Conversational Agent Experience: How to Create Good Alexa Skill

by

Xunan Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Information
(School of Information)
in The University of Michigan
2020

Thesis Committee:

Eytan Adar, Chair
Assistant Professor Steve Oney

For all the people

ACKNOWLEDGEMENTS

Thanks to my family, especially my grandparents Lingen Ge and Liqing Jiang who consistently support me to chase my dream. I also want to express my gratitude to my advisor Eytan Adar who helps me to make this study happen.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

<u>**Table**</u>

LIST OF ABBREVIATIONS

CA Conversational Agent

MTurk Amazon Mechanical Turk Workers

RER Response Error Rate

GTTS Google Text To Speech

Pyttsx Python Text to Speech Synthesis

ABSTRACT

Conversational Agent Experience: How to Create Good Alexa Skill

by

Xunan Zhou

Chair: Eytan Adar

Conversational Design guidelines offer recommendations on how to lead the user-agent conversation, how to help customers achieve their goals, and how to handle the mistakes caused by each side. However, the effective methodology to evaluate the experience of user-agent conversation is unclear. Here we show a data pipeline that evaluates the user-agent experience on a variety of scenarios. We found that the coherence of Alexa's response has a positive impact on user's experience, which is based on the categories of skills, the number of slots in utterances, and the goals that users are trying to achieve. Furthermore, our study shows a gap between the theoretical conversational design guideline and the needs for practical testing for Conversational Agent (CA). Our data pipeline demonstrates the importance of testing experience by measurements that cast positive or negative affect on conversational experience. We anticipate our study to be a starting point for a more robust user experience evaluating system for CA and related applications.

CHAPTER I

Introduction

Our study aspires to find an effective supplement to Alexa Design guidelines. We are interested in how skills can be designed with better user experience, and what metrics, which are practical and convenient, should be employed for measurement.

We see the increasing volume of shipping of CA. There is a 44.9% annual growth of smart speaker shipments in 2019 across the world, in which Amazon has been pulling ahead of the pack by shipping 10.4 million smart speakers after merging from a Prime Day Performance in Q3 2019. However, the daily usage of the conversational agents is surprisingly low. A study from the *Deloitte* (2016) presented that Smart speakers are mainly used for playing music and listening to weather updates. This contradiction suggests that CAs are not not driving user engagement and could be improved.

Amazon tried multiple ways to improve their CA's user experience, like conducting Alexa Prize Award and proposing *Guideline* (2020). There are 50 very detailed recommendations with examples for developers when designing an Alexa skill. However, except paraphrasing the responses a bit to users' vocal requests (utterances), Alexa skills are unable to follow most of the other guidelines. Moreover, the skills with more sophisticated needs of dialogues like natural language, the poor customer satisfaction they have.

Different from the specific guidelines presented by *Guideline* (2020), we developed a data pipeline to quantitatively test skills' user experience by different metrics. We are inspired by existing metrics used for cognitive models such as Flexibility and Consistency (chapter IV). We also referred to the metrics used to evaluate CA by the *Radziwill and Benton* (2017). By increasing the granularity of user-CA dialogues, we are able to perform correlation testing between the skill design and associate user experience.

We anticipate our study can facilitate the Alexa support team and Alexa skill developers. Our data pipeline could be further developed into in-place user experience testing toolkits taking designed utterances and responses that promptly return feedback on potential usability issues.

We tried to evaluate the skills by different metrics (Chapter IV) such as Personalization, Flexibility, Coherence, etc. We found some properties of a skill, such as Flexibility and Personalization , don't significantly influence user experience. However, customer satisfactions are influenced by the coherence of skills. Coherence is defined by *Radziwill and Benton* (2017) that how relevant a response is to an utterance. Moreover, coherence levels are affected by the number of slots of skill utterances and different types of conversational interfaces of skills. We expect that Alexa skills would provide better user experience if developers design skills considering the common use cases for associate conversational interface and the feedback from our proposed usability testing.

CHAPTER II

Problem Definition and Data

2.1 Problem Definition

The primary goal of our study is to find what quantitative metrics should be used for tracking user experience, i.e. which metrics of conversational interface bring user positive experience (We will elaborate the classifications of conversational interface in Section Methodology). Next, we will build an evaluation system that is able to quantitatively measure the user experience as a supplement to design guidelines.

We classified the Alexa skills interface-wise. For example, one-shot skill, such as Sleep Sound, plays peaceful music when users say "Alexa, play Sleep Sound". Conversational-Service Skill, such as Uber, can order a ride for customers via multiple rounds of dialogues. Although there are a number of factors that affect users' experience such as price of purchase, convenience of installment, stability of networks, entertainment, etc, we focused on the conversational interface per se. In the methodology section, we will elaborate more about five conversational interface categories we defined.

2.2 Data Sourcing

We crawled data from three different perspectives. First, we sampled skills randomly from Amazon.com which have at least 30 user comments. Second, we designed experiments to simulate dialogues between users and Alexa using Python. Third, we utilized crowd-sourcing workers to annotate the conversations by metrics of user experience discussed by *Ram et al.* (2018)

We sampled 200 Alexa skills from Amazon.com that have more than 30 customer reviews randomly. After filtering out those skills are one-shot, need hardware device connection or payment information. We have 38 valid skills to conduct user-agent conversations. Also, we crawled 3,800 customer reviews and user ratings for these skills.

Moreover, we designed a series of experiments to simulate the dialogues with Alexa using Python. We simulated 3 sets of experiments. First, we simulated the dialogues with Alexa about the primary intents such as Open Intent, Help Intent, Stop Intent. Second, we simulated the dialogues by uttering vocal requests with one slot of keywords. Third, we uttered with two slots of keywords. Overall, we got around 7600 records of responses of Alexa skills.

Finally, we used the Collective Intelligence Amazon Mechanical Turk Workers (MTurk) to annotate the dialogues between users and Alexa skills by metrics proposed by Cognitive Model and *Venkatesh et al.* (2018). Each record of dialogues is rated by five MTurk workers to minimize bias. We fetched 22,800 data points of coherence evaluation.

CHAPTER III

Related Work

In order to create better CA experience, the CA development teams need to go through an entire process of design, development, and testing. Many efforts are done in the first two phases, such as improving the accuracy of voice recognition, designing more intuitive prompts, or optimizing the coherence of CA's responses, while an objective and reusable usability testing framework for CA is still absent. There are three distinctive approaches that researchers are studying to test the experience of CA. First, testing CA that whether they are following the design guideline such as *Guideline* (2020). Second, testing CA by whether they could meet certain linguistic metrics such as coherence and Response Error Rate (RER). Third, testing CA by conducting analysis on customer reviews and other qualitative data.

Guideline (2020) proposed more than 50 guidelines for developers to design better Alexa skills. There are 11 guidelines for Adaptiveness (how smart a skill is), 15 guidelines for Availability (how considerate a skill is), 9 guidelines for Customization (how personalized a skill is), 15 guidelines for Relativeness (how natural a skill is). However, the design guideline does not provide hand-on instructions for different types of skills because the flow and complexity of conversations differ from one another. The specific examples in Design Guidelines unable to be massively applied by developers when designing utterances of skills or iterating skills for better user experience. For example, the *Guideline* (2020) suggests that a skill should be able to handle over answering of users, which means the skill could understand the exceeded information provided by users. In reality, the RER of replying to a user's utterances significantly increases by 20% on average when user queries a skill with more slot of keyword.

Han and Yeh (2019) evaluated voice skills by design guidelines using an automatic voice crawler. They picked 8 guidelines from those proposed by Amazon and tested how Alexa skills responses comply with them. This study introduced an automatic approach to massively utter vocal intents to Alexa skills, while the study is limited by the simplicity of intent. For example, they evaluated Alexa by using variants of open intent, such as "Alexa, open Ted Talk and play talk about nature topic" and "Alexa, ask Ted Talk to play the topic of nature".

Venkatesh et al. (2018) systematically evaluated the importance of various measurements for CA. They found coherence and RER (Response Error Rate) have a significant influence on user rating, which are in line with our study. But their study did not cover the conversational interface between Alexa skills and users, and how to design a better skills, or applications, for CA is still unclear.

Kinsella (2018) quantitatively analyzed the customer reviews and illustrated the gap between the reviews and Alexa skills usage. The Alexa skills, different from the usual product on Amazon.com, not necessarily need to be reviewed by users before enabled. Even though the Alexa skill with higher reviews is more likely to be used and enabled by users, there is a good number of skills with zero reviews activated by users too. Moreover, user interacting with the skills is different from the commercial product on the website i.e., voice, which increase the difficulty for research to get access to users' opinions.

Moreover, there are other factors that affect the fidelity of testing. We listed several studies that trying to mitigate the external factors that could affect usability testing.

A conversational interface is influenced by the noise along with users' utterance. *Feng et al.* (2017) proposed Vauth, a system that provides continuous authentication for voice assistant and guarantees the voice assistant to execute only the commands that originate from the voice of the owner.

Distinctive issues across different domains are challenging for CA testing. *Goh et al.* (2007) discusses that testing the domain-specific CAs are becoming a real issue, illustrating the inappropriateness of the existing measures for response quality evaluation and the call for new standard measures and related considerations. *Tsai et al.* (2018) summarized and analyzed the classification of Alexa commands by analyzing the history of 82 users with totaling 193,664 commands. This work creates a categorization for the type of tasks and commands asked of Alexa, with ten categories. They also standardized the assignments and command classification, which may benefit the community for use as a springboard for both feature development across different domains.

The axis of time and physical placement also has influence on CA user experience. *Sciuto et al.* (2018) presents how households integrate Alexa into their lives, behaviors around purchasing and acclimating to Alexa, in the number and physical placement of devices, and in daily use patterns. They also uniformly described interactions between children and Alexa.

CHAPTER IV

Methodology

We took multiple data sources to build the CA evaluation data pipeline. First, we sampled 200 skills from Amazon randomly and kept 38 qualified skills for testing after skill selection (Chapter 4.1). These skills are with more than 30 customer reviews and ratings, and they could have multiple round of dialogues along with the interaction. Second, we crawled customer reviews and ratings of skills. Third, we generated and designed intent and utterances for skills based on their information provided on landing page (Table 4.2, Utterance Design).

As for the CA dialogue crawling, we created an automatic voice crawler that records Alexa responses to our designed utterances (Chapter 4.4 Conversation Crawling). We tested the responses of skills by different metrics (Chapter 4.5 Experience Metrics) to find the most robust and effective metrics for experience testing.

In short, we sourced data from 1) Amazon skill web pages, 2) automatic audio crawler, and 3) MTurk. Also, we set up the metrics for evaluation from three perspectives: 1) the types of conversational interface, 2) the variety of utterances, and 3) the coherence of skill responses graded by MTurk.

4.1 Skill Selection

Classifying skills by conversational interface allow us to study the user experiences of skills from the Human Computer Interaction perspective. Alexa skills are originally grouped by topics on Amazon web pages, while our approaches reorganized the classification of skills in order to keep an eye towards conversational interface. For example, Hardcore History and Sleep Sound are both labeled as 'Music & Audio' by Amazon. Among this topic, Hardcore History is a skill that answer user's questions about history via a variety of utterances (we classified it as Media-player skill), while Sleep Sound play peaceful music by user's Open Intent utterance ('Alexa, open Sleep Sound'), so we classify it as One-shot skill.

Therefore, we categorized the skills by conversational interface, each category has its distinctive manner of dialogues with users to achieve tasks. We picked skills of Media Player and Conversational Service because they can give us the most profound data points for analysis.

Because we envision that CAs are designed to interface with users for a variety of goals via natural language conversation, we didn't track the skills of One-shot conversational interface because they do not provide sufficient data. Likewise, we did not study Flash Briefings skills because they are a special kind of one-shot skill that closely work with Alexa, which also don't provide profound data. For example, the day in history top story follows a similar fashion as one-shot skills and doesn't involve further dialogues with users. We didn't research Game skills because this kind of skills involve too many possibilities of conversation flow, which is too hard to be massively crawled within our study time frame.

Type of Skill	Description	Sample Skill	Sample Conversation
One-shot	A kind of skill that is designed specifically for certain tasks	4AFart	**User:** "Alexa, Ask For A Fart" **Alexa** plays a fart sound
Media Player	A kind of skill that plays podcast, lecture, episodes, or instructions per users' queries or vocal commands	Ted Talks	**User:** "Alexa, ask TED Talks to find talks about nature" **Alexa:** "Here is 'How to grow a forest in your backyard' by Shubhendu Sharma, around 20 minutes, to skip to the next talk; say 'Next'" **Alexa** plays the Ted Talk lecture

Conversational Service	A kind of skill that can accomplish more than one task	Best Buy	**User:** ”Alexa, open Best Buy” **Alexa:** “Welcome to Best Buy, we provide safe and convenient service either you visit in-store, pick-up or delivery to your home. How can I help you today?” **User:** ”Search for iPad” **Alexa:** ”Sure, here are the results for iPad. Results on the top of the list for iPad: Apple iPad...would you like to hear more about comments or next result?” **User:** ”Read the comments” **Alexa:** ”[Reads the comment] Do you want to hear more or do you want to search for another item?” Conversation keeps moving forward
Flash Briefing	A kind of built-in skills that provide information updates such as News or lectures	CNN	**User:** ”Alexa, open CNN” **Alexa:** ”Welcome back to CNN News” **Alexa** plays news/updates

Game	A type of skill that involve one or multiple players to play a game via conversational interactions	Jeopardy	**User:** "Alexa, open Jeopardy" **Alexa** "[Music] This is Jeopardy! This is the host Alex Trebek. Haven't seen you for a while, thank you for coming back...the first category is American History...[Question]" **User:** "What's the color?" **Alexa** "That's not correct, the correct response is [Answer]" Jeopardy keeps moving on

Table 4.1: Categorize skills by conversational interface
i.e., by how users interface skills

Media Player skills provide audio content by allowing users to query the podcast they are interested in; Ted Talk is one of the representative skills. They basically work as a radio yet covers a fixed range of topics. Conversational Service skills, like the apps built on Alexa system, providing similar features as the counterparts on web or mobile platform. For example, there are Uber, Best Buy, Dominos, etc. These two classes of skills, which are capable of achieving a variety of tasks for users and conducting extensive conversations with users. They are relatively closer to the generic imagination of what a CA could do.

4.2 Customer Reviews and Ratings

Although the customer rating depends on various factors, for the sake of time resources, we used it as a proxy of overall feedback to Alexa skills. We found skills with easier usage are more popular considering customer reviews and ratings. *Deloitte* (2016) presented that the top three use cases of CA is for playing music, weather updates, and setting alarms. Those skills used for playing music, flash briefings, and other audios such as 'funny sound' kind of audio receive more reviews and positive ratings than those skills involve multiple rounds of dialogues with users. Based on our skill sampling, the median customer rating for skills of flash briefing, game, and one-shot, is 4.1, 4.0, and 3.92 respectively. Moreover, the customer reviews of game skills and one-shot skills are more popular than other types of skills (Conversational Service, Game, Flash briefing).

4.3 Utterance Design

Our study presents the relationship between the variety of an utterance and the user experience of CA. We extend the variety of utterance by changing the number of **Slot** of keyword within it and creating synonymous utterances.

As for natural language speaking, we communicate information via various utterances in which have a number of keywords. While interfacing with Alexa skills, a user has to communicate keywords, which is called **Slot**, in the manner of pre-designed utterances. Poorly-designed utterances for skills could cause that users are barely understood by Alexa and bad user experience. For example, users' intuition of utterance speaking is different from what is designed by developers.

	1 slot	2 slots	Multiple slots

Utterance Alternative #1	What's the store hours of [city_name]? **Example:** What's the store hours of Ann Arbor?	How about going to [city_name] in [season]? **Example:** How about going to Shanghai in Fall?	When will it [weather_condition] [day] in [city_name]? **Example:** When will it rain tomorrow in Ann Arbor?
Utterance Alternative #2	What's the store hours of [zip_code]? **Example:** What's the store hours of 48104?	How much does it cost to stay in [city_name] this [season]? **Example:** How much does it cost to stay in Shanghai this summer?	Would [city_name] [weather_condition] [day]? **Example:** Would Seattle rain the day after tomorrow?
Utterance Alternative #3	Search for store hours in [city_name]? **Example:** Search for store hours in San Francisco?	Plan a trip to [city_name] this [season]? **Example:** Plan a trip to Shanghai this summer?	What's the weather forecast of [day] [daytime] in city? **Example:** What's the weather forecast of tomorrow afternoon in Detroit?

Table 4.2: Utterances of skills with different number of slots

We designed utterances for the most typical intent of a skill based on its affiliated introduction. For example, Expedia skill primarily responds to users with the hotel and flight information of a destination. Then we designed utterances with different number of slots. (1 slot, 2 slots, and slots in multiple rounds). Slot is defined as a place holder in an utterance where the keyword resides. Speaking natural language in reality, it's normal to utter more than 5 keywords in a sentence. However, most of the skills don't support utterances with more than two slots and can barely follow up with deep conversations. It's frustrating for users to make efforts to utter utterances to Alexa that are different from natural language. While users are willing to learn the usage of a skill by uttering Help Intent, *Han and Yeh* (2019)'s research showed 28.72% of the skills would memorize previous support, the tedious duplication of prompts just make the experience worse.

We also create the synonymous utterances for an intent by paraphrasing. It's because, first, according to the Alexa Design Guidelines, a skill should match a variety of utterances to an intent. Second, being inspired by the cognitive model of programming languages, we are interested in the consistency of the conversational interfaces; how the rest utterances of an intent of a skill can be inferred when some of the utterances are learned by users. We set up a library for 32 common types of slots, such as week day, city, and genre, with on average 5 different options. We randomly selected slots for paraphrased utterances to extend variety.

Slot Type	Example 1	Example 2	Example 3	Example 4	Example 5
podcast	Doctor Laura program	The Joe Rogan Experience	The Daily	New York times The Daily	The Rachel Maddow Show
weather	sunny	rainy	snow	windy	
weather condition	wind speed	humidity			

room	living room	kitchen	bathroom	home	bedroom
week day	Monday	Weekend	Thursday	Friday	Sunday
day	today	tomorrow	yesterday	the day after tomorrow	
part of day	afternoon	night	morning	evening	
media action	skip	next	previous	pause	resume
news headline	cornoa virus	new york city	president	Google	Italy
celebrity name	Jeff Bezos	Nicholas Negroponte	Warren Buffett	Trevor Noah	Larry page

Table 4.3: In order to extend the variety of our utterances, we developed a library to provide alternatives of slot for utterances

4.4 Conversation Crawling

In order to programatically speak to Alexa and record the conversations, we created an automatic audio crawler following *Han and Yeh* (2019)'s works using Python Text to Speech Synthesis (Pyttsx) to articulate utterances and track the dialogues using pyAudio. We tried to conduct deep conversations by leveraging Google Text To Speech (GTTS). But we did not successfully go through because of the issues of transcribing speed. Although the accuracy of transcribing is satisfying for single dialogue (the average Jaccard distance is 0.056 (best:0, worst:1) between the GTTS

Algorithm 1 Alexa Response Crawling

```
for skill = s1, s2, s3 ... do
    for intent = open, help, ..., intents do
        for time = range(2) do
            speech ← Pyttsx.generate(intent)
            play speech
            record ← Start listening
            if Silence < 7s then
                record ← keep listening
            else if Length < 20s then          ▷ Prevent skills playing forever
                record ← keep listening
            else
                save record
            stopSpeech ← Pyttsx.generate(StopIntent)          ▷ Media player
skills need to be stopped first then be terminated
            record ← listen
            save record
            stopSpeech ← Pyttsx.generate(ExitIntent)
            record ← listen
            save record
```

Figure 4.1: This script presents the workflow of our crawler; how we programmatically simulate conversations with different Alexa skills

transcript and human transcript on 30 random responses of skills), transcribing by GTTS is not fast enough that allows us to extract keywords and respond to Alexa skills for further conversation. In other words, our crawler cannot finish transcribing in time before the Alexa skill 'lost its patience' and starts repeating itself. For those skills need to be interfaced with multiple rounds of dialogue, we manually designed utterances for different dialogues throughout the conversation in advance.

4.5 Metrics of User Experience

We tried two different sets of metrics to evaluate the experience of CA. The metrics should directly relate to the CA interface-wise instead of external factors such as the content of podcast or hardware settings. For instance, in contrast to a user is satisfied with the topic of a podcast or a user is annoyed by the slow response of a skill because of the poor WiFi connection, the user is happy about the fluent experience of using the CA. For instance, the coherence directly reflects how comprehensible and relevant responses are to a user's utterance, while the experience of users connecting their personal account to a skill is not included in our study. Being inspired by the Selenium Cognitive Model, we tested two metrics, Flexibility and State of CA. Coherence and Response Error Rate introduced by *Venkatesh et al.* (2018) are also referred to evaluate the user experience.

4.5.1 Metrics for Cognitive Model

Flexibility: We defined flexibility of CA as the capability of paraphrasing responses to provide more natural dialogues. By giving utterances for skills multiple times, we crawled variants of response to a same utterance. We found most skills are able to reply users with some degree of paraphrasing.

State of a skill: A system is described as stateful if it is designed to remember preceding events or user interactions. Our study analyzed how skills continuously process preceding utterance for later use so as to improve user experience.

4.5.2 Metrics for Conversational Chat bot

Coherence: *Ram et al.* (2018) defined that a coherent response indicates a comprehensible and relevant response to a user's request. RER is used to quantitatively measure the coherence of Alexa's responses.

$$\text{RER} = \frac{Number\ of\ incoherent\ responses}{Total\ number\ of\ utterances}$$

A skill's response is incoherent if it is not logically answering user's questions. For instance, a skill reply to a user with the information of weather when the user is asking about traveling in another city. In special cases, a response that clarifies utterances such as "I did not catch that, can you say it again" are equally coherent and incoherent responses by MTurk. Based on the coherence rate annotated by MTurks, we are able to learn the RER of responses of a skill. We set the responses whose coherence rates equal to or less than 2 as incorrect, irrelevant, or inappropriate responses to users' utterances. As for those responses whose coherence rates equal to 3, they are not specifically answering user's utterance but they are to some extent moving the conversation forward. Those responses with higher coherence rate of 4 are specifically replying to users' utterances or asking clarifying questions with keywords in the context.

Metric	Definition	Example
Flexibility	The capability to provide synonymous responses to an intent	User: Ask Domino's to order a pizza. **1)** Domino's: Welcome back to Domino's. What' kind of pizza would you like today? **2)** Domino's: Thank you for using Domino's. How can I help you today?

Coherence	Ratings of responses given by MTurk workers that reflect how relevant a response is to an utterance ranging from completely coherent to completely incoherent	**Completely coherent:** User: Will it rain tomorrow in Ann Arbor? The weather channel: From 6am to 12 pm tomorrow, there is a 30 percent chance of rain in Ann Arbor. Overall, you can expect a cloudy day tomorrow. **Completely incoherent**: Users: What's the office hour of the closest Best Buy store? Best buy: On the top of result list, Apple iPad pro, black, $749, I have 20 customer reviews, do you want me read it?
Number of slots	Number of keywords that Alexa can extract from an utterance so as to fulfill an intent	1-slot utterance: I want to go to **Seattle**. 2-slot utterance: How much is it flying from **Detroit** to **Seattle**? 3-slot utterance: How much it **cost** to travel to **Seattle**, **this summer**?

Personalization	Capability of replying to users based on their personal information saved in advance	**Know personal information by linking account:** Uber: Thanks for using Uber. What's your destination? (Uber in advance linked user account for personal information so that the user doesn't need to say basic information again). **Save personal information by previous interaction:** Song Quiz: Welcome back to Song Quiz! You have enjoyed 70s Classic Rock last time, do you want to resume?
State of a skill	Statefull skills or process are those can be returned to again and again, like online banking or email. They're performed with the context of previous transition and the current interaction may be affected by what happened during previously	Kayak: It's about $230 to fly from Detroit to **Seattle**. User: How much is it to rent a car in that city? Kayak: In **Seattle**, it's around $41 per day to rent a Ford Fiesta and $53 per day to rent a Toyota Camery. Do you want to hear more information about car renting?

Table 4.4: The metrics that are used for evaluating user experience

CHAPTER V

Evaluation

It's challenging to design a universal guideline for skills regarding the fact that skills are serving various users, needs, and scenarios. There are more than 50 qualitative metrics recommended by *Guideline* (2020) considering adaptiveness, availability, personalization, reliability. Although the *Guideline* (2020) gives a number of of recommendations, the customer ratings across skills vary significantly by different conversational interfaces, which suggests that the *Guideline* (2020) is not applied to all skills.

5.1 Customer ratings and comment for skills with different conversational interfaces

Alexa users intend to engage in giving comments and higher ratings for those skills that are easier to use. On the other hand, conversational-services type of skills, such as Uber and Best buy, that provide similar features as their Mobile or Web counterparts, are less likely to engage customers.

Customer Ratings: From our study, we found that skills employ straightforward conversational interface, as in few rounds of dialogue or have very specific design of questions such as "yes or no", are more likely to get higher customer ratings (customer satisfaction). Flash briefing skills, such as New York Times, Fox News, and CNN,

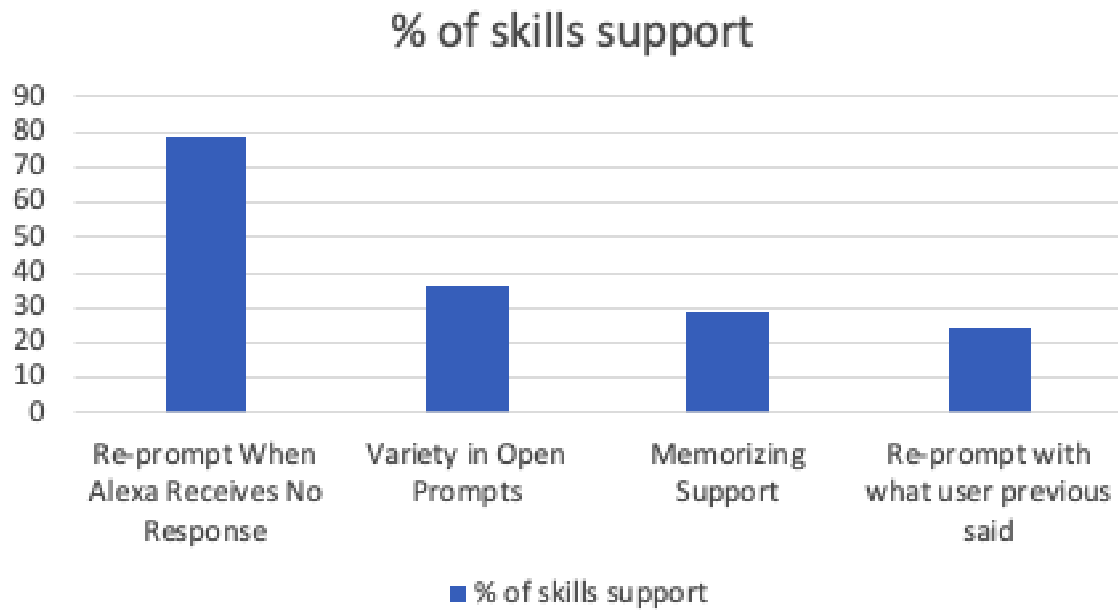

Figure 5.1: Performance of sampled Alexa skills when being benchmarked with the *Guideline* (2020). According to our study, 80% of Alexa skills are able to re-prompt if the user doesn't keep the dialogue going, while only 20% of Alexa skills re-prompt with the previous keywords uttered by users

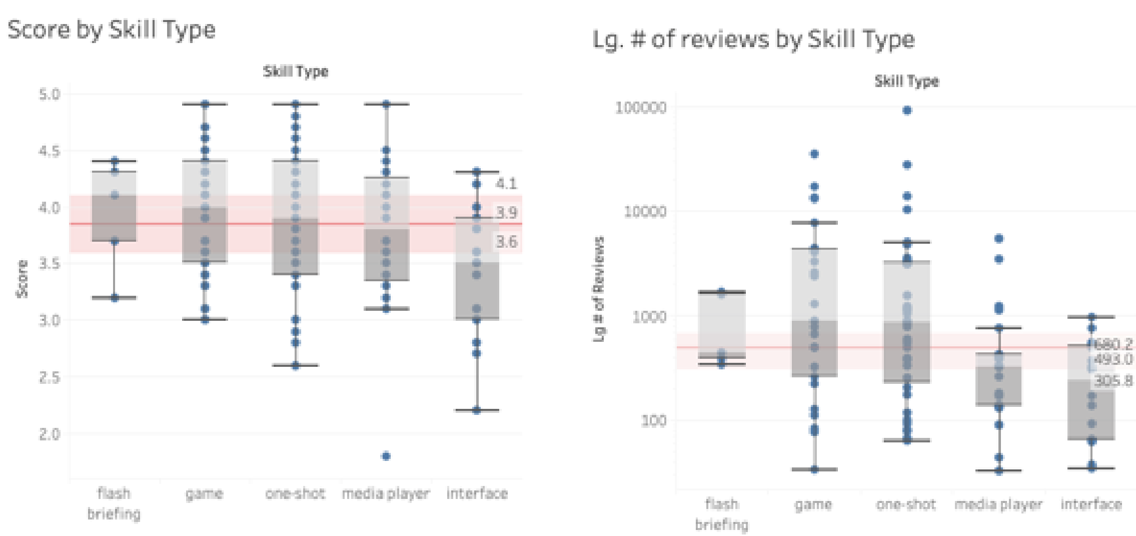

Figure 5.2: Skills of game, one shot, and flash briefings receive higher ratings. Skills of game and one-shot engage more customers in commenting.

have the highest medium customer rating of 4.1 (out of 5), followed by game skills with medium rating of 4.0. But conversational service skills, such as Domino's, Best Buy, and Uber, have the lowest medium rating as 3.6. Among our randomly selected skills, the skill Word of the Day Flash Briefing won the competition among the flash briefing skills with 4.3 starts out of 5, while All Recipes, as one of the conversational service skills which provides cooking instructions for users, got the lowest rating from customers as 2.2 out of 5.

Customer Reviews: Skills with straight forward conversational interfaces have higher customer engagement. Game skills and one-shot skills are more likely to have more customer feedback. For example, Jeopardy!, one of the most popular game skills in our study, got the highest number of customer reviews of 35,839. Interestingly, 4AFart is a one-shot skill, which only plays a fart sound with Open Intent, not only got the highest number of reviews (91,804 reviews) among one-shot skills but also among all selected skills in our study. The skill, Uber, with the highest log number of customer reviews as 758, is close to the median log number of reviews of one-shot skill (814 reviews).

5.2 Metrics of Cognitive Model

Our study demonstrated important metrics that could help developers to evaluate and iterate their skill developments. We find coherence is important for customer satisfaction and response error rate are related to some skills' customer satisfaction, while flexibility and state of skills don't have equally important influence.

Type of Response	Conversational Interface	Number of Skills
Direct response(AutoLaunch)	One-shot	18
	Media Player	4
Flexible Responses	One-Shot	15

	Media Player	6
	Conversational App	6
Inflexible Responses	One-Shot	63
	Media Player	20
	Smart Home	20
	Flash Briefings	11
	Conversational App	72

Table 5.1: Inflexibility of responses occurred more across different type of skills ranging from Alexa skills of One-shot to Conversational App.

Flexibility of Skill Responses: There are three types of response to users' utterance (vocal request). First, the skill directly proceeds by a user's open intent, such as playing a music or a podcast. Second, the skill always responds with the same content. Third, the skill would response with paraphrasing to the same intent over time. We found skills in which intent are fulfilled directly without further interaction are higher than those whose responses are either flexible (p-value < 0.05 using Student's t-test as 0.023) or inflexible (p-value < 0.05 using Student's t-test as 0.042). We also found that customer ratings for skills respond with paraphrasing (flexible responses) are not significantly higher than those who are not. Moreover, the median customer ratings for skills with flexible responses are even slightly lower than those who are not.

State of Skills: The influence of state of a skill on customer rating is similar to the flexibility of skill responses; stateful skills' customer rating isn't significantly higher than those who are stateless (p-value < 0.05 using Student's t-test as 0.151).

These results could attribute to that the more fixed a conversational interface is,

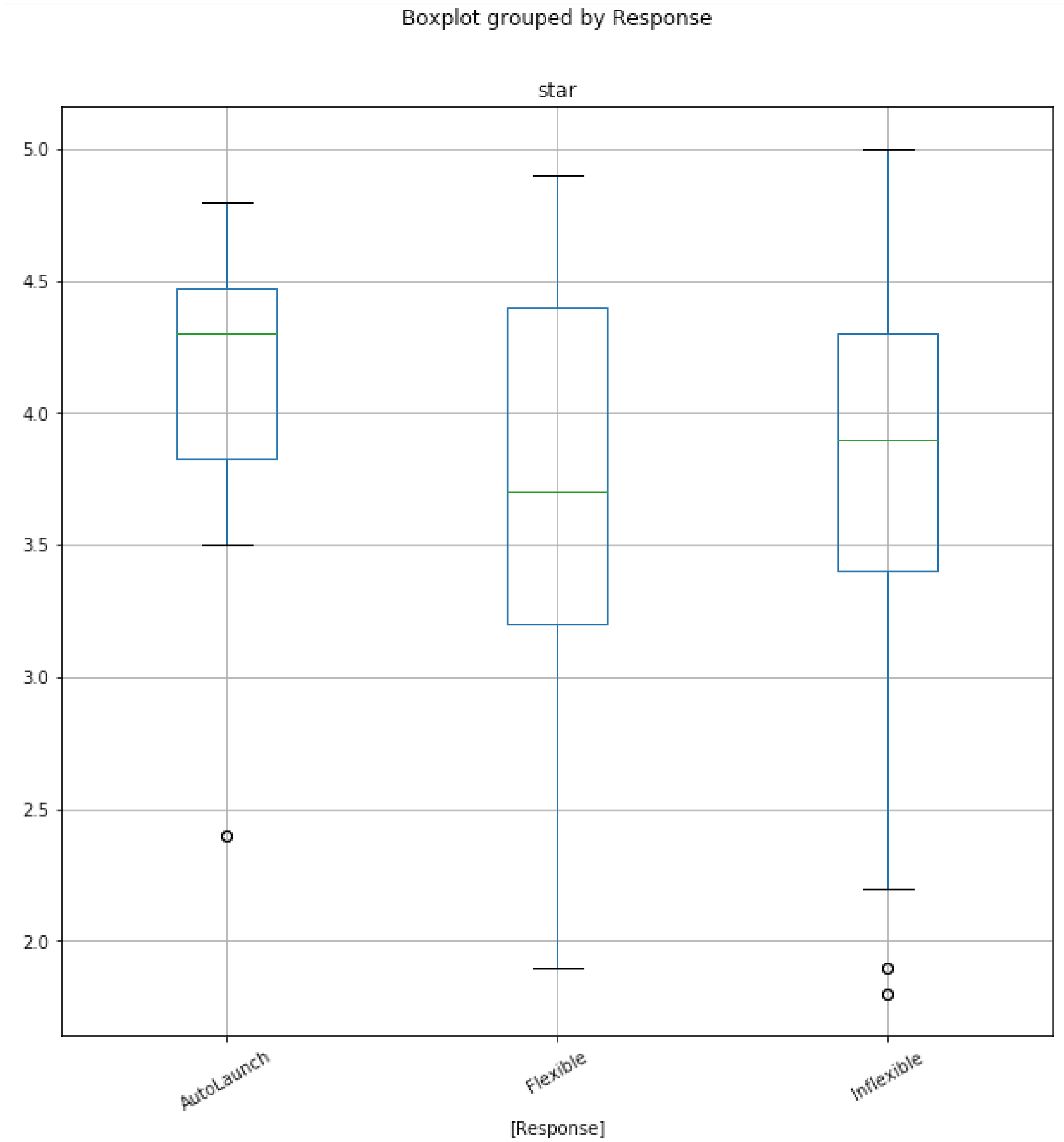

Figure 5.3: Alexa skills which directly fulfill users' intents without further conversational interactions (AutoLaunch) receive the highest median customer ratings.

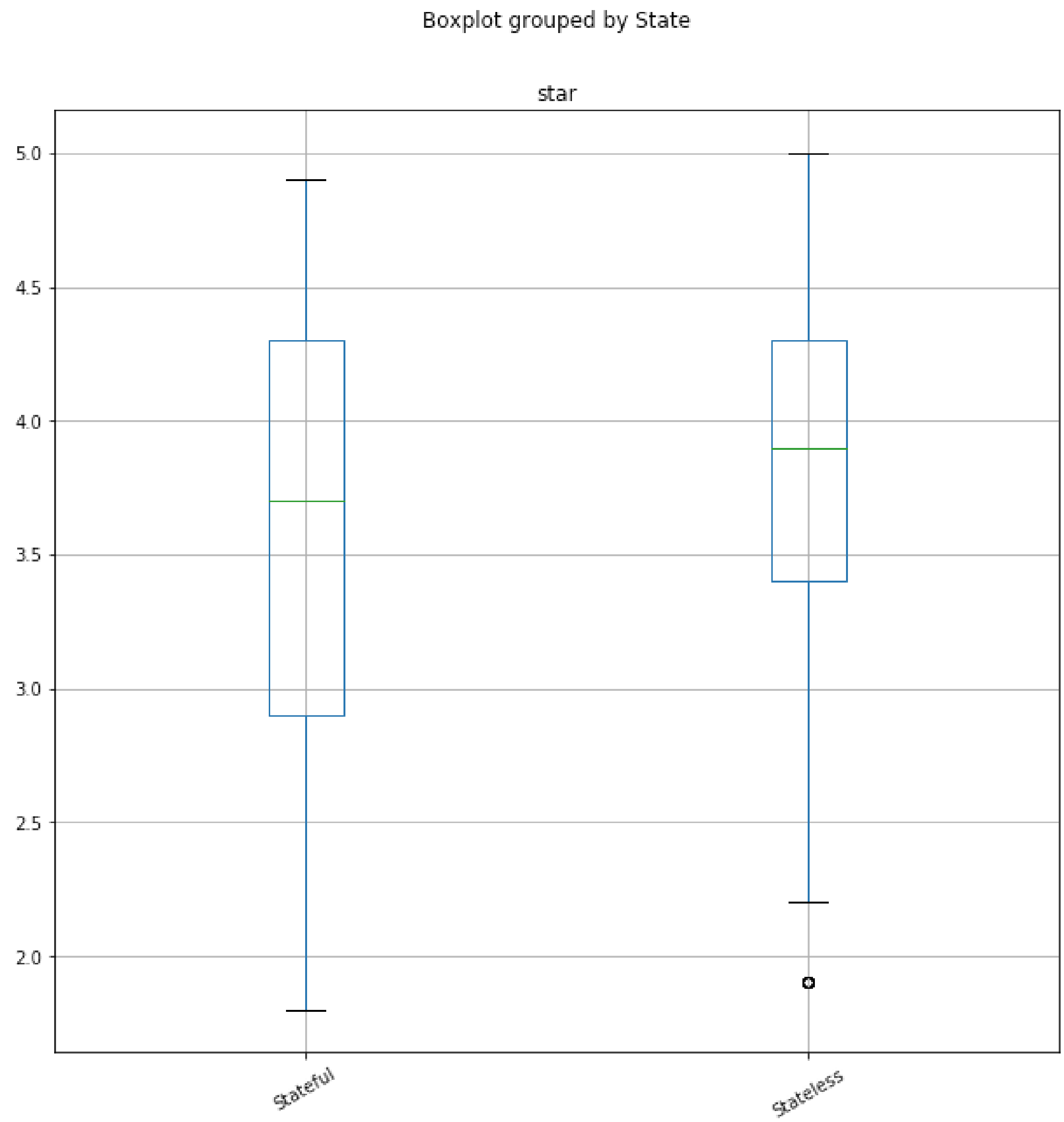

Figure 5.4: Stateless Alexa skills receive higher customer ratings than those are stateful, which presents the positive relationship between the customer satisfaction and simplicity of conversational interfaces. Stateful skills are commonly classified as conversational service, which need deep conversations.

the less options a user could go down for (less utterances a user could say), and less errors and usability issues could happen throughout the conversation, and hence the more positive experience the user would have. Even though users have limited options while interfacing conversational interfaces, which is inconsistent with the experience of using natural language, the good experience of successfully fulfilling intents of users overwhelms the constrains.

Besides cognitive model, we are interested in the experience of CA from conversational interface perspective. We elaborate the experience testing by adding two variables. First, the variation of utterance. We elaborate the variation of utterances by plugging in different number of slots and creating synonymous utterances. Second, the Response Error Rate (RER) of responses to utterances.

As slots and utterances are the fundamental units to fulfill an intent, which help build the entire conversational interfaces, We found different types of interfaces handle slots to different degrees. For example, Expedia can take two slots in one utterance such as city and price, while Ted Talk can only take one slot at a time such as the speaker's name or the name of a topic. Moreover, we are interested in how user experience is influenced by different numbers of slots in an utterance. Does the skills support more number of slots at a time provide better experience? How about those skills that are able to recognize synonymous utterances for an intent?

According to the study of *Radziwill and Benton* (2017), researchers see a high negative correlation with RER, leading to the conclusion that users give poor ratings if responses are incoherent. Based on their conclusion, we asked MTurk workers to grade the responses to utterances we designed (utterances with one slot, with two slots). And we define the incoherent responses as 'Poor - Mostly incoherent' and 'Bad - Completely incoherent' by MTurk workers.

We found the increasing number of slots significantly influences RER. Media-player skills and Conversational-service skills performs similarly under this condition (figure 5.5). Even though users are used to speaking multiple keywords (slots) in day-to-day conversations, the Alexa skills are bad at recognizing multiple keywords at the same time. According to our study, in order to fulfill a intent, Media-player skills got 42% RER when requested by one-slot utterances, 62% RER to two-slot ones. (0% means completely coherent responses, 100% means completely incoherent responses). Likewise, Conversational-service skills got 37% RER and 58% RER for one-slot utterances and two-slot utterances respectively.

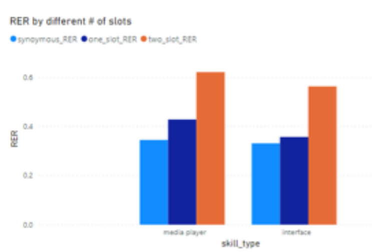

Figure 5.5: Media Player Skills v. Conversational Service Skills

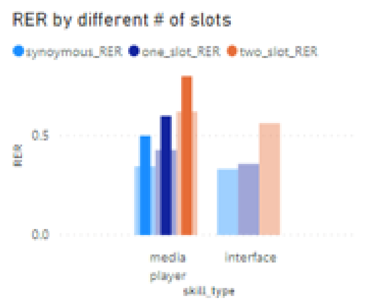

Figure 5.6: ESPN has higher RER than other Media Player skills

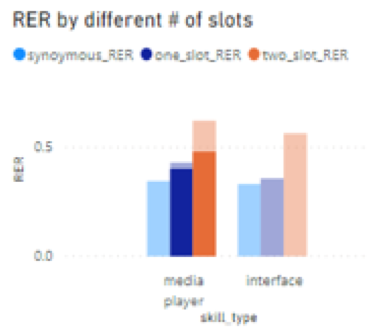

Figure 5.7: Big Sky has lower RER than other Conversational Service skills

For example, ESPN (figure 5.6), a Media-player skill that plays sports podcasts has much higher RER than the average performance among the group. Its RER for different utterances is higher than the average Media Player skills' RER by 12%. On the other hand, Big Sky (figure 5.7) is able to handle multiple slots in utterances. Its one-slot utterance RER is 40% and two-slot utterance RER is 49%, which are lower than the average RER of Media Player skills.

We think Conversational-service skills perform better than Media-player skills (figure 5.5) regarding RER because Conversational-service skills are closer to natural language speaking as in taking multiple slots of keywords to fulfill an intent. However, lower RER does not necessarily equal to customer satisfaction. Conversational-service skills' customer rating, of which median rating is 3.5, is lower than all other kinds

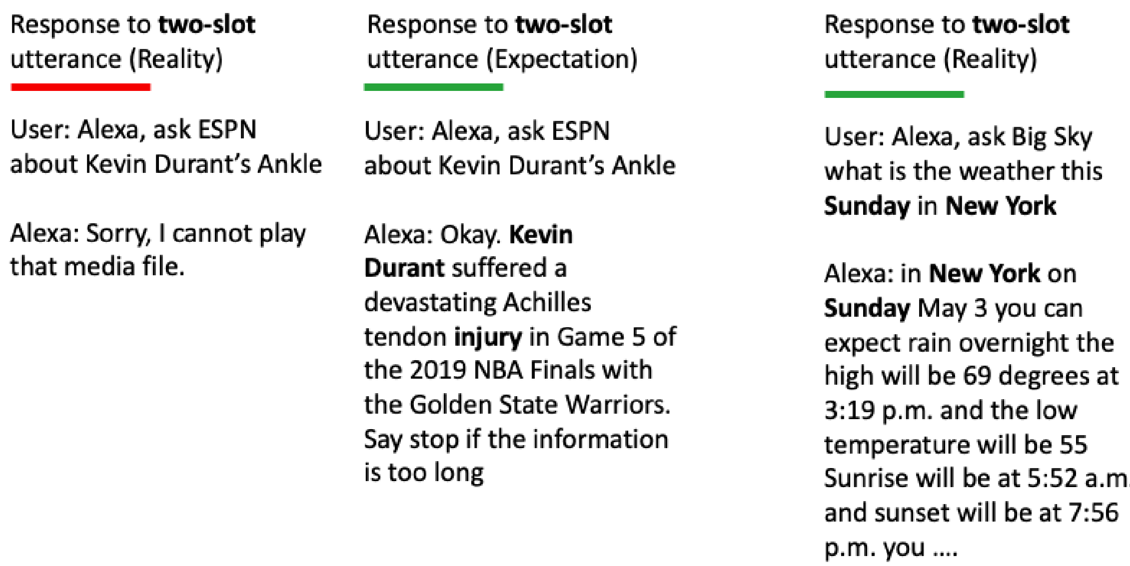

Figure 5.8: ESPN badly interprets the slots in user's utterance, which affects the user experience, while Big Sky updates weather properly when users utter multiple slots.

of skills. We think it's attribute to increasing RER throughout deeper conversations. That is, users getting worse experience as talking further to conversational-service skills.

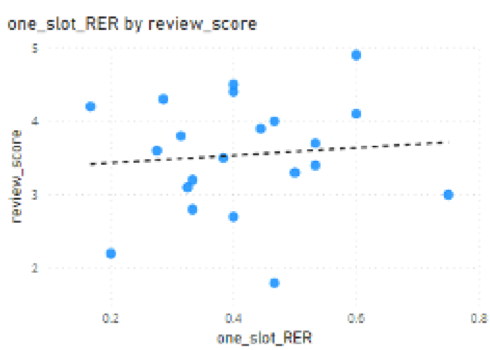

Figure 5.9: Customer ratings improve as the RER to one-slot utterances increases.

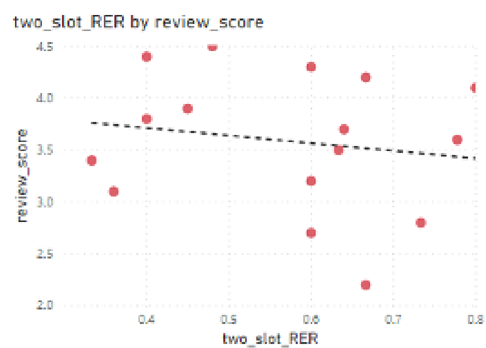

Figure 5.10: Customer ratings drop as the RER to two-slot utterances decrease.

Our study found that there is a positive relationship between RER to one-slot utterance and customer ratings, while RER to two-slot utterance shows negative correlation. This could attribute to that if a skill is supposed to support two-slot utterances, then it should also be able to support a similar manner of speaking using natural language. When the skills are not coherent enough, i.e., high RER to two-slot utterances, to meet people's expectations, then people would be frustrated and give lower ratings. On the other hand, skills with fixed dialogue flow in which force users to utter specific utterances, such as one-slot utterances or binary answers (yes or no), would have less chances to cause usability issues or return incoherent responses. Therefore, even though the RER to one-slot utterances increase, the customer satisfaction improves. We thinks it's probably because that skills mainly support one-slot utterances are Media-Player skills. Regardless of the increasing RER, the podcasts are attractive enough to users or the issue of fulfilling an intent can be fixed easily, which provide users better experience compared with skills could support two-slot utterances.

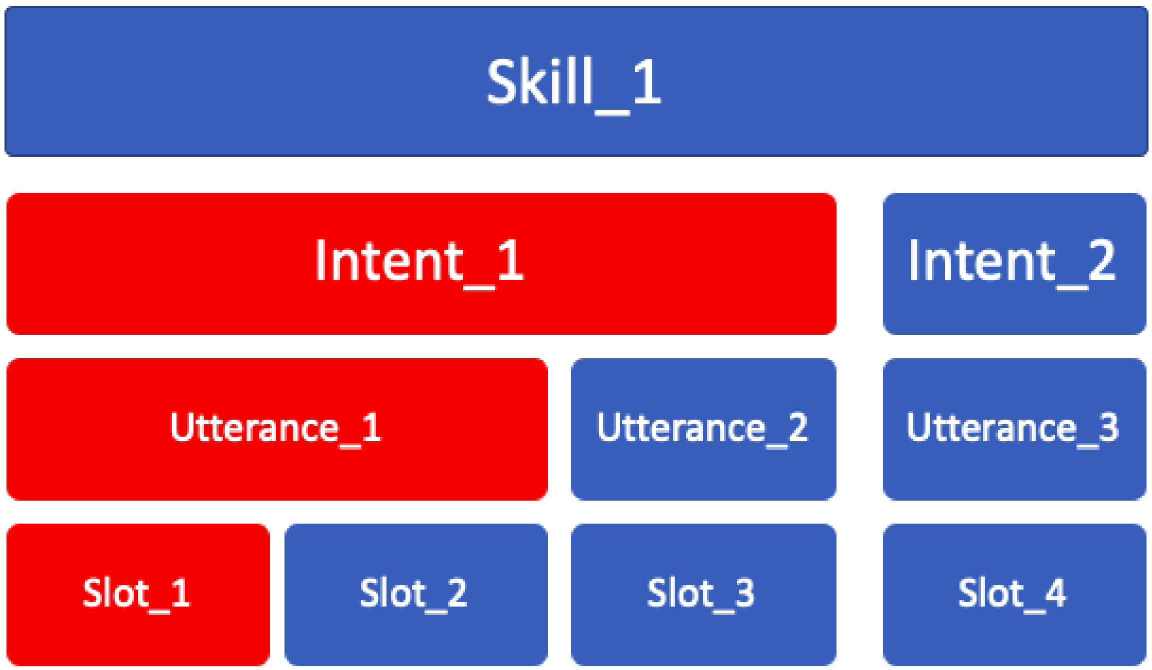

Figure 5.11: The more slots a user includes in an utterance, the more likely the skill could run into problems when fulfilling the intent. Paradoxically, the more fixed utterances or small amount of slots that users are allowed to use for an intent, the less naturally CA is. However, the more flexibility users have to Alexa skills, the more likely the Alexa skill would fail to achieve the user's intent and hence cause poor user experience.

CHAPTER VI

Discussion

Throughout the course of our study, we found that even though some recommendations of the *Guideline* (2020) are followed by Alexa skills, such as accommodating a variety of utterances for an intent, the others are too aggressive to be realized. There are both internal and external limitations for skills to practice design guidelines, such as the network infrastructure, accuracy of utterance recognition, logistics issues.

We also found that the metrics of the cognitive model are not effective enough. Although *Guideline* (2020) recommends skills to paraphrase responses to the same intent, personalize responses based on users' information, and handle user's over answering gracefully. Those stateful or flexible responses of skills do not show significant positive influence on customer rating or comment engagement. It suggests skills reply to users personally or flexibly do not necessarily prompt users to engage more. Our study presents user engagement and satisfaction are affected by conversational interface and response coherence.

Moreover, our study demonstrates utterance with different number of slots could be a direction for developers to consider while designing Alexa skills. Developers are suggested to aware the difference of conversational interfaces when designing skills. For instance, what number of slots users might expect to use in an utterance during the conversation? As for a skill, what is the probability of failure of fulfilling an intent when using the utterances with multiple slots? Some Media-player skills only support utterance with one slot, whereas users are tended to say using multiple keywords. How to design dialogue flow based on that? As for Conversational-service skills, they are supposed to handle users' utterances with multiple slots more frequently because fulfill an intent using deep conversation smoothly more aligned with users' intuition. Being unaware of the nature of speaking and distinctive attributes between conversational interfaces cause unsatisfactory user experience to customers.

In summary, we anticipate our study to be a start point of an experience evaluation toolkit for CA developers. This toolkit would aggregates Alexa skills/conversational applications with similar types of conversational interfaces and topics. The toolkit would provide insights of response coherence, conversations, and customer feedback that other similar skills have with the users. It allows developers to optimize their own skills by preparing for potential utterances and conversations that happened with other similar skills. For example, Expedia and Kayak could both be asked for questions about traveling from city A to city B and the price of hotels in the destination. The skill iteration could be expedited when developers of Expedia are able to learn the utterances for Kayak and vice versa.

CHAPTER VII

Conclusion

Our research proposed a data pipeline that programmatically crawl user-agent conversations, grade them in terms of coherence, and perform hypothesis testing. Although there is increasing shipments of CA and associate Design Guidelines for them, they are facing various challenges.

We found some of the reasons behind the unsatisfactory user experience for CAs and demonstrated effective metrics to evaluate it. By tracking down the fundamental components of user-agent interactions–slot and utterance, we found that the number of slots in an utterance and the type of conversational interface could significantly influence customer satisfaction and user engagement. We anticipate our works to be a starting point of building a more robust and widely-applicable experience testing toolkit for CA in the future.

BIBLIOGRAPHY

BIBLIOGRAPHY

Deloitte, L. (2016), Global mobile consumer survey 2016: Uk cut.

Feng, H., K. Fawaz, and K. G. Shin (2017), Continuous authentication for voice assistants, in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pp. 343–355.

Goh, O. S., C. Ardil, W. Wong, and C. C. Fung (2007), A black-box approach for response quality evaluation of conversational agent systems, *International Journal of Computational Intelligence*, *3*(3), 195–203.

Guideline, A. A. D. (2020), Amazon alexa design guideline.

Han, X., and T. Yeh (2019), Evaluating voice skills by design guidelines using an automatic voice crawler, *arXiv preprint arXiv:1906.01122*.

Kinsella, B. (2018), 61% of alexa skills still have no ratings and only 1% have more than 100, *Voicebot.ai*.

Radziwill, N. M., and M. C. Benton (2017), Evaluating quality of chatbots and intelligent conversational agents, *arXiv preprint arXiv:1704.04579*.

Ram, A., et al. (2018), Conversational ai: The science behind the alexa prize, *arXiv preprint arXiv:1801.03604*.

Sciuto, A., A. Saini, J. Forlizzi, and J. I. Hong (2018), " hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage, in *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 857–868.

Tsai, J. Y., T. Ammari, A. Wallin, and J. Kaye (2018), Alexa, play some music: Categorization of alexa commands, *Voice-based Conversational UX Studies and Design Wokrshop at CHI. ACM.*

Venkatesh, A., et al. (2018), On evaluating and comparing conversational agents, *arXiv preprint arXiv:1801.03625*, *4*, 60–68.