Direct and Mediated Interaction with a Holocaust Survivor

Bethany Lycan and Ron Artstein

Abstract The *New Dimensions in Testimony* dialogue system was placed in two museums under two distinct conditions: docent-led group interaction, and free interaction with visitors. Analysis of the resulting conversations shows that docent-led interactions have a lower vocabulary and a higher proportion of user utterances that directly relate to the system's subject matter, while free interaction is more personal in nature. Under docent-led interaction the system gives a higher proportion of direct appropriate responses, but overall correct system behavior is about the same in both conditions because the free interaction condition has more instances where the correct system behavior is to avoid a direct response.

1 Introduction

Conversational agents that serve as museum exhibits interact with two populations: museum visitors, and museum docents who may show the system to visitors. Sometimes, docent-led interactions can be useful in initial stages of deployment, before the system is ready for interacting with visitors. For example, the Virtual Museum Guides at the Museum of Science in Boston (Swartout et al, 2010) were initially operated by docents; the system was later extended to enable direct interaction with the public, though even this direct interaction was partly constrained by posting a list of suggested questions, which accounted for 30% of all visitor utterances (Traum et al, 2012). But when a system is designed for direct public interaction, several questions arise about the role of museum docents. Is docent-mediated interaction helpful? Does it enhance the visitor experience or detract from it? And how should the needs of museum docents affect the design of the system?

Bethany Lycan
California State University, Long Beach, e-mail: lycanbg@yahoo.com

Ron Artstein
USC Institute for Creative Technologies, Playa Vista, California e-mail: artstein@ict.usc.edu

This paper presents a natural experiment, where the same dialogue system was placed in two museums, under two different conditions – docent-led and open to the public. *New Dimensions in Testimony* (Traum et al, 2015a,b) is a dialogue system that replicates conversation with Holocaust survivor Pinchas Gutter. Users talk to a persistent representation of Mr. Gutter presented on a large (almost life-size) video screen, and the system selects and plays pre-recorded video clips of the survivor in response to user utterances. The result is much like an ordinary conversation between the user and the survivor. The system was designed from the outset for direct interaction: an extensive testing process resulted in a library of over 1600 video clips, including responses to the most common user questions and well as utterances designed for maintaining coherence and continuity. The system was installed in the Illinois Holocaust Museum and Education Center in Skokie on March 4, 2015, and is still in use as of the time of this writing. Due in part to properties of the physical location and in part to the museum's choice, the exhibit in Illinois is primarily docent-led. Between April 24, 2016 and September 5, 2016, a copy of the system was also installed in the United States Holocaust Memorial Museum in Washington, DC (USHMM). The exhibit at USHMM was built as a booth where individual museum visitors could come up to the system and start a conversation. Comparing the system's operation in the two installations allows us to study both differences in how users interact with the system, and how the system performs in these two distinct settings.

2 Method

The analysis is based on interaction logs from the museums, which contain time-stamped user utterance texts and system response IDs and texts. The user utterance texts are automatic transcriptions by Google Chrome ASR[1], as logged by the system in real time; previous testing has shown that this ASR has a word error rate of about 5% on this domain (Traum et al, 2015a), so we relied on the ASR output for analysis rather than transcribe the recorded audio. System response IDs identify the video clip used for each response, and the system response texts are the words spoken by Mr. Gutter in the video clip. A sample of the interaction logs from contiguous time periods was selected for quantitative analysis, covering about 2000 user-system interchanges from each museum (Table 1).

Table 1 Data selected for quantitative analysis

Museum	Dates	User Utterances	System Responses
Illinois	2015-09-16 – 2015-11-20	2030	2003
USHMM	2016-05-09 – 2016-06-08	2025	1995

[1] https://www.google.com/intl/en/chrome/demos/speech.html

While the systems in Illinois and at USHMM are identical, the settings are different. Interaction with *New Dimensions in Testimony* in Illinois is primarily in groups, where communication between visitors and the system is mediated by a museum docent. In a typical interaction the docent will demonstrate a conversation with the survivor, and relay questions from the audience. In contrast, visitors at USHMM talked directly into the microphone connected to the system. Museum docents were available to give background information and offer suggestions in case a visitor needed help, but the docents were specifically instructed to not interfere with the user's conversation and not to make suggestions unless absolutely necessary. In order to encourage natural conversation, users were not provided with any written examples of things they might say to the system. At both museums, the only data recorded were direct audio inputs to the system and the system's action in response; interactions between the docents and the visitors were not recorded. (A separate evaluation observed a sample of interactions and collected user feedback, but these data are not available to us.)

The interaction logs were inspected manually to identify common patterns of interaction. Lexical differences between the user utterances in the two museums were analyzed using the *AntConc* software (Anthony, 2014). User utterances were also annotated by the first author to code consecutive repetitions of (essentially) the same question.

The system's responses were annotated for appropriateness by the first author according to the following scheme. *On-topic responses* are selected by the system when it believes it has found an appropriate response to the user utterance; these were rated on a scale of 1–4, with 1 being irrelevant and 4 being relevant. *Off-topic responses* are utterances used by the system to indicate non-understanding when it is not able to identify a direct response to the user's utterance (for example "please repeat that" or "I don't understand"); these were coded as to whether the decision to use an off-topic was correct (the system does not have a direct response) or an error (the system has a direct response which was not identified).

3 Results

Interactions from USHMM show users relating to Mr. Gutter on a more personal level. Visitors introduce themselves (e.g. *My name is Sheila, I have three granddaughters with me…*), apologize conversationally (*I'm sorry to interrupt you*), and react emotionally to stories told by the survivor (*I'm so sorry to hear that*). In the one instance at USHMM where the survivor asks the visitor why they came to listen to him, the user replies with a long and detailed answer.

The user utterances in Illinois are more tailored to Mr. Gutter's story than those at USHMM. Table 2 shows the most frequent user utterances in the sample from each museum. Many are the same; of those that differ, the questions in Illinois relate more to specific aspects of Mr. Gutter's story, while those asked at USHMM are more interpersonal in nature.

Table 2 Most frequent user utterances (boldfaced utterances appear in only one column)

Illinois		USHMM	
Utterance	N	Utterance	N
Testing	24	Where were you born?	19
Hello	22	How old are you?	17
How old are you?	19	How did you survive?	15
Where were you born?	17	Hello	14
Hello Pincus[a]	16	**Where do you live now?**	14
How are you?	14	Thank you	13
How many languages do you speak?	13	**What happened to your family?**	12
Tell us about your childhood	10	Good morning	10
Can you hear me?	9	**Can you tell us about yourself?**	9
What was life like in the Warsaw Ghetto?	8	How are you today?	8
How did you survive?	8	Do you have any regrets?	8
Do you have any regrets?	8	How are you?	8
Good morning	7	**What's your name?**	8
What was life like before the war?	7	**Hi Pincus**[a]	7
Why didn't the Jews fight back?	7	Hello Pincus[a]	7
How are you today?	6	**Tell me a joke**	6
How did you meet your wife?	6	**What's your favorite color?**	6
Do you have children?	6	**What is your name?**	6
Thank you	6	**Can you tell me a joke?**	6

[a]Mistranscription of Pinchas

A similar observation can be made by looking at the words whose frequency differs the most between the two data sets (Table 3). The top words – *us* in Illinois and *I* at USHMM – reflect the difference between a docent-led group setting and an individual interaction. Other frequent words from Illinois reflect the docents' familiarity with Mr. Gutter's specific story (*Majdanek*, *liberation*, *England*, *Warsaw*), whereas USHMM shows higher relative frequency for *concentration [camp]*, a generic descriptor associated with the Holocaust, as well as words that connect on a personal level (*joke*, *favorite*). Lexical variation is higher at USHMM, with a vocabulary size of 1,386 and density of 10.5 (total tokens divided by vocabulary size), while Illinois

Table 3 Words which differ the most in frequency between the two corpora; values are the "Keyness" feature from *AntConc* (Anthony, 2014)

Illinois	us	165.9	danik[a]	39.7	war	38.4	liberation	36.7	hear	31.4
	what's	30.8	tell	26.9	testing	26.4	didn't	24.9	sing	21.6
	England	20.6	why	20.0	life	19.4	about	19.0	after	18.5
	Warsaw	17.1	I'm	16.2	rap	16.2	cats	14.7	it's	14.7
USHMM	I	54.7	it	49.2	concentration	37.8	kosher[b]	27.7	yourself	23.8
	don't	23.7	me	23.2	joke	22.2	is	21.2	if	20.1
	or	18.1	much	16.9	and	16.8	favorite	16.4	now	16.1
	that	15.5	they	13.8	summary	13.7	eat	13.6	experienced	13.2

[a]Mistranscription of *Majdanek* [b]Mostly from one visitor

Table 4 Appropriateness of responses to user questions

	On-topic responses				Off-topic resposnes	
	1	2	3	4	OK	Err
Illinois	311	68	65	1346	163	50
USHMM	365	83	99	1169	243	36

has a vocabulary size of 961 and density of 14.2, indicating that docents in Illinois are more likely to stick to familiar topics.

Repetitions of user utterances do not show differences between the two museums. Most repetitions happen after the system gives an inappropriate or off-topic response. In Illinois, instances of user repetition after a seemingly appropriate response give the impression that the docent is trying to elicit a specific utterance they had in mind; however, similar user behavior was observed at USHMM as well.

The results of the response annotations are shown in Table 4. Illinois has a higher proportion of on-topic responses than USHMM ($\chi^2 = 10$, $df = 1$, $p < 0.005$), which receive overall higher ratings for relevance ($\chi^2 = 24$, $df = 3$, $p < 0.001$). Among the off-topic responses, Illinois has a higher proportion of utterances that should have received a direct response ($\chi^2 = 8.6$, $df = 1$, $p < 0.005$). All of this is expected if the docents in Illinois have a tendency to use familiar utterances – more of these would be recognized, those that are recognized are more likely to lead to an appropriate response, and those that are not recognized are more likely to be misrecognitions by the system rather than questions that cannot be addressed directly. According to these measures, the *New Dimensions in Testimony* system is performing better with the docents, since it yields a higher proportion of appropriate on-topic responses.

However, the difference in performance between the sites is lower if we consider all errors together. Comparing all the appropriate responses (on-topic rated 3–4 and off-topic rated "OK") to the inappropriate ones (on-topic rated 1–2 and off-topic rated "Err"), USHMM data still have a slightly higher proportion of errors (24% compared to 21% at Illinois), but the difference is not highly significant ($\chi^2 = 4.4$, $df = 1$, $p = 0.035$). This is because the higher proportion of off-topic responses at USHMM represents a correct behavior of the system when faced with user utterances it cannot address directly.

4 Discussion

The main difference between museum visitors and museum docents is familiarity with the dialogue system: docents know the system and have some expectations from it, whereas visitors are likely interacting with the system for the first time. We therefore expect the docents to tailor their utterances to elicit survivor stories they know and like, which is exactly the behavior we observe.

Docents in Illinois are not a passive channel between the visitors and the system, but rather active participants in the dialogue. Do they enhance or detract from the visitor experience? The more interpersonal nature of the USHMM interactions suggests that at least in some respects, a docent-led interaction is inferior, as it interferes with the direct connection between the visitor and the survivor.

Finally, it is interesting to note that overall correct system behavior is about the same in both museums. Shouldn't we expect docent-led interactions, with more limited inputs and better familiarity with the content, to result in better performance? We suspect that the requirements imposed by direct interaction might lead to suboptimal performance in docent-led interaction. A challenge would be to design a dialogue system that could best cater for the needs of both visitors and docents.

Acknowledgements We are grateful to two anonymous reviewers for their comments. The first author was supported by the National Science Foundation under grant 1560426, "REU Site: Research in Interactive Virtual Experiences" (PI: Evan Suma). The second author was supported in part by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

This work is based on materials from *New Dimensions in Testimony* – a collaboration between the USC Institute for Creative Technologies, the USC Shoah Foundation, and Conscience Display, which was made possible by generous donations from private foundations and individuals. We are extremely grateful to The Pears Foundation, Louis F. Smith, and two anonymous donors for their support, and to the Illinois Holocaust Museum and Education Center and the United States Holocaust Memorial Museum for hosting the exhibit. We owe special thanks to Pinchas Gutter for sharing his story, and for his tireless efforts to educate the world about the Holocaust.

References

Anthony L (2014) AntConc (version 3.4.3) [computer software]. URL http://www.laurenceanthony.net/

Swartout W, Traum D, Artstein R, Noren D, Debevec P, Bronnenkant K, Williams J, Leuski A, Narayanan S, Piepol D, Lane C, Morie J, Aggarwal P, Liewer M, Chiang JY, Gerten J, Chu S, White K (2010) Ada and Grace: Toward realistic and engaging virtual museum guides. In: Intelligent Virtual Agents, LNAI, vol 6356, Springer, Heidelberg, pp 286–300

Traum D, Aggarwal P, Artstein R, Foutz S, Gerten J, Katsamanis A, Leuski A, Noren D, Swartout W (2012) Ada and Grace: Direct interaction with museum visitors. In: Intelligent Virtual Agents, LNAI, vol 7502, Springer, Heidelberg, pp 245–251

Traum D, Georgila K, Artstein R, Leuski A (2015a) Evaluating spoken dialogue processing for time-offset interaction. In: Proceedings of SIGDIAL, Prague, Czech Republic, pp 199–208

Traum D, Jones A, Hays K, Maio H, Alexander O, Artstein R, Debevec P, Gainer A, Georgila K, Haase K, Jungblut K, Leuski A, Smith S, Swartout W (2015b) New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In: Interactive Storytelling, LNCS, vol 9445, Springer, pp 269–281