Semi-formal Evaluation of Conversational Characters

Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum

Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Rey, CA 90292 USA

Abstract. Conversational dialogue systems cannot be evaluated in a fully formal manner, because dialogue is heavily dependent on context and current dialogue theory is not precise enough to specify a target output ahead of time. Instead, we evaluate dialogue systems in a semi-formal manner, using human judges to rate the coherence of a conversational character and correlating these judgments with measures extracted from within the system. We present a series of three evaluations of a single conversational character over the course of a year, demonstrating how this kind of evaluation helps bring about an improvement in overall dialogue coherence.

1 Introduction

In the past two decades, the field of Computational Linguistics has placed an increasing emphasis on formal evaluation of systems and system components; typically, this involves creating a target reference ("gold standard"), and measuring system performance against this reference. The availability of standard targets – most notably the Penn Treebank [1] – has greatly facilitated the use of machine learning for computational linguistic tasks, and formal system evaluation has opened up the field to competitions between systems working on a shared task.

While formal system evaluation is responsible for much of the progress made in Computational Linguistics in recent years, it has limitations. In order to conduct a formal evaluation, the desired target performance needs to be defined prior to the task, and this gets more difficult as we move from the surface of an utterance to more abstract levels of representation. Thus it is fairly straightforward to specify the desired output of a speech recognizer for a set of input utterances; specifying the desired part-of-speech labels or syntactic parse is somewhat more controversial, and defining a desired semantic representation or translation to a different language is even more difficult. The present study is concerned with dialogue systems, where the abstraction continues to climb with dialogue acts, and reaches the most abstract level with response selection – the decision *what* to say in response to a user utterance. For conversational characters this is essentially an open-ended problem.

Even in the absence of a predefined target, some dialogue systems can be evaluated formally. The performance of task-oriented systems can be measured

O. Grumberg et al. (Eds.): Francez Festschrift, LNCS 5533, pp. 22–35, 2009.

by the task success rate, length of the dialogues (assuming it is desirable to complete the task quickly), and other proxies for user satisfaction. Such measures are useful not only for evaluation and comparison but also for machine learning of dialogue strategies that optimize system performance according to the chosen criteria [2,3]. In contrast, conversational characters are not designed to help a user accomplish a specific goal; the criteria for successful dialogue are "soft" ones such as user satisfaction and tutoring outcomes, and at present we do not have proxies for these criteria that can be measured automatically.

We present a conversational character, Sergeant Star, who answers questions about the U.S. Army. He appears at conventions and conferences together with live exhibitors, and his purpose is to generate interest and engage the audience rather than to conduct efficient dialogues aimed at achieving a specific task. SGT Star's components include a number of statistical subsystems such as a speech recognizer and a response classifier; these components are formally trained and evaluated using large sets of data mapping inputs to outputs. But we have no way to formally evaluate the overall, end-to-end performance of SGT Star. Instead, we use what we call a "semi-formal" approach: we get the "soft" performance metrics by using human judges to rate the coherence of SGT Star's responses from actual field deployments, and then use these ratings together with measures taken from within the system in order to gain a better understanding of SGT Star's performance. The combination of ratings with system data allows us to see patterns in the overall behavior that would be difficult to detect in a detailed item-by-item qualitative analysis, and this influences the continued development of SGT Star.

We describe the SGT Star system in section 2 and our rating studies in section 3. Results and analysis, correlating the ratings with system data, are in sections 4 and 5. The conclusions in section 6 describe how the results are used in the authoring process to improve SGT Star's performance from one iteration to the next.

2 Sergeant Star

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, animated voice answers (Figure 1). SGT Star is based on technology similar to that used in previous efforts [4,5], which treats question-answering as an information retrieval problem: given a natural-language question, the character should retrieve the most appropriate response from a predefined list. An Automatic Speech Recognition (ASR) module transforms the user's question to text, and then a statistical classifier trained on question-response pairs ranks possible responses according to their similarity to a language model derived from the user's utterance; the top-ranked response is SGT Star's best guess about the correct answer to the user's question. The size of SGT Star's domain is about 200 responses, and the training

Fig. 1. SGT Star

data contain a few hundred questions that link to the responses in a many-to-many mapping. The classifier has a few simple dialogue management capabilities, such as detecting when the best response has a low score (and thus might not be appropriate), avoiding repetitive responses, and prompting the user to ask a relevant 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.

Evaluation of SGT Star is based on performance in three actual field deployments: the National Future Farmers of America Convention (Indianapolis, October 2007 and 2008) and the National Leadership and Skills Conference (Kansas City, June 2008).[1] The main advantage of a field test is that the interactions being evaluated are real, rather than simulated interactions from the lab (Ai et al. [8] show that dialogues with real users have different characteristics from dialogues with lab subjects). The two main challenges presented by field evaluation are the lack of experimental controls and the demands of interacting with a live audience, which take precedence over experimental needs.

Initially, the field studies were intended as a general evaluation, but the results quickly turned these into a detailed study of SGT Star's "off-topic" responses. SGT Star is a simple question-answering character who does not have a dialogue manager to keep track of the state of the dialogue, the commitments and obligations of the various participants, his own goals and desires, and so on. Instead of a separate manager, dialogue management capabilities are incorporated

[1] Evaluations of the October 2007 and June 2008 deployments were reported in [6,7].

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 specified 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?"). The threshold is set automatically during training in order to find an optimal balance between false positives (inappropriate responses above threshold) and false negatives (appropriate responses below threshold). We should note that the labels "on-topic" and "off-topic" characterize SGT Star's responses, not the user's questions: an in-domain question can receive an off-topic response (e.g., if it was not properly recognized), and such a response may well turn out to be coherent in the context of the dialogue; an out-of-domain question can also receive an on-topic response, though this usually indicates that SGT Star misunderstood the question and therefore the response is typically not appropriate.

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 [4,9]. 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: in the October 2007 study, the vast majority of the on-topic responses (80.7%) received the maximum coherence rating of 5, whereas 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 [10], 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.

Clarify: Ask the user to repeat the question, for example:

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?

Stall: Wait for user initiative, for example:
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: Decline to answer, for example:
Can we talk about what I want to talk about?
Uh, next question.
You can get answers at GoArmy dot com.
Prompt: Direct the user to a new topic, for example:
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?

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 "clarify" type; if the following user utterance is also not understood, SGT Star produces a "stall" 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. Due to an unrelated installation problem, we were not able to use the off-topic selection policy in the June 2008 study, and we had to revert to a version of the software which selects off-topic responses at random. This gave us an unintended experimental control, since we had the re-authored off-topic response classes, but without the policy. The October 2008 study used the same off-topic responses as the June 2008 study, with the full selection policy implemented.

3 Coherence Rating and Reliability

In the absence of a predetermined target performance, we used a post-hoc rating of SGT Star's responses. Raters were asked 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.

Following each field deployment, we transcribed all the user utterances from the recordings, and then created full transcripts of the dialogues from the transcribed user utterances and the character's responses obtained from system logs. These transcripts were presented as web pages on which judges rated each of SGT Star's responses, in the context of the immediately preceding dialogue, on a scale of 1 to 5 (Figure 2). In order to reduce the burden on the judges, we automatically identified the cases where the transcribed user utterance was identical

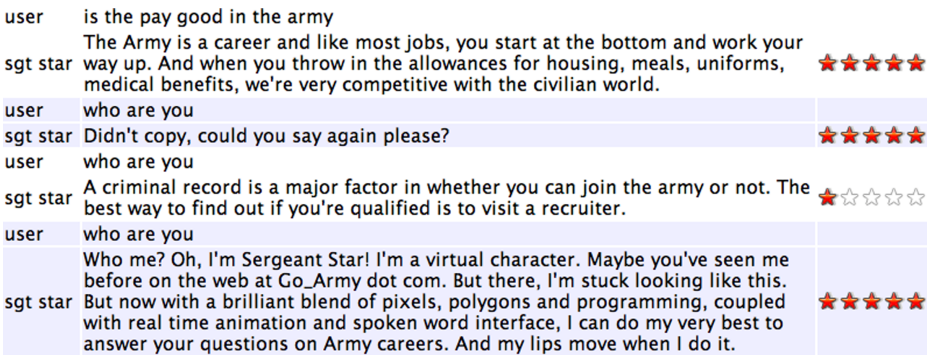


Fig. 2. The rating interface

Table 1. Number of responses

	Total	Perfect match (On-topic)		Rated					
				On-topic		Off-topic		Prompt	
		N	%	N	%	N	%	N	%
October 2007	3216	703	22	1283	40	1230	38	—	
June 2008	2095	578	28	704	34	632	30	181	9
October 2008	1321	282	21	705	53	281	21	51	4

to a training question and the response was linked to that question in the training data, and these were automatically rated as 5; the remaining responses were rated by the judges. Table 1 gives the number of responses rated in each study.

To ensure the ratings were meaningful we calculated inter-rater reliability using Krippendorff's α [11].[2] Four raters participated in each reliability study: the first author participated in all studies, the third author in the October 2008 and November 2008 studies, and the remaining judges were student annotators (one of whom had also transcribed the utterances). In each study, at least 2 of the raters marked all of the responses while the remaining raters only marked a portion for calculating reliability. The results of the reliability studies are shown in Table 2.

[2] Krippendorff's α is a chance-corrected agreement coefficient, similar to the more familiar K statistic [12]. 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 [13].

Table 2. Reliability of rater judgments (Krippendorff's α)

	All responses			On-topic	Off-topic	Prompt
	All Raters	Excluding Outlier	Range[a]	All Raters	All Raters	All Raters
October 2007	0.786	0.886	0.676–0.901	0.794	0.097	—
June 2008	0.583	0.655	0.351–0.680	0.842	0.017	0.080[b]
October 2008	0.699	0.757	0.614–0.763	0.841	0.219	0.155

[a] Reliability for the most discordant and most concordant pairs of coders.
[b] Value reported for the two main judges only, because of the small number of prompts rated by the control judges.

Reliability of ratings for on-topic responses was $\alpha = 0.8$ or above in all studies, demonstrating that the coders share an understanding of the task and are able to apply it consistently. In contrast, reliability for off-topic responses and prompts was essentially at chance level for the October 2007 and June 2008 studies, and only slightly above chance for the October 2008 study; this reflects the fact that evaluating the coherence of an off-topic response is much more difficult than evaluating the coherence of an on-topic response. The improvement in reliability for off-topic and prompt ratings in the October 2008 study is statistically significant, and may be attributed to improved instructions, or to improved system performance which makes the off-topic responses better overall, allowing for better discrimination among coherent and incoherent ones (in other words, the fact that the character's overall coherence has improved makes it easier to rate his coherence).

Overall reliability decreased from the October 2007 to the June 2008 study. The reason for the drop in reliability is the improvement in the actual ratings of off-topic responses. In both studies the raters showed little ability to agree with each other on the ratings of individual off-topic responses, but in the October 2007 study these ratings were all very low (80.1% of the off-topic responses were rated between 1 and 2) and thus had little effect on the overall reliability; in the June 2008 study, off-topic responses were not ranked so low, making them less distinct from the on-topic ratings and therefore reducing overall reliability. Overall reliability improved with the October 2008 study despite the fact that an additional improvement in the off-topic ratings made them even more similar to on-topic ratings. This improvement in overall reliability is probably due to the improvement in the reliability of off-topic ratings.

We also calculated confidence intervals for α following a bootstrapping method similar to that of [14]; however, we found that variation in reliability among different subsets of raters was typically larger than the 95% confidence interval for a specific set of raters. We therefore report in Table 2 only the variation among subsets of raters (the range of reliability scores displayed by different rater subsets). We also note that in each study, removing one outlying rater bumps up reliability by 6 to 10 percentage points (the outlier was not always

the same person). Since overall reliability was close to acceptable in all studies, and since we do not have reason to believe that the outlying raters are less correct than the others, we continued our analysis using the mean rating for each response (mean of all available scores in October 2007 and October 2008, mean of the two main raters but not the controls in June 2008).

4 Response Ratings

A straightforward way to measure system performance is to look at the coherence ratings. The overall mean is not very telling, because the distribution is far from normal. Instead, Figure 3 shows histograms of all the response ratings (including those that were automatically rated as 5). The three studies show a similar

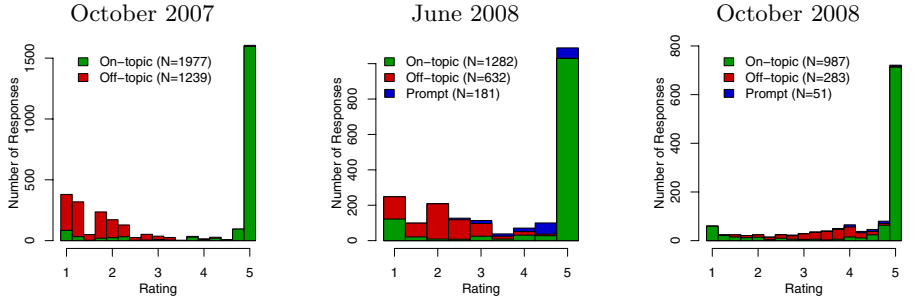


Fig. 3. Individual response ratings

pattern for on-topic response ratings. There is a very strong concentration of responses that are rated very high – in each of the studies, more than 80% of the on-topic responses received a mean rating of 4.5 or above; this means that when SGT Star's response score is above threshold, the response is usually very appropriate. There is also a discernible (though much smaller) bump at the lower end of the scale, which shows that when SGT Star chooses a wrong on-topic response, it is usually very inappropriate. In contrast to the stable on-topic rating pattern, the off-topic response ratings show a consistent improvement – the improvement from October 2007 to June 2008 is due to rewriting the responses in the three off-topic classes, and the improvement from June 2008 to October 2008 is due to the implementation of the off-topic selection policy.

We gain additional insight by looking at the individual response types, comparing their ratings with the frequency in which they occur in the dialogues (Figure 4). Again, the pattern for on-topic responses is the same in all studies: the frequent responses are more highly rated. The likely explanation, as we proposed in [6], is that the handlers are aware of which responses are easy to elicit, and target their questions to elicit these responses. The pattern thus demonstrates an interplay between the inherent capabilities of the system and the human handlers who are working to maximize its performance in a live show. The

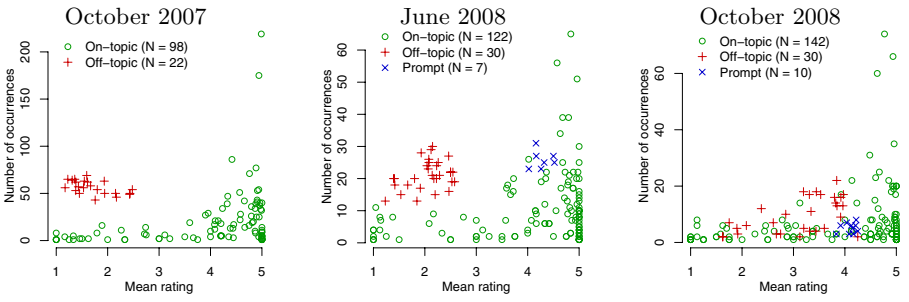
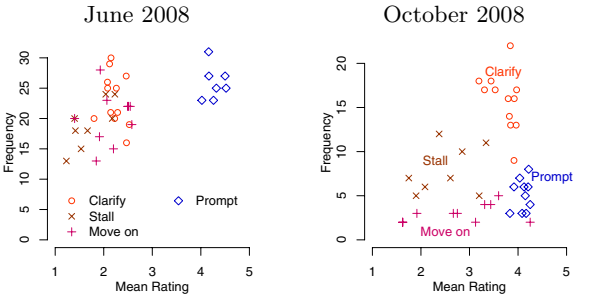


Fig. 4. Rating and frequency of response types

off-topic pattern, on the other hand, shows substantial variation. The October 2007 study shows a negative correlation between response rating and frequency ($r = -0.55, p < 0.01, df = 20$). The reason for this, as we argued in [6], is that some off-topic responses were linked to out-of-domain questions in the training data – for example, the question "so do you have a girlfriend?" was linked to the response "ha ha, you're a bad man". This boosted the frequency of the linked responses, but these turned out to be lower rated than clarification requests like "I didn't hear that, could you repeat the question?", which were typically not linked to any question. This observation led to reauthoring the off-topic responses, dividing them into classes, and unlinking all of them in the training data. The result in the June 2008 study is the absence of a significant correlation between rating and frequency for off-topics ($r = 0.35, p = 0.06, df = 28$), since they all appear with similar frequency. Finally, the implementation of the off-topic selection policy in the October 2008 study resulted in a positive correlation between rating and frequency ($r = 0.53, p < 0.005, df = 28$), reflecting the fact that the more coherent clarification requests are now also the most frequent, due to the fact that the off-topic policy chooses them as the first response. The last panel of Figure 4 shows that in the October 2008 study, the off-topic responses fall into the same pattern as on-topic responses with respect to the rating-frequency relation, and no longer stand out as a distinct cluster.

The question remains, whether the improvement in the scores of the frequent off-topic responses is due only to the fact that the new policy ensures that clarification requests are more frequent, or whether part of the improvement can be attributed to the actual sequencing. In other words: would placing clarification requests as the second and third responses in a sequence of off-topics improve coherence or degrade it? The data do not provide a clear answer. Figure 5 is a rating-frequency plot like Figure 4, highlighting the separate off-topic response classes. The different response classes have similar frequencies in the June 2008 study but different frequencies in the October 2008 study – this is a direct result of implementing the off-topic selection policy. In both studies, clarification questions are rated as the most coherent, though the difference between "clarify" and "move on" is not significant in the June 2008 study. Note that the October 2008 study had fewer data points overall, which makes reaching significance more



	Mean rating	
	June 2008	Oct 2008
Clarify	2.21	3.69
Stall	1.77	2.58
Move on	2.16	2.94

	t	
	June 2008	Oct 2008
Clarify–Stall	6.16 ***	9.94 ***
Clarify–Move on	0.59 n/s	3.59 **
Stall–Move on	−3.99 ***	−1.59 n/s

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Fig. 5. Rating and frequency of off-topic responses

difficult, especially for the less frequent classes. We do not have an explanation for the overall increase in ratings for all classes of off-topics from June 2008 to October 2008; the responses were identical in the two studies, the only difference being the selection policy. It would be nice to attribute the increase in coherence to the implementation of the policy, though another possibility is that the raters may have interpreted the task differently, perhaps due to an emphasis shift in the instructions (these were revised between the two studies). The ratings for the October 2008 study are probably more trustworthy, since inter-rater reliability was somewhat higher.

5 Speech Recognition and Classifier Scores

Automatic speech recognition (ASR) affects performance [4]: 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 in October 2007, 0.365 in June 2008, and 0.428 in October 2008; the results are not directly comparable because the language models were re-trained for each study, and the acoustic environments differed. Figure 6 shows the distribution of utterance word error rates.

In all three studies we found a highly significant negative correlation (ranging from $r = -40$ to $r = -0.47$) between the rating of SGT Star's on-topic responses

[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.

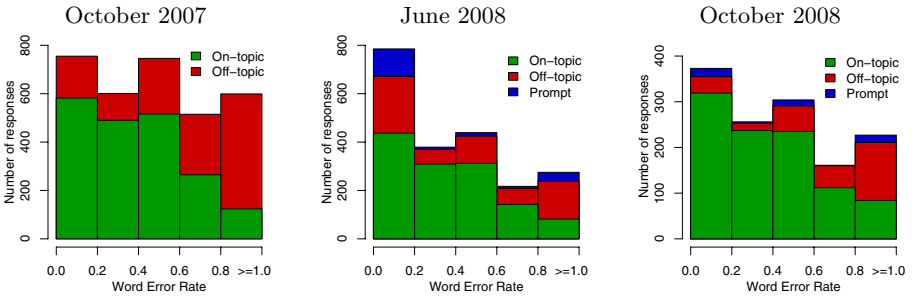


Fig. 6. Word error rates and the responses they triggered

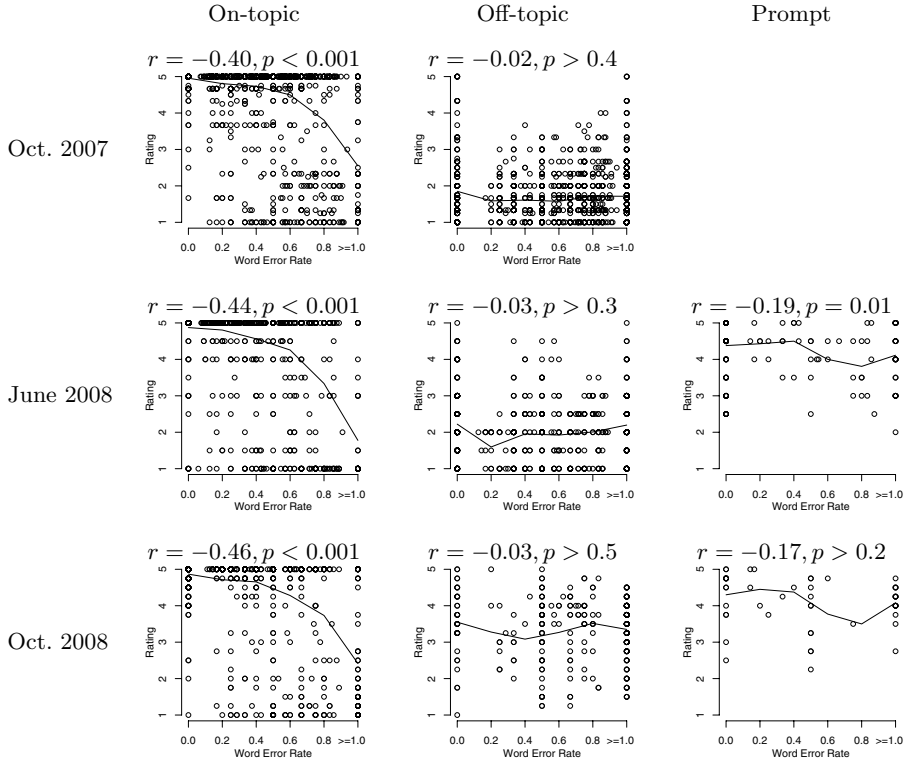


Fig. 7. Word error rates and ratings: the lines show the mean rating for each WER band

and the word error rate of the immediately preceding user utterances; off-topic responses and prompts typically did not exhibit such a correlation (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

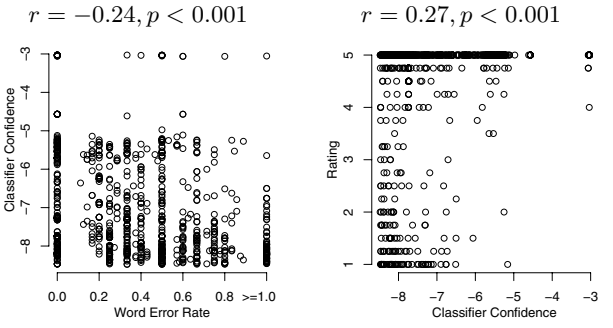


Fig. 8. Relation between speech recognition, classifier confidence and rating

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.

Our final measures concern the classifier confidence. We mentioned above that the decision whether to utter an on-topic or off-topic response depends on the classifier's confidence in the adequacy of the top-ranked response: if the confidence exceeds a specified threshold (determined during training), SGT Star utters that response, otherwise he gives an off-topic. With the rating study we can check how effective this strategy is. Figure 8 plots the word error rate of each user utterance against the classifier's confidence in the top-ranked response for that utterance, and the classifier's confidence against the rating (we only have data for the on-topic responses for the October 2008 study). The results are what we expect. Classifier confidence shows a negative correlation with word error rate, because noisy input is less similar to the input on which the classifier is trained. Confidence is positively correlated with coherence, meaning that the measure the classifier uses – similarity between language models – is similar to human judgment of the appropriateness of the responses. To evaluate the suitability of the threshold chosen by the system we will need to collect additional data, namely confidence and ratings for top-ranked responses that fall below the off-topic threshold.

6 Conclusion

We choose to call our method of evaluation "semi-formal" because it combines "hard" numbers taken from within the system with the "soft" numbers of the human rating study. The analysis is quantitative, but the conclusions are qualitative: The numbers are used to identify which classes of responses work in different dialogue contexts, and eventually which individual responses are good and which need to be improved upon. We believe that this sort of analysis allows better insight into the operation of SGT Star than a simple qualitative analysis of the dialogues, because the numbers reveal patterns such as the relative success of the various off-topic response classes, or the rating-frequency correlation for

on-topic responses which exposes the effect that the handlers have on steering SGT Star towards his more popular (and more successful) responses. Additionally, the ratings can be used to evaluate aggregations of specific responses from different contexts, which is an important tool for improved authoring.

In a system where responses are selected from a fixed pool rather than generated on the fly, authoring plays an important role: the improvement of off-topic responses from October 2007 to June 2008 is due mainly to the re-authoring of these responses, since the selection policy was not implemented until October 2008. But authoring appropriate responses for an interactive system that can respond to any user input is a difficult task. SGT Star's ability to respond appropriately depends on his training data – a list of questions, a list of responses, and links between the two. Developing the training corpus is an iterative process, whereby new questions are added based on actual user data, responses get added as the character's domain expands and are modified based on their acceptance by the audience, and links keep getting tweaked in order to improve the character's performance. The information from the semi-formal analysis is helpful at the level of the individual response, but more importantly it exposes global patterns in SGT Star's performance which influence the directions of future character development.

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. Thanks to Jacob Cannon, Joe Henderer, Ryan Kim and Jordan Peterson 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.

References

1. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2), 313–330 (1993)
2. Levin, E., Pieraccini, R., Eckert, W.: A stochastic model of human–machine interaction for learning dialog strategies. IEEE Transactions on Speech and Audio Processing 8(1), 11–23 (2000)
3. Walker, M.A.: An application of reinforcement learning to dialogue strategy selection in a spoked dialogue system for email. Journal of Artificial Intelligence Research 12, 387–416 (2000)
4. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, Association for Computational Linguistics, July 2006, pp. 18–27 (2006)

5. Leuski, A., Traum, D.: A statistical approach for text processing in virtual humans. In: 26th Army Science Conference, Orlando, Florida (December 2008)
6. Artstein, R., Gandhe, S., Leuski, A., Traum, D.: Field testing of an interactive question-answering character. In: ELRA Workshop on Evaluation, Marrakech, Morocco, May 2008, pp. 36–40 (2008)
7. Artstein, R., Cannon, J., Gandhe, S., Gerten, J., Henderer, J., Leuski, A., Traum, D.: Coherence of off-topic responses for a virtual character. In: 26th Army Science Conference, Orlando, Florida (December 2008)
8. Ai, H., Raux, A., Bohus, D., Eskenazi, M., Litman, D.: Comparing spoken dialog corpora collected with recruited subjects versus real users. In: Keizer, S., Bunt, H., Paek, T. (eds.) Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007, pp. 124–131. Association for Computational Linguistics (2007)
9. 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, Marrakech, Morocco (May 2008)
10. Patel, R., Leuski, A., Traum, D.: Dealing with out of domain questions in virtual characters. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS, vol. 4133, pp. 121–131. Springer, Heidelberg (2006)
11. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, ch. 12, pp. 129–154. Sage, Beverly Hills (1980)
12. Siegel, S., Castellan Jr., N.J.: Nonparametric Statistics for the Behavioral Sciences, 2nd edn., ch. 9.8, pp. 284–291. McGraw-Hill, New York (1988)
13. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4), 555–596 (2008)
14. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 1(1), 77–89 (2007)