A Case Study of User Communication Styles with Customer Service Agents versus Intelligent Virtual Agents

Timothy Hewitt
Verint - Next IT
Spokane Valley, WA USA
timothy.hewitt@verint.com

Ian Beaver
Verint - Next IT
Spokane Valley, WA USA
ian.beaver@verint.com

Abstract

We investigate differences in user communication with live chat agents versus a commercial Intelligent Virtual Agent (IVA). This case study compares the two types of interactions in the same domain for the same company filling the same purposes. We compared 16,794 human-to-human conversations and 27,674 conversations with the IVA. Of those IVA conversations, 8,324 escalated to human live chat agents. We then investigated how human-to-human communication strategies change when users first communicate with an IVA in the same conversation thread. We measured quantity, quality, and diversity of language, and analyzed complexity using numerous features.

We find that while the complexity of language did not significantly change between modes, the quantity and some quality metrics did vary significantly. This fair comparison provides unique insight into how humans interact with commercial IVAs and how IVA and chatbot designers might better curate training data when automating customer service tasks.

1 Introduction

An intelligent virtual agent (IVA) is a subset of chatbots designed for the commercial enterprise realm to mimic a human customer service agent. A popular use case for IVAs is live chat deflection, where they are trained to handle the most common interactions while still allowing for escalation to a human agent when required or requested.

As a company that has designed and built IVAs for enterprise applications for many years, we had intuition that the language we saw in live chat interaction was different from the language we saw coming into the IVA, but the difference had not yet been quantified. After using live chat data for training an IVA, we were occasionally surprised at the gaps in understanding it presented once in production, even though the training data originated from the same company the IVA was built for. In order to improve analysis and training, we sought a standard to create and gather data more consistent with actual IVA usage and filter out some of the non-representative live chat data.

We also wanted to investigate how the IVA was affecting conversations with live chat operators. While there are differences, a user behaves consistently when he/she is chatting with a human, similarly they are consistent when chatting with an IVA. In this paper we demonstrate that chatting with an IVA has significant impact on language beyond what has been documented by human-to-human computer mediated conversation such as instant messenger or live chat.

The IVA and live chat corpora used in this study originated from a financial services company where customers are interacting with the IVA and live chat on their website for the same purposes making the comparisons extremely relevant. However, due to data use agreements with the financial services company, the identification of the origin and corpora cannot be made public.

2 Related Works

Hill et al. (2015) have done comparisons between inter-human and "toy" chatbot conversations. However, in this comparison the conversations were sampled from completely unrelated domains making the comparison less valid.

While tools to improve the training process of IVAs from live chat or call center transcriptions exist such as (Bouraoui et al., 2019), there has not been a focused linguistic study on the difference in communication styles between human-human and human-machine in service dialogs. Such a study could inform such tools where specific samples may or may not make good training samples due to projected communication differences with IVAs. To our knowledge this is the first study to compare real world language of users with IVAs and live

, pages 283 - 289

chat from the same origin.

3 Method

The IVA for this research was originally trained on live chat conversations from the financial sector and continuously refined while in production. It was designed to understand frequently asked questions and conversational work flows around the largest business use case: waiving fees (for example conversations see Appendix B). Besides business intents, the IVA also responds to persona (e.g. asking if the IVA is married), common courtesy, and profanity. Escalation points were designed where human involvement was desired (e.g. credit limit changes, account closure). There was no dynamic response delay, no avatar, and users were informed at the beginning of the conversation that they were speaking with an IVA (see Appendix A).

For our corpus, we selected 16,794 conversations with live chat agents from June through October 2017 and 27,674 conversations with the IVA that occurred in January 2020. Within the IVA conversations, 19,350 conversations were completed with IVA only while 8,324 escalated at some point to a live chat agent.

For the purpose of this work we only looked at the user language and actions and not the IVA or live agent responses. The IVA was launched in 2017 on the company website along side a live chat option. After 2017, access to live chat without first talking to the IVA was disabled due to the success of the IVA at automating a continuously expanding set of use cases. We chose to sample IVA data from 2020 to allow for adequate refinement time to present statistics representative of communicating with a mature IVA implementation.

3.1 Conversational Clicks

When we discuss turn-taking in conversation with a multi-modal IVA, we must consider that there are different methods than typing to elicit more information. For instance, clicking on suggested topic or answer links presented by the IVA will continue the conversation as though the user had typed the text of the link. In our domain, specific actions need to occur if a credit card is stolen. If a user goes to either an IVA or a live chat operator and says, “I need to get a replacement card,” the operator might respond with a “Was the card stolen?” whereas the IVA might present two conversational clicks, <Replace a lost or damaged card> <Replace a stolen card>. There were a few considerations for counting these interactions in respect to word counts and user turns.

Remove conversational clicks as a word level metric. This metric allows for direct comparison of the complexity of typed user inputs, but hinders the ability to compare at a conversational level. Both IVA and live chat operators can ask a yes or no questions, but if we drop the click of a “yes” response link to the IVA we lose the comparison to the “yes” response in live chat.

Count clicks as a one word turn. In our example, if we assume a conversational click would only solicit a single piece of information a single word turn would be a fair metric. However, conversational clicks are not always of this type. Some present additional information (such as what to do if a stolen card is found) or other suggested topics (such as upsell opportunities).

Count the language in the link text as the user input. In our example, the same information is required, but the method of eliciting that information has changed the user’s interaction from a single word typed input, “yes,” to a four word conversational click.

For any of these metrics, the count would not be representative of the language a user might input if the conversational click was not present.

For all options considered, there were sufficient concerns that any metrics provided on this data set would be implementation dependent, so we chose to present the statistics for all three options outlined so the reader can understand where the differences lie and to what extent noise exists within our IVA data set from conversational link clicks. To control for question complexity between environments, we measured the frequency of yes/no questions and found that they occurred 8% more often in live chat than IVA conversations.

3.2 Turns

For the purposes of this study, if the user clicks on a suggestion by the IVA that advances the conversation (that is, it returns a response in the IVA), it will count as a turn. IVA turns are ABAB, that is, the user (A) takes a turn and the IVA (B) follows. Live chat turns can extend over multiple inputs, such as, ABAAAB. In such cases, these will be joined into a single turn. In other words, we will treat ABAAAB as ABAB.

	Live Chat	IVA Only	Mixed Sessions	Mixed - IVA	Mixed - Live Chat
User Words/Session	68.83 (61.90)	27.91 (22.10)	114.88 (84.67)	32.45 (23.50)	82.51 (80.31)
Words/Session (links = 0 words)	n/a	23.70 (20.97)	107.61 (84.50)	25.16 (22.34)	n/a
Words/Session (links = 1 word)	n/a	24.72 (21.16)	109.15 (84.52)	26.71 (22.58)	n/a
User Turns/Session	5.12 (3.81)	3.06 (2.16)	10.65 (5.56)	4.03 (2.04)	6.62 (2.04)
Type/Token	0.77 (0.11)	0.82 (0.07)	0.79 (0.07)	0.81 (0.06)	0.78 (0.10)

Table 1: Means and standard deviation of session level analysis. Words/Session is raw words including link click text, links = 0 ignores link clicks, and links = 1 treats link clicks as single word inputs. Type/Token is the ratio of unique words over all words in the session.

3.3 Sentences

Successful conversation over chat does not require full, grammatically complete sentences and IVAs are frequently used as keyword searches. Sentence boundaries and punctuation are many times missing or grammatically misused. As such, we ignore sentence-level metrics.

3.4 Metrics

There are 3 session types: Live Chat (human to human conversation), IVA Only (human to IVA conversation) and Mixed Session (sessions that started with the IVA and escalated to a human live chat operator). A mixed session has two subtypes: Mixed - IVA (user inputs to the IVA in a Mixed Session) and Mixed - Live Chat (user inputs to the human live chat operator in a Mixed Session).

We used the L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010) to measure complexity. However, we will not be using any of L2SCA's sentence based metrics for the reasons discussed in 3.3. We also ran the user turns through our own measures for quality and quantity.

For quality, we selected some of the variables selected by Lortie and Guitton (2011) and Hill et al. (2015) from LIWC (Pennebaker et al., 2015) and included a metric for politeness. However, we did not to use LIWC due to data security policies. For fair comparison, we used word lists from closed class words and opted out of the more subjective open class word based features, other than profanity. The variables of quality we investigated were misspellings, words with more than 6 characters, pronouns, articles, profanity, and politeness.

Misspellings compared tokens against a list of company products and services first, and, if the token was not found there, it was then spell-checked against the English aspell wordlist[1].

[1]`https://ftp.gnu.org/gnu/aspell/dict/en`

Gratitude is a count of the variations of *thank* in a turn. We considered only expressions of gratitude as politeness for this study to reduce potential classification error from approaches such as (Yildirim et al., 2005).

Profanity was checked using a regex of common swear word phrases. There is substantial variation in how people manage to misspell a profane word. The regular expressions are not exhaustive, but broad enough to ensure a quality sample.

Tokens are counted by splitting on white space. Thus punctuation won't count as unique tokens and contractions will only count as a single token.

Type/Token is the ratio of unique words over all words in a turn or session.

Sentiment was measured using the NLTK implementation of VADER (Hutto and Gilbert, 2014) and is normalized from -1 to 1.

4 Analysis

Conversation Level: We begin with the full conversation level metrics shown in Table 1. Each conversation which escalated to live chat involves a link click where the link text was 4 words. This extra click is included in the IVA session.

Live chat conversations take 1.7 times more turns with more than 2.5 times more words. Where escalation is not required, a user can achieve a more efficient resolution with the IVA. However, if the IVA is in fact deflecting the easier to handle issues this could explain some of the differences.

On the other hand, the user experience for escalation is substantially less efficient. First the user has an average length IVA conversation and then escalates to the human agent for a more involved conversation with an average of 1.5 more turns and 14 more words than the live chat sessions alone. This indicates the user's tasks presented to the IVA are not being properly reviewed by the live chat agents, requiring substantial additional effort

	Live Chat	IVA Only	Mixed - IVA	Mixed - Live Chat
Tokens	14.33 (14.29)	9.11 (8.00)	8.05 (7.03)	12.54 (13.44)
Tokens (links = 0 words)	n/a	7.74 (9.00)	6.24 (8.16)	n/a
Tokens (links = 1 word)	n/a	8.07 (8.72)	6.62 (7.87)	n/a
Type/Token	0.79 (0.13)	0.80 (0.12)	0.78 (0.12)	0.75 (0.17)
Misspellings	0.61 (1.16)	0.18 (0.53)	0.13 (0.44)	0.58 (1.06)
Six Character Words	3.08 (3.69)	2.45 (2.29)	2.21 (2.20)	2.44 (3.51)
Profanity	0.00 (0.02)	0 (0)	0 (0)	0.00 (0.02)
Gratitude	0.20 (0.41)	0.05 (0.22)	0.01 (0.09)	0.11 (0.35)
Sentiment	0.20 (0.33)	0.08 (0.27)	0.06 (0.24)	0.20 (0.30)

Table 2: Means and standard deviation of language quality metrics per turn. Tokens includes link click text, links = 0 ignores link clicks, and links = 1 treats link clicks as single word inputs. Type/Token is the ratio of unique words over all words in a turn.

on the part of the user to restate them.

Turn Level: Table 2 gives the turn level metrics. Users type substantially shorter inputs (between 1.5 and 1.8 times) when speaking with an IVA. It appears that beside being more concise with IVA, users are also more careful as there were 4.2% of tokens misspelled in live chat vs 2.0% when interacting with the IVA only and 1.6% when interacting with the IVA prior to speaking with a live chat agent. After communicating with an IVA, users increased to a 4.6% misspelling rate.

Human-to-human gratitude is significantly more frequent than with an IVA. However, after continuing to a human after the IVA, gratitude is almost halved. This reflects the more difficult conversations when live chat is tier 2 support.

Sentiment for human-to-human was significantly more positive. IVA turns were neutral. IVA-only turns averaged 0.08 where as live chat conversation turns averaged 0.2. One would expect the live chat conversations that were preceded by the IVA to be more negative reflecting the decrease in user efficiency discussed in the previous section. However, sentiment for live chat after IVA actually remained at 0.2, perhaps indicating that live chat was usually leading to a reasonable (if not always satisfactory) resolution or the additional effort seemed justified to the users as they were in a sense restarting the conversation with a new party.

Hill et al. (2015) showed significant profanity in chatbot language and Burton and Gaskin (2019) showed a self-reported tendency to be less polite to digital assistants. In our data, only live chat sessions had any profanity to speak of. We speculate that the overall lack of profanity has to do with the professional setting of the customer service environment where previous studies were on open domain chatbots and personal assistants such as Amazon's Alexa.

Pronouns: Live Chat users were almost 2.9 times more likely to refer to the human as 'you' than they were with an IVA (Table 3). When a user escalates to live chat, the pronouns increase, but in general pronoun use is less in conversations that escalated. This implies that when a user knows that they aren't chatting with a human, they remove any references to it as a person, consistent with Burton and Gaskin (2019).

L2SCA returned results that could be explained by shorter turns and fewer words shown between live chat and IVA (Table 3). However, there were two increases worth mentioning in IVA-only conversations. Complex nominals per T-unit (CN/T) increased in IVA usage from a mean of 0.64 to 0.70. The other is mean length of clauses which increased from 5.34 to 5.78. Given the decrease in T-units and Clause/T-unit, this may indicate a tendency of IVA users to rely on conveying information through noun phrases than complete verb phrases. However, these increases were not reflected in users who escalated to Live Chat, the reason for this is unclear.

L2SCA did show that live chat language after IVA was less complex across every measure. This may be part of the explanation for the reduction in gratitude in those conversations: they were less polite because they were more concise. It may be that as the conversation is less efficient, the language becomes more efficient to compensate, but more research is needed to prove this hypothesis.

5 Application and Conclusion

When designing an IVA and when given live chat data for training, it's tempting to tag random inputs indiscriminately for training. However, indiscriminately adding longer inputs more

	Live Chat	IVA Only	Mixed - IVA	Mixed - Live Chat
Pronouns	1.92 (2.22)	1.60 (1.60)	1.33 (1.55)	1.46 (2.06)
1st Person	1.31 (1.77)	1.28 (1.35)	1.09 (1.30)	0.88 (1.53)
2nd Person	0.31 (0.59)	0.13 (0.37)	0.08 (0.30)	0.30 (0.65)
3rd Person	0.30 (0.59)	0.19 (0.49)	0.16 (0.45)	0.28 (0.72)
Articles	0.73 (1.19)	0.58 (0.88)	0.50 (0.82)	0.64 (1.20)
Verb Phrase (VP)	2.37 (2.54)	1.93 (1.70)	1.67 (1.58)	1.96 (2.26)
Clause (C)	1.92 (1.97)	1.52 (1.30)	1.34 (1.21)	1.60 (1.74)
T-Unit (T)	1.26 (1.13)	1.10 (0.79)	0.99 (0.75)	1.09 (1.02)
Dependent Clause (DC)	0.62 (1.12)	0.40 (0.75)	0.32 (0.68)	4.03 (2.04)
Complex T-Unit (CT)	0.40 (0.70)	0.29 (0.53)	0.23 (0.49)	0.31 (0.61)
Coordinate Phrase (CP)	0.17 (0.46)	0.13 (0.37)	0.10 (0.33)	0.12 (0.42)
Complex Nominal (CN)	1.14 (1.63)	0.88 (1.12)	0.74 (1.03)	0.93 (1.42)
Mean Length of T	8.13 (7.94)	7.90 (6.40)	6.95 (6.26)	7.42 (7.65)
Mean Length of C	5.34 (4.10)	5.78 (3.97)	5.20 (3.96)	5.10 (4.15)
VP/T	1.50 (1.43)	1.46 (1.20)	1.32 (1.14)	1.38 (1.39)
C/T	1.22 (1.10)	1.15 (0.88)	1.05 (0.84)	1.13 (1.05)
DC/C	0.18 (0.27)	0.14 (0.25)	0.12 (0.24)	0.16 (0.27)
DC/T	0.40 (0.78)	0.29 (0.61)	0.24 (0.57)	0.34 (0.74)
CT/T	0.24 (0.39)	0.21 (0.39)	0.18 (0.34)	0.21 (0.39)
CP/T	0.10 (0.31)	0.09 (0.29)	0.08 (0.27)	0.09 (0.29)
CP/C	0.07 (0.22)	0.06 (0.22)	0.05 (0.20)	0.06 (0.21)
CN/T	0.70 (1.06)	0.64 (0.90)	0.55 (0.84)	0.63 (1.00)
CN/C	0.41 (0.54)	0.43 (0.56)	0.38 (0.55)	0.39 (0.55)

Table 3: Means and standard deviation of pronoun and article usage and the results of L2SCA per turn.

common in live chat may introduce unnecessary noise to the data. Given our observations, we recommend that training language be more focused to the task and rely on more direct language. We also recommend designers do not neglect to add training samples in the form of keyword searches for the users who still view the IVA as a search tool.

When live chat data is not available, a synthetic strategy must take place. One such strategy outlined by Leuski et al. (2006) is to give a human a sample input and ask them to synthesize new data. A better plan would be to give the user a task of retrieving information and then asking them what questions they would use to get that information. This would encourage the simple direct language that IVAs are more likely to see in the wild instead of forcing the human to be creative, which may result in language unlikely to be seen by an IVA.

Users are currently more likely to be concise with IVAs than human live chat operators in the same domain. This is an advantage to the user as an IVA can respond more quickly and get them the needed information with less language production.

Modern contact centers use various performance metrics to rate contact center agents, which has a direct impact on their compensation and recognition (Cheong et al., 2008). A successful IVA will significantly reduce the number of conversations coming into a contact center, and the conversations that do will largely consist of more difficult cases. As these types of conversations now make up a much larger part of the performance metrics, and the operators no longer get the positive feedback from the easy cases, they appear less effective than before the IVA was implemented. This decrease in performance should be expected and these metrics adjusted, perhaps by weighting by the difficultly of the task, so that human contact center agents are not punished by the deployment of IVAs alongside them.

There exists substantial research regarding how to measure the performance of a dialog system, but the study of how people communicate through language with artificial intelligence in the wild is still in its infancy. The nature of the data originating from commercial IVAs means that corpora are seldom shared, making the research more challenging. In spite of these sharing restrictions, in this paper we have presented an approach to analyze the nature of language use between humans and IVAs compared to that of human chat operators in a way that still allows the research community to understand in what way humans currently communicate differently with IVAs than other humans in the same domain. If others with commercially deployed IVAs repeat such experiments we can observe how humans adapt to IVAs over time in the wild, and change the way IVA conversations are designed accordingly.

References

Jean-Léon Bouraoui, Sonia Le Meitour, Romain Carbou, Lina M Rojas Barahona, and Vincent Lemaire. 2019. Graph2bots, unsupervised assistance for designing chatbots. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 114–117.

Nathan G Burton and James Gaskin. 2019. "Thank you, Siri": Politeness and intelligent digital assistants. In *AMCIS 2019*.

K Cheong, J Kim, and S So. 2008. A study of strategic call center management: relationship between key performance indicators and customer satisfaction. *European Journal of Social Sciences*, 6(2):268–276.

Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92:539–548.

Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2018. Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. In *26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change, ECIS 2018*, page 143975.

Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia. Association for Computational Linguistics.

Catherine L Lortie and Matthieu J Guitton. 2011. Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One*, 6(9).

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Serdar Yildirim, Chul Min Lee, Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 2005. Detecting politeness and frustration state of a child in a conversational computer game. In *Ninth European Conference on Speech Communication and Technology*.

A Implementation Considerations

Some research on human-chatbot interactions suggests that dynamic response times increased overall satisfaction with the chatbot (Gnewuch et al., 2018). Ciechanowski et al. (2019) recommended avoiding avatars or otherwise making the chatbot appear human visually. It is unknown how these, or any other measures, will impact user language.

B Example Dialogs

User: hi i did not mean to make a late payment i was in the hospital, can you please waive the charge for me :)

IVA integration checks against user account and eligibility for fee waiver

IVA: Your account is eligible for a fee waiver. Customers can have this fee waived when <list of business conditions>. Are you sure you want to apply this fee waiver now?
Yes (link)
No (link)

User: yes please

IVA: The fee has been waived. Can I help you with anything else?

User closes window

Live Chat Agent: Hello <user>. How can I help you today

User: Hi. My bill came in and I saw a fee.

User: What is that all about?

Live Chat Agent: Let me take a look at that.

Live Chat Agent: It looks like this fee was applied to your account for <business reasons>.

User: I didn't know. Is there any way you can waive the fee this time?

User: That never has happened before.

Live Chat Agent: It looks like your account is eligible for a one time waiver. You will not be eligible again until <business requirements>. Would you like to apply that now?

User: Yes. Please.

Live Chat Agent: Ok. It's done. Is there anything else I can do for you?

User: No Thanks.

User: Bye.

Live Chat Agent: Thank you for contacting <company>. Have a great day.

User: thanks bye.

Figure 1: An example conversation with an IVA (top) and an example with a human live chat operator (bottom) completing the same task of waiving a fee. Company-specific information has been sanitized.

User: waive the fee

IVA integration checks against user account and eligibility for fee waiver

IVA: Your account is eligible for a fee waiver. Customers can have this fee waived when <list of business conditions>. Are you sure you want to apply this fee waiver now?
Yes (link)
No (link)

User: *clicks link "Yes"*

IVA processes fee waiver and returns with confirmed status

IVA: The fee has been waived. Can I help you with anything else?

User closes window

Figure 2: An example conversation showing the user advancing the conversation through conversational link clicks.

User: waive the fee

IVA integration checks against user account and eligibility for fee waiver

IVA: Your account is not eligible for a fee waiver. Customers can have this fee waived when <list of business conditions>.

User: Is there any leniency for a one time exception

IVA: Exceptions can be made in the cases of <list of business conditions>.

IVA integration checks if a live chat operator is available

IVA: In order to do this, we need to connect you to a live chat operator. Wait time is currently 2 minutes.
Connect to Live Chat (link)

User: *clicks link "Connect to Live Chat"*

IVA hands off conversation to live chat operator

Operator: Hello <user>, I see you are asking to waive a fee on your account?

User: Yes. Can you do that for me?

Operator: Give me just a second while I look up your account

Operator: I'm sorry, it looks like your account is not currently eligible for a fee waiver right now.

User: The thing before told me you could make an exception for <business condition>.

User: Can you do that for me?

Conversation continues with live chat operator

Figure 3: An example conversation showing integration points for waiving a fee and escalation to live chat, as well as the use of conversational links of more than one word. See Section 3.1 for a discussion on the various ways to count such click interactions.