Field Testing of an Interactive Question-Answering Character

Ron Artstein, Sudeep Gandhe, Anton Leuski and David Traum

Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292, USA
`artstein,gandhe,leuski,traum [at] ict.usc.edu`

Abstract

We tested a life-size embodied question-answering character at a convention where he responded to questions from the audience. The character's responses were then rated for coherence. The ratings, combined with speech transcripts, speech recognition results and the character's responses, allowed us to identify where the character needs to improve, namely in speech recognition and providing off-topic responses.

Figure 1: SGT Star

Figure 2: SGT Star in a live demo

1. Background

We created Sergeant Star, a virtual question-answering character, for the U.S. Army Recruiting Command as a hi-tech attraction and an information source about the Army. He is a life-size character built for use in mobile exhibits, who accepts speech input 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; Robinson et al., 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. A user's question is transformed to text through an Automatic Speech Recognition (ASR) module, and possible responses are ranked by a statistical classifier trained on a set of questions linked to responses. If the top-ranked response exceeds a preset threshold, the character utters that response; if no response reaches the threshold, the character makes a random choice among a set of predefined "off-topic" responses (like "Take the gum out of your mouth"). The system also includes provisions for detecting repetitive questions, and prompting the user if they are unable to come up with questions that the character can answer.

2. Setting

SGT Star was designed to be part of a mobile exhibit, so we evaluated his performance on the road, at the National Future Farmers of America Convention on 24–27 October 2007 in Indianapolis, where SGT Star was part of the U.S. Army exhibit (Figure 2 shows a similar installation at Fort Knox, Kentucky). Since SGT Star is demonstration technology, convention attendees did not talk to SGT Star directly, but passed their questions to a human handler who talked into the microphone. In order to get the best speech recogni-

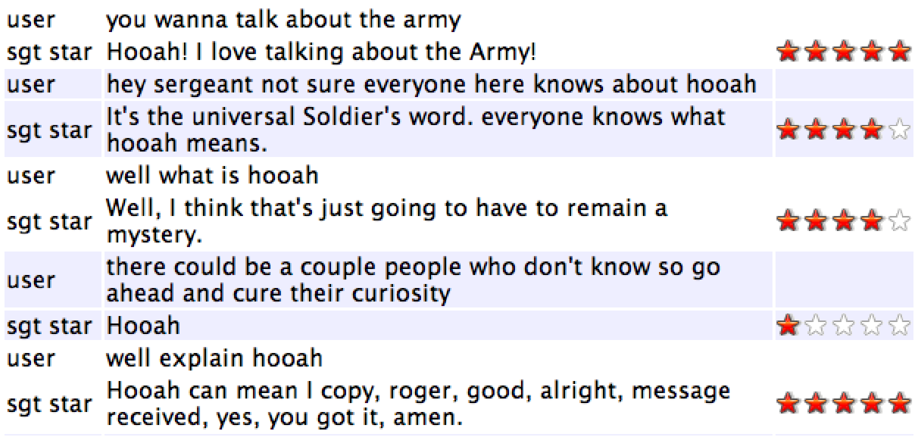

Figure 3: The rating interface

tion in a noisy convention environment, the acoustic models were tuned to the three individuals who did most of the handling.

To motivate convention attendees to explore the Army exhibit, the Army devised a task which required attendees to gather specific pieces of information from various parts of the exhibit in order to win a prize; the information they had to get from SGT Star was the meaning of "hooah" (a U.S. Army expression) and his Military Occupational Specialty (MOS). As a result, many of the dialogues from the convention are extremely short, and a disproportionate number of questions ask about "hooah" and SGT Star's MOS (approximately 17% and 13%, respectively).

3. Rating study

SGT Star's mission is to generate interest in learning about the Army and possible careers in it, but we have no way to measure the amount of interest generated. We assessed SGT Star's *coherence*, that is the appropriateness of his responses; the idea is that the more coherent a character is, the better he can engage the audience and create interest. 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 doesn't have a straight answer. We conducted a rating study in order to identify where SGT Star's coherence could be improved, to make him a more believable and engaging character.

SGT Star's performance resulted in a total of 3216 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 3). In 703 cases, the transcribed user utterance was identical to a training question and the response was linked to that question, and these were automatically rated as 5; the remaining 2513 responses were rated by the judges.

To ensure the ratings were meaningful we calculated inter-rater reliability using α (Krippendorff, 1980).[1] Three judges rated all 2513 responses, and a fourth judge (the first author) rated 474 of these. Overall reliability for the four judges was $\alpha = 0.789$; reliability for sub-groups of judges ranged from $\alpha = 0.901$ for the most concordant pair of judges to $\alpha = 0.676$ for the most discordant pair. Since overall reliability was close to the accepted threshold of 0.800, we continued the analysis by assigning each response the mean of all available ratings. Broken down by response type, reliability was high for on-topic responses ($\alpha = 0.794$) but barely better than chance for off-topic responses ($\alpha = 0.097$).

4. Response ratings

SGT Star has a total of 152 possible responses, of which 22 are tagged as off-topic. Off-topic responses are intended to be suitable both for genuine out-of-domain questions, for which SGT Star does not have

[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).

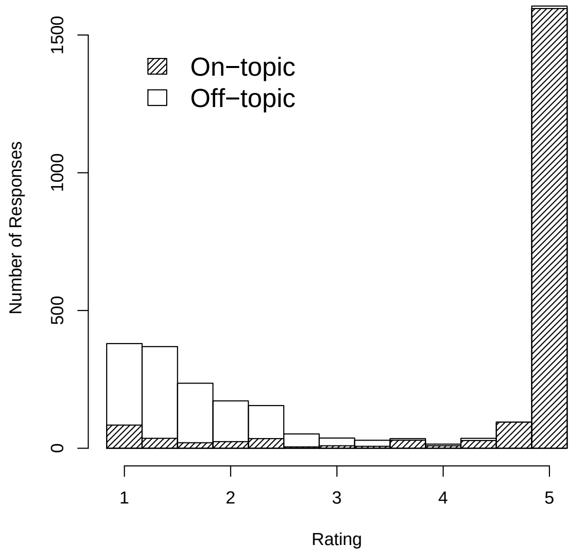

Figure 4: On-topic and off-topic ratings

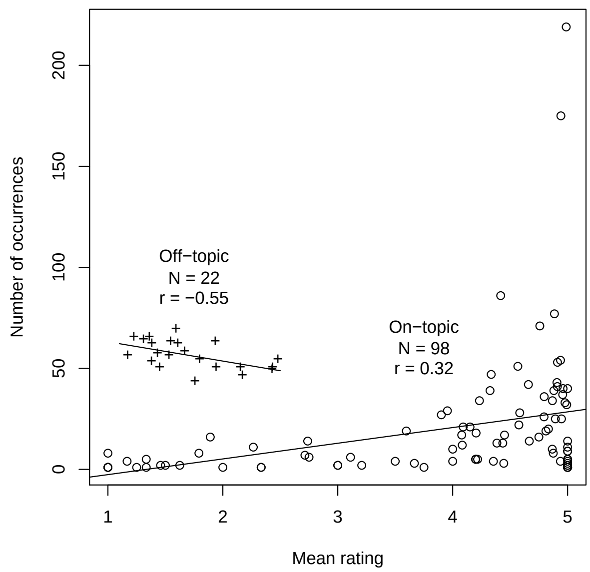

Figure 5: Rating and frequency correlations

an appropriate on-topic response, as well as for classifier failures due to factors like speech recognition errors or insufficient training data. The handlers at the convention were very familiar with SGT Star's range of responses and as a consequence there were very few out-of-domain questions; the vast majority of off-topic responses were a result of classifier failure.

The different responses were not all used to the same extent: in the testing, SGT Star produced 120 different responses (including all 22 off-topics), and their distribution was not even. This skewing is due to the uneven distribution of questions: The two most frequent responses by far, used 175 and 219 times, answer questions about "hooah" and SGT Star's MOS, brought about by the convention attendees' task.

The mean rating of SGT Star's responses was 3.47, but very few responses were close to the mean: most responses were either very good or very bad (first quartile 1.67, median 4.75). About 57% of the responses were rated above 3 and 43% below 3; this split roughly correlates with the difference between on-topic responses (61.5%), of which 80.7% received the maximum rating of 5, and off-topic responses (38.5%), of which 80.1% were rated 2 or less (Figure 4). There was also a clear separation in the frequency of individual responses. The off-topic responses were all used with similar frequency (ranging from 43 to 69) and received mean ratings of less than 2.5. In contrast, the low-rated on-topic responses appeared much less frequently (maximum frequency 16 for ratings under 3.5), while frequent on-topic responses were rated much higher (Figure 5). There is a positive correlation between rating and frequency for on-topic responses ($r = 0.32, p < 0.002, df = 96$),[2] whereas for off-topic responses the correlation is negative ($r = -0.55, p < 0.01, df = 20$).

The correlation between rating and frequency for on-topic responses remains robust even when we remove questions about the more common topics such as "hooah" and SGT Star's MOS. The reason is probably that the handlers quickly learned which responses were easy to elicit and popular with the crowd, and then targeted their questions to elicit these responses. The result was a selection of question topics narrower than SGT Star's full repertoire, which led to an overall good performance.

The negative correlation between rating and frequency for the off-topic responses was unexpected, since agreement on off-topics was low and individual off-topic responses are chosen at random. However, some off-topic responses are also linked to out-of-domain questions in the training data (for example, the response "ha ha, you're a bad man" is linked to the question "so do you have a girlfriend?"). The linked responses are expected to occur more frequently. As it turns out, requests for repetition ("I didn't hear that, could you repeat the question?") are usually not linked to any question, but these received higher ratings than the linked off-topic responses.

[2]The correlation is stronger if we use log frequencies: $r = 0.48, p < 0.001, df = 96$.

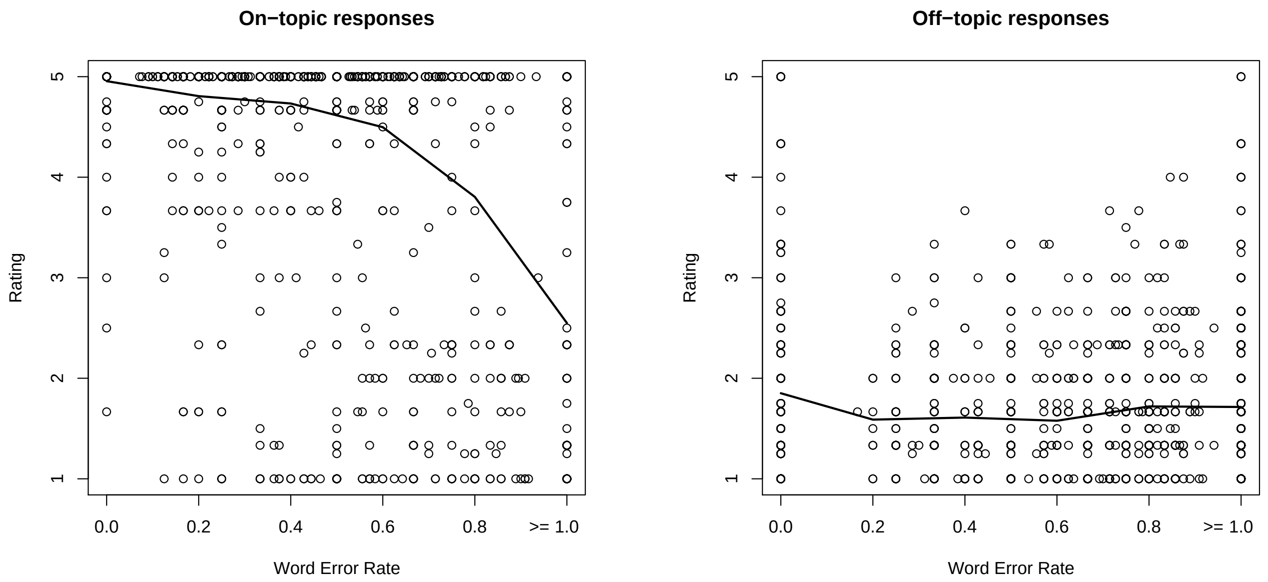

Figure 7: Word error rates and ratings: the lines show the mean rating for each WER band.

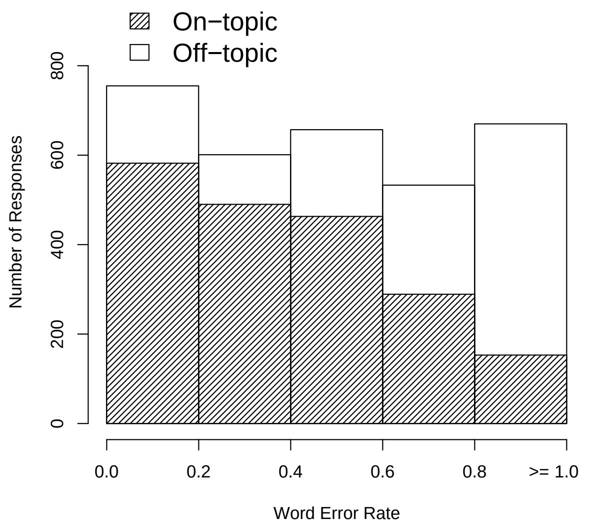

Figure 6: Word error rates

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 0.469, with an approximately uniform distribution; higher word error rates were more likely to trigger off-topic responses (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 ($r = -0.47, p < 0.001, df = 3214$). This is partly due to the large block of off-topic responses with low ratings and high word error rates; however, the on-topic responses on their own also exhibit a (slightly weaker) negative correlation between response rating and word error rate ($r = -0.40, p < 0.001, df = 1975$). The off-topic responses do not show a similar correlation ($r = -0.02, p > 0.4, df = 1237$). 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 is expected: the less SGT Star understands the spoken utterance, the less likely he is to come up with a suitable on-topic response. Off-topic responses should not degrade with the mismatch between actual and recognized user utterance. One might even expect to find an improvement: due to the statistical language modeling in the ASR component, misrecognition of spoken words is more likely for out-of-domain questions, and SGT Star's off-topic responses should be more appropriate for those. We have not found this kind of effect, possibly because there were few out-of-domain questions.

6. Conclusions

The rating study of data gathered in SGT Star's field deployment allowed us to study his functioning in

[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.

the situation for which he was designed, though with somewhat different parameters, namely being repeatedly asked for two pieces of information. The results show an interplay between SGT Star and his handlers, who are working to help the virtual character give his best performance. It is clear that SGT Star would have performed very differently if arbitrary users were allowed to ask unrestricted questions; dealing with such users and out-of-domain questions is the focus of another study, SGT Blackwell (Robinson et al., 2008). The study confirmed that speech recognition is a major obstacle – this is a difficult problem in the noisy environment where SGT Star operates. The study also identified off-topic responses as a place with substantial room for improvement, perhaps along the lines of Patel et al. (2006).

The rating study combined data extracted from system logs (ASR results and SGT Star's responses) with manual transcription, a human rating study, statistical testing and qualitative assessment. A question that comes up naturally is whether this method of evaluation can be automated or made less human-intensive. There is definitely some room for saving – for example, once we have established that the ratings are reliable, it is sufficient to have just one judge rate each response. However, rating the responses is not where most of the human effort went. All user utterances need to be manually transcribed, because the appropriateness of responses needs to be judged relative to the actual user utterance (this manual transcription is independently needed in order to improve performance of the highly domain-specific speech recognition models). But the most labor-intensive part is probably the analysis of individual responses. This is because we are not merely interested in a score that reports SGT Star's performance, but are also seeking to improve it for future exhibits. SGT Star's ability to respond appropriately depends on his training data, which consist of a list of questions, a list of responses, and links between the two. The questions come from actual user data, the responses reflect what we want SGT Star to be able to talk about, and the links come from a careful analysis of appropriateness which can only be achieved by manually examining actual conversation transcripts.

7. Acknowledgments

SGT Star is loosely based on a character created by Next IT for the Army's recruiting web site.[4] Thanks to two anonymous reviewers, to Kip Haynes for taking SGT Star on the road, and to Jacob Cannon, Jillian Gerten and Joe Henderer for rating SGT Star's utterances. 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.

8. References

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`.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

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, Sydney, Australia, July. Association for Computational Linguistics.

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*, pages 121–131. Springer, Berlin.

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, May.

Sidney Siegel and N. John Castellan, Jr, 1988. *Nonparametric Statistics for the Behavioral Sciences*, chapter 9.8, pages 284–291. McGraw-Hill, New York, second edition.

[4]`http://www.GoArmy.com/`