International Journal of Human-Computer Studies 119 (2018) 12–27

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

The effect of response complexity and media on user restatement with multimodal virtual assistants

C. Freeman*, I. Beaver*

Verint Intelligent Self Service, 12809 E. Mirabeau Pkwy, Spokane, WA 99216, USA

ARTICLE INFO

Keywords:
Intelligent virtual agents
Multimodal automated assistants
Multimodal conversation design

ABSTRACT

We investigate the occurrence of user restatement when there is no apparent error in Intelligent Virtual Assistant (IVA) understanding in a multimodal customer service environment. Analysis was performed on 2998 sessions with a live customer service IVA deployed on a major airline company website and mobile application. Restatements of a user's previous turn in a conversation are separated into two classes by evidence of dissatisfaction in the IVA response. We consider combinations of response media and linguistic complexity features and determine through detailed statistical analysis which combinations of features should be minimized to improve user comprehension. A discussion of these findings follows demonstrating how the presentation of response in addition to the formulation of response text in a multimodal environment can have an effect on user understanding. Through this analysis, [1] we derive guidelines when crafting responses for designers of multimodal IVAs.

1. Introduction

With the continuing rise of Intelligent Virtual Assistants (IVA) (Marois, 2013) and analysts predicting that human customer service agents will be altogether replaced by IVAs in the near future (Auxbreak, 2015), discovering means to optimize human-computer interactions is necessary. As a company that builds IVAs primarily for customer service, we are interested in cases where there is no apparent misunderstanding on the part of the IVA, but the user continues to restate their query.

Restatements in conversation are problematic because they can break the *principle of least collaborative effort*: both the user and system want the dialogue to be finished as efficiently as possible and with success (Clark and Wilkes-Gibbs, 1986). Restatements are a type of error correction mechanism employed by users when they sense the conversation is not progressing as it should (Aberdeen and Ferro, 2003; Krahmer et al., 1999). Even if the IVA understood the user's query, restatements may follow because the answer was not specific enough, the user did not fully read or understand the response, or the response was presented in a format that did not appeal to the user. In the latter case, as these IVAs are increasingly multimodal (Jiang et al., 2015; Johnston et al., 2014), we theorize that not only is the formulation of the response important, but so is the media it is presented on.

For example, a user may prefer the IVA to answer their query directly in the response text instead of displaying a web page with the answer contained in it. If such a user were to ask an airline IVA the maximum carry-on dimensions, and in response, the IVA displays a web page containing all of the airline's baggage policies instead of directly answering the query in text form, the user may restate the query hoping for a more direct answer. Even though the IVA understood the user's request and displayed the correct page containing the answer, the user restated because they would prefer a direct answer from the IVA. This additional back and forth to resolve a query can lead to user dissatisfaction in the IVA and increase the time required to resolve customer support issues. Therefore, we are motivated to discover the causes of such restatements so that we can design IVA responses in a way that minimizes them.

As users have different preferences in communication, we concede that there is no one "right" way to formulate a response. Following the above example, a different user may actually prefer to see the web page with the entire carry-on policy as it would provide more detailed information. In light of this, we resort to statistical analysis to determine which features of a response are correlated to user restatements so they can be taken under consideration when designing responses. Knowing which features do and do not have an effect on user restatement gives dialog designers the tools to make more informed design choices.

* Corresponding authors.
E-mail addresses: cynthia.freeman@verint.com (C. Freeman), ian.beaver@verint.com (I. Beaver).

[1] This work is an extension of a paper appearing in SLT2016 Beaver and Freeman (2016). Through additional annotations and deeper analysis, many new and valuable conclusions that were missing in the SLT paper are presented here.

https://doi.org/10.1016/j.ijhcs.2018.06.002
Received 28 July 2017; Received in revised form 7 June 2018; Accepted 9 June 2018
Available online 15 June 2018

In this article, we explore interactions with an IVA that communicates with customers over embedded live chat on a large company website as well as the company's mobile application. In both cases, the IVA is exposed on multimodal interfaces that use audio, text, images, User Interface (UI) controls, and web content as media. After tagging numerous features in these interactions, we perform statistical analysis to determine why the IVA response can appear acceptable to a reviewer (where "acceptable" means a reviewer has deemed the IVA response to have answered the user's question) but still fail to satisfy the user. Our contribution is to provide designers of multimodal IVAs guidance for intelligently selecting the media to present information to the user and the linguistic features of response text to minimize in order to reduce confusion.

2. Related works

In this section, we briefly discuss the relevant literature which includes initial work on multiple input modalities, evaluation and error spotting, and analysis of the sequences of user behavior patterns in human and IVA interactions.

The authors in Dybkjaer et al. (2004) provide a comprehensive overview on the evaluation and usability of spoken language dialogue systems (SLDS). The authors state that the inner workings and evaluation of commercial SLDSs are typically kept secret, but note that it is a well-known fact that a high rate of transaction success does not guarantee happy users. Test subjects may judge differently and more positively than real users. Thus, in our work, we not only determine if an IVA's response is acceptable to reviewers, but also consider *why* a user may repeat his or her request multiple times despite a reviewer's positive judgment call.

Although initial studies conclude empirically that multiple input and output modalities go well together for the user (Cohen et al., 1997; Oviatt, 1997; Roth et al., 1997), we determine in our paper that it is not so simple; combinations of media and textual linguistic complexity need to be simultaneously considered for user satisfaction. This is also a very difficult problem given that users may score the same system very differently; one user may prefer related topic links over web content, but another may just prefer text.

Similar research involving the analysis of user behavior for the evaluation of IVAs exists such as in Jiang et al. (2015) where sequences of user behavior patterns (commanding, selecting, or confirming actions) are used to determine user satisfaction. However, user repetition is not analyzed. In addition, although this work involves interactions with intelligent assistants, the authors note that their approach works best on device function tasks (making phone calls, checking calendar events, etc.) and the worst on chat tasks.

Error spotting in conversation between IVAs and humans is well covered in Krahmer et al. (1999), Aberdeen and Ferro (2003), Hirschberg et al. (2001) and Bohus and Rudnicky (2005). However, these works focus on the detection and recovery of errors resulting from misunderstanding or non-understanding on the part of the IVA, not the user. In addition, they do not take into account possible effects of modality as all IVAs involved were SLDSs, and therefore, communicating over a single media. Aberdeen and Ferro (2003), in particular, does highlight user-initiated error correction mechanisms and breaks user restatement into four specific actions: user repeats command, user repeats info, user rephrases info, and user rewords query. There is no discussion of the circumstances that these actions occur in as the data set consisted of only 40 dialogs, and there were few occurrences of each type of restatement.

While restatements can be an indication of conversational error correction, they may be motivated by ignorance on the part of the user to design a proper query (Clark et al., 1991). A restatement as a result of a poor query would be the user trying to make his or her question more specific. For example, the authors in Ehlen and Johnston (2010) discuss the problem of geolocation with Speak4it, a consumer-oriented application that uses multimodal input and output to help users search for local business information. Typically, local search systems assume the device's location for queries when the location is not explicitly stated by the user. The authors discover that users repeat queries and add locations to overcome errors arising from this basic assumption. Some information search systems also offer query suggestions, and this can be another source of repetition. Query suggestions can help users execute searches when it is difficult to formulate a query, especially if the user does not know what kind of vocabulary to use (Niu and Kelly, 2014). The authors in Jansen et al. (2009) discovered that reformulation and system query assistance accounted for almost 45 percent of query reformulation actions. Users may make a series of small queries instead of one large one in the hopes of obtaining the best search results (Teevan et al., 2004). Thus, one cannot assume that all user restatements are automatically detrimental to the conversation.

To the best of our knowledge, ours is the only work involving the direct analysis of combinations of response media and linguistic complexity on user restatement. As mentioned previously, we consider *why* a user may repeat his or her request multiple times despite a reviewer determining that the IVA has correctly answered the user's questions.

3. Methods

Verint Intelligent Self Service designs and builds IVAs on behalf of other companies and organizations, typically for customer service automation. This unique position allows access to a large number of IVA-human conversations that vary widely in scope and language domain. During routine review of these conversations to improve IVA understanding, we frequently noticed that conversations would be flagged for review due to user restatements within them, but upon further inspection, they were not due to errors in IVA understanding. If the IVA correctly understands the user but the query is immediately restated, we reasoned that the response presentation must be somehow unacceptable to the user. To better understand why this phenomenon was occurring, we conducted the following experiment to determine what features of the IVA response can lead to these restatements.

3.1. IVA selection

We reviewed our multimodal IVAs and selected a large international airline IVA for our analysis. The IVA interacts with users on the airline's website and mobile application, providing general travel advice such as flight status information, baggage and security rules, and even helps with the booking process. This particular assistant was selected as user interactions are a good middle ground between an Information Retrieval agent, as it must fetch flight status and travel documents, and a dialogue system, as it contains several tasks such as collecting everything needed to book a flight or transfer award miles between accounts. In addition, it is a very active IVA with a diverse user base. On average, it responds to 4.6 user inputs per second and engages in 115.5 unique conversations per minute with users located around the world. It supports mixed-initiative conversational dialog and can recognize 1230 unique *user intentions*, which, in the context of Natural Language Processing, are interpretations of a user input or action that allows one to formulate the *best* response. The intentions are used as a class label within the IVA. Once the IVA determines the user intention (classification), the response associated with that intention is returned.

The input media supported by this agent are voice, text, UI elements, and web page events. Voice service is provided by the speech application programming interfaces available on the mobile device or browser; therefore, we have no access to Automatic Speech Recognition features or original audio. We are simply given the resulting text translation. Example UI elements may be additional links provided by the agent as suggestions of related topics or drop down selection boxes used for tasks like indicating a country code. Web page events may be clicking on a help icon next to text on a webpage which will launch the