Sounding Board – University of Washington's Alexa Prize Submission

Hao Fang[1], **Hao Cheng**[1], **Elizabeth Clark**[2], **Ariel Holtzman**[2], **Maarten Sap**[2], **Mari Ostendorf**[1], **Yejin Choi**[2], **and Noah A. Smith**[2]

[1] Department of Electrical Engineering, University of Washington, Seattle
{hfang,chenghao,ostendor}@uw.edu
[2] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle
{eaclark7,ahai,msap,yejin,nasmith}@cs.washington.edu

Abstract

This paper introduces the University of Washington's Alexa Prize socialbot, Sounding Board, which is designed to engage users with a wide variety of content. The system models the user utterance using a multi-dimensional representation. A hierarchical dialogue manager is employed where a master manages the overall conversation and a collection of miniskills manage different conversation segments. The system constructs responses using speech acts selected by the dialogue manager, where each speech act is instantiated with randomness to introduce variation to the conversation. Further, we analyze the impact of miniskill variety, user personality, and speech recognition performance on user interaction ratings.

1 Introduction

The University of Washington (UW) socialbot, Sounding Board, is a conversational agent that is addressing the Alexa Prize challenge, which is to engage users in discussions about topics of their choosing. While there have been previous studies exploring development of socialbots in an open domain setting [1, 2, 3, 4, 5, 6], these have primarily involved "chit-chat" conversations that are considered successful when generating responses that are reasonable in context. With Sounding Board, we chose to treat the socialbot as a more task-oriented problem, where the goal is to identify informative content and generate responses that are sensitive to user interests. In this paper, we outline the Sounding Board design philosophy, provide details of the system architecture, and present initial analyses of user interactions.

The UW team approached the Alexa Prize challenge starting from scratch – the team had no existing dialogue system to build on, nor were there corpora available that were suitable for training a neural sequence-to-sequence model [7] for open domain, information sharing conversations.[1] As a result, the Sounding Board system evolved substantially over the course of the competition. The initial system was entirely rule-based and had only one mechanism for identifying discussion content. As we have added capabilities and collected conversations with real users, the system architecture has become more sophisticated and components are being replaced with versions that leverage machine learning. Since the early versions of the system were necessarily rudimentary, the focus of this paper is on documenting the system operation in the final stages of the competition.

The Sounding Board design philosophy is primarily reflected in the conversation strategy and the system engineering approach. The conversation strategy has two key features. First, it is content

[1] We did explore movie scripts and discussion forum text, but the text was just too different in style and content from utterances we observed from our system users.

1st Proceedings of Alexa Prize (Alexa Prize 2017).

driven; we want to engage the user by providing them with information that they may not already know or perspectives that they may not have heard. Thus, information retrieval is important in our system. To cover a range of topics and user interests, we draw from different information sources using miniskills that can be thought of as a "panel of experts". We include some chit-chat, but it mainly plays a role in dialogue transition points. Second, the dialogue policy is highly user driven, and the system attempts to track the user mental state to adjust topic choice and interaction strategy. This goal impacts the system in several ways, including a multi-dimensional representation of the user utterance that includes sentiment and stance as well as utterance intent, using a personality quiz to help guide topic selection, and detecting user frustration to initiate topic change. In system development, we also put a high priority on accuracy of user intent recognition. Analysis of user behaviors and tester feedback influenced many of the architecture decisions.

The system engineering strategy is driven by the lack of appropriate conversational data for training and the plan to use multiple content sources, which together motivate a modular architecture with a hierarchical dialogue management strategy. As new capabilities are developed, it is relatively straightforward to add them to the system, and the modular architecture also easily scales to handle more miniskills and facilitates updating components as more useful data becomes available for data-driven learning. The generation strategy is also modular, with different components for different broad categories of speech acts. In order to make the conversational data currently being collected more useful for dialogue policy learning (and to make the system less monotonous), the response generation module includes several mechanisms for randomly introducing variation, and the randomness associated with content availability for different topics also leads to potentially useful variation in the conversations.

The complete system is described in detail in §2 and §3. In §4, we describe a few insights gleaned from analyzing conversations at the end of the competition period, and in §5, we discuss related work. Finally, §6 summarizes the main features of the system and outlines directions for future work.

2 System Architecture

Sounding Board interacts with users through Amazon's Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) APIs included in the Alexa Skill Kit (ASK)[2] which acts as the front-end of Sounding Board. The system is deployed as an AWS Lambda service triggered by ASK events, which contain the ASR hypotheses for a user utterance and the output of the voice user interface (VUI) which identifies selected user intents. The AWS Lambda service acts as the middle-end of Sounding Board. As shown in Fig. 1, it consists of three major system modules of Sounding Board: a natural language understanding (NLU) module, a dialogue management (DM) module, and a natural language generation (NLG) module. These system modules also communicate with different back-end services, including: a Stanford CoreNLP server [8] deployed at AWS EC2 for providing the parsing service, AWS DynamoDB tables where topic-indexed contents are stored, and Evi.com[3] for providing question answering and joke services.

Upon receiving an ASK event, Sounding Board goes through the three major system modules to produce a response to be returned to the ASK. First, the NLU module produces an input frame for the current event by analyzing the ASR hypotheses, the VUI output, and the dialogue state. Then, the DM module examines the input frame, executes the dialogue policy, and updates the dialogue state. The DM operates at two levels, with a master to manage the overall conversation and a collection of miniskills to handle different types of conversation segments (discourse segments). Finally, the NLG module uses the speech act and content selected by the DM module to build the response, which is returned to the ASK and also stored as part of the conversation context in the DM module.

The capabilities of the three modules and the available miniskills have evolved over the course of the competition. In the following subsections, we describe the three major modules as realized during the last 8 days of the semifinals (August 8–15, 2017). The miniskills are described in §3.

[2] `https://developer.amazon.com/alexa-skills-kit`

[3] `https://www.evi.com`

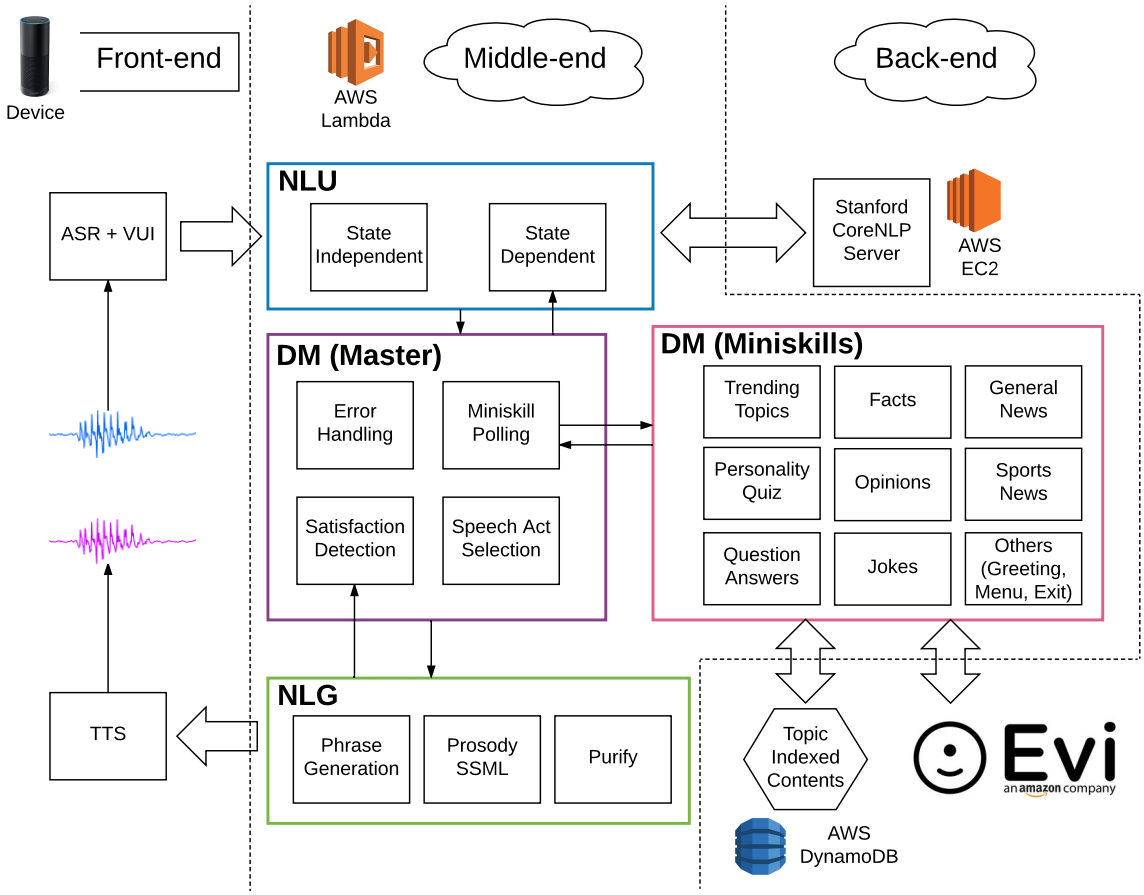

Figure 1: System architecture.

2.1 Natural language understanding

In order to appropriately respond to a user in a Sounding Board conversation, the NLU module needs to extract multiple types of information from a user utterance, including 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. Accordingly, we design a multi-dimensional input frame for Sounding Board as shown in Table 1, which defines the output of the NLU module. The attributes of the input frame are described further below.

To populate the attributes of the input frame, the NLU module uses the ASR hypotheses and the VUI output, as well as the dialogue state. Amazon's VUI is trainable from sample utterances, and it is used as a first-stage module for identifying input frame attributes. A second stage of processing refines the attributes, using parsing results and dialogue state information in a set of text classifiers targeting different attributes, described below. The dialogue state is useful for cases where the system has asked a question with constraints on the expected response, e.g. "continue" vs. "new topic".

primaryIntent: The `primaryIntent` attribute distinguishes between 22 intents that require different conversation strategies (and miniskill invocations) from the bot. These intents correspond to four broad classes: content retrieval commands, navigation commands, common Alexa commands, and a converse intent. The NLU module represents 9 categories of *Content Retrieval Commands*, including popular topics, facts, opinions, jokes, general news, sports news, personality quiz, question answers, and unspecified. Unlike human-to-human conversations, *Navigation Commands* are common in human-to-machine conversations, e.g., help, repeat, next, cancel, etc. Given that users are accustomed to standard Alexa commands, the navigation commands captured in the NLU module include 8 of Amazon's built-in intents, as well as "continue" and "change" intents. Sometimes users say *Common Alexa Commands* to call up skills that we cannot handle within Sounding Board, particularly "play music" and "read books". We automatically detect such commands to enable responding with an explanation of our limitations and instructions for exiting Sounding Board. The NLU module assigns all other utterances to the category of *Converse Intent*, including informing decisions/answers,

Table 1: Attributes in an input frame

Attribute	Description
`primaryIntent`	primary intent of the user utterance (a specific command or converse)
`questionType`	the question type of the user utterance (null if it is not a question)
`primaryTopic`	the primary topic inferred from the user utterance
`candidateTopics`	all possible topics extracted from the user utterance
`userReaction`	confirmation decision, opinion stance, sentiment

expressing opinions/feelings, asking questions, sharing experiences, providing content, and back-channeling. Most *Converse Intent* utterances are context-dependent in that they reference information in previous turns.

questionType: Analysis of user interactions showed that most questions could be classified as one of four types, which are handled with different strategies or miniskills in the DM. Specifically, Sounding Board distinguishes the following types of questions to further characterize user intent: 1) command, which is usually a polite way of making commands, e.g., "Can we talk about the Mars mission?"; 2) backstory/personal question, where the user seeks information about the socialbot's persona, e.g., name, birthday, hobby, etc.; 3) factual question, e.g., "Who is the president of the United States?"; 4) questions on sensitive topics (e.g., related to sex, violence, or drugs) or advice questions, e.g., "Which stock should I purchase?". We also include a fifth category for utterances that are not questions. While many questions are related to content retrieval, questions can be associated with any of the `primaryIntent` types.

candidateTopics: In Sounding Board, a topic can be any noun phrase[4] in the utterance. As more than one noun phrase can occur in an utterance, we store all such phrases in `candidateTopics`, filtering the list of detected noun phrases to remove invalid topics (such as "this", "yep", "something") and sensitive topics.

primaryTopic: The primary topic is chosen as the VUI identified topic, if available, and otherwise it is taken to be the longest noun phrase in the utterance.

userReaction: After the socialbot presents content, users sometimes express emotion in responding to a comment, or they may react by taking a positive or negative stance about the content. The NLU module of Sounding Board is equipped with three user reaction classifiers, each of which focuses on a specific dimension of user reactions, including: i) the user's decision on confirmation questions (approve, reject, unsure, null), ii) the user's stance on an opinion piece (agree, disagree, unsure/neutral, null), and iii) the user's sentiment about a fact or joke (like, dislike, neutral, null). The "null" case is used when the user utterance does not match the context of the classifier, e.g., when the previous utterance is not a confirmation question for (i) or when there is a recognizer error that makes the utterance uninterpretable.

2.2 Dialogue management

Sounding Board uses a hierarchically structured, state-based dialogue model, where the state includes a discrete set of interaction types, the result of the last personality quiz (if any), and a memory of previously discussed content. Within the DM module, there is a master processing sequence that manages the conversation as a whole, and a collection of miniskills that manage conversation segments for specific types of interactions, which we refer to as conversation modes. The hierarchical architecture simplifies the process of updating or adding new capabilities, and it is useful for handling the frequent high-level conversation mode changes that we observe in user interactions with Sounding Board.

At each turn, the DM module executes a sequence of processing steps that aim to identify a response strategy that addresses the user intent and meets the constraints on the conversation topic, if any. At the master processing level, the goal is to identify the conversation mode and the appropriate miniskill to respond to it. First, a state-independent processing step tries to identify cases that clearly initiate a new conversation segment, as for an explicit topic request or other command types. If

[4]We also include nbar segments, nouns, and lemmas to increase the hit rate, since we use "exact" topic indexing.

such cases are not found in the state-independent process, a second processing stage is used where state-dependent dialogue policies are executed. Both of these processing stages poll miniskills to identify which ones are able to satisfy constraints of user intent and/or topic. Miniskills with the most detailed topic match are prioritized, but otherwise miniskills are selected randomly, trying to avoid the same miniskill for consecutive turns. The miniskill manages the conversation segment of a specific conversation mode, and different miniskills have different dialogue policies and emit different system actions. Continuation on the same topic involves providing content from different sources; there is no understanding implemented for the content sources to allow diving deeper into a particular story. In addition, a state-dependent processing stage has capabilities to detect and take action on errors and other problems in the conversation, including negative sentiment and negative satisfaction (i.e. need for a topic change). Ultimately, the DM module produces a list of speech acts and corresponding content to be used by the NLG module, and then updates the dialogue state.

Content retrieval: Both state-independent and state-dependent processing stages often involve actions that retrieve content from back-end services, which result from the user making an explicit content retrieval command or the DM module choosing to push the conversation forward by presenting new content. To reduce the chance of content retrieval failure, the following backoff strategy is used. First, the DM miniskills try to retrieve content that satisfies constraints specified in the input frame, including the `primaryIntent` attribute which encodes the content type and the `primaryTopic` attribute. When no content is available, the DM module first removes the content-type constraint, and then subsequently relaxes the topic constraint to allow anything in `candidateTopics`. The DM module emits a content retrieval failure action if no content is retrieved from the above actions, which will result in an action to inform the user that the bot has nothing (more) to say on this topic.

Error handling: Sounding Board handles two types of errors, i.e., system errors and understanding errors. System errors include service exceptions (e.g., request time-out) and software failure due to bugs. In such cases, we use a soft exception handler that resets the dialogue state, restarts the conversation with a proper apology, and sends an email notification to the system developers. Understanding errors are caused by ASR errors, unanticipated user intents, and language processing errors. The system detects such errors when the input frame misses attributes required by the dialogue policy. When an understanding error is detected, the system initially responds by acknowledging a misunderstanding and providing suggestions to the user for continuing the conversation. If the system detects a second understanding error, the DM module chooses a new topic or miniskill to push the conversation forward, since we found that repeated requests for rephrasing were more annoying than random topic changes.

Satisfaction detection: A different type of problem in a conversation occurs when the system provides content that the user finds offensive, unpleasant or simply boring. While the user could respond by asking for a topic change, analyses of user interactions indicate that they much more often respond by expressing discontent, in which case the system needs to detect a problem and initiate a topic change. To address this problem, we designed a simple binary classifier, which was trained based on data annotated to indicate whether or not the system should change the topic, since this framing led to more consistent annotations. User satisfaction was associated with no topic change needed. The training data included 2381 hand-annotated examples from early Sounding Board conversations. Due to the limited data, we used a logistic regression classifier with n-gram features ($n \in \{1, 2, 3\}$) from both user and agent utterances. Using 10-fold cross-validation and optimizing for F1, the system achieves 87% accuracy (vs. 83% for never changing topics), and recall of 47% at precision of 65%.

Speech act selection: Handling a user turn can involve a sequence of actions, depending on the miniskill, but the end result requires selection of the speech act(s) to be used in the response to the user. A response can include multiple speech acts that reflect the goals of conversation management and providing information to the user, including four types: grounding, inform, request, and instruction. For purposes of grounding the conversation, the DM can specify one of 6 broad categories of speech acts: back-channel, echo of user request for confirmation, three forms of problem acknowledgement (misunderstanding, lack of content, user challenge), and gratitude. Previous research has shown that such conversational feedback is important [9]. Additionally, such feedback can communicate the agent's (potentially erroneous) understanding of the user's utterance, thereby preparing the user for potential non-sequiters. The grounding acts are included to acknowledge the user's utterance, and are primarily determined by the input frame produced by the NLU module, content retrieval results, topic change detection, and error handling. A lack of content occurs when all miniskills polled fail to

return content that meets topic constraints, as well as when the user asks follow-up questions that require coreference analysis (not yet handled in the system) or a deeper understanding of the content (e.g., "What is the reason for that?", "How did they do that?"). The category of "user challenge" is chosen based on detecting common phrases observed in the Sounding Board conversations associated with skepticism (e.g., "I don't believe that.", "I think that's a lie."), and repetition (e.g., "You just told me that."). The "inform" speech act is used when content is to be provided to the user and is coupled with the content. The request speech acts include confirmation questions, offers for the user to comment, and open requests for topics. Sounding Board usually pairs an inform act with a request act, which helps encourage users to provide a follow-up comment in response to the system utterance. Sounding Board's strategy uses minimal explicit confirmation, based on feedback from test users that frequent confirmation can quickly become irritating. The instruction speech acts are help messages depending on the dialogue state and error detection.

2.3 Natural language generation

The Sounding Board NLG module takes as input the speech acts and content provided by the DM module, and constructs a response by generating and then combining the specified response components. The response can contain up to three speech acts from the four broad categories: grounding, inform, request, and instruction. As required by the Amazon TTS API, the response is split into two parts: *message* and *reprompt*. The device always reads the *message*; the *reprompt* is optionally used when the device "hears" nothing from the user for a given duration. The grounding act is usually the beginning of the response, and instructions are usually placed in the *reprompt*.

The grounding acts are generated by randomly choosing from collections of transition phrases/sentences associated with the specific category. Examples include: back-channelling (e.g., "I see.", "Cool."), user request echoing (e.g., "Looks like you want to talk about news.", "I heard you ask, where is the University of Washington."), misunderstanding apology (e.g., "Sorry, I'm having trouble understanding what you said."), unanswerable user follow-up questions (e.g. "I'm sorry. I don't remember the details."), and gratitude (e.g., "I'm happy you like it."). The inform acts are generated using simple templates that combine a randomly chosen introductory phrase (e.g. "Someone on Reddit said" or "My friend in the cloud told me that") with content provided by the DM module. The request acts are templates for requesting input from the user with slot-level variation that again is chosen randomly. The instruction acts are comprised of a collection of context-sensitive help messages that have minimal variation.

We make extensive use of ASK SSML for prosody and pronunciation to better convey the information our bot wishes to communicate. We use it to improve 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.

Finally, the constructed response goes through an utterance purifier which replaces profanity words/phrases with a non-offensive word chosen randomly from a list of innocuous nouns. The purifier is needed since the constructed response may contain part of the recognized user utterance and contents retrieved from online sources, either of which may include profanity. Some of the word replacements have an amusing result.

3 Miniskills

Sounding Board is equipped with several different *miniskills*, each of which manages a set of dialogue states and is responsible for conversation segment coherence. This section describes the three types of miniskills used in Sounding Board: content-oriented miniskills, a personality quiz, and general miniskills.

3.1 Content-oriented miniskills

Content acquisition and management are two important steps towards implementing a successful content-oriented miniskill. Sounding Board acquires content from multiple sources, including Amazon-provided trending topics, online user-generated content from Reddit,[5] news articles from

[5]Reddit posts come with user votes, which we use to identify content that is of higher interests.

the Washington Post, and question answers and jokes from Evi. We implement several different content-oriented miniskills, including trending topics, facts, opinions, general news, sport news, jokes, and question answers. The specific sources are chosen because they provided news or commentary of broad interest and in a style that was reasonably well suited to spoken conversations. In addition, since individual exchanges need to be relatively short, we choose sources for which it is easy to extract snippets of information that are informative and require little context to understand. The text extracted is filtered to remove content with profanity and content covering controversial or offensive topics. Simple text normalization post-processing is used to ensure that the content was TTS-friendly (e.g., urls are avoided).

Trending Topics: This miniskill recommends topics for the user to choose from. Trending topics are provided by Amazon. We categorize them using pre-computed topic-personality associations from [10], as described next in §3.2. At a single topic suggestion turn, two topics are selected based on the topic suggestion history, content availability, and personality assessment results if available.

Facts: We crawl a large collection of interesting facts from the `TodayILearned` subreddit. Most posts in this subreddit have an informative and well-formatted title. We index these post titles by all possible topics appearing in the title. The August of Sounding Board was equipped with more than 60K entries covering a wide range of topics.

Opinions: Opinion pieces are crawled from the `ChangeMyView` subreddit, where the post titles are usually arguable statements. Similar to the `TodayILearned` subreddit, post titles in `ChangeMyView` also have a structure that is easy to parse. In August, Sounding Board had obtained around 5K opinion pieces. Considering that the opinion pieces are usually controversial, rather than making the opinion sound like from the socialbot itself, we explicitly tell the user that this is a Reddit post, using a template like "I'm curious what you would say about this Reddit post. *postTitle*.".

General News: We crawl general news from the `UpliftingNews` subreddit in hopes of retrieving more positive content than standard news outlets, though not all posts in `UpliftingNews` sound positive to everyone. Effectively presenting a news article to users in a conversation is challenging. Unlike post titles in `TodayILearned` and `ChangeMyView`, news titles can be less informative (to encourage users to read the full article), but a complete news article is too lengthy for a single turn. Sounding Board presents a general news article to users in a two-turn fashion, first reading the news title and asking whether the user is interested in a summary of the news. Upon the user's approval, Sounding Board reads a summary (up to 4 sentences) in the following turn. The summary is obtained using the unsupervised TextRank algorithm [11, 12] implemented in Gensim [13].

Sports News: Sports are a big part of many people's lives, and users want the socialbot to be able to talk meaningfully about the latest developments. Since this is a time-sensitive subject, we scrape recent sports events using the Washington Post API. However, reciting a sports article is too lengthy to be acceptable in conversation. We find that the headline combined with the "blurb" given in the Washington Post articles provide good coverage, without being too verbose for conversation. However, many of these excerpts are missing some context, so we manually annotated a hundred such excerpts for coherence. This enabled us to develop a filter with greater than 90% accuracy.

Jokes: Sounding Board requests jokes from Evi. We insert a short break between the set-up and the punchline using SSML to improve the joke delivery.

Question Answering: We redirect most questions to the question answering engine – Evi. After presenting the answer we retrieved, we suggest the next miniskill based on whether it has content related to the topic(s) that appeared in the question. The act of miniskill suggestion seems to improve the user experience, especially when Evi fails to provide an answer. Common questions related to Alexa's backstory (e.g., name, birthday), sensitive topics, and those seeking advice (e.g., financial, legal) are handled by Sounding Board with a deflection strategy. For these types of questions, Sounding Board follows up by offering trending topics as the next miniskill.

3.2 Personality Assessment

Keeping different types of users engaged benefits from knowing something about them. A short sequence of questions is used to categorize people into 4 personality quadrants. We use the "Extraversion" and "Openness" dimensions from the Big 5 personality model [14], and take the corresponding

Table 2: Personality type to Disney character mappings.

Openness	Extraversion	Character
−	−	Aladdin, Snow White
−	+	Kristoff (*Frozen*), Marlin (*Nemo*)
+	−	Elsa (*Frozen*), Belle (*Beauty and the Beast*)
+	+	Ariel (*The Little Mermaid*), Anna (*Frozen*)

questions from the mini-IPIP scale [15]. Extroversion aims to capture how talkative and social a person is, and openness relates to how intellectually and artistically curious a person is.

We interleave personality questions with hand-made "goofy" questions to keep the conversation engaging, and give users the option to get their personality results after five questions (or keep going until all 8 personality questions have been asked). For each question answered by the user, we provide our own response before asking the next question. We find through test user feedback that having our own responses greatly enhances the user experience.

Each personality question loads either positively or negatively onto their respective dimension, which we use to score users to their personality quadrants.[6] We then assign the user to a Disney character (using mappings described in Table 2).

Users who answer the personality questions will subsequently get topics tailored to their personality based on a what each personality type is likely to be interested in [10]. We further explore how different personalities respond to our bot in §4.3.

3.3 General miniskills

Greeting: This miniskill is used at the opening and is only used once in a conversation. It initiates the conversation with a how-are-you question and empathizes with the user's answer accordingly.

Menu: This miniskill introduces Sounding Board's functions to the user. It is invoked when the system cannot decide the next content-oriented miniskill for pushing the conversation forward.

Exit: This miniskill instructs the user to *explicitly* say "stop" to exit Sounding Board. Depending on the invocation condition, Sounding Board adds different grounding speech acts before the instruction. When the miniskill is invoked because the user issues a common Alexa command, the added speech act explains the limitation of the system. When it is invoked because the user makes an implicit stop command (e.g., "good night", "I need to leave now"), the added speech act thanks the user for chatting.

4 Rating insights

As we develop Sounding Board, we want to shed light on what users like about our approach. We extract features at the conversation level and use the user ratings to shed light into what makes our system better. Although our system is evolving as it is being rated, we aggregate insight over conversations from the entire semi-finals competition (2017-07-01 to 2017-08-15).[7]

4.1 Miniskills performance

We wish to understand how each miniskill contributes to a user's rating of the conversation. We correlate the percentage of utterances made by our agent in each of our miniskills with the end user rating (Table 3). Jokes are associated with higher user ratings, as expected, and the personality quiz miniskill has the highest correlation with better ratings. We hypothesize that users enjoy the playful aspects of those miniskills, but the more agent-driven nature of these conversation modes could also play a role. Most other miniskills are not significantly correlated with user ratings, likely the nature of the content presented is more indicative of enjoyment than the miniskill itself.

[6] We acknowledge that this is an over-simplification of a user's personality, but the binary format was better suited for a socialbot.

[7] Some miniskills were introduced 2017-08-01, so their analyses only cover conversations after that day.

Table 3: Correlation between ratings and percentage of conversation spent in particular miniskills. Number of conversations differ based on when the miniskill was rolled out. $^{*}: p < .05, ^{**}: p < .001$, where p-values were Holm [16] corrected for multiple comparisons.

Miniskill	Pearson r	# conv.
Trending topics	-0.071^{**}	9820
Opinions	*not sig.*	2291
Sports news	*not sig.*	2291
Facts	*not sig.*	9820
General news	*not sig.*	2291
Jokes	0.077^{**}	9820
Personality quiz	0.123^{**}	9820

Table 4: Personality types associations with conversation metrics. Note that while most users end up classified as extroverted or open, 40.6% are one but not the other. Reported Pearson r coefficients are all significant: $^{*}: p < .05, ^{**}: p < .001$.

	Openness	Extraversion
% users	89.2%	61.8%
conversation rating	-0.037^{*}	0.089^{**}
# talk turns	*not sig.*	0.030^{*}
avg. utterance length	*not sig.*	0.036^{*}

We hypothesized that having more miniskills, and thus a greater variety of content, would increase user satisfaction. This motivated us to add more functionalities to our bot, rolled out on 08-01, which led to a significant increase in user ratings, i.e., from 3.24 during [07-22, 07-31] to 3.37 during [08-01, 08-05], $t(2892) = -2.3342, p < 0.02$.

4.2 Impact of ASR

Amazon's ASR was updated on 2017-07-20, which we anticipated would lead to improved ratings for all systems. In order to decouple the impact of ASR changes from subsequent changes to our bot, we ran two analyses for the following time ranges: [07-01, 07-19] and [07-22, 07-31]. User rating scores from before the ASR update ($M = 3.25$, $SD = 1.34$) and after the update ($M = 3.245$, $SD = 1.41$) did not differ significantly; $t(7036) = 0.283$, $p = 0.77$. We also looked at correlation of ASR confidences with user ratings. Specifically, we used the average length-normalized confidence of the top ASR hypotheses, computed as the product of the token-level confidences. For both time-periods, higher confidence scores correlate with higher user ratings ($p < 0.005$), but the correlation in the period after the ASR changes is higher (0.085 vs. 0.042). If confidence is a good proxy for accuracy, this would indicate that higher ASR accuracy does lead to higher user ratings, but the effect is small.

4.3 Personality insights

We investigate how different personality types respond to our system. Specifically, we ask whether particular personality types rate our system higher, talk to the system for longer, or speak longer utterances (results in Table 4). Note that we use results from the first quiz if users took the quiz multiple times (on average, the quiz was taken 1.26 times per conversation).

We find that extroverted users (according to our quiz) tend to like our bot better, whereas openness slightly correlates with lower ratings. These findings hold when controlling for number of turns in the conversation, a variable itself correlated with higher ratings ($r = 0.139$, $p < .001$).

From our analyses, it seems extraverted users tend to speak more at each talk turn (avg. utterance length). This is a nice validation of social psychology theory, which directly associates extraversion with talkativeness [14]. Interestingly, extraversion is only slightly correlated with the length of the conversation (# talk turns).

We take these insights with a grain of salt, as the mini-IPIP personality scale has imperfect reliability [15]. Additionally, we have no guarantee that users are being truthful when answering questions.

5 Related Work

There has been a substantial amount of work on conversational dialog systems that roughly falls into two groups, depending on whether they address a problem involving open-domain chit-chat vs. goal-oriented dialog. Open-domain chit-chat systems aim at generating responses that are contextually appropriate, but they are generally content free. Many variants of the sequence-to-sequence model have been proposed for this task [3, 4, 5, 6, 17, 18, 19], and experiments leverage data sources such as Twitter, movie scripts, Reddit, etc. Most goal-oriented systems involve constrained-domain information seeking tasks, e.g., restaurant information or question answering [20, 21, 22, 23] . These systems generally rely on obtaining information from a knowledge base, in which case response generation is often conditioned on a semantic frame. (An exception is the help desk task, where the "knowledge" is represented in terms of prior conversational interactions, as in the Ubuntu chat corpus [24] for which sequence-to sequence models are often used.) Recent studies with both types of systems often leverage reinforcement learning for training the dialog policy or the overall end-to-end system.

Sounding Board addresses a problem that has aspects of both tasks. It assumes that the user is interested in information but does not generally have specific questions that they want answered, and some chit-chat is useful to manage the open-ended nature of the conversation. Thus, the objective of generation is that the response be informative and sensitive to user interests, as well as appropriate for the context. Because Sounding Board aims to present novel and evolving content, existing corpora for sequence-to-sequence modeling are not useful, and because the user seeks information broadly, there is no single right answer that can be used as a reference for supervised training. Thus, our generation strategy cannot benefit from much of the prior work and instead emphasizes selection of speech acts and presentation of content with a conversational style.

An important aspect of Sounding Board is a focus on user personality and satisfaction. Generation work has addressed persona of the agent [25, 26], but there has been less attention to the user. Some work has looked at user engagement [27, 28, 29, 30], which is related but somewhat different from our notion of user satisfaction, and often relies on audio/video cues not current available from Alexa. Other relevant is work in [31], which showed that user satisfaction can be predicted from dialog features. Like prior work, Sounding Board leverages user attitude/satisfaction in the dialog policy, but uses a multi-dimensional representation. Sounding Board identifies user personality directly through explicit questions (vs. indirectly through their speaking style), and uses that information to drive content presentation. We do not vary the persona of the bot.

6 Conclusion

Sounding Board approaches the Alexa Prize challenge as a content-oriented dialogue system. In this paper, we describe the system architecture and the design philosophy of Sounding Board. Sounding Board engages users through content-oriented conversation segments, each of which is managed by a specific miniskill. A user-driven dialogue policy is used, attempting to adjust topic choice and interaction strategy according to the user mental state and personality. The hierarchical dialog management architecture facilitates adding and updating capabilities, and the number of mini-skills continues to expand. We carry out analyses on user ratings and discuss their correlation with miniskill variety, user personality, and ASR performance.

There are several areas where the system could be improved. Most notably, we have not used data-driven strategy for learning the dialog policy due to the emphasis on first adding capabilities to the system. With the new capabilities added and associated user interactions, it would be interesting to use reinforcement learning to improve mini-skill selection and content planning aspects of the dialog policy. For pursuing deeper conversations, we need to implement coreference and language understanding of the content sources. We would also like to make use of the personality and satisfaction history of the user in ranking miniskills. Finally, it is likely that the types of error states will change as the user-bot interactions become more natural, necessitating different approaches to error handling and response generation.

Acknowledgements

In addition to Amazon's financial and cloud computing support, this work was supported in part by the National Science Foundation via the Graduate Research Fellowship awarded to Elizabeth Clark, and in part by DARPA grant "Communicating Intelligently with Computers" (Award #: 67102239 Amd 1 PTE: W911NF-15-1-0543) which funded some of Maarten Sap and Ari Holtzman's time. The conclusions and findings are those of the authors and do not necessarily reflect the views of sponsors.

References

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

[2] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proc. NAACL*, pages 196–205, 2015.

[3] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proc. ACL*, pages 1577–1586, 2015.

[4] Oriol Vinyals and Quoc V. Le. A neural conversational model. In *Proc. ICML*, 2015.

[5] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI*, 2016.

[6] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proc. NAACL-HLT*, 2016.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[8] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL: System Demonstrations*, pages 55–60, 2014.

[9] Maria Stubbe. Are you listening? cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3):257–289, 1998.

[10] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[11] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Proc. EMNLP*, pages 404–411, 2004.

[12] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of TextRank for automated summarization. *arXiv:1602.03606 [cs.CL]*, 2016.

[13] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proc. LREC Workshop New Challenges for NLP Frameworks*, pages 45–50, 2010.

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

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

[16] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[17] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proc. ACL*, page 2431–2441, 2016.

[18] Jiwei Li, Will Monroe, Alan Ritter, Jianfeng Gao Michel Galley, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv:1606.01541 [cs.CL]*, 2017.

[19] Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proc. EMNLP*, 2017.

[20] Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proc. ACL*, 2015.

[21] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proc. EMNLP*, 2015.

[22] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proc. EACL*, 2017.

[23] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proc. ACL*, 2017.

[24] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proc. SIGDIAL*, 2015.

[25] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proc. ACL*, 2016.

[26] Marilyn A. Walker, Ricky Grant, Jennifer Sawyer, Grace I. Lin, Noah Wardrip-Fruin, and Michael Buell. Perceived or not perceived: Film character models for expressive nlg. In *Proc. International Conference on Interactive Digital Storytelling*, pages 109–121, 2011.

[27] Dan Bohus and Eric Horvitz. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proc. SIGDIAL*, 2015.

[28] Kate Forbes-Riley and Diane Litman. Adapting to multiple affective states in spoken dialogue. In *Proc. SIGDIAL*, pages 217–226, 2012.

[29] Zhou Yu, Alexandros Pangelis, and Alexander Rudnicky. TickTock: a non-goal-oriented multimodal dialog system with engagement awareness. In *Proc. AAAI Symposium*, 2015.

[30] Zhou Yu, Dan Bohus, and Eric Horvitz. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *Proc. SIGDIAL*, 2015.

[31] Syaheerah Lebai Lutfi, Fernando Fernández-Martínez, Juan Manuel Lucas-Cuesta, Lorena López-Lebón, and Juan Manuel Montero. A satisfaction-based model for affect recognition from conversational features in spoken dialog systems. *Speech Communication*, 55(7):825 – 840, 2013.