Observations of a New Chatbot

Drawing Conclusions from Early Interactions with Users

Lisa N. Michaud
Aspect Software

This article shares insights gathered from the design and implementation of an SMS chatbot-based virtual assistant to hotel guests in London. The author discusses challenges and first outcomes, and makes recommendations based on this experience toward best practices for approaching the design of chatbots in the customer service domain.

The ascendance of virtual assistant technology has occurred side by side with tremendous advances in the science of natural language processing (NLP), which is required to connect human language input to its intended meaning. Although speech and text interfaces are hardly new, older systems often kept their tasks simple by retaining the initiative in the interaction. They prompted the user to express what he or she wanted in the form of limited answers to directed questions that kept the scope of possible responses narrowly defined. In virtual assistant environments, however, the prompt is both extremely broad and implied rather than spoken: *What do you need?* The user therefore effectively initiates the interaction, and our technologies face a much more difficult job because the onus of anticipating a dizzyingly open-ended horizon of possible inputs now rests on the shoulders of the designer.[1]

Thankfully, the state of the art has given rise to many commercial frameworks for attempting this daunting task. Most of these identify that the key objective is to determine the intent of the user's sentence. What question is the user asking? What activity is he or she requesting? What use case does he or she wish to initiate? The horizon is still vast, however, and while these frameworks are powerful, the technologies are young, and the best practices behind designing virtual assistants are still little known.

In early 2016, we collaborated with a high-end London hotel chain to design and introduce a new kind of virtual assistant for the hospitality domain. The resulting "virtual host," Edward, uses SMS to communicate with hotel guests who have registered a mobile phone number with their reservation. Edward possesses the capability to provide answers to frequently asked questions and to directly summon requested items and services from appropriate housekeeping and maintenance staff, deflecting both activities from front desk staff so they can concentrate on more complex interactions with guests. Edward originally debuted with the ability to recognize and

Published by the IEEE Computer Society
1520-9202/18/$33.00 ©2018 IEEE

respond to 180 different customer intents, and has since grown through continual revision and improvement.

The benefits of deploying Edward were immediate and easily measured by the hotel management. Edward automatically tracks every incoming contact, escalating issues to management if a resolution does not occur within a certain interval of time, resulting in detailed records of incoming requests and a much higher rate of prompt issue resolution. Edward is also a model member of staff—because a chatbot is only new once, it can only get better with time. From the customer-facing perspective, Edward never has a bad day, and can treat any request with the same respect and unfeigned politeness. A chatbot like Edward can also be extended to function in multiple languages, making it possible to serve customers naturally without requiring multilingual staff members.

This article reviews the data obtained from the first two months of Edward's interactions with real customers, and provides reflections on the lessons learned in terms of informing best practices for virtual assistant design in the customer service domain.

From the customer-facing perspective, Edward never has a bad day, and can treat any request with the same respect and unfeigned politeness.

THE DATASET: A CORPUS OF USER REQUESTS

The data collected from the first 1,023 texts sent to Edward comprise 53 days of interaction with 491 distinct guest accounts. These texts contain a total of 1,258 different sentences (as determined by the detection of sentence-ending punctuation).

Going into this project, one unknown was the way in which customers would choose to express themselves in this channel. Human-to-human conversations over SMS are often highly abbreviated and succinct, containing many non-standard spellings and shortcuts. For example, one study found that college-aged users in the UK averaged more than five deliberate shortenings per text message.[2] Other studies have found occurrences of out-of-vocabulary (OOV) terms to be as high as 15 percent in SMS and Twitter.[3] However, there is also great individual variation in these practices—where some users can be minimalist texters, others might use fully inflected, complex sentences with standardized spelling. In some of those cases, this can result from the fact that many users prefer to compose SMS messages via voice recognition rather than by typing.

Diverse styles are represented in our dataset. Multiple sentences were found in 12 percent of the texts; 28 percent of the customer-typed "sentences" contained only one word and 22 percent contained 8 words or more. These longer sentences still focused on single intents, however; although even the earliest versions of Edward were designed to recognize multiple intents occurring in a single input, only 1 percent of the total sentences attempted multiple tasks at once. In this SMS corpus, only 1 percent of the sentences contained a word that was not correctly spelled.

The data was hand-labeled with the user's original intent by a human reviewer, using both the original list of 180 recognizable intents and the more abstract categories discussed next.

CUSTOMER SERVICE DIALOGUE ACTS

From this corpus, we created a hierarchy of dialogue acts that reflect the more general categories into which application-specific intents could be placed. This hierarchy—which could be applied to virtually any user input in the general customer service domain—is shown in Figure 1. The figure has been color coded to show that some of these acts distinguish between those that can be serviced entirely by a virtual agent (green) and those that are likely to require human intervention (blue). The idea behind these dialogue acts is that they represent abstractions from specific intents in which distinctions are made based on the way in which the system will respond.

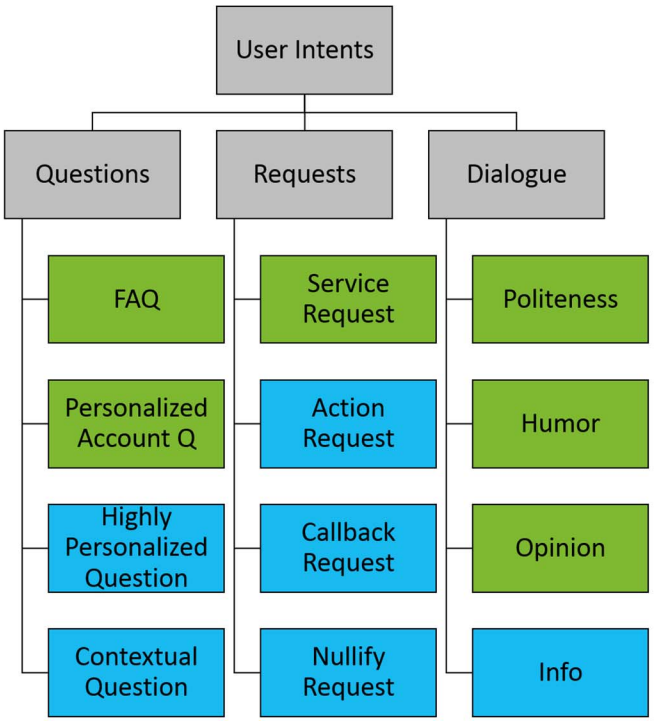

Figure 1. A hierarchy of user intent dialogue acts.

Within the "questions" part of the hierarchy, there are two subtypes of questions that a well-designed system should be able to automate: the "frequently asked question" (where the answer is independent of the person asking the question) and the "personalized account question" (where the answer is individual to the asker, but is easily constructed via a database lookup).

> Is there a spa? I would like a massage. *Frequently asked question*
>
> Can you confirm which room I am in? *Personalized account question*

Two other types of questions are likely to require a human. "Highly personalized questions" require information that is not available in a database (and might require additional interaction with the user), and "contextualized questions" require a sophisticated level of dialogue understanding to interpret the intent.

> How much would a taxi be to the museum? *Highly personalized question*
>
> Will I get this without further charges? *Contextualized question*

In particular, contextualized questions use expressions called *anaphora*—pronouns and other linguistic indicators (such as "this") that refer to content from previous sentences in the conversation. Contextualized questions would ideally be possible to automate under a framework with sophisticated dialogue-state tracking and anaphora-resolution capabilities, but this is not currently common among chatbot frameworks.

Under the "requests" hierarchy, we also have a division between automated and human-reliant intents. A "service request" asks the virtual assistant to perform a service that can be completely automated, an "action request" requires a human to perform the action, a "callback request" initiates contact with a human staff member, and a "nullify request" asks that the previous request be cancelled.

> I need a copy of my bill emailed to me, please. *Service request*
>
> Could you bring me some sugar and a cup of tea? *Action request*

Connect me to a human, please. *Callback request*

Ignore that—I just found it! *Nullify request*

Under "dialogue," we place actions taken by the user that are natural parts of having a conversation but not necessarily directly connected to the core functionality of the assistant. There are "politeness" expressions, in which the user says *hello*, *goodbye*, or *thanks*. There are also "humor" inputs, as many users (and their children) are aware that various "Easter eggs" (a hidden message or secret feature) can be found when you ask the right question. "Opinion" texts are those in which the user is expressing an opinion about their experience as a customer (which can be further broken down into "praise" or "complaints"). Finally, the "information" category exists to acknowledge that some sentences contain information that support or contextualize a question or request without containing a question or request of their own.

Hi, Edward. *Politeness*

Can you tell me a joke? *Humor*

Room not up to standard. *Opinion*

I checked in this morning. *Information*

The complete list of intent categories used in our analysis of the Edward data is shown in Table 1, and the distribution of sentences in our data set between the different categories is shown in Figure 2. Note that in the specific domain of hospitality, we had a subcategory of "action request"—the "object request"—in which a person has asked that a particular object be brought to the guest room (as opposed to asking for something like turndown service). There was also a subcategory of "service request": "opt-out," in which the user requests that the system stop sending texts.

Table 1. The complete list of intent categories.

Action request	ARQ
Complaint	COM
Callback request	CRQ
Contextualized question	CXQ
Frequently asked question	FAQ
Highly personalized question	HPQ
Humor	HUM
Information	INF
Nullify request	NUL
Opt-out	OPT
Object request	ORQ
Personalized account question	PAQ
Politeness	POL
Praise	PRS
Service request	SRQ
Unknown	UNK

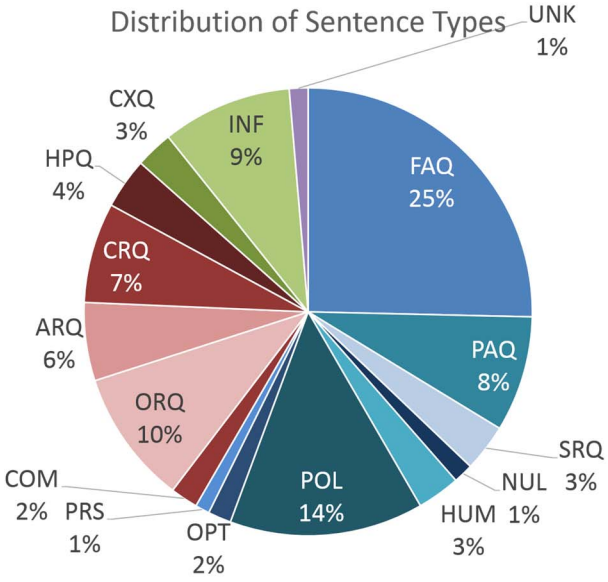

Figure 2. The distribution of sentences in the data between the different intent categories.

One of the striking observations that can be made from this data is the importance of realizing that a chatbot should have at least some capability to actually *chat*—this is underscored by how much of the communicative acts (14 percent) were simply being polite. Surprisingly, a user started the text with "hi" or "hello" 140 times, even knowing that he or she was speaking to a virtual assistant. Several times, the user thanked the bot for performing a task. We assert that a bot must be able to respond appropriately to these "purely dialogue" interactions to maintain an optimal user experience.

Other surface observations include that 59 percent of the sentences communicated dialogue acts that were in self-service categories and did not require human staff at all. When including those acts that involved the automated alerting of housekeeping and maintenance staff without the involvement of the hotel's front desk, the portion of deflected contacts increases to 76 percent. This is clear proof of the potential return on investment for customer service chatbots in this domain, in addition to the benefits discussed earlier.

LESSONS LEARNED

Following are some additional lessons learned from both the design processes around Edward's implementation and a deeper analysis of the data.

Never Underestimate the Importance of Data

Edward's list of expected requests that formed the basis for the original 180 intents was the product of a brainstorming session with senior staff in the environment in which the bot was to be deployed. The objective was to come up with a list of the most common requests; however, actual usage during the first two months was different than expected. Nearly half (42 percent) of the sentences communicated to Edward during that time did not contain one of those 180 intents, and more than half of the intents that had been integrated into Edward's design were never asked for in that initial time.

The lesson to be learned from this is the seriousness of the challenge of determining which intents to design a virtual assistant to recognize. Ideally, a preliminary data-gathering period using a similar channel (phone or an SMS/chat contact that is manned by human staff) can be leveraged to provide concrete data on the set of intents that would provide the greatest impact. Review of the distribution of disposition codes from these past contacts could show which intents

are the most common. Without the availability of such data, another approach would be to deploy a bot that was pre-engineered to handle very few intents, sending almost all incoming requests to a human backup. After this bot had seen thousands of incoming interactions, the distribution of the intents could be reviewed with an eye for engineering the full chatbot.

The ideal result of this data gathering would be to discover a distribution of incoming requests evoking Pareto's principle, so that one might be able to cover 80 percent of incoming requests by implementing only 20 percent of the different intents.

Keyword-Based Approaches Are Likely to Fail When Sentences Are Complex

When we compared the performance of a mostly keyword-based intent classifier against the length of the sentences in the data, we saw that there was a statistically significant difference in the length of those sentences whose intent was correctly identified versus those whose intent was incorrectly identified, as illustrated in Figure 3 below.

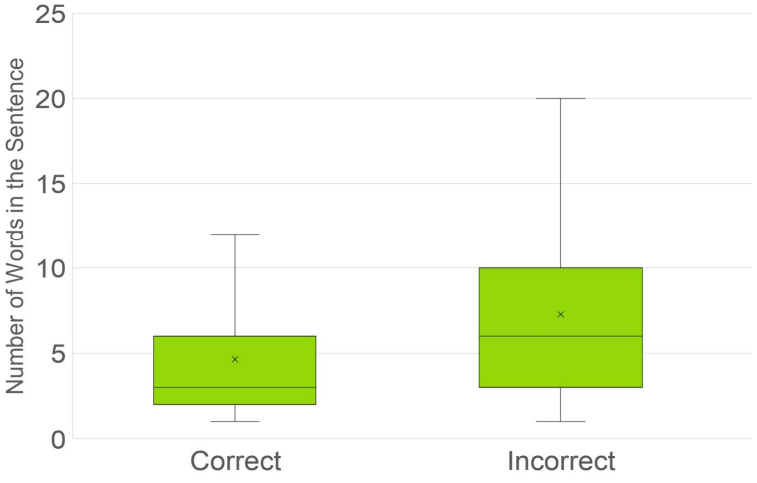

Figure 3. The success of a keyword-based approach as affected by the length of a sentence.

This reinforces two intuitions regarding virtual assistant design. First, keyword-spotting techniques (by over-simplifying the intent-recognition process) are likely to fail when longer sentences provide key contextual information. Second, it might be important for virtual assistants to recognize that no matter what technique they use, the longer and more complex a sentence is, the more likely there is information the system is failing to interpret. We advocate the practice of estimating sentence complexity and passing complex user inputs on to human agents rather than attempting a task that is likely to result in failure.

Of course, the cost balance between task failure and employing a human can vary greatly depending on the specific environment in which the bot is deployed. A conservative design approach might result in an implementation that strives for high accuracy in its responses, at the risk of not understanding (or passing to a human) requests that it should have understood. For example, a bot that has been designed to respond to *I want [X]* will correctly respond to:

> I want **towels**.
> *Some will be brought to your room.*

But it will incorrectly reject this input:

> More **towels** would be great.
> *I'm sorry; I didn't understand that.*

We call this a *false negative*, in that the intent classifier is not only mapping an incoming text to a specific intent but also deciding whether it is in scope (positive) or out of scope (negative). In this instance, a negative judgement was incorrect. An alternative approach to design would be more open and risk *false positives*:

> Could someone come and change my **sheets**?
> *Housekeeping will be right there.*
>
> These **sheets** are fantastic.
> *Housekeeping will be right there.*

The comparative risk of false negatives versus false positives depends on the specific application and the level of acceptability of false interpretations if more automation results in desired cost reduction. In the Edward project, with human backup easily available and the customer experience at the high-end hotel chain being of paramount importance, false positives were far more costly than false negatives. The bot needed to be precise and highly accurate, at the risk of requiring human help on inputs that it had missed. This is an environment in which the use of a complexity metric and cutoff would be particularly useful. When misanswering a question is less costly than connecting to an agent, a broader design might be more successful in deflection at the risk of occasionally requiring the user to try again.

In many bot frameworks where the intent classification is created by providing example sentences, the "conservative" approach would take the form of providing a small selection of sentences of very similar construction. This would result in a learned classifier that recognizes that particular construction (*I want [X]*) and avoids false positives. A "broad" approach would be to use sentences that are varied, such as the two examples above. This might make the system able to avoid the false negatives, but runs the risk that the system learns that what unifies those two sentences and makes them distinct from other intents is only the word "sheets."

It's a Dialogue, Not Just Serving Answers

As mentioned above, 12 percent of the texts in this dataset involved multiple sentences. It is also the case that 9 percent of the sentences were providing context for a question that preceded them or would follow. Edward's design did not include any type of maintained dialogue state or pronoun resolution, nor do those of many automatic question-answering systems. It is typical to view the task as simply serving an answer to a given question; however, context can be vitally important when questions are underspecified,[4] and interactions with a virtual assistant that span multiple turns are highly likely to result in users making reference to the dialogue context. A typical flow might be to start with a common question, and then drill down with follow-up questions pertaining to that subject. An example interaction from this domain might be:

> Is breakfast included with my room?
>
> *No, it is not.*
>
> Can I add that to my reservation?
>
> *Yes, you can add it at any time.*
>
> How much?
>
> *It costs [X].*

While solutions for pronoun resolution do exist, they are not infallible. The task of identifying entities in utterances and determining the mostly likely antecedent to match a pronoun can be a challenge. In the context of performing natural language understanding on the sentence, it might require replacing the pronoun with the identified entity in an attempt to make the sentence autonomous. A simpler model we are currently investigating for handling pronouns such as "it," "that," and "this" and under-specified questions would be to keep track of the most recent major topic introduced by past questions and leverage that information in the interpretation of those immediately following.

Chatbots Take Time to Mature

No matter what framework you use, recognize that chatbot design is an iterative process. Do not underestimate the time and resources needed for a bot to ripen to its full potential for performance. Despite the marketing hype, this is not a technology that operates out of the box. "A lot of pundits say that chatbots aren't good; that's only because people have decided to make them not good," noted one of Aspect Software's customer strategy executives. "You might have to put some effort into it." Collecting data is part of it, but you must also expect to invest time in tuning performance once real users are involved. This might create a challenge if you want to evaluate multiple platforms against each other; it could be more useful to watch the literature as side-by-side comparisons on accuracy and effectiveness are executed and reviewed.

CONCLUSION

Interaction data provides a wealth of knowledge for the design, implementation, and improvement of virtual assistants. Collecting data both before the implementation of a system and during the early stages of its deployment can inform critical design decisions, including which intents are most likely to occur, and the many different forms a given intent is likely to take.

Human language is infinitely diverse—no chatbot design can anticipate everything on that wide horizon. It can only ever approximate the understanding power of a human. However, the more you know in the early stages, the better your chances will be that you will capture enough of the diversity to have a significant impact. Above all, always continue to gather data, and then feed that data back into the design in subsequent stages for continual improvement.

REFERENCES

1. R. Laroche et al., "D6.4: Final Evaluation of CLASSiC TownInfo and Appointment Scheduling Systems," *Computational Learning in Adaptive Systems for Spoken Conversation*, technical report, CLASSiC, April 2011; www.microsoft.com/en-us/research/wp-content/uploads/2017/07/D6_4_Final_evaluation_of_CLASSiC_TownInfo_and_Appo.pdf.
2. M. Bieswanger, "2 abbrevi8 or not 2 abbrevi8: A Contrastive Analysis of Different Space- and Time-Saving Strategies in English and German Text Messages," *Texas Linguistic Forum*, vol. 50, 2007; http://didattica.uniroma2.it/assets/uploads/corsi/39543/smsgermanenglish.pdf.
3. B. Han, P. Cook, and T. Baldwin, "Lexical Normalization for Social Media Text," *ACM Trans. Intelligent Systems and Technology*, vol. 4, no. 1, 2013.
4. R. Fernández, J. Ginzburg, and S. Lappin, "Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach," *Computational Linguistics*, vol. 33, no. 3, 2007, pp. 397–427.

ABOUT THE AUTHOR

Lisa N. Michaud is a leader in natural language R&D at Aspect Software. Her research interests include natural language processing and computational linguistics, human–computer interaction, user modeling, dialogue systems, parsing, and the analysis of non-grammatical text. Michaud received a PhD in computer science from the University of Delaware and has been published in multiple international journals, workshops, and conferences. Contact her at lisa.michaud@aspect.com.