COHERENCE OF OFF-TOPIC RESPONSES FOR A VIRTUAL CHARACTER

Ron Artstein, Jacob Cannon, Sudeep Gandhe, Jillian Gerten, Joe Henderer, Anton Leuski, and David Traum
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292, USA
<lastname>@ict.usc.edu

ABSTRACT

We demonstrate three classes of off-topic responses which allow a virtual question-answering character to handle cases where it does not understand the user's input: ask for clarification, indicate misunderstanding, and move on with the conversation. While falling short of full dialogue management, a combination of such responses together with prompts to change the topic can improve overall dialogue coherence.

1. BACKGROUND

Sergeant Star is a virtual question-answering character developed for the U.S. Army Recruiting Command as a high-tech attraction and information source about the Army. He is a life-size character built for demos in mobile exhibits, who listens to human speech and responds with pre-recorded voice answers (Figure 1). SGT Star is based on technology similar to that used in previous efforts (Leuski et al., 2006; Leuski and Traum, 2008), which treats question-answering as an information retrieval problem: given a natural-language question, the character should retrieve the most appropriate answer from a list of available responses. An Automatic Speech Recognition (ASR) module transforms the user's question to text, and then a statistical classifier ranks possible responses according to the similarity of the user's utterance to associated questions in the training data. The top-ranked response is SGT Star's best guess about the correct answer to the user's question.

In a typical exhibit setting, SGT Star interacts with groups of attendees. Since SGT Star can only talk to one person at a time, conversation with the group is mediated by a human handler, who uses SGT Star to create a two-man show. There is a small group of handlers who demonstrate SGT Star at various shows, and acoustic models of the speech recognition component are tuned to their voices in order to get the best recognition in noisy convention environments.

2. DIALOGUE MANAGEMENT

Conversational characters typically employ *dialogue managers* – software components which keep track of the state of the dialogue, the commitments and obligations of the various participants, the character's goals and desires, hypotheses about the mental states of other participants, and so on. SGT Star is a simple character who does not have a representation of the world or the ability to reason; instead of a separate manager, dialogue management capabilities are incorporated into the classifier. The most important capability is detecting when the best response is not good enough: if the score of the top-ranked classifier output falls below a preset threshold, SGT Star does not produce that output, but instead chooses among a set of predefined "off-topic" responses (e.g. "Sorry, I didn't catch that. Could you say it again?"). Additional capabilities include retaining a short history of SGT Star's own utterances (so that he doesn't keep repeating himself) and the ability to prompt the user to ask questions that the character can answer.

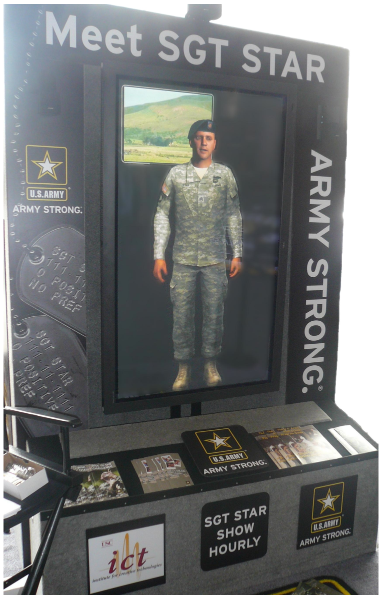


Figure 1: SGT Star at SKILLS 2008

The off-topic strategy for dealing with classification failures has been successful for other efforts such as SGT Blackwell – a general-domain question-answering character who interacts with visitors in a museum setting (Leuski et al., 2006; Robinson et al., 2008). The environment in which SGT Star is deployed differs from that of SGT Blackwell in two important ways: speech input to SGT Star typically comes from trained handlers rather than from the general public, and the handlers try to engage SGT Star for a conversation consisting of a greeting phase, a few information exchanges, and a closing routine. Since handlers are trained, few user utterances are genuine out-of-domain questions, and most of SGT Star's classifier failures are caused by faulty speech recognition or insufficient training data; since interactions are fairly long (compared to SGT Blackwell), random off-topic interruptions are very disruptive. Initial versions of SGT Star were very successful at providing on-topic responses, but rather poor when an off-topic response was called for: Artstein et al. (2008) conducted an evaluation of SGT Star's responses at the National Future Farmers of America Convention in Indianapolis in October 2007, and while the vast majority of the on-topic responses (80.7%) received the maximum coherence rating of 5, the majority of off-topic responses (80.1%) were rated between 1 and 2. An individual analysis of the off-topic responses showed that requests for repetition were usually ranked as more coherent than other types of off-topic responses.

To improve the coherence of off-topic responses we re-authored many of the responses, and implemented a new off-topic selection policy. We were not able to use a separate classifier trained on out-of-domain questions (Patel, Leuski, and Traum, 2006), because very few of the questions SGT Star gets are truly outside his domain. Instead, we designed a strategy based on the knowledge that the vast majority of SGT Star's off-topic responses are triggered by speech recognition errors and classification failures. If SGT Star fails to find an answer, then in all likelihood he either misheard the user's utterance or misunderstood it. We therefore authored off-topic responses for SGT Star in the following four classes (see examples in Table 1).

Didn't hear: SGT Star indicates that he did not hear the user's utterance and asks the user to repeat it.

Didn't understand: SGT Star implies that he did not understand the user's utterance, leaving room for the user to repeat or change the topic.

Move on: SGT Star declines to answer, and asks the user to move on.

Prompt: SGT Star suggests a question to the user.

The off-topic response classes were designed to fit into a simple strategy: the first off-topic response after a successful dialogue is always of the "didn't hear" type; if the following user utterance is also not understood, SGT Star produces a "didn't understand" utterance, and if a third successive off-topic response is needed, SGT Star produces a "move on" utterance followed immediately by a prompt. The rationale behind this strategy is to ensure that the majority of SGT Star's off-topic responses are requests for clarification (which have been shown to be more coherent), while avoiding repetitiveness and progressively encouraging the user to move the conversation to a new and more fruitful direction.

Table 1: Example Off-topic Responses

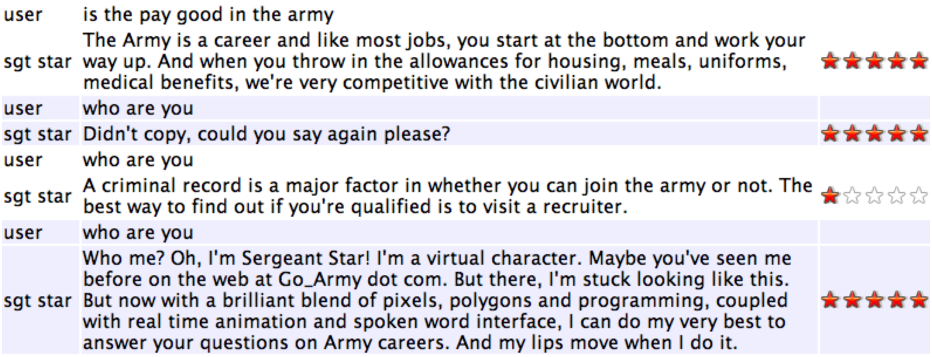
Didn't hear
Could you throw that at me again?
I didn't copy that. Could you repeat that?
Sorry, I didn't catch that. Could you say it again?
Didn't understand
Aw this feels too much like school. I didn't study last night.
Sometimes I think you're just testing me.
You know, sometimes you ask a lot of questions.
Move on
Can we talk about what I want to talk about?
Uh, next question.
You can get answers at GoArmy dot com.
Prompt
Ask me how I can see in the dark.
Hey why don't you ask me about my badges?
Why don't you ask me about becoming an MP in the army?

3. RATING STUDY

The new off-topic responses were prepared for SGT Star's appearance at the National Leadership and Skills Conference in Kansas City in June 2008. Due to implementation issues on-site we were not able to test the full progressive strategy, but only the individual response classes.

We conducted a rating study in order to assess the coherence of SGT Star's responses, similar to the study reported in Artstein et al. (2008). The instructions explicitly told the judges to rate the *coherence* (appropriateness) of SGT Star's responses rather than their correctness. The idea is that the more coherent a character is, the better he can engage the audience. An appropriate response to a question does not have to be a direct answer: a question or off-topic comment may sometimes be more appropriate, and SGT Star's off-topic responses were designed to allow him to hold a coherent conversation when he does not have a straight answer.

SGT Star's performance resulted in a total of 2095 responses, and our study judged the appropriateness of these responses in context. The user utterances were transcribed individually, and entire dialogues (user utterances and SGT Star's responses) were presented as web pages on which judges rated each of SGT Star's responses on a scale of 1 to 5 (Figure 2). In 578 cases, the transcribed user utterance was identical to a training question and the response



user	is the pay good in the army	
sgt star	The Army is a career and like most jobs, you start at the bottom and work your way up. And when you throw in the allowances for housing, meals, uniforms, medical benefits, we're very competitive with the civilian world.	★★★★★
user	who are you	
sgt star	Didn't copy, could you say again please?	★★★★★
user	who are you	
sgt star	A criminal record is a major factor in whether you can join the army or not. The best way to find out if you're qualified is to visit a recruiter.	★☆☆☆☆
user	who are you	
sgt star	Who me? Oh, I'm Sergeant Star! I'm a virtual character. Maybe you've seen me before on the web at Go_Army dot com. But there, I'm stuck looking like this. But now with a brilliant blend of pixels, polygons and programming, coupled with real time animation and spoken word interface, I can do my very best to answer your questions on Army careers. And my lips move when I do it.	★★★★★

Figure 2: The rating interface

was linked to that question, and these were automatically rated as 5; the remaining 1517 responses were rated by the judges.

To ensure the ratings were meaningful we calculated inter-rater reliability using α (Krippendorff, 1980).[1] Two judges rated all 1517 responses, and two additional judges (including the first author) rated portions of the responses in order to get more robust reliability figures. Reliability for on-topic responses was $\alpha = 0.834$, slightly higher than the value of $\alpha = 0.794$ reported in Artstein et al. (2008) (the difference is not significant). Reliability for off-topic responses and prompts was essentially at chance level ($\alpha = 0.017$ and $\alpha = 0.080$),[2] reflecting the fact that evaluating the coherence of an off-topic response is much more difficult than evaluating the coherence of an on-topic response. Overall reliability was $\alpha = 0.583$, ranging from $\alpha = 0.680$ for the most concordant pair of judges to $\alpha = 0.351$ for the most discordant pair. This figure is significantly lower than the value of $\alpha = 0.789$ reported in Artstein et al. (2008); the reason for the drop in overall reliability is the *improvement* in the ratings of off-topic responses, making them less distinct from on-topic ratings. For the following analysis we associated each response with the mean of the ratings of the two main judges.

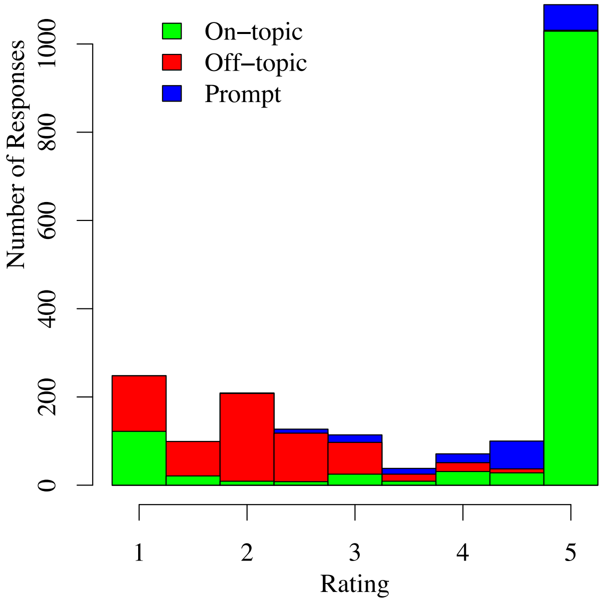


Figure 3: On-topic and off-topic ratings

4. RESPONSE RATINGS

SGT Star has a total of 213 possible responses, including 10 each of the 3 off-topic categories "didn't hear", "didn't understand", and "move on", and 7 prompts. The mean rating of SGT Star's responses was 3.72, but the distribution is far from normal (first quartile 2, median 5). About 61% of the responses are on-topic, and these most of these are either very good (82% rated 4.5 or 5) or very bad (11% rated 1 or 1.5). The remaining responses are off-topics (30%), with a mean rating of 2.01, and prompts (9%), with a mean rating of 4.28 (Figure 3).

The different responses were not all used to the same extent: in the testing, SGT Star produced 159 different responses (including all 30 off-topics and 7 prompts), and their distribution was not even. Some on-topic responses

[1] Krippendorff's α is a chance-corrected agreement coefficient, similar to the more familiar K statistic (Siegel and Castellan, 1988). Like K, α ranges from -1 to 1, where 1 signifies perfect agreement, 0 obtains when agreement is at chance level, and negative values show systematic disagreement. The main difference between α and K is that α takes into account the magnitudes of the individual disagreements, whereas K treats all disagreements as equivalent; α is more appropriate for our study because the ratings are numerical, and the disagreement between ratings of 2 and 3, for example, is clearly lower than between 2 and 5. For additional background, definitions and discussion of agreement coefficients, see Artstein and Poesio, to appear.

[2] Reliability for prompts is reported for the two main judges only, because of the small number of prompts rated by the control judges.

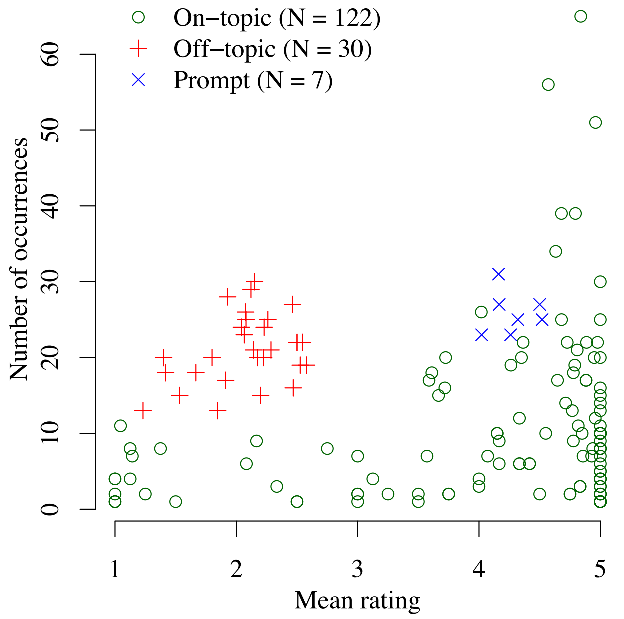


Figure 4: Rating and frequency correlations

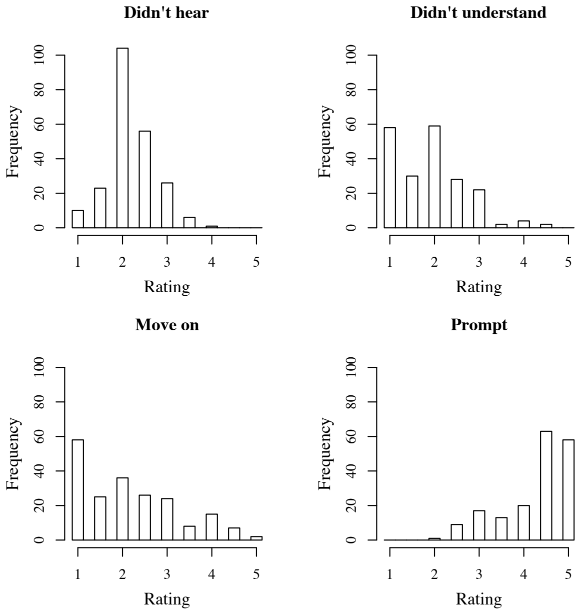


Figure 5: Ratings for the different response classes

were more common than others because they answered more popular questions; the mean rating of a response type correlated positively with its frequency, so that the more frequent responses were also the better rated ($r = 0.26, p < 0.01, df = 120$). This pattern is similar to that of Artstein et al. (2008), and the explanation is the same – the handlers are aware of which responses are easy to elicit, and target their questions to elicit these responses. Off-topics and prompts were more evenly distributed, with frequencies ranging from 13 to 30 for off-topics and from 23 to 31 for prompts; they do not show a significant correlation between rating and frequency (off-topic: $r = 0.35, p = 0.06, df = 28$; prompt: $r = 0.02, p > 0.9, df = 5$). Off-topic responses are generally more frequent than similarly rated on-topic responses, but the separation between the classes of responses was less clear than in Artstein et al. (2008) (Figure 4).

There were differences in the ratings for the individual off-topic response classes: “didn't understand” (1.90), “move on” (2.16), “didn't hear” (2.19); only the difference between “didn't understand” and “didn't hear” is significant according to a Wilcoxon rank-sum test ($W = 29430, p < 0.001$). The prompts received substantially higher ratings (4.28), with highly significant differences from each class of off-topics (Figure 5). SGT Star did not follow an explicit strategy for using the three off-topic classes; prompts always immediately followed a previous utterance without waiting for user input. While the differences among the classes are not significant, the ratings are consistent with our previous observation that clarification requests are the most coherent off-topic responses (previous versions of SGT Star did not have prompts).

5. SPEECH RECOGNITION

Automatic speech recognition (ASR) affects performance (Leuski et al., 2006): if what SGT Star hears doesn't match what the user said, then SGT Star's response is more likely to be inappropriate. We computed the word error rate for each user utterance by comparing the ASR output with the transcribed speech.[3] Mean word error rate was 0.365, a substantial improvement over the 0.469 word error rate of Artstein et al. (2008); this improvement is probably due to a combination of better trained speech models and an upgraded wireless head-mounted microphone (Sennheiser EW 352 G2). About 27% of the utterances had a word error rate of 0, with the remainder spread rather uniformly across the range from 0.05 to 1. Contrary to Artstein et al. (2008), we did not find word error rate to be a good predictor of whether SGT Star's response would be on-topic or off-topic (Figure 6).

We found a negative correlation between the rating of SGT Star's response and the word error rate of the immediately preceding user utterance ($r = -0.30, p < 0.001, df = 2093$). The correlation is stronger for on-topic responses ($r = -0.44, p < 0.001, df = 1280$), rather weak (but still significantly non-zero) for prompts ($r = -0.19, p = 0.01, df = 179$), and non-significant for off-

[3] Word error rate is the number of substitutions, deletions and insertions needed to transform one string into the other, divided by the number of words in the actual (transcribed) speech; values above 1 were recorded as 1.

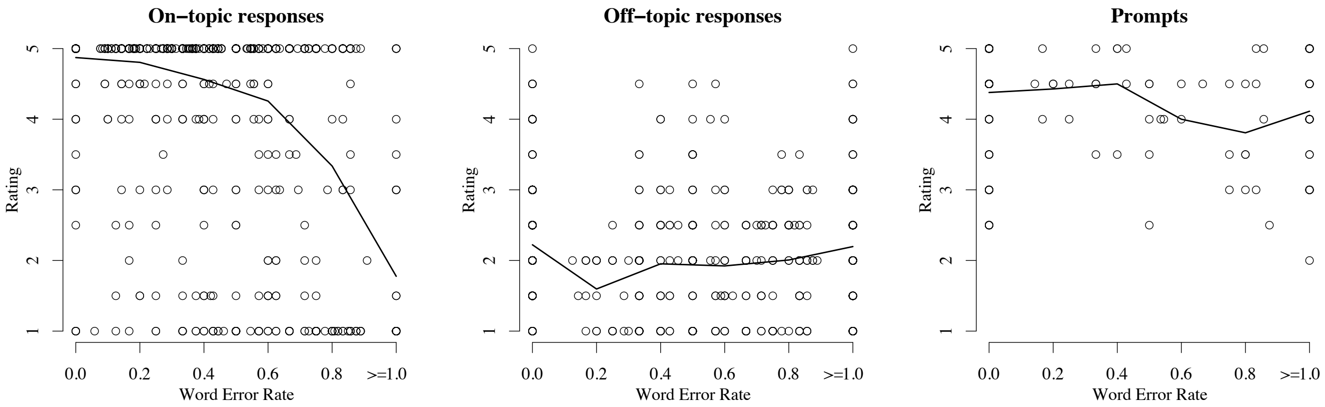


Figure 7: Word error rates and ratings: the lines show the mean rating for each WER band.

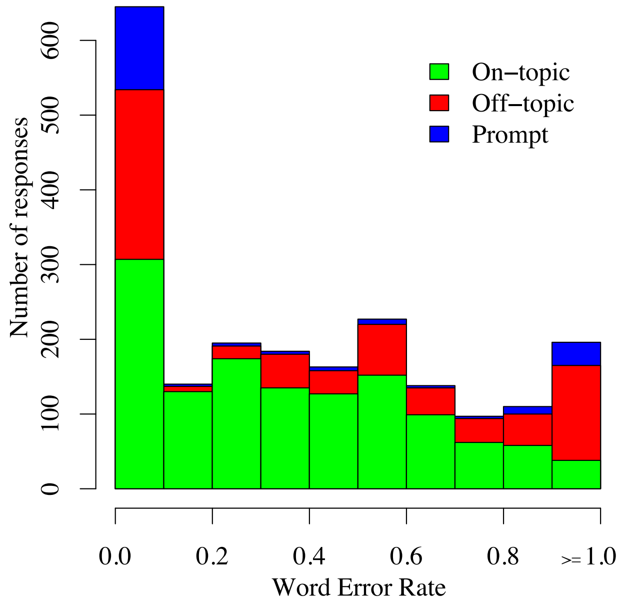


Figure 6: Word error rates and the responses they triggered

topic responses ($r = -0.03, p = 0.39, df = 630$). The relations between response rating and word error rate of the preceding utterance are shown in Figure 7.

The negative correlation between rating and word error rate for on-topic responses is expected: the less SGT Star understands the spoken utterance, the less likely he is to come up with a suitable on-topic response, so if an on-topic response is selected it is more likely to be inappropriate. Off-topic responses and prompts are not expected to degrade with the mismatch between actual and recognized user utterance.

6. CONCLUSIONS

Maintaining a coherent dialogue when faced with uncertainty about the meaning of the user's utterance requires dialogue management capabilities. We have shown that substantial improvements can be made with a simple off-topic policy by authoring off-topic responses in appropriate classes. The proposed strategy for advancing between these classes is expected to bring even more gains in the SGT Star's handling of difficult speech situations.

The off-topic strategy can be thought of as a primitive form of dialogue management, which can be handled within the classifier. We have found that other characters, which started out as classifiers, needed more sophisticated dialogue management techniques as their domains and their need to sustain conversation grew (Gandhe et al., 2008). The demands on SGT Star are similarly growing: current development calls for him to have some (limited) awareness of the physical world around him, and his recent installation in USAAC's Adventure Semis brings him closer to more audiences. The current strategy is pushing the limits of dialogue management without a manager; it is therefore likely that in the near future we will need to add a dedicated dialogue management component, to allow

SGT Star to keep up and support the Army's mission.

ACKNOWLEDGMENTS

SGT Star is loosely based on a character created by Next IT for the Army's recruiting web site, http://www.GoArmy.com/. SGT Star's body and animations were developed by the Mixed Reality team at ICT: Diane Piepol, Kip Haynes, Matt liewer, Josh Williams, and Tae Yoon.

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). 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.

REFERENCES

Ron Artstein, Sudeep Gandhe, Anton Leuski, and David Traum. 2008. Field testing of an interactive question-answering character. In *ELRA Workshop on Evaluation*, Marrakech, Morocco.

Ron Artstein and Massimo Poesio. to appear. Inter-coder agreement for computational linguistics. *Computational Linguistics*. Pre-publication draft at http://cswww.essex.ac.uk/Research/nle/arrau/icagr-short.pdf.

Sudeep Gandhe, David DeVault, Antonio Roque, Bilyana Martinovski, Ron Artstein, Anton Leuski, Jillian Gerten, and David Traum. 2008. From domain specification to virtual humans: An integrated approach to authoring tactical questioning characters. In *Interspeech 2008*, Brisbane, Australia.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, California.

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, Association for Computational Linguistics, Sydney, Australia.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *26th Army Science Conference*, Orlando, Florida.

Ronakkumar Patel, Anton Leuski, and David Traum. 2006. Dealing with out of domain questions in virtual characters. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier, editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, pages 121–131.

Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *LREC 2008 Proceedings*, Marrakech, Morocco.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, second edition, chapter 9.8. McGraw-Hill, New York.