Sounding Board: A User-Centric and Content-Driven Social Chatbot

Hao Fang* **Hao Cheng*** **Maarten Sap**† **Elizabeth Clark**† **Ari Holtzman**†
Yejin Choi† **Noah A. Smith**† **Mari Ostendorf***
University of Washington
*{hfang,chenghao,ostendor}@uw.edu
†{msap,eaclark7,ahai,yejin,nasmith}@cs.washington.edu

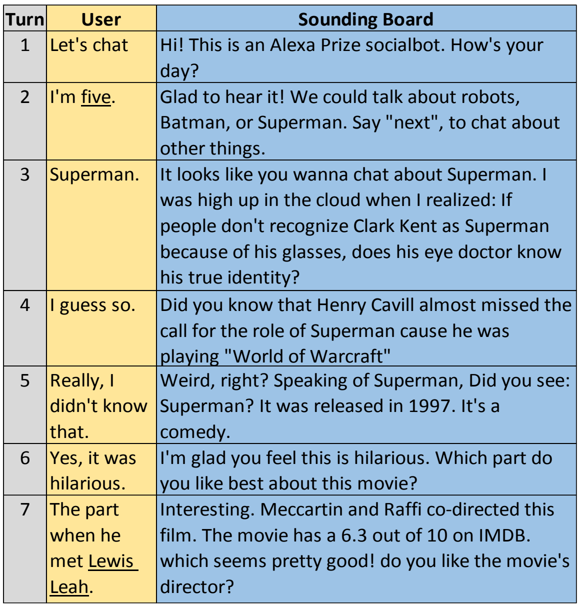
Abstract

We present Sounding Board, a social chatbot that won the 2017 Amazon Alexa Prize. The system architecture consists of several components including spoken language processing, dialogue management, language generation, and content management, with emphasis on *user-centric* and *content-driven* design. We also share insights gained from large-scale online logs based on 160,000 conversations with real-world users.

1 Introduction

Researchers in artificial intelligence (AI) have long been interested in the challenge of developing a system that can have a coherent conversation with humans: early systems include Parry (Colby et al., 1971), Eliza (Weizenbaum, 1966), and Alice (Wallace, 2009). The Loebner Prize[1] is a Turing test for conversational AI that has challenged researchers since 1990. Recent work has addressed tasks where passing the Turing test is not a concern. Goal-oriented conversational systems facilitate natural user interaction with devices via text and spoken language. These AI assistants typically focus on short interactions, as in commercial products such as Amazon Alexa, Microsoft Cortana, Google Assistant, and Apple Siri. General conversational systems, called chatbots, have constrained social interaction capabilities but have difficulty generating conversations with long-term coherence (Serban et al., 2017; Sato et al., 2017; Shao et al., 2017; Tian et al., 2017; Ghazvininejad et al., 2018).

The Alexa Prize sets forth a new challenge: creating a system that can hold a coherent and engaging conversation on current events and popular topics such as sports, politics, entertainment, fashion and technology (Ram et al., 2017). Our system, Sounding Board,[2] demonstrates that it is feasible to build an agent that can engage in long-term conversation when backed by rich content and knowledge of the user obtained through interaction. Sounding Board won the inaugural Amazon Alexa Prize with an average score of 3.17 on a 5-point scale and an average conversation duration of 10:22, evaluated by a panel of independent judges.[3]



Turn	User	Sounding Board
1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
2	I'm five.	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
3	Superman.	It looks like you wanna chat about Superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his true identity?
4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
5	Really, I didn't know that.	Weird, right? Speaking of Superman, Did you see: Superman? It was released in 1997. It's a comedy.
6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part do you like best about this movie?
7	The part when he met Lewis Leah.	Interesting. Meccartin and Raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?

Figure 1: A sample dialog. Suspected speech recognition errors in the user utterances are underlined.

There are two key design objectives of Sounding Board: to be *user-centric* and *content-driven*. Our system is *user-centric* in that users can control the topic of conversation, while the system adapts responses to the user's likely interests by gauging the user's personality. Sounding Board is also *content-driven*, as it continually supplies interesting and relevant information to continue the conversation,

---

[1] http://aisb.org.uk/events/loebner-prize

[2] https://sounding-board.github.io

[3] https://developer.amazon.com/alexaprize/2017-alexa-prize

enabled by a rich content collection that it updates daily. It is this content that can engage users for a long period of time and drive the conversation forward. A sample conversation is shown in Fig. 1.

We describe the system architecture in §2, share our insights based on large scale conversation logs in §3, and conclude in §4.

2 System Architecture

Sounding Board uses a modular framework as shown in Fig. 2. When a user speaks, the system produces a response using three modules: natural language understanding (NLU), dialog manager (DM), and natural language generation (NLG). The NLU produces a representation of the current event by analyzing the user's speech given the current dialog state (§2.1). Then, based on this representation, the DM executes the dialog policy and decides the next dialog state (§2.2). Finally, the NLG uses the content selected by the DM to build the response (§2.3), which is returned to the user and stored as context in the DM. During the conversation, the DM also communicates with a knowledge graph that is stored in the back-end and updated daily by the content management module (§2.4).

2.1 Natural Language Understanding

Given a user's utterance, the NLU module extracts the speaker's intent or goals, the desired topic or potential subtopics of conversation, and the stance or sentiment of a user's reaction to a system comment. We store this information in a multidimensional frame which defines the NLU output.

To populate the attributes of the frame, the NLU module uses ASR hypotheses and the voice user interface output (Kumar et al., 2017), as well as the dialog state. The dialog state is useful for cases where the system has asked a question with constraints on the expected response. A second stage of processing uses parsing results and dialog state in a set of text classifiers to refine the attributes.

2.2 Dialog Management

We designed the DM according to three high-level objectives: engagement, coherence, and user experience. The DM takes into account user **engagement** based on components of the NLU output and tries to maintain user interest by promoting diversity of interaction strategies (conversation modes). Each conversation mode is managed by a miniskill that handles a specific type of conversation segment. The DM tries to maintain dialog **coherence** by choosing content on the same or a related topic within a conversation segment, and it does not present topics or content that were already shared with the user. To enhance the **user experience**, the DM uses conversation grounding acts to explain (either explicitly or implicitly) the system's action and to instruct the user with available options.

The DM uses a hierarchically-structured, state-based dialog model operating at two levels: a master that manages the overall conversation, and a collection of miniskills that handle different types of conversation segments. This hierarchy enables variety within specific topic segments. In the Fig. 1 dialog, Turn 3 was produced using the *Thoughts* miniskill, Turn 4 using the *Facts* miniskill, and Turns 5–7 using the *Movies* miniskill. The hierarchical architecture simplifies updating and adding new capabilities. It is also useful for handling high-level conversation mode changes that are frequent in user interactions with socialbots.

At each conversation turn, a sequence of processing steps are executed to identify a response strategy that addresses the user's intent and meets the constraints on the conversation topic, if any. First, a state-independent processing step checks if the speaker is initiating a new conversation segment (e.g., requesting a new topic). If not, a second processing stage executes state-dependent dialog policies. Both of these processing stages poll miniskills to identify which ones are able to satisfy constraints of user intent and/or topic. Ultimately, the DM produces a list of speech acts and corresponding content to be used for NLG, and then updates the dialog state.

2.3 Natural Language Generation

The NLG module takes as input the speech acts and content provided by the DM and constructs a response by generating and combining the response components.

Phrase Generation: The response consists of speech acts from four broad categories: grounding, inform, request, and instruction. For instance, the system response at Turn 7 contains three speech acts: grounding (*"Interesting."*), inform (the IMDB rating), and request (*"do you like the movie's director?"*). As required by the hosting platform, the response is split into a message and a reprompt. The device always reads the message; the reprompt is optionally used if the device does not detect a

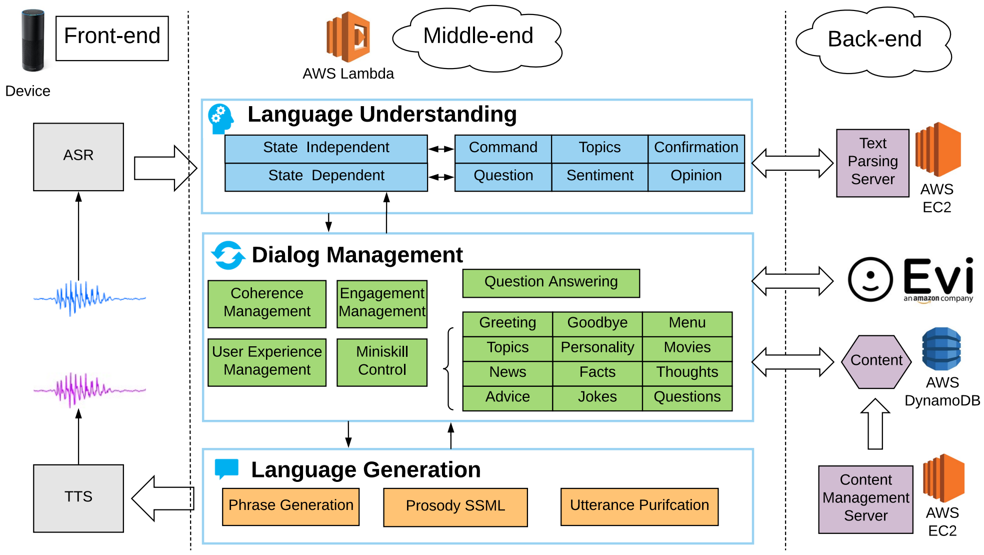


Figure 2: System architecture. **Front-end**: Amazon's Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) APIs. **Middle-end**: NLU, DM and NLG modules implemented using the AWS Lambda service. **Back-end**: External services and AWS DynamoDB tables for storing the knowledge graph.

response from the user. The instruction speech acts are usually placed in the reprompt.

Prosody: We make extensive use of speech synthesis markup language (SSML) for prosody and pronunciation to convey information more clearly. to communicate. We use it to improve the naturalness of concatenated speech acts, to emphasize suggested topics, to deliver jokes more effectively, to apologize or backchannel in a more natural-sounding way, and to more appropriately pronounce unusual words.

Utterance Purification: The constructed response (which may repeat a user statement) goes through an utterance purifier that replaces profanity with a non-offensive word chosen randomly from a list of innocuous nouns, often to a humorous effect.

2.4 Content Management

Content is stored in a knowledge graph at the back-end, which is updated daily. The knowledge graph is organized based on miniskills so that query and recommendation can be carried out efficiently by the DM. The DM drives the conversation forward and generates responses by either traversing links between content nodes associated with the same topic or through relation edges to content nodes on a relevant new topic. The relation edges are compiled based on existing knowledge bases (e.g., Wikipedia and IMDB) and entity co-occurrence between content nodes.

Because Sounding Board is accessible to a wide range of users, the system needs to provide content and topics that are appropriate for a general audience. This requires filtering out inappropriate and controversial material. Much of this content is removed using regular expressions to catch profanity. However, we also filtered content containing phrases related to sensitive topics or phrases that were not inherently inappropriate but were often found in potentially offensive statements (e.g., "your mother"). Content that is not well suited in style to casual conversation (e.g., URLs and lengthy content) is either removed or simplified.

3 Evaluation and Analysis

To analyze system performance, we study conversation data collected from Sounding Board over a one month period (Nov. 24–Dec. 24, 2017). In this period, Sounding Board had 160,210 conversations with users that lasted 3 or more turns. (We omit the shorter sessions, since many involve cases where the user did not intend to invoke the system.) At the end of each conversation, the Alexa Prize platform collects a rating from the user by asking *"on a scale of 1 to 5, how do you feel about speaking with this socialbot again?"* (Ram et al., 2017). In this data, 43% were rated by the user, with a mean score of 3.65 ($\sigma = 1.40$). Of the rated conversations, 23% received a score of 1 or 2, 37% received a score of 3 or 4, and 40% received a score of 5.[4] The data are used to analyze how different personality types interact with the system (§3.1) and length, depth, and

---

[4] Some users give a fractional number score. These scores are rounded down to the next smallest integer.

	ope	con	ext	agr	neu
% users	80.02%	51.70%	61.59%	79.50%	42.50%
# turns	0.048**	*not sig.*	0.075**	0.085**	*not sig.*
rating	0.108**	*not sig.*	0.199**	0.198**	*not sig.*

Table 1: Association statistics between personality traits (**ope**nness, **con**scientiousness, **ext**raversion, **agr**eeableness, **neu**roticism) and z-scored conversation metrics. "% users" shows the proportion of users scoring positively on a trait. "# turns" shows correlation between the trait and the number of turns, and "rating" the correlation between the trait and the conversation rating, controlled for number of turns. Significance level (Holm corrected for multiple comparisons): $^{**}p < 0.001$.

breadth characteristics of the conversations (§3.2).

3.1 Personality Analysis

The *Personality* miniskill in Sounding Board calibrates user personality based on the Five Factor model (McCrae and John, 1992) through exchanging answers on a set of personality probing questions adapted from the mini-IPIP questionnaire (Donnellan et al., 2006).

We present an analysis of how different personality traits interact with Sounding Board, as seen in Table 1. We find that personality only very slightly correlates with length of conversation (# turns). However, when accounting for the number of turns, personality correlates moderately with the conversation rating. Specifically, we find users who are more extraverted, agreeable, or open to experience tend to rate our socialbot higher. This falls in line with psychology findings (McCrae and John, 1992), which associate extraversion with talkativeness, agreeableness with cooperativeness, and openness with intellectual curiosity.[5]

3.2 Content Analysis

Most Sounding Board conversations were short (43% consist of fewer than 10 turns), but the length distribution has a long tail. The longest conversation consisted of 772 turns, and the average conversation length was 19.4 turns. As seen in Fig. 3, longer conversations tended to get higher ratings.

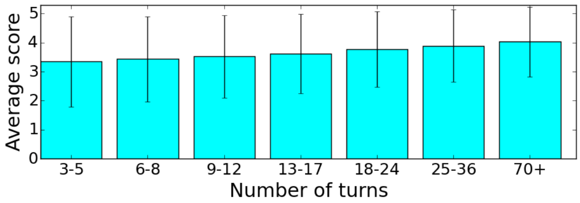


Figure 3: Average conversation score by conversation length. Each bar represents conversations that contain the number of turns in the range listed beneath them and is marked with the standard deviation.

While conversation length is an important factor, it alone is not enough to assess the conversation quality, as evidenced by the low correlation with user ratings ($r = 0.14$) and because some turns (e.g., repairs) may have a negative impact. Therefore, we also study the breadth and depth of the sub-dialogs within conversations of roughly equal length (36–50) with high (5) vs. low (1–2) ratings. We automatically segment the conversations into sub-dialogs based on the system-identified topic, and annotate each sub-dialog as engaged or not depending on the number of turns where the system detects that the user is engaged. The breadth of the conversation can be roughly characterized by the number and percentage of engaged sub-dialogs; depth is characterized by the average number of turns in a sub-dialog. We found that the average topic engagement percentages differ significantly (62.5% for high scoring vs. 28.6% for low), but the number of engaged sub-dialogs were similar (4.2 for high vs. 4.1 for low). Consistent with this, the average depth of the sub-dialog was higher for the high conversations (4.0 vs. 3.8 turns).

4 Conclusion

We presented Sounding Board, a social chatbot that has won the inaugural Alexa Prize Challenge. As key design principles, our system focuses on providing conversation experience that is both user-centric and content-driven. Potential avenues for future research include increasing the success rate of the topic suggestion and improving the engagements via better analysis of user personality and topic-engagement patterns across users.

Acknowledgements

In addition to the Alexa Prize financial and cloud computing support, this work was supported in part by NSF Graduate Research Fellowship (awarded to E. Clark), NSF (IIS-1524371), and DARPA CwC program through ARO (W911NF-15-1-0543). The conclusions and findings are those of the authors and do not necessarily reflect the views of sponsors.

---

[5] These insights should be taken with a grain of salt, both because the mini-IPIP personality scale has imperfect reliability (Donnellan et al., 2006) and user responses in such a casual scenario can be noisy.

References

Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. 2006. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment*, 18(2):192.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proc. AAAI*.

Anjishnu Kumar, Arpit Gupta, Julian Chan, Sam Tucker, Bjorn Hoffmeister, and Markus Dreyer. 2017. Just ASK: Building an architecture for extensible self-service spoken language understanding. In *Proc. NIPS Workshop Conversational AI*.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational AI: The science behind the Alexa Prize. In *Proc. Alexa Prize 2017*.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proc. ACL Student Research Workshop*, pages 120–127.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proc. EMNLP*, pages 2210–2219.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? An empirical study on context-aware neural conversational models. In *Proc. ACL*, pages 231–236.

Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.*, chapter Parsing the Turing Test. Springer, Dordrecht.

Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.