Athena 2.0: Contextualized Dialogue Management for an Alexa Prize SocialBot

Juraj Juraska, Kevin K. Bowden, Lena Reed, Vrindavan Harrison,
Wen Cui, Omkar Patil, Rishi Rajasekaran, Angela Ramirez, Cecilia Li,
Eduardo Zamora, Phillip Lee, Jeshwanth Bheemanpally, Rohan Pandey,
Adwait Ratnaparkhi and **Marilyn Walker**
University of California Santa Cruz, Santa Cruz, CA
`jjuraska, kkbowden, lireed, vharriso, wcui7, opatil,`
`rrajasek, aramir62, yli331, ezamora9, plee35, jbheeman,`
`rspandey, adratnap, mawalker@ucsc.edu`

Abstract

Athena 2.0 is an Alexa Prize SocialBot that has been a finalist in the last two Alexa Prize Grand Challenges. One reason for Athena's success is its novel dialogue management strategy, which allows it to dynamically construct dialogues and responses from component modules, leading to novel conversations with every interaction. Here we describe Athena's system design and performance in the Alexa Prize during the 20/21 competition. A live demo of Athena as well as video recordings will provoke discussion on the state of the art in conversational AI.

1 Introduction

There has been tremendous progress over the last 10 years on conversational AI, and a number of practical systems have been deployed. The Alexa Prize competition seeks to stimulate research and development on conversational AI for open-domain topic-oriented dialogue (Fang et al., 2018; Liang et al., 2020; Finch et al., 2020; Harrison et al., 2020; Pichl et al., 2020; Curry et al., 2018). However, the longstanding tension between hand-scripting the dialogue interaction, and producing systems that scale to new domains and types of interaction still remains (Eric and Manning, 2017; Cervone et al., 2019) Neural end-to-end spoken dialogue systems are not yet at a point where they perform well in interactions with real users (Paranjape et al., 2020; Gopalakrishnan et al., 2020; Dinan et al., 2019).

Athena's dialogue management architecture aims to be scalable and dynamic, by supporting many different interactions for every topic, and by constructing system utterances by concatenating multiple dialogue acts that achieve different purposes (Stent, 2000). A key aspect of Athena is the existence of multiple Response Generators (RGs) for each topic, which can be flexibly interleaved during a particular interaction, as illustrated in Figure 1.[1] This approach contrasts with the commonly used approach of handcrafting conversation flow-graphs for each topic, a static directed graph where the nodes are the system utterances, and outgoing edges are represent possible user replies. This approach has not changed for over 20 years (Seneff et al., 1998; Glass and Weinstein, 2001; Buntschuh et al., 1998), and its strengths and limitations are well known. Flow-graphs are hand constructed and thus do not scale well. However, each system response can assume a fixed prior context, which allows it to support fluent and coherent dialogues with sufficient handcrafting.

In contrast, the ability of Athena's Dialogue Manager (DM) to interleave RGs allows Athena to dynamically construct conversations that never follow the same path. However, this more flexible approach requires RGs to pay the overhead cost of continuously adapting to the current context, as described in Section 3. By eschewing a graph-based representation of dialogue state, Athena's DM is flexible enough to use RG responses in contexts that were not planned out prior to the conversation starting, and that do not need to follow rigid guidelines. We believe this modular dialogue management approach promises to scale to deeper and richer conversations, while at the same time allow new conversational topics to be easily added to and integrated into the system.

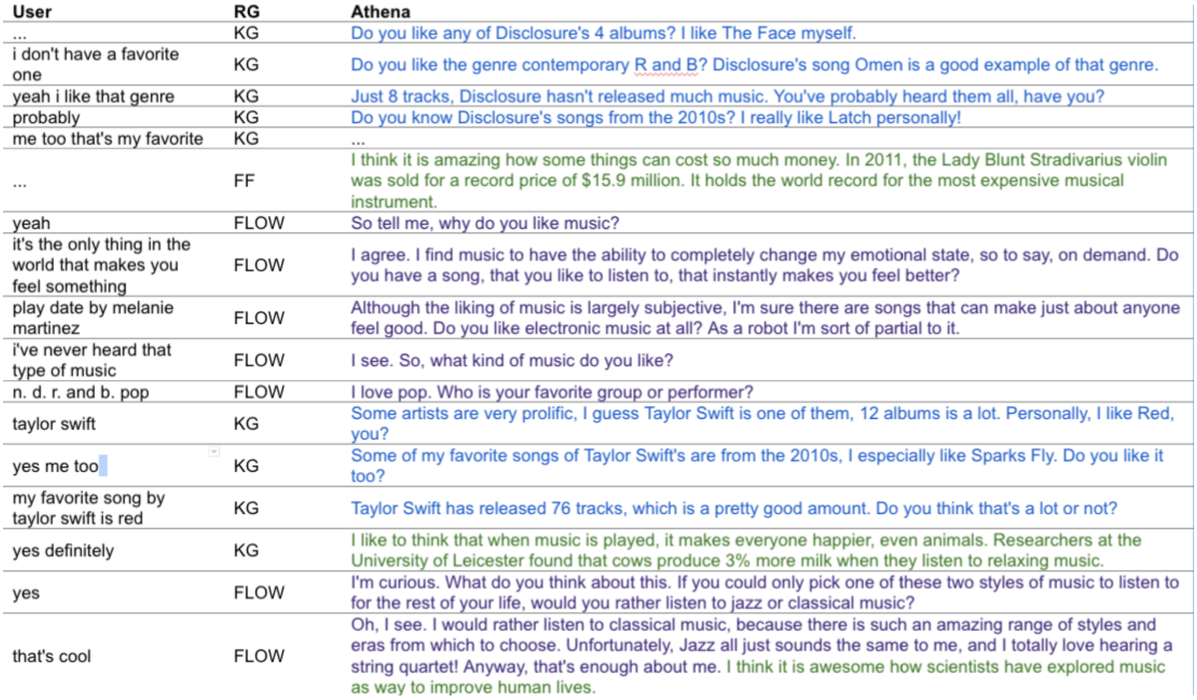
2 Athena Architecture and Overview

Figure 2 details Athena's architecture. Athena is built using the Cobot framework provided by Amazon (Khatri et al., 2018). It runs as an on-demand application that is initiated by an "Alexa, let's chat" user request to any Alexa-enabled device, such as

---

[1] In accordance with the Alexa Prize rules on user privacy, these conversations are between team members and Athena.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 124–133
November 7–11, 2021. ©2021 Association for Computational Linguistics



User	RG	Athena
...	KG	Do you like any of Disclosure's 4 albums? I like The Face myself.
i don't have a favorite one	KG	Do you like the genre contemporary R and B? Disclosure's song Omen is a good example of that genre.
yeah i like that genre	KG	Just 8 tracks, Disclosure hasn't released much music. You've probably heard them all, have you?
probably	KG	Do you know Disclosure's songs from the 2010s? I really like Latch personally!
me too that's my favorite	KG	...
...	FF	I think it is amazing how some things can cost so much money. In 2011, the Lady Blunt Stradivarius violin was sold for a record price of $15.9 million. It holds the world record for the most expensive musical instrument.
yeah	FLOW	So tell me, why do you like music?
it's the only thing in the world that makes you feel something	FLOW	I agree. I find music to have the ability to completely change my emotional state, so to say, on demand. Do you have a song, that you like to listen to, that instantly makes you feel better?
play date by melanie martinez	FLOW	Although the liking of music is largely subjective, I'm sure there are songs that can make just about anyone feel good. Do you like electronic music at all? As a robot I'm sort of partial to it.
i've never heard that type of music	FLOW	I see. So, what kind of music do you like?
n. d. r. and b. pop	FLOW	I love pop. Who is your favorite group or performer?
taylor swift	KG	Some artists are very prolific, I guess Taylor Swift is one of them, 12 albums is a lot. Personally, I like Red, you?
yes me too	KG	Some of my favorite songs of Taylor Swift's are from the 2010s, I especially like Sparks Fly. Do you like it too?
my favorite song by taylor swift is red	KG	Taylor Swift has released 76 tracks, which is a pretty good amount. Do you think that's a lot or not?
yes definitely	KG	I like to think that when music is played, it makes everyone happier, even animals. Researchers at the University of Leicester found that cows produce 3% more milk when they listen to relaxing music.
yes	FLOW	I'm curious. What do you think about this. If you could only pick one of these two styles of music to listen to for the rest of your life, would you rather listen to jazz or classical music?
that's cool	FLOW	Oh, I see. I would rather listen to classical music, because there is such an amazing range of styles and eras from which to choose. Unfortunately, Jazz all just sounds the same to me, and I totally love hearing a string quartet! Anyway, that's enough about me. I think it is awesome how scientists have explored music as way to improve human lives.

Figure 1: A subdialogue in which Athena discusses music with the user, interweaving the responses of 3 different RGs. Responses from the KG-based RG are highlighted in blue, responses from the entity-based indexing RG (fun facts) in green, and those from the flow-based RG in purple.

an Amazon Echo or the Alexa app installed on a phone. During the Alexa Prize, Athena participates in about 9K conversations a week. The Cobot framework provides support for automatically scaling to large volumes of user traffic.

The inputs to Athena are the ASR hypotheses for a user's turn from Amazon, and a conversation ID that is used to retrieve the conversation history and state information from a back-end database. The ASR hypothesis is fed into a natural language understanding (NLU) pipeline that produces a set of NLU features for the user utterance and conversation context. The NLU consists of Cobot's module for topic classification, and Athena modules for utterance segmentation, dialogue act tagging, named entity recognition and linking, and coreference resolution (Harrison et al., 2020; Patil et al., 2021). The right-hand side of Figure 2 indicates how Athena's RGs use knowledge bases and fun facts databases organized by topic and named entity. Athena uses the Wikidata Knowledge Graph to aid in Named Entity Resolution and for Knowledge-Graph based RGs. These are essential for creating an intelligent and versatile conversational agent (Fang et al., 2018; Chen et al., 2018).

Based on the NLU features and conversation context, the Dialogue Manager (DM) calls specific Response Generators (RGs) to populate a response pool. The DM then applies a trained neural response ranker to select from the response pool generated by the RGs. Finally, Athena's responses are spoken by Amazon's text-to-speech service.

3 Dialogue Management

A Dialogue Manager (DM) for open-domain conversation faces a particularly challenging task due to the universe of possible valid responses at each point of a conversation. While goal-oriented dialogues have a clear task completion objective which the DM can optimize when making decisions (Walker et al., 2001, 1997; Walker, 2000), the DM for open-domain dialogues does not have an obvious way to measure the appropriateness of possible candidate responses.

Athena's DM architecture can be decomposed into a number of sub-components, corresponding to phases of dialogue management, oriented as a pipeline. The DM sub-modules in Figure 3 are described in more detail in Harrison et al. (2020).

The Topic Manager in Figure 3 is responsible for classifying user utterances into topics, and the implementation of the DM's topic hierarchy. The topic hierarchy is a partially ordered list of topics in order of predicted "goodness" learned from past conversations, using a scoring function that combines user ratings and the number of turns per topic

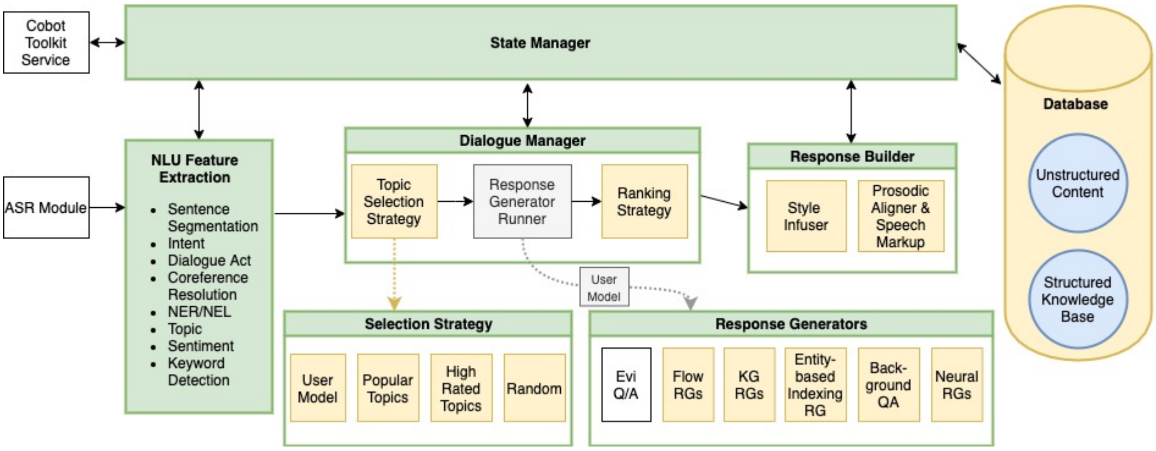


Figure 2: Athena's system architecture.

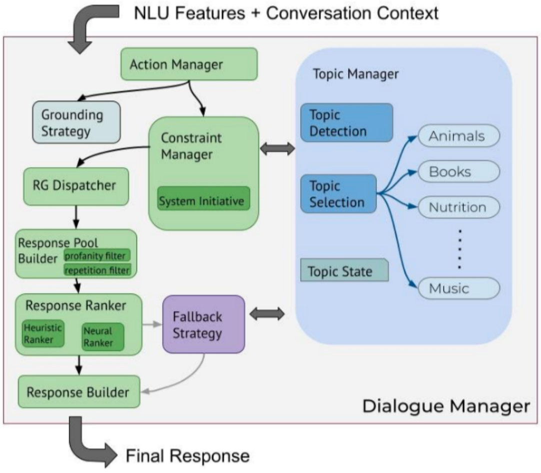


Figure 3: Dialogue manager architecture.

per conversation, as described in Section 5. The topic hierarchy is a parameter for system-initiative topic initiations as well as suggesting topics for users to initiate. This makes it extremely easy to change which topics are promoted at any time, e.g., for collecting more data on a particular topic. It can also be personalized for each user. For example, if when asked about weekend activities, the user describes playing in a baseball league, we can prioritize talking about sports. This information persists across conversations. If the user is also an avid painter, but our system did not get a chance to discuss painting in the previous conversation, we will prioritize it when the user returns.

The interface between the DM and the RGs in Figure 3, is a contract-based approach. The DM passes a set of response conditions to the RGs, which the RGs must meet for their response to be considered. This approach allows Athena to have many RG types (see Section 4).

The Response Ranker is based on a BERT-based ranker fine-tuned on hand-annotated Alexa Prize conversation data (Wolf et al., 2019; Devlin et al., 2018). The current tuning set size is ~10K utterances. Annotation involves ranking candidate responses within a context of five turns. We have repeatedly annotated additional data and retrained our response ranker, which is useful when, for example, new RGs are added to Athena.

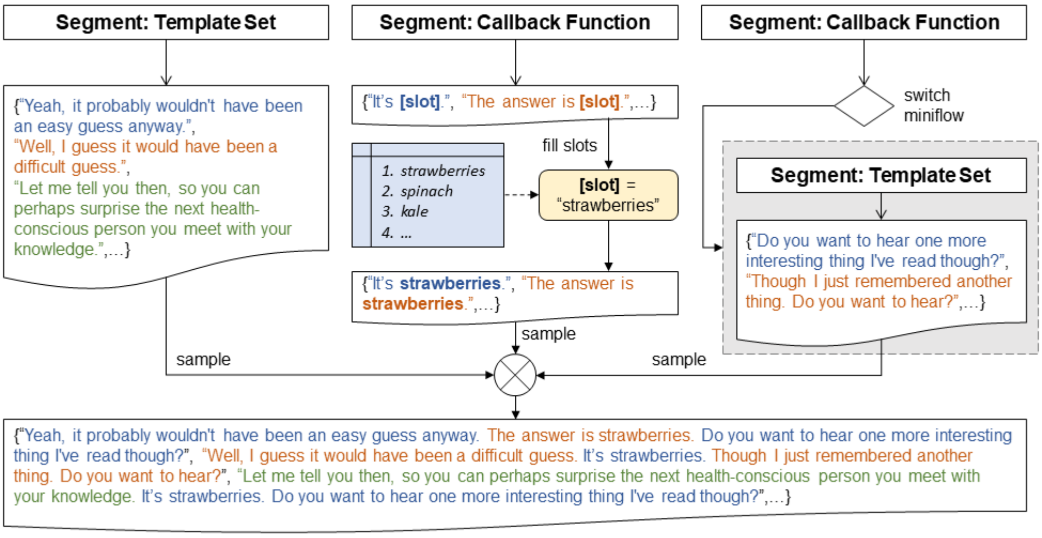
4 Response Generation

Athena uses four types of RGs: Flow-RGs, Knowledge-Graph RGs, Entity-Based Indexing RGs, and Neural NLG RGs.

4.1 Flow-RG

Flow-RG is a framework that we developed with the objective of creating robust and modular flow-based RGs. This is still the most reliable way to provide the DM with a pool of possible responses at each turn of the dialogue, even though such flows have to be handcrafted. Flow-based RGs exhibit context-awareness and fluency superior to other RG types, such as retrieval-based or neural. This RG design naturally has a rather limited support for user initiative, which we make up for with other RGs in Athena, and by ensuring the responses from different RGs get smoothly interwoven across multiple turns, as well as within a single turn.

An RG defined in this framework has three components. First, a *flow graph* consisting of nodes specifying the responses, and edges determining which node of the flow to move on to given the current user utterance and dialogue state. Flow-RG enforces each next turn in the flow graph to be conditioned on the dialogue act(s) of the user utterance, while other features of the utterance – such as its sentiment, or the presence of a named entity or a particular keyword – are deemed secondary and



{"Yeah, it probably wouldn't have been an easy guess anyway.", "Well, I guess it would have been a difficult guess.", "Let me tell you then, so you can perhaps surprise the next health-conscious person you meet with your knowledge.",...}

{"It's **[slot]**.", "The answer is **[slot]**.",...}

{"It's **strawberries**.", "The answer is **strawberries**.",...}

{"Do you want to hear one more interesting thing I've read though?", "Though I just remembered another thing. Do you want to hear?",...}

{"Yeah, it probably wouldn't have been an easy guess anyway. The answer is strawberries. Do you want to hear one more interesting thing I've read though?", "Well, I guess it would have been a difficult guess. It's strawberries. Though I just remembered another thing. Do you want to hear?", "Let me tell you then, so you can perhaps surprise the next health-conscious person you meet with your knowledge. It's strawberries. Do you want to hear one more interesting thing I've read though?",...}

Figure 4: Illustration of response composition in Flow-RG.

are optional in branching conditions.[2] This reduces the chance of Athena's subsequent response ignoring the user's intent, which can be anything from expressing an opinion, to requesting information, to merely acknowledging Athena's response in the previous turn. The second component comprises *response segment templates*, while the third component is a set of *callback functions* that generate more context-dependent response segments.

A flow graph can be broken down into smaller *miniflows* that are independent and can possibly be executed in an arbitrary order. Each RG then typically handles a single topic, with multiple miniflows being responsible for different subtopics. An example of multiple miniflows forming a cohesive dialogue can be seen in Appendix A.

Response Composition. The response in each turn is assembled from one or more *segments* specified in the corresponding node. Each segment is defined either (1) in the form of a set of templates, or (2) as a callback function that returns a set of templates. While both offer an easy way to use paraphrases for increased diversity of the responses, the latter is more robust in that it can use the previous context and more of the NLU information about the user utterance. Figure 4 shows the process of a response being assembled from three segments, two of which are different types of callback function: one fills a template slot with a value from the associated knowledge source, while the other initiates a new miniflow and composes the response text recursively, which ultimately corresponds to the last segment in the example.

When composing a response, each segment's final set of texts is sampled from, and all of them are concatenated. This is repeated until up to five different response candidates are composed. These are eventually all returned to the DM, which picks one of them that is not too similar to any of Athena's previous responses.

Interweaving with Other RGs. Every topic in Athena has a corresponding Flow-RG, and most topics also have one or two other RGs that can interact with its Flow-RG to dynamically construct a topical sub-dialogue. In line with the DM's way of response building, the final response in Flow-RG is split into three parts: an opener, a body, and a hand-off. This response structure is particularly useful for creating seamless transitions whether it is between miniflows, or between two RGs. To this end, Flow-RG sets the response from an ending miniflow as the opener (typically, some form of acknowledgement of the user's response, or a short answer), and the body and hand-off parts are reserved for the intro response provided by a new miniflow. The same mechanism is used for certain transitions from Flow-RG to a different RG, mainly: 1) when the flow's content is exhausted, and transitions thus to a fallback response chosen by the DM that initiates a new topic, and 2) when a leaf node of the miniflow is reached, and the DM decides to switch to a different RG on the same topic. The latter is utilized in the DM's interweaving strategy, wherein a flow-based RG takes turns with an entity-centric or fun fact-based RG in han-

---

[2] The other NLU features, such as named entities or anaphora, often only affect the choice of content in the next system response rather than its type/DA, and they do not have a value at each turn, unlike the DA feature.

dling a subdialogue on the same topic.

Flow-RG makes it possible for a flow to resume after a few turns handled by a different RG on the same topic. The flow can simply begin a new miniflow, if there is at least one miniflow that has not yet been visited. Resumption is also possible in the middle of a miniflow, which allows a different RG to chime in for up to two turns (such as EVI answering an on-topic factual question that the flow has no answer prepared for), and then have the miniflow pick up where it left off.

ATHENA	Okay, can't wait to get to know a thing or two about you. Valentine's day is on Sunday. Do you have a Valentine?
USER	yes my wife
ATHENA	Cool! I wish I could meet a socialbot that I could fall in love with; then they'd be my Valentine. Anyway, I've always wanted to go on a vacation and to travel. What's a place that you've always wanted to travel to?
USER	paris sounds great
ATHENA	The Eiffel Tower is a masterpiece of architecture. For me, I think the great wall of china is at the top of my list of travel destinations. It seems like it would be amazing to walk along it. So, Is there a particular reason why you want to go to paris?
USER	the food is really bomb
ATHENA	Delicious food is always a good reason to go somewhere. Hey, I'm curious, Do you prefer flying? or road trips?
USER	I like flying generally
ATHENA	That's an interesting answer. I'd like to go on the road. You can see so much during a roadtrip, it would be cool to watch the world go by. Hmm, Do you like to relax during vacations?

Figure 5: Introduction Flow-RG subdialogue, just before Valentine's day.

Introduction RG. The Introduction Flow-RG, which every user experiences, has a strong effect on the user's overall experience (see Figure 5). The Introduction front-loads the conversation with getting-to-know-you content, by learning the user's name and asking icebreaker questions, such as favorite travel destinations, and weekend activities. The Introduction also brings up relevant current events, such as holidays, and gives the user a chance to ask Athena questions. Some of these turns will be the same for most users, e.g., asking for their name. Other content will change based on proximity to significant events in the year or the current day of the week, while some content changes randomly, for example, asking different questions related to vacation preferences. Content related to particular holidays as illustrated in Figure 5 are set up on a calendar and automatically started and stopped. The introduction also changes significantly for repeat users to indicate that we remember them, and provide a novel experience.

4.2 Knowledge Graph-Based RGs

The goal of the Knowledge Graph-based RGs (KG RGs) is to create deep knowledge-grounded conversations, where Athena always has more to say, by traversing relations in the Wikidata knowledge graph. Athena has four KG RGs covering movies, music, sports and TV, with conversations anchored around KG nodes (named entities). Each topic attempts to continue the conversation by either responding with a fact about an entity in context, or by selecting an entity from a set of fallback entities. When the system has either run out of facts on a particular entity, or has been on the same entity for a number of turns above a threshold, the RG attempts to traverse one or more relation edges, to a related entity, to continue the conversation. An example for the TV KG RG is in Figure 1.

ATHENA	Is there a television show that you like?
USER	**Friends** is one of my favorites
	Entity linking: Friends -> entity: Friends, wiki_id: Q79784
ATHENA	Friends is a great show. Why do you like it?
USER	It is really funny
ATHENA	What character do you like the most?
USER	I love Rachel
ATHENA	I really like how Jennifer Aniston plays Rachel Green. Did you know that she won a Primetime Emmy Award for Outstanding Lead Actress in a Comedy Series for her role?
USER	No, that's interesting

Table 1: Conversation using the knowledge graph.

Each topic has one to three entity types which the RG can respond about and each entity has a set of relations that can be used to generate responses. Each relation can only be used once for a particular entity, but can be reused when the RG has switched to a new entity. One limitation of the KG RGs is the need to select "interesting" relations and write templates by hand (Moon et al., 2019).

4.3 Entity-Based Indexing RGs

Entity-Based Indexing RGs are topical retrieval-based generators where the focus of the response is on "fun facts" for entities in a topic. Table 2 indicates how many fun facts these RGs have for each topic, and provides examples.

4.4 Neural NLG RGs

We have also developed and experimented with several different neural NLGs, including neural

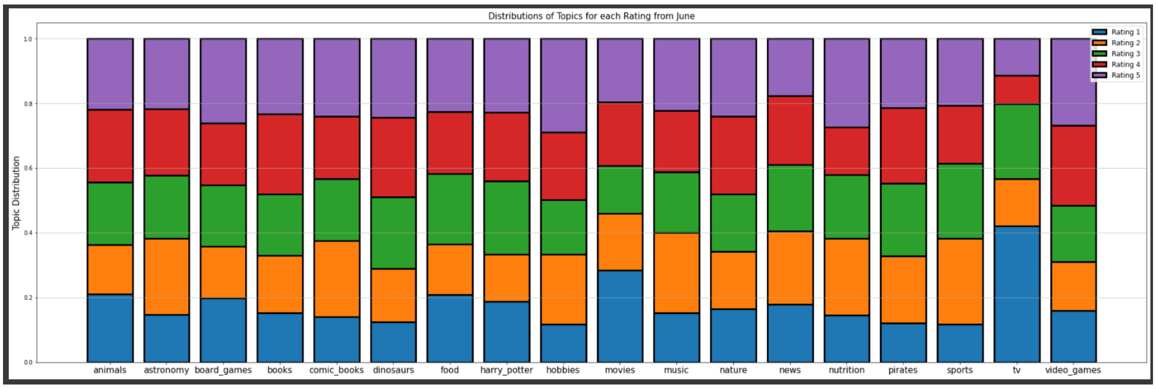


Figure 6: Distribution of ratings for each topic for the period from January 1st to June 16th.

Topic	#Facts	Example
Animals	90	I read this surprising fact about koalas. The fingerprints of a koala are so indistinguishable from humans that they have on occasion been confused at a crime scene. Imagine having your fingerprints confused with a koala, how strange!
Comic Books	26	Batman and Robin are the best superhero sidekick team. Once, after Batman and Robin rescued it from a slaughterhouse, DC comics included a Bat-cow.
Harry Potter	21	Fred and George Weasley were such tricksters. When Fred and George Weasley bewitched snowballs to hit Professor Quirrell's turban, they were unwittingly hitting Voldemort in the face.
Movies	54	One of my favorite movie series of all time are the James Bond movies. Before signing on as James Bond, Daniel Craig wasn't sure he wanted to play the role.
Music	31	I like to think that when music is played, it makes everyone happier, even animals. Researchers at the University of Leicester found that cows produce 3% more milk when they listen to relaxing music.
Nature	15	I like learning more about nature. It's actually really dark in the Amazon Forest! The forest is so thick that only 1% of sunlight can make it through.
Video Games	20	Here's a fact I discovered recently about World of Warcraft. A lot of famous people played World of Warcraft, including Vin Diesel, Mila Kunis and even Robin Williams. Isn't that cool?

Table 2: Fun facts for popular entity-based RG topics.

NLGs that generate from meaning representations and are thus topic specific (Juraska et al., 2019; Harrison et al., 2019; Oraby et al., 2019).

We also developed a neural NLG that we call Discourse-Driven NRG (DD-NRG) that generates directly from the conversation context and can be used for any topic (Rajasekaran, 2020; Tosh, 2020). We also systematically tested two topic-agnostic neural NLGs provided by Amazon, the PD-NRG (Hedayatnia et al., 2020) and a model called Topical-NRG that was trained on the Alexa Prize conversations of all finalists in the 19/20 competition. We found that it was difficult to control the quality of the neural RG outputs and guarantee their coherence, so we only deployed them to collect experimental data for short periods. We are currently experimenting with methods for controllable generation for these RGs (Reed et al., 2020; Harrison et al., 2019; Juraska and Walker, 2021).

5 Evaluation and Analysis

The two criteria that are specified in the Alexa Prize Grand Challenge that systems aim to optimize are length of conversation and user ratings. The Grand Prize will go to a system that achieves conversations of **at least 20 minutes** with **average ratings of 4.0** on a scale of 1 to 5.[3]

Over the 4 years our team has been in the competition, we have found that interactions with users are vulnerable to noise due to the competition setup (Bowden et al., 2019a,b; Harrison et al., 2020). Users often get into the Alexa Prize skill by accident leading to many conversations of only 1 or 2 turns (Shalyminov et al., 2018). Surprisingly, even for single turn conversations, some users still provide ratings. To improve our analysis of system performance, we remove these very short conversations from the data. Table 3 show the ratings, lengths in turns, and durations, during the semi-finals and the finals. On June 25th, before entering the finals, the average rating across all the systems in the semi-finals was 3.41 and the median duration was 2.12.

	Ratings		Turns		Time
	Mean	Median	Mean	Median	Median
Semifinals	3.62	4.0	17	24	2.46
Finals	3.71	4.0	18	24	2.01

Table 3: Athena's performance during the semi-finals and the finals for rating, length and duration.

Obviously, user's interactions with different RGs and topics affect their conversations and therefore their ratings. While only about 20% of users actually provide ratings, over the course of this year, we collected about 38K conversations with ratings. The distribution of ratings by topic presence in con-

---

[3] https://www.amazon.science/academic-engagements/alexa-prize-socialbot-grand-challenge-4-finalists-announced

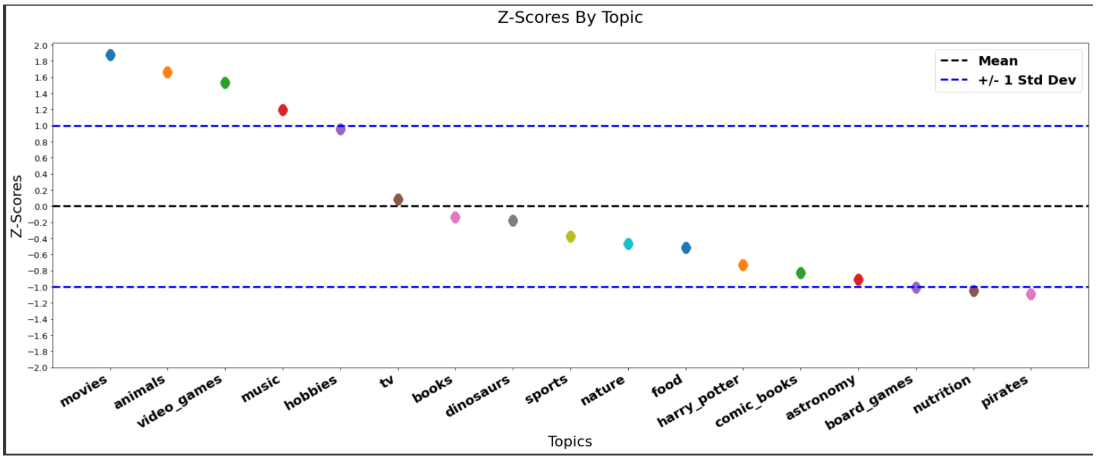


Figure 7: Z-scores for Athena topics for the period from June 1st to June 16th.

versations from January to June are in Figure 6. The purple and red bars indicate proportions of the topic that occur in conversations with ratings of 4 and 5 respectively. This suggests that the highest performing topics include animals, comic books, Harry Potter, hobbies, and video games, and that only a few topics are actually performing poorly, such as dinosaurs, news and sports.

However, presence in a conversation is a rather imprecise indicator of topic quality. In order to better understand the contribution of each topic to Athena's overall ratings, we developed a novel scoring function that aims to optimize topic selection over the prizes' user ratings and conversation duration criteria. Thus, our scoring function gives credit based on the number of utterances in a conversation that are contributed by each topic in the conversation. The number of utterances is multiplied by the conversation rating and summed for each topic over all rated conversations. This sum is then normalized to produce its Z-score. The plot shown in Figure 7 indicates for each topic, how many standard deviations its performance is above or below the Athena's mean performance for two weeks in June, 2021. This plot indicates that the topics that contribute most to long dialogues that are more highly rated include movies, animals, video games, music and hobbies, all of which are 1 or more standard deviations above the mean in performance. Walker et al. (2021) discusses the impact of topic selection and dialogue behaviors on user ratings and conversation length.

We also investigated the impact of personalization on the dialogue manager with an A/B study run over 22 days. The A system has the fully personalized dialogue manager, while B utilizes our previously existing heuristic topic selection strategy. We selected conversations of 6 turns or more to allow personalization to have an effect and also filter out conversations that never perform topic promotion; thus, we only inspect conversations that differ due to the topic selection strategy. After filtering, personalization occurred in roughly 25% of conversations. Our results, shown in Table 4, indicate a statistically significant increase in both user rating and conversation length.

	personalized	heuristic	p-value
convs	1066	614	—
rating	4.02	3.85	.007
length	38.07	35.07	.016

Table 4: Personalized topic selection vs. heuristic topic selection over 22 days.

6 Conclusion

This paper describes Athena, and its design and performance in the Alexa Prize Grand Challenge 4. Our system demonstration will illustrate both the strengths and the weaknesses of Athena and provide the basis for discussion about the state of the art of open-domain conversational AI.

Acknowledgements

We would like to thank Christian Benavidez, Yaqing Cao, James Graupera, Colin Harmon, Venkatesh Nagubandi, Meltem Ozcan, Diego Pedro, Navya Rao, Stephanie Rich, Jasiel Rivera-Trinadad and Aditya Tarde for helping with fun facts, Wikidata queries and prosody markup.

References

Kevin K Bowden, Jiaqi Wu, Wen Cui, Juraj Juraska, Vrindavan Harrison, Brian Schwarzmann, Nicholas

Santer, Steve Whittaker, and Marilyn Walker. 2019a. Entertaining and opinionated but too controlling: a large-scale user study of an open domain alexa prize system. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–10.

Kevin K Bowden, Jiaqi Wu, Wen Cui, Juraj Juraska, Vrindavan Harrison, Brian Schwarzmann, Nick Santer, and Marilyn Walker. 2019b. Slugbot: Developing a computational model and framework of a novel dialogue genre. *arXiv preprint arXiv:1907.10658*.

Bruce Buntschuh, C. Kamm, G. DiFabbrizio, A. Abella, M. Mohri, S. Narayanan, I. Zeljkovic, R.D. Sharp, J. Wright, S. Marcus, J. Shaffer, R. Duncan, and J.G. Wilpon. 1998. VPQ: A spoken language interface to large scale directory information. In *Proc. of ICSLP*, Sydney, Australia.

Alessandra Cervone, Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Anu Venkatesh, Dilek Hakkani-Tur, and Raefer Gabriel. 2019. Natural language generation at scale: A case study for open domain question answering. In *arXiv preprint arXiv:1903.08097*.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, et al. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa Prize Proceedings*.

Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *SIGDIAL 2017: Session on Natural Language Generation for Dialog Systems*.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. *NAACL HLT 2018*, page 96.

Sarah E Finch, James D Finch, Ali Ahmadvand, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, Jinho D Choi, et al. 2020. Emora: An inquisitive social chatbot who cares for you. *Alexa Prize Proceedings*.

James Glass and Eugene Weinstein. 2001. Speech-Builder: Facilitating spoken dialogue system development. In *Eurospeech*, pages 1335–1338.

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tür. 2020. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study. *Proc. Interspeech 2020*, pages 911–915.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, et al. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *Alexa Prize Proceedings*.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in nlg: Personality variation and discourse contrast. *DSNNLG 2019*, page 1.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421.

Juraj Juraska, Kevin K Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Juraj Juraska and Marilyn Walker. 2021. Attention is indeed all you need: Semantically attention-guided decoding for data-to-text nlg. In *International Conference on Natural Language Generation, INLG*.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A user adaptive social conversational system. *Proceedings of the Alexa Prize: Social Bot Grand Challenge 3*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and method for joint control of semantics and style in neural nlg. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5938–5951.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. *Proceedings of the Alexa Prize: Social Bot Grand Challenge 3*.

Omkar Patil, Lena Reed, Kevin K. Bowden, Juraj Juraska, Wen Cui, Vrindavan Harrison, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Phillip Lee, Eduardo Zamora, Jeshwanth Bheemanpally, Rohan Pandey, Adwait Ratnaparkhi, and Marilyn Walker. 2021. Athena 2.0: Discourse and user modeling in open domain dialogue. *Proceedings of the Alexa Prize*.

Jan Pichl, CTU FEE, Petr Marek, Jakub Konrád, Martin Matulík, and Jan Šedivỳ. 2020. Alquist 2.0: Alexa prize socialbot based on sub-dialogue models. *Proceedings of the Alexa Prize: Social Bot Grand Challenge 3*.

Rishi Rajasekaran. 2020. Transfer-learning approaches towards controllable response generation for conversational agents. *PDF Download*.

Lena Reed, Vrindavan Harrison, Shereen Oraby, Dilek Hakkani-Tur, and Marilyn Walker. 2020. Learning from mistakes: Combining ontologies via self-training for dialogue generation. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2020)*.

Stephanie Seneff, Ed Hurley, Raymond Lau, Christine Pao, Philipp Schmid, and Victor Zue. 1998. Galaxy-ii: A reference architecture for conversational system development. In *Fifth International Conference on Spoken Language Processing*.

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8.

A. Stent. 2000. Rhetorical structure in dialog. In *Proc. of the first international conference on Natural language generation-Volume 14*, pages 247–252. Association for Computational Linguistics.

Zachary Tosh. 2020. Developing knowledge-grounded natural language generation models for the dstc9 competition. *PDF Download*.

M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. 2001. DARPA communicator dialog travel planning systems: The june 2000 data collection. In *EUROSPEECH 2001*.

M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics, ACL/EACL 97*, pages 271–280.

Marilyn Walker, Colin Harmon, James Graupera, Davan Harrison, and Steve Whittaker. 2021. Developing paradise for open domain dialogue systems. In *International Workshop on Spoken Dialogue System Technology*. Springer.

Marilyn A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Appendix

A: I'd like to hear your ideas on this. If you could choose one food to live on for your entire life, which food would it be? Why?
U: i think nachos
A: That's an interesting answer! If I was a human, I would eat Sweet potatoes. Sweet potatoes are a really nutritious food. You can eat them in a variety of ways like baked, fries, or twice baked. Anyway, that's my thoughts on the matter.
A: Okay. I was curious about your opinion on this. If you had the option, would you rather spend time in the mountains, or, at the beach
U: i like the beach more
A: Choosing the beach is a good choice! If I was a human, I would spend time at the ocean - I would walk on the beach, search for shells, and relax in the sun. Anyhow, that's where my minds at, let's move forward.

Table 5: Two examples of Athena's attempt to increase topical depth by asking different types of playful questions. We crowdsourced 2.5k playful questions, which were then curated for quality. All of this content was also updated to match Athena's self-aware personality, e.g., *If I was human*. "A" denotes Athena and "U" the user turns.

A: But most people seem to find dinosaurs really interesting. There are dinosaur movies and whole floors of natural history museums dedicated to dinosaurs. Personally, I find it fascinating that modern birds are descended from a particular family of dinosaurs.
U: hella cool
A: Definitely. I was surprised to hear how much work went into creating the movie Land Before Time. Over 600 background paintings were made for the film. The artists had to create a believable environment in which there was almost no foliage. Most paintings depicted beautiful but barren wastelands.
U: surprising
A: Right? My favorite dinosaur is the Tyrannosaurus Rex. It's huge and ferocious, but the little arms make it somewhat adorable. What about you? What's your favorite dinosaur?
U: i'm not sure i have to say a t. rex
A: Okay. There are lots of movies about dinosaurs. One of the most famous dinosaur movies is Jurassic Park, old but classic. Are there any dinosaur movies that you love?
U: i have to say jurassic park because that's the only one i've seen
A: Awesome! That one is well known! I seem to remember people saying the dinosaurs were very vivid
U: i never knew that
A: yeah, I love to collect these fun facts! But anyways, I'm interested in your opinion. Would you rather ride on the back of a pterodactyl or brontosaurus as your primary means of transportation?
U: i have to say brontosaurus and i hate heights
A: Choosing a Brontosaurus you say, great pick! I would rather ride a pterodactyl so I could experience flight as well as move around more efficiently. Riding a brontosaurus would be thrilling because of their size but less practical than a pterodactyl. So, okay. Do you know why dinosaurs are extinct?
U: yeah cause i'm eat your head it right
A: Ah nice, you must be super smart. Let me tell you what I heard.The mass extinction of dinosaurs is a scientific mystery, but most hypotheses include an extraterrestrial impact or a massive bout of volcanism. What do you think happened?
U: i'd say aliens
A: I'd like to hear your ideas on this. Let's say you are trapped on Jurassic World alone with the dinosaurs, what would you do to survive?
U: i'm not sure i just say find sheltering something and start there
A: It's a hard question! I would try to become closer to the younger dinosaurs so that they would grow up with me and feel more comfortable with me. I would also try to help them by finding food and shelter for them. At any rate, Dinosaurs had a lot of different ways to attack other dinosaurs and to defend themselves. To help fight meat eaters such as the Allosaurus or Spinosaurus, many plant eaters had natural weapons at their disposal.

Table 6: An example conversation about Dinosaurs, one of our more niche topics. We use a variety of different miniflows to increase our topical depth. "A" denotes Athena and "U" the user turns.

A: One thing that really interests me is people's hobbies. Can you give me some info about swimming?
U: it's pretty fun
A: Right? Sounds totally tubular. Why do you like swimming?
U: it's because i can kind of just feel like i can move around and do a lot of stuff i can't do on land
A: Cool! How did you first get started with swimming?
U: i took swimming lessons
A: I realize you might be totally in love with swimming, but I'm curious. Is there any part of the hobby you don't like?
U: tough lessons
A: Well swimming sounds interesting, thanks for telling me all about your hobby. A lot of people have cool hobbies, but what about me? What kind of hobby do you think I would like?
U: maybe writing
A: Alright. Why do you think writing is a good fit for me?
U: because you like expressing yourself seems like
A: Ok. I see. Do you need any special equipment or skills to get into writing?
U: just paper and pen and imagination
A: I see. Thank you for recommending writing to me, I'll try to find out more about the hobby in the future!

Table 7: An example of Athena engaging the user about their hobby, swimming, which we previously learned about in the Introduction RG. "A" denotes Athena and "U" the user turns.