

# How Do People Talk with a Virtual Philosopher: Log Analysis of a Real-World Application

Xuan Wang and Ryohei Nakatsu

Keio-NUS CUTE Center, Interactive and Digital Media Institute  
NUS Graduate School for Integrative Sciences and Engineering  
National University of Singapore  
{wangxuan, e1enr}@nus.edu.sg

**Abstract.** Conversation with computers is an important form of human computer interaction. Inappropriately designed conversational agents can easily lead to unsatisfying user experience and even frustration, and this is especially true when the application is deployed in the real world. Currently, research on casual non-task oriented systems and our understanding in how people interact with such agents are still limited. To gain more insights on this issue, we carried out both quantitative and qualitative content analysis of conversation logs collected from a real-world application, featuring a non-task oriented conversational agent as a virtual philosopher. We construct a taxonomy of user utterances to the agent and discuss a few strategies that an agent might employ to provide a better user experience.

**Keywords:** non-task oriented conversational agent, log analysis, virtual philosopher.

## 1 Introduction

A conversational agent (CA) is a computer system that is able to carry out natural language conversation with a human user. As an intuitive interface, CA has been used in a variety of ways, such as helping people to accomplish tasks, obtain information and learn about things [1–3]. When a CA is deployed in the real world, it can often face user utterances that are untrained before. The ability to handle the so-called “out-of-domain” questions is key to the lifelikeness of an agent [4], and coherence/appropriateness of the answer is important in maintaining engagement from the user [5], as user frustration can be caused otherwise. This is especially true for non-task oriented CAs, where there is no clear common user-agent goal [6], which means the user utterances are largely unbounded. As a result, common dialogue strategies in task-oriented CAs are no longer effective. In order to develop good strategies for non-task oriented CA to deal with open conversation, we need to have good understanding of how users talk to the agent first. This paper attempts to shed some light on this issue by presenting a qualitative and quantitative analysis of the conversation logs between users and a non-task oriented conversational agent in a real world

setting. We situate our study in the case of a virtual philosopher, which is a conversational agent that mimics Confucius, a famous philosopher in ancient China.

## 2 Related Work

As compared to task-oriented agents where there is a well-defined task or well-circumscribed body of information sought by the users, non-task oriented conversational agents are designed to simulate human conversation for entertainment, learning, etc., usually within a certain domain. Some examples are virtual museum guides [2, 7–9], virtual historical persons[4, 10]. When evaluating such systems, besides user’s subjective report through questionnaires, interviews, log analysis is a very informative way to understand the actual performance of an agent, and it can also help us to have a good understanding of how users behave when interacting with the system. In [11], Aggarwal et al. presents a large corpus of spoken conversation with two virtual museum guides, Ada and Grace, where common questions are identified and added to the characters’ training data to improve their conversation ability. Kopp et al. carried out a more detailed content analysis on the conversation logs during the museum guide Max’s employment in the Nixdorf Museum [2]. They categorized the utterances into “Greeting”, “Farewell”, “Flaming”, “Feedback to agent”, “Questions”, “Answers”, “Request to do something”, and showed that people use human-like communication strategies when interacting with the system [2]. Robinson et al. also analyzed the content of dialogues carried out between visitors and the virtual character Sgt Blackwell in the Cooper Hewitt Museum [12]. They categorized user utterances into user initiated and reacted utterances, with finer categories within each category. By presenting a very detailed description of the utterance content, they provide a good picture of user’s preference, which points out directions for the expansion of the agent’s knowledge base. In this paper, we conduct a similar analysis in the context of a virtual philosopher, to see whether users have similar behavior and expectations, and also inform our design to handle the user utterances appropriately.

## 3 Content Analysis

### 3.1 Data Collection

The application used for data collection is called iSage Confucius. It is a conversational agent that mimics Confucius, the renowned philosopher and educator in ancient China. It models the problem of finding answers as an information retrieval task, and uses a combination of scripts and semantic analysis methods. More details of the system can be found in [13]. We published the application in Google Play Store for free download since September 2011. All conversations are logged anonymously, and users are notified the first time they open the application. Each user is identified by a UUID (Universally Unique Identifier). As

of the end of 2012, the application has collected over 1500 conversation turns. A turn is defined here as one utterance from the user and one response from the virtual Confucius. We take the log files collected during the year of 2012 for the analysis, with a total of 1029 turns coming from of 115 different user IDs (which means on average 8.95 turns per user, assuming each user ID corresponds to one user).

### 3.2 Method

Following the method in [2], we also took a data-driven approach in the content analysis of log files. Different from [2], we distinguish the categories based on the nature and content of user utterances. The rationale for such classification is that the agent needs to be equipped with different techniques to respond to user utterances of different type and scope. Knowing the distribution of utterances can inform algorithm design and improve the agent’s ability in carrying out open dialogues. To derive the classification scheme, two coders skimmed through all the log data and jointly came out with the initial categories. Then the coders coded the first 100 turns, discussed and modified the categories if necessary. After the categories were finalized, the coders independently coded the first 200 turns, and the inter-coder agreement was checked. A Cohen’s Kappa value of 0.642 was computed, indicating satisfactory inter-coder agreement. The coders then proceeded to code the rest of the log files, each responsible for half of the data.

### 3.3 Classification Scheme

The resulting categories, selected examples and their frequencies are shown in Table 1. There are altogether 6 high-level categories. Some of the high-level categories are further classified into lower-level categories to reveal more details. The meaning of each of the high-level category is explained as follows:

1. *Confucius and Confucianism*: This category contains user utterances that are directly related to the agent. This is an important part of the intended usage of the application - to know more about Confucius and Confucianism through conversation. Many of the conversational agents reviewed in the related work section are designed mainly to handle this kind of questions.
2. *Factoid*: This category contains factoid questions that have a definite answer (which have been the subject of study in many question answering systems, e.g. [14]). Such questions are not related to the agent and are considered out-of-domain.
3. *Subjective*: As opposed to factoid questions, this category contains questions that do not have a definite answer. This is also an important part of the intended usage of the application - to get Confucius’s subjective opinion on things. For utterances that contain only a single noun or noun phrases, we assume that users expect Confucius can give some comments on it. So we include such utterances in this category as well.

4. *Conversational*: This category contains user utterances which serve the purpose of starting a conversation and maintaining the flow. Such utterances are a main indicator of social interaction.
5. *Nonsense*: This category contains utterances that do not have any meaning or contribution to the conversation, as well as obscene language.
6. *Others*: Two of the utterances about the app itself turn out to fall out of the above 5 categories, so we put them in a separate category called others.

**Table 1.** A taxonomy of user utterances in the iSage corpus

No.	Categories	Examples	Count (%)
1	<b>Confucius and Confucianism</b>		<b>172 (16.7%)</b>
	1A. Biographical info	Were you ever married?	168 (16.3%)
	1B. Domain specific terms	What is the Way?	4 (0.4%)
2	<b>Factoid</b>	How many miles is it to Jupiter.	<b>47 (4.6%)</b>
3	<b>Subjective</b>		<b>546 (53.1%)</b>
	3A. Definition/explanation of abstract concepts	What is wisdom?	115 (11.2%)
	3B. Reasons	Why is life so complicated?	36 (3.5%)
	3C. Judgments, Predictions	Can machines love?	114 (11.1%)
	3D. Advice	How can I be successful?	105 (10.2%)
	3E. Statements	I am depressed.	61 (5.9%)
	3F. Single word, word phrase	love	115 (11.2%)
4	<b>Conversational</b>		<b>147 (14.3%)</b>
	4A. Feedback to agent	This is amazing.	91 (8.8%)
	4B. Greeting, goodbye	Hi.	56 (5.4%)
5	<b>Nonsense</b>		<b>115 (11.2%)</b>
	5A. Mistyped, meaningless	How many studenrs you have?	95 (9.2%)
	5B. Pornographic/abuse	Obscene words	20 (1.9%)
6	<b>Others</b>		<b>2 (0.2%)</b>
		What is the point of this app?	
		Can I go to main menu?	
<b>Total</b>			<b>1029 (100%)</b>

## 4 Discussions

From the frequency count of the categories, we can see, in general, users are willing to suspend their disbelief and act like talking to the real Confucius, asking questions regarding Confucius's life (16.3%), posing difficult philosophical questions, revealing their problems and seeking advice (53.1%). Conversational utterances can also be observed (14.3%). In accordance with findings in other studies [2, 12], we also observed flaming (obscene language, random keystrokes, meaningless utterances)(11.2%), or testing (by asking factoid questions) (4.6%).

From user's replies, we can have a partial idea of whether they enjoyed using the system. Some users appear to be satisfied with the answer given by the virtual Confucius, as we can observe them replying "*Oh I see.*", "*This is amazing.*". Enjoyment can also be observed: "*Hahaha*", "*Funny*". On the other hand, we also observe frustrations when the system was not able to give a good answer, or completely misunderstood user's question. Users expressed their unhappiness using utterances like "*You are not making any sense*". It is observed that usually after getting a few irrelevant responses, users would stop using the system. Currently, when the agent is not able to find any good answer, its strategy is to honestly tell that he does not understand and ask the user to rephrase the question. The log analysis reveals that it is not effective. Very few people would rephrase the question as the agent requested.

Users of different CAs may have differing needs and expectations [12]. Different from the results found in the case of virtual museum guides [2] and virtual soldier [12], where users tend to ask questions regarding the agent itself, we found that for a virtual philosopher, users tend to share more about their own experiences or problems with the agent, expecting the agent to clear their doubts or provide guidance. It seems that the identity of the agent plays a large role in the types of questions users would ask. As the main purpose of using a virtual philosopher system is information-seeking or advice-seeking, traditional dialogue acts are less useful in this case. Therefore, when designing strategies for answering, it may be a good choice to refer to the taxonomy constructed from real user data. It is important for the agent to know the type of input utterance, so that it can either invoke appropriate answer finding technique, or provide an appropriate reply when no answer can be found in the knowledge base. A classifier can be trained to classify the utterances using the collected data, and corresponding strategies can be designed to respond to these utterances. For example, to improve the agent's ability to handle category 1 utterances, we can either ask domain experts to expand the knowledge base, or look for information online using QA engines. Category 2 questions can also be handled by consulting existing QA systems, or designing appropriate default responses to inform the user that it is not supposed to answer this kind of questions. Category 3 utterances are more difficult - users tend to share their own experience, or ask about subjects of their own interest, so we cannot predict the questions and prepare answers beforehand. Therefore, we need better semantic and pragmatic understanding to select the best answer from the repository of candidate answers. Category 4 utterances is also important, as it is essential in making the agent human-like. A simple and good solution is to use standard chatbot's techniques.

We believe that the taxonomy of utterances derived from real user data can help us to have a better understanding of how users talk to a virtual philosopher system in an unrestricted setting, and such understanding can inform the algorithm design for a more human-like agent. In future work, we plan to implement the discussed strategies and conduct experimentally controlled studies to find the optimal strategy for handling open conversation with a virtual philosopher.

**Acknowledgments.** This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

## References

1. Crockett, K., O'Shea, J., Bandar, Z.: Goal Orientated Conversational Agents: Applications to Benefit Society. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS, vol. 6682, pp. 16–25. Springer, Heidelberg (2011)
2. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide – design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
3. D'Mello, S., Graesser, A., King, B.: Toward Spoken Human-Computer Tutorial Dialogues. *Human-Computer Interaction* 25(4), 289–323 (2010)
4. Marinelli, D., Stevens, S.: Synthetic interviews: the art of creating a 'dyad' between humans and machine-based characters. In: Proceedings 1998 IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications, IVTTA 1998 (Cat. No.98TH8376), pp. 43–48. IEEE (1998)
5. Artstein, R., Cannon, J., Gandhe, S., Gerten, J., Henderer, J., Leuski, A., Traum, D., Rey, M.: Coherence of Off-Topic Response for a Virtual Character. In: Proceedings of the 26th Army Science Conference (2008)
6. Bernsen, N.O., Dybkjar, L.: Evaluation of spoken multimodal conversation. In: Proceedings of the 6th International Conference on Multimodal Interfaces - ICMI 2004, p. 38. ACM Press, New York (2004)
7. Swartout, W., et al.: Ada and grace: Toward realistic and engaging virtual museum guides. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 286–300. Springer, Heidelberg (2010)
8. Bell, L., Gustafson, J.: Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system. In: EUROSpeech 2003, pp. 613–616 (2003)
9. Richards, D.: Agent-based museum and tour guides. In: Proceedings of The 8th Australasian Conference on Interactive Entertainment Playing the System - IE 2012. ACM Press, New York (2012)
10. Bernsen, N., Dybkjæ, R.L.: Domain-oriented conversation with H. C. Andersen. In: Proc. of the Workshop on Affective Dialogue Systems, pp. 142–153. Springer (2004)
11. Aggarwal, P., Artstein, R., Gerten, J., Katsamanis, A., Narayanan, S., Nazarian, A., Traum, D.: The Twins Corpus of Museum Visitor Questions. In: LREC 2012, vol. 3, pp. 2355–2361 (2012)
12. Robinson, S., Traum, D., Ittycheriah, M., Henderer, J.: What would you ask a Conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. In: LREC 2008 Proceedings, pp. 1125–1131 (2008)
13. Wang, X., Khoo, E.T., Siriwardana, S., Iroshan, H., Nakatsu, R.: Philosophy Meets Entertainment: Designing an Interactive Virtual Philosopher. In: Herrlich, M., Malaka, R., Masuch, M. (eds.) ICEC 2012. LNCS, vol. 7522, pp. 100–113. Springer, Heidelberg (2012)
14. Voorhees, E.M., Dang, T.H.: Overview of the TREC 2005 Question Answering Track. In: Proc. of TREC 2005 (2005)