Analysis of User Interaction with Service Oriented Chatbot Systems

Marie-Claire Jenkins, Richard Churchill, Stephen Cox, and Dan Smith

University of East-Anglia
School of Computer Science
Norwich
UK
mcjenkins@gmail.com, mail@richardchurchill.co.uk,
sjc@cmp.uea.ac.uk, djs@uea.ac.uk

Abstract. Service oriented chatbot systems are designed to help users access information from a website more easily. The system uses natural language responses to deliver the relevant information, acting like a customer service representative. In order to understand what users expect from such a system and how they interact with it we carried out two experiments which highlighted different aspects of interaction. We observed the communication between humans and the chatbots, and then between humans, applying the same methods in both cases. These findings have enabled us to focus on aspects of the system which directly affect the user, meaning that we can further develop a realistic and helpful chatbot.

Keywords: human-computer interaction, chatbot, question-answering, communication, intelligent system, natural language, dialogue.

1 Introduction

Service oriented chatbot systems are used to enable customers to find information on large complex websites, which are difficult to navigate. Norwich Union [1] is a very large insurance company offering a full range of insurance products. Their website attracts 50,000 visits a day, with over 1,500 pages making up the website. Many users find it difficult to discover the information they need from website search engine results, the site being saturated with information. The service-oriented chatbot acts as an automated customer service representative, giving natural language answers, and offering more targeted information in the course of a conversation with the user. This virtual agent is also designed to help with general queries regarding products. This is a potential solution for online business, as it is time saving for customers, and allows the company to have an active part in the sale.

Internet users have gradually embraced the internet since 1995, and the internet itself has changed a great deal since then. Email and other forms of online communication such as the messenger programs, chat rooms and forums have become widely spread and accepted. This would indicate that the methods of communication

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 76–83, 2007.

involving typing are quite well integrated in online user habits. A chatbot is presented in the same way. Programs such as Windows "messenger" [2] involve a text box for input and another where the conversation is displayed. Despite the simplicity of this interface, experiments have shown that people are unsure as to how to use the system.

Despite the resemblance to the messenger system, commercial chatbots are not widespread at this time, and although they are gradually being integrated in large company websites, they do not hold a prominent role there, being more of an interactive tool or a curiosity rather than a trustworthy and effective way to go about business on the site. Our experiments show that there is an issue with the way that people perceive the chatbot. Many cannot understand the concept of talking to a computer, and so are put off by such a technology. Others do not believe that a computer can fill this kind of role and so are not enthusiastic, largely due to disillusionment with previous and existing telephone and computer technology. Another reason may be that they fear that they may be led to a product by the company, in order to encourage them to buy it. In order to conduct a realistic and useful dialogue with the user, the system must be able to establish rapport, acquire the desired information and guide the user to the correct part of the website, as well as using the appropriate language and having a human-like behaviour. Some systems, such as ours, also display a visual representation of the system in the form of a picture (or an avatar), which is sometimes animated in an effort to be more human-like and engaging. Our research however shows that this is not of prime importance to users. Users expect the chatbot to be intelligent, and expect them to also be accurate in their information delivery and use of language.

In this paper we describe an experiment which involved testing user behaviour with chatbots and comparing this to their behaviour with a human. We discuss the results of this experiment and the feedback from the users.

Our findings suggest that our research must not only consider the artificial intelligence aspect of the system which involves information extraction, knowledgebase management and creation, and utterance production, but also the HCI element, which features strongly in these types of system.

2 Description of the Chatbot System

The system which we named KIA (Knowledge Interaction Agent) was built specifically for the task of monitoring human interaction with such a system. It was built using simple natural language processing techniques. We used the same method used in the ALICE [3] social chatbot system which involves seeking for patterns in the knowledgebase using the AIML technique [4]. AIML (artificial intelligence markup language) is a method based on XML. The AIML method uses templated to generate a response in as natural a way as possible. The templates are populated with patterns commonly found in the possible responses. The keywords are migrated into the appropriate pattern identified in the template. The limitation of this method is that there is not enough variety in the possible answers. The knowledge base was drawn from the Norwich Union website. We then manually corrected errors and wrote in a "chat" section to the knowledge base from which the more informal, conversational

utterances could be drawn. The nouns and proper nouns served as identifiers for the utterances and were initiated by the user utterance.

The chatbot was programmed to deliver responses in a friendly, natural way. We incorporated emotive-like cues such as using exclamation marks, interjections, and utterances which were constructed so as to be friendly in tone. "Soft" content was included in the knowledge base giving information on health issues like pregnancy, blood pressure and other such topics which it was hoped would be of personal interest to users. The information on services and products was also delivered using as far as possible the same human-like type of language as for the "soft" content language.

The interface was a window consisting of a text area to display the conversation as it unfolded and a smaller text box for the user to enter text. An "Ask me" button allowed for utterances to be submitted to the chatbot. For testing purposes the "section" link was to be clicked on when the user was ready to change the topic of the discussion as the brief was set in sections. We also incorporated the picture of a woman smiling in order to encourage some discussion around visual avatars. The simplicity of the interface was designed to encouraged user imagination and discussion; it was in no way presented as an interface design solution.

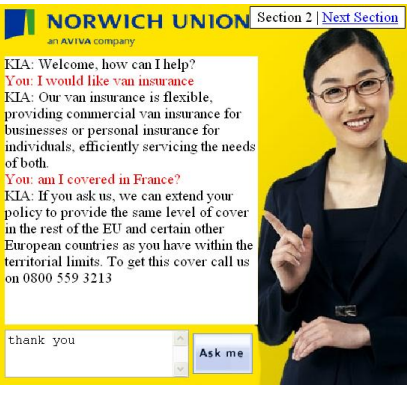

Fig. 1. The chatbot interface design

3 Description of the Experiment and Results

Users were given several different tasks to perform using the chatbot. They conversed with the system for an average of 30 minutes and then completed a feedback questionnaire which focused on their feelings, and reactions to the experience. The same framework was used to conduct "Wizard of Oz" experiments to provide a benchmark set of reactions, in which a human took the customer representative role instead of the chatbot. We refer to this chatbot as the "Human chatbot" (HC).

We conducted the study on 40 users with a full range of computer experience and exposure to chat systems. Users were given a number of to fulfill using the chatbot. These tasks were formulated after an analysis of Norwich Union's customer service system. They included such matters as including a young driver on car insurance, traveling abroad, etc…The users were asked to fill in a questionnaire at the end of the

test to give their impressions on the performance of the chatbot and volunteer any other thoughts. We also prompted them to provide feedback on the quality and quantity of the information provided by the chatbot, the degree of emotion in the responses, whether an avatar would help, whether the tone was adequate and whether the chatbot was able to carry out a conversation in general.

We also conducted an experiment whereby one human acted as the chatbot and another acted as the customer and communication was spoken rather than typed. We collected 15 such conversations. The users were given the same scenarios as those used in the human-chatbot experiment. They were also issued with same feedback forms.

3.1 Results of the Experiments

The conversation between the human and HC flowed well as would be expected, and the overall tone was casual but business like on the part of the HC, again as would be expected from a customer service representative. The conversation between chatbot and human was also flowed well, the language being informal but business-like.

1.1 User language

- Keywords were often used to establish the topic clearly such as "I want car insurance", rather than launching into a monologue about car problems. The HC repeated these keyword, often more than once in the response. The HC will also sometimes use words in the same semantic field (e.g. "travels" instead of "holiday").
- The user tends to revert to his/her own keyword during the first few exchanges but then uses the words proposed by the HC. Reeves and Nass [5] state that users respond well to imitation. In this case the user comes to imitate the HC. There are sometimes places in the conversation where at times the keyword is dropped altogether such as "so I'll be covered, right?". This means that the conversation comes to rely on anaphora. In the case of the chatbot-human conversation, the user was reluctant to repeat keywords (perhaps due to the effort of re-typing them) and relied very much on anaphora, which makes the utterance resolution more difficult.

The result of this was that the information provided by the HC was at times incomplete or incorrect and at times there was no answer given at all. The human reacted well to this and reported no frustration or impatience. Rather, they were prepared to work with the HC to try and find the required information.

1.2 User reactions

- Users did however report frustration, annoyance, impatience with the chatbot when it was also unable to provide a clear response or a response at all. It was interesting to observe a difference in users' reaction to similar responses from the HC and the chatbot. If neither was unable to find an answer to their query after several attempts, users became frustrated. However this behaviour was exhibited more slowly with the HC than with the chatbot. This may be because users were aware that

they were dealing with a machine and saw no reason to feign politeness, although we do see evidence of politeness in greetings for example.

1.3 Question-answering

- The HC provided not only an answer to the question, where possible, but also where the information was located on the website and a short summary of the relevant page. The user reported that this was very useful and helped them be further guided to more specific information.
- The HC was also able to pre-empt what information the user would find interesting, such as guiding them to a quote form when the discussion related to prices for example, which the chatbot was unable to do. The quantity of information was deemed acceptable for both the HC and the chatbot. The chatbot gave the location of the information but a shorter summary than that of the HC.
- Some questions were of a general nature, such as "I don't like bananas but I like apples and oranges are these all good or are some better than others?" which was volunteered by one user. As well as the difficulty of parsing this complex sentence, the chatbot needs to be able to draw on real-world knowledge of fruit, nutrition etc…To answer such questions requires the use of a large knowledgebase of real-world knowledge as well as methods for organizing and interpreting this information.
- The users in both experiments sometimes asked multiple questions in a single utterance. This led both the chatbot and the HC to be confused or unable to provide all of the information required at the same time.
- Excessive information is sometimes volunteered by the user, e.g. as explaining how the mood swings of a pregnant wife are affecting the fathers' life at this time. A machine has no understanding of these human problems and so would need to grasp these additional concepts in order to tailor a response for the user. This did not occur in the HC dialogues. This may be because users are less likely to voice their concerns to a stranger, than an anonymous machine. There is also the possibility that they were testing the chatbot.

Users may also feel that giving the chatbot the complete information required to answer their question in a single turn is acceptable to a computer system but not acceptable to a human, using either text or speech.

1.4 Style of interaction

- Eighteen users found the chatbot answers succinct and three long-winded. Other users described them as in between, not having enough detail in them or being generic. The majority of users were happy with finding the answer in the sentence rather than in the paragraph as Lin [6] found during his experiments with encyclopedic material. In order to please the majority of users it may be advisable to include the option of finding out more about a particular topic. In the case of the HC, the responses were considered to be succinct and containing the right amount of information. However some users reported that there was too much information.

- Users engaged in chitchat with the chatbot. They thank it for its time and also sometimes wish it “Good afternoon” and “Good morning”. Certain users tell the chatbot that they are bored with the conversation. Others tell the system that this ”feels like talking to a robot”. Reeves and Nass [5] found, that the user expects such a system to have human qualities. Interestingly the language of the HC was also described as “robotic” at times by the human. This may be due to the dryness of the information being made available; however it is noticeable that the repetition of keywords in the answers contributes to this notion.□□□

3.2 Feedback Forms

The feedback forms from the experiment showed that users described in an open text field the tone of the conversation with the chatbot as ”polite”, ”blunt”, ”irritating”, ”condescending”, “too formal”, ”relaxed” and ”dumb”. This is a clear indication of the user reacting to the chatbot. The chatbot is conversational therefore they expect a certain quality of exchange with the machine. They react emotionally to this and show this explicitly by using emotive terms to qualify their experience. The HC was also accused of this in some instances.

The users were asked to rate how trustworthy they found the system to be using a scale of 10 for very trustworthy to 0 for not trustworthy. The outcome was an average rating of 5.80 out of 10. Two users rated the system as trustworthy even though they rated their overall experience as not very good. They stated that the system kept answering the same thing or was poor with specifics. One user found the experience completely frustrating but still awarded it a trust rating of 8/10. The HC had a trustworthiness score of 10/10.

3.3 Results Specific to the Human-Chatbot Experiment

Fifteen users volunteered without elicitation alternative interface designs. Ten of these all included a conversation window, a query box, which are the core components of such a system. Seven included room for additional links to be displayed. Four of the drawings include an additional window for the inclusion of ”useful information”. 1 design included space for web links. One design included disability options such as the choice of text color and font size to be customizable. 5 designs included an avatar. One design included a button for intervention by a human customer service representative.

A common feature suggested was to allow more room for each of the windows and between responses so that these could be clearer. The conversation logs showed many instances of users attacking the KIA persona, which was in this instance the static picture of a lady pointing to the conversation box. This distracted them from the conversation.

3.4 The Avatar

Seven users stated that having an avatar would enhance the conversation and would prove more engaging. Four users agreed that there was no real need for an avatar as the emphasis was placed on the conversation and finding information. Ten stated that

having an avatar present would be beneficial, making the experience more engaging and human-like. Thirteen reported that having an avatar was of no real use. Two individuals stated that the avatar could cause "embarrassment", and may be "annoying". Two users stated that they thought that having a virtual agent would not help actually included them in their diagrams.

When asked to compare their experience with that of surfing the website for such information, the majority responded that they found the chatbot useful. One user compared it to Google and found it to be "no better". Other users stated that the system was too laborious to use. Search engines provide a list of results which then need to be sorted by the user into useful or not useful sites. One user stated that surfing the web was actually harder but it was possible to obtain more detailed results that way. Others said that they found it hard to start with general keywords and find specific information. They found that they needed to adapt to the computer's language. Most users found it to be fast and efficient and generally just as good as a search engine although a few stated that they would rather use the search engine option if it was available. One user clearly stated that the act of asking was preferable to the act of searching. Interestingly a few said that they would have preferred the answer to be included in a paragraph rather than a concise answer.

The overall experience rating ranged from very good to terrible. Common complaints were that the system was frustrating, kept giving the same answers, and was average and annoying. On the other hand some users described it as pleasant, interesting, fun, and informative. Both types of user gave similar accounts and ratings throughout the rest of the feedback having experienced the common complaints.

The system was designed with a minimal amount of emotive behavior.

It used exclamation marks at some points, and more often than not simply offered sentences available on the website, or which were made vaguely human-like. Users had strong feedback on this matter calling the system "impolite", "rude", "cheeky", "professional", "warm", and "human-like". One user thought that the system had a low IQ. This shows that users do expect something which converses with them to exhibit some emotive behavior. Although they had very similar conversations with the system, their ratings varied quite significantly. This may be due to their own personal expectations. The findings correlate with the work of Reeves and Nass [5]: people are associating human qualities to a machine. It is unreasonable to say that a computer is cheeky or warm for example, as it has no feelings.

Table 1. Results of the feedback scores from the chatbot –human experiment

Experience	0.46	Useful answers	0.37
Tone	0.37	Unexpected things	0.2
Turn-taking	0.46	Better than site surfing	0.43
Links useful	0.91	quality	0.16
emotion	0.23	Interst shown	0.33
Conversation rating	0.58	Simple to use	0.7
Succinct responses	0.66	Need for an avatar	0.28
Clear answers	0.66		

Translating all of the feedback into numerical values between 0 and 1, using 0 as a negative answer, 0.5 as a middle ground answer and 1 as a positive answer, we can clearly see the results. The usefulness of links was voted very positive with a score of 0.91, and tone used (0.65), sentence complexity (0.7), clarity (0.66) and general conversation (0.58) all scored above average. The quality of the bot received the lowest score at 0.16.

4 Conclusion

The most important finding from this work are: that users expect chatbot systems to behave and communicate like humans. If the chatbot is seen to be "acting like a machine", it is deemed to be below standard. It is required to have the same tone, sensitivity and behaviour than a human but at the same time users expect it to process much more information than the human. It is also expected to deliver useful and required information, just as a search engine does. The information needs to be delivered in a way which enables the user to extract a simple answer as well as having the opportunity to "drill down" if necessary. Different types of information need to be volunteered such as the URL where further information or more detailed information can be found, the answer, and the conversation itself. The presence of "chitchat" in the conversations with both the human and the chatbot show that there is a strong demand for social interaction as well as a demand for knowledge.

5 Future Work

It is not clear from this experiment whether an avatar can help the chatbot appear more human-like or make for a stronger human-chatbot relationship. It would also be interesting to compare the use of search engines to that of the chatbot.

It would be interesting to compare the ease of use of the chatbot with a conventional search engine. Many users found making queries in the context of a dialogue useful, but the quality and precision of the answers returned by the chatbot may be lower than what they could obtain from a standard search engine. This is a subject for further research.

Acknowledgements. We would like to thank Norwich Union for their support of this work.

References

1. Norwich Union, an AVIVA company: http://www.norwichunion.com
2. Microsoft Windows Messenger: http://messenger.msn.com
3. Wallace, R.: ALICE chatbot, http://www.alicebot.org
4. Wallace, R.: The anatomy of ALICE. Artificial Intelligence Foundation
5. Reeves, B., Nass, C.: The media equation: how people treat computers, television and new media like real people and places. Cambridge University press, Cambridge (1996)
6. Lin, J., Quan, D., Bakshi, K., Huynh, D., Katz, B., Karger, D.: What makes a good answer? The role of context in question-answering. INTERACT (2003)