprint arXiv:1605.06069

Luan, Y., Ji, Y., & Ostendorf, M. (2016). LSTM based Conversation Models. *arXiv preprint arXiv:1603.09457*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Guo, F., Metallinou, A., Khatri, C., Raju, A., Venkatesh, A., & Ram, A. (2018). Topic-based evaluation for conversational bots. *In NIPS, 2017.*

Khatri, C., Goel, R., Hedayatnia, B., Venkatesh, A., Gabriel, R., & Mandal, A. (2018b). Detecting Offensive Content in Open-domain Conversations using Two Stage Semi-supervision.

Yi, S., Khatri, C., Goel, R., Hedayatnia, B., Venkatesh, A., Gabriel, R., & Mandal, A. (2018). Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A. & Goel, R. (2018). On Evaluating and Comparing Conversational Agents. *In NIPS, 2017.*

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017-1024).

Bosselut A., Rashkin H., Sap M., Malaviya C., Celikyilmaz A., & Choi Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. a*rXiv preprint arXiv:1906.05317*

Sap M., LeBras R., Allaway E., Bhagavatula C., Lourie N., Rashkin H., Roof B., Smith N., and Choi Y. (2019). Atomic: An atlas of machine commonsense for if-then reasoning. In AAAI.

Radford A., Narasimhan K., Salimans R., and Sutskever I. (2018). Improving language un-derstanding with unsupervised learning. Technical report, OpenAI.

Radford A., Wu J., Child R., Luan D., Amodei D., and Sutskever I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Curry, A. C., Papaioannou, I., Suglia, A., Agarwal, S., Shalyminov, I., Xu, X., Dušek, O., Eshghi, A.Yu, Y., & Lemon, O. (2018). Alana v2: Entertaining and Informative Open-domain Social Dialogue using Ontologies and Entity Linking. *Alexa Prize Proceedings, 2018*.

Pichl, J., Marek, P., Konrád, J., Matulík, M., & Šedivý, J. (2018). Alquist 2.0: Alexa Prize Socialbot Based on Sub-Dialogue Model. *Alexa Prize Proceedings, 2018*.

Fulda, N., Etchart, T., Myers, W., Ricks, D., Brown, Z., Szendre, J., Murdoch, B., Carr, A & Wingate, D. (2018). BYU-EVE: Mixed Initiative Dialog via Structured Knowledge Graph Traversal and Conversational Scaffolding. *Alexa Prize Proceedings, 2018*.

Jonell, P., Bystedt, M., Dŏgan, F. I., Fallgren, Per., Ivarsson, J., Slukova, M., Wennberg, U., Lopes, José., Boye, Johan & Skantze, G. (2018). Fantom: A Crowdsourced Social Chatbot using an Evolving Dialog Graph. *Alexa Prize Proceedings, 2018*.

Chen, C., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., Iyer, S., Sreenivasulu, G., Cheng, R., Bhandare, A & Yu, Z. (2018). Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *Alexa Prize Proceedings, 2018*.

Ahmadvand, A., Choi, I., Sahijwani, H., Schmidt, J., Sun, M., Volokhin, S., Wang, Z & Agichtein, E. (2018). Emory IrisBot: An Open-Domain Conversational Bot for Personalized Information Access. *Alexa Prize Proceedings, 2018*.

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., Santer, N & Walker, M. (2018). SlugBot: Developing a Computational Model and Framework of a Novel Dialogue Genre. *Alexa Prize Proceedings, 2018*.

Larionov, G., Kaden, Z., Dureddy, H. V., Kalejaiye, G. B. T., Kale, M., Potharaju, S. P., Shah, A. P., & Rudnicky, A. I. (2018). Tartan: A retrieval-based socialbot powered by a dynamic finite-state machine architecture. *Alexa Prize Proceedings, 2018*.

Pichl, J., Marek, P., Konrad, J., Lorenc, P., Ta, V. D., Sedivy, J. (2020). Alquist 3.0: Alexa Prize Bot Using Using Conversational Knowledge Graph. *Alexa Prize Proceedings, 2020.*

Harrison, V., Juraska, J., Cui, W., Reed, L., Bowden, K. K., Wu, J., Schwarzmann, B., Ebrahimi, A., Rajasekaran, R., Varghese, N., Wechsler-Azen, M., Whittaker, S., Flanigan, J., Walker, M. (2020). Athena: Constructing Dialogues Dynamically with Discourse Constraints. *Alexa Prize Proceedings, 2020.*

Hong, C. H., Liang, Y., Roy, S. S., Jain, A., Agarwal, V., Draves, R., Zhou, Z., Chen, W., Liu, Y., Miracky, M., Ge, L, Banovic, N., Jurgens, D. (2020). Audrey: A Personalized Open Domain Conversational Bot. *Alexa Prize Proceedings, 2020.*

Majumder, B. P., Li, S., Ni, J., Mao, H. H., Sun, S., McAuley, J. (2020). Bernard: A Stateful Neural Open-domain Socialbot. *Alexa Prize Proceedings, 2020.*

Paranjape, A., See, A., Kenealy, K., Li, H., Hardy, A., Qi, P., Sadagopan, K. R., Phu, N. M., Soylu, D., Manning, C. D. (2020). Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations. *Alexa Prize Proceedings, 2020.*

Kuratov, Y., Yusupov, I., Baymurzina, D., Kuznetsov, D., Cherniavskii, D., Dmitrievskiy, A., Ermakova, E., Ignatov, F., Karpov, D., Kornev, D., Le, T. A., Pugin, P., Burtsev, M. (2020). DREAM technical report for the Alexa Prize 2019. *Alexa Prize Proceedings, 2020.*

Finch, S. E., Finch, J. D., Ahmadvand A., Choi, I., Dong, X., Qi, R., Sahijwani, H., Volokhin, S., Wang, Z., Wang, Z., Choi, J. D. (2020). Emora: An Inquisitive Social Chatbot Who Cares For You. *Alexa Prize Proceedings, 2020.*

Liang, K., Chau, A., Li, Y., Lu, X., Yu, D., Zhou, M., Jain, I., Davidson, S., Arnold, J., Nguyen, M., Yu, Z. (2020). Gunrock 2.0: A User Adaptive Social Conversational System. *Alexa Prize Proceedings, 2020.*

Chen, F., Chi, T.-C., Lyu, S., Gong, J., Parekh, T., Joshi, R., Kaushik, A., Rudnicky, A. (2020). Tartan: A Two-Tiered Dialog Framework For Multi-Domain Social Chitchat. *Alexa Prize Proceedings, 2020.*

Schallock, W., Agress, D., Du, Y., Dua, D., Hu, L., Matsubara, Y., Singh, S. (2020). ZOTBOT: Using Reading Comprehension and Commonsense Reasoning in Conversational Agents. *Alexa Prize Proceedings, 2020.*