Evaluating the Effectiveness of Dialogue for an Automated Spoken Questionnaire

Stephen Sutton, Brian Hansen, Terri Lander, David G. Novick, Ronald Cole

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
20000 NW Walker Rd.
PO Box 91000
Portland, OR 97291-1000
Email: sutton@cse.ogi.edu

Abstract

We present and apply an empirical methodology for evaluating the effectiveness of dialogues in spoken language systems. This methodology is suitable in particular for evaluation of dialogue-based systems that collect information from the user, such as an automated spoken questionnaire. Our method for assessing effectiveness involves coding answers from users for responsiveness. For this effort, we developed a behavioral coding scheme tailored to the requirements of automated spoken questionnaires interacting via the telephone. The codes cover a range of behavior from "Concise" to "No response."

We have used this evaluation methodology in the development of an automated spoken questionnaire. In connection with this project, we collected over 4,000 telephone calls responding to the questionnaire. A sample of the calls was transcribed and coded using our behavioral coding scheme. We then used the data from the codes to choose among alternative protocols for the dialogue and to evaluate differences in system voice, such as natural versus synthetic and male versus female. In particular, we illustrate the utility of our methodology by testing the hypothesis that a synthesized system voice would elicit more constrained user responses than a human voice and report the evaluation results.

Introduction

In this paper, we present an evaluation methodology that was used in the development of a prototype automated spoken questionnaire (ASQ) for the Year 2000 Census (Cole et al., 1994). Our motivation arises from the need for an evaluation metric and a systematic procedure with which to measure the effectiveness of a dialogue. This, in turn, would provide a basis for making dialogue design decisions, both in terms of establishing an appropriate system protocol and exploring the effects of varying the system's voice (e.g., human or synthesized).

Briefly, the prototype ASQ is designed to capture census information from a telephone dialogue. Typical information includes name, sex, date of birth, marital status, origin and race.[1] The nature of interaction is predominantly single-initiative and is led by the system. Information is elicited through a series of questions and follow-up questions (or sub-dialogues). For example, a person's marital status is determined as follows:

- Have you ever been married? Please say yes or no.

  If "yes" then:

- Which of the following best describes your marital status: now married, widowed, divorced or separated?

The project is currently entering an evaluation phase in which the system will be assessed as part of the larger 1995 Census Test. Approximately 200,000 callers from three selected test cities will be given the option of providing Census information by filling out (and returning by mail) a Census questionnaire, or via the telephone through interaction with a human operator or the ASQ.

Despite the many advances in spoken language system technology, state-of-the-art systems must still find ways to limit the expected vocabulary of user responses, especially in a task such as the Census where the system is faced with a wide variety of regional accents. The problem is further exacerbated by the periodic nature of the census task: the user cannot be expected to learn the capabilities of the system given that they interact with the system once every ten years. These factors require that our systems elicit only the most concise and recognizable answers to questions, and that we empirically refine the questions in order to elicit those concise answers. There is a danger however, in making conciseness our sole criterion; a protocol that relied entirely on yes/no answers would be easy for the recognizer but frustrating for users. Thus, the design process involves exploring the space of possible dialogues in search of one that is relatively natural without compromising the overall effectiveness of the system.

1. The specific information requirements were laid out by the U.S. Bureau of the Census in accordance with the contents of the written Census questionnaire.

What do we mean by effectiveness? The census task has a clearly defined objective, namely to elicit the requisite information from callers. Thus, we choose the responsiveness of callers' utterances as our primary evaluation metric. To this end, we developed a behavioral coding scheme for characterizing callers' utterances. We report on the replicability of coding results by multiple judges and illustrate the utility the coding scheme by testing the hypothesis that a synthesized system voice would elicit more constrained user responses than a human voice.

Behavioral coding

Behavioral coding is an analytical technique for classifying users' responses according to type. The behavioral coding scheme provides a set of features that characterize aspects of the response deemed useful for ASQ evaluation purposes. Particular emphasis was placed on capturing the responsiveness of utterances.

The behavioral coding scheme is an adaptation of a coding scheme used by the Bureau of the Census (Esposito & Rothgeb, 1992) for the purposes of pretesting a questionnaire. It was intended to identify particular interviewer and respondent behavior indicative of problems with a question or sections of the questionnaire. We made a number of key modifications to improve its suitability for use with an ASQ. For instance, the original scheme included a single class for "adequate answers," which we expanded into three separate classes that were significant for expected accuracy of recognition. Also, the single class for "inadequate answers" was expanded into two classes that would distinguish utterances useful for dialogue design. The precise nature of these new classes will be introduced shortly. These extensions provide additional detail useful for improving the system.

The behavioral coding scheme consists of eleven classes. Each class has an associated code which is used during transcription to label subjects' responses. Broadly speaking, the classes can be organized into four groups: (1) adequate answers, (2) inadequate answers, (3) "meta" codes, and (4) multiple codes. In the

Code	Response Class	Description	Example
AA1	Adequate Answer 1	Answer is concise and responsive.	S: Have you ever been married? U: Yes
AA2	Adequate Answer 2	Answer is usable but not concise.	S: Have you ever been married? U: No I haven't
AA3	Adequate Answer 3	Answer is responsive but not usable.	S: Have you ever been married? U: Unfortunately
IA1	Inadequate Answer 1	Answer does not appear to be responsive.	S: What is your sex, female or male? U: Neither
IA2	Inadequate Answer 2	User says nothing at all.	S: What is your sex, female or male? U: <silence>
QA	Qualified Answer	User expresses uncertainty in an otherwise adequate answer.	S: What year were you born? U: Nineteen fifty five I think
RC	Request for Clarification	User requests clarification as to the meaning of a concept or survey question.	S: Are you black, white or other? U: What do you mean?
IN	Interruption	User interrupts the speaking of the question.	S: What year were you born? U: *teen fifty five
DK	Don't Know	User responds "I don't know" or some other equivalent formulation.	S: Are you black, white or other? U: I'm not sure
RF	Refusal	User refuses to answer.	S: What year were you born? U: I'm not telling you
O	Other	User behavior that is not captured by the codes listed above.	S: What year were you born? U: Thirty two <noise>

Table 1: Summary of behavioral coding scheme

following sections, we describe the general characteristics of each group and individual classes within a group. Table 1 gives a summary of the behavioral coding scheme including example behavioral code classifications.[2]

Adequate answers

Adequate answers are responses that contain the required information. Adequacy is a measure of responsiveness with respect to the intent of the question rather than the intent of the answer. That is, even though the user may have attempted to provide a responsive answer, if she misunderstood the question and gave the wrong information it would be regarded as "inadequate" according to this definition.

We distinguish among three levels of adequate answers. `Adequate answer 1` (AA1) is a concise response in which the answer contains precisely the information sought. `Adequate answer 2` (AA2) is a response considered usable but not concise; an AA2 contains the sought-after information along with additional, usually predictable, words. An `Adequate Answer 3` (AA3) is responsive but is problematic in some aspect. Typical AA3 characteristics include:

- *Verbosity*. The speaker is verbose in an unpredictable way, such as "Yeah, we were married once but then I left her" in response to the question "Have you ever been married?"
- *Truncation*. The speaker gets interrupted by the system before she is done speaking. For instance, the speaker may begin to say "thirty four" but get cut off after uttering "Thirty-fo*"
- *Additional information*. The user provides more information than was specifically requested. For example, the user says both first name and last name in response to the request "Please say your last name."
- *Self correction*. The person provides multiple answers, as a result of correcting an apparent mistake, such as "yes, I mean no."
- *Paraphrasing*. The response is phrased in such a way that the meaning is still apparent given adequate knowledge and inference capabilities. For example, the response "I'm sure not female" in response to the question "What is your sex, female or male?"

From an ASQ perspective, AA1 responses consist entirely of target words that appear in a minimal recognition vocabulary. These are the preferred kind of response. AA2 responses consist of target words in a predictable context and require word-spotting speech recognition capabilities. AA3 responses typically require natural language processing, although the tendency for unpredictable words may exceed speech recognition capabilities and as such may often be unusable.

Inadequate answers

Inadequate answers are responses that do not contain the required information. We distinguish between two kinds of inadequate response. `Inadequate answer 1` (IA1) is a case where the user says something other than the information required. Responses falling into this class may arise for a number of reasons, including misunderstanding of the question and general uncooperative behavior. `Inadequate answer 2` (IA2) is a case where the user says nothing at all. There are a number of possible reasons for observing such behavior, including users choosing not to respond, users not realizing they were expected to respond, or users having hung up.

From an ASQ perspective, the challenge is to distinguish IA1 responses from other classes of response. Of particular concern is the problem of false positives: misrecognizing an inadequate response as an adequate response. IA2 responses present less of a problem than IA1 responses, although high levels of background noise can cause difficulties. Also, it is useful to monitor the drop-out rate of callers since this can provide some indication of user frustration levels.

Meta codes

In many ways, meta codes are related to the group of inadequate answers considered earlier. What sets them apart, however, is some apparent reason for their inadequacy. We distinguish between three kinds of Meta responses. `Request for clarification` (RC) responses are where the user, instead of providing an answer, seeks to clarify the question. There are various kinds of clarification requests possible, including general request for repetition, specific request for repetition, specific request for confirmation, and specific request for specification (Lloyd, 1992). `Don't know` (DK) responses are cases where the user indicates that the information is not available. `Refusal` (RF) responses are where the user explicitly declines to provide the requested information.

Multiple codes

Three codes are intended for use in certain restricted combinations with codes already described. The

2. See Lander (1994) for a more detailed account of the behavioral coding scheme, including labeling conventions.

Qualified answer (QA) code signifies uncertainty in the response. It should be used in conjunction with the AA3 code, such that an otherwise concise response (normally coded as AA1) becomes "AA3+QA" when qualified.

The *Interruption* (IN) code is used to label responses that interrupt system prompts. In systems that do not support overlapping speech, a user's interruption will result in part or all of the response being cut off. Note that this code is intended for use when the user interrupts the system rather than the other way around. The IN code is used in conjunction with either the AA3 or IA1 codes depending on the level of intelligibility of the response.

The *Other* (O) code denotes "other respondent behavior" and is used to classify behaviors such as extraneous speech not directed at the computer, laughs, coughs, sneezes and excessive background noise.

Evaluation

We conducted a data collection effort in which callers interacted with an early system prototype that incorporated recognition capabilities only at decision points in the dialogue. At all other places, the user's response was simply recorded for analysis by a human transcriber. This configuration enabled us to evaluate alternative protocols and to experiment with varying speaker characteristics (e.g., synthesized or human voice, male or female). At the same time, it enabled us to collect data to train task-specific speech recognizers. The data collection effort yielded approximately 4,000 calls. In the following sections, we describe two experiments that make use of these data and demonstrate the practicality of the behavioral coding scheme.

Inter-rater agreement

We conducted an experiment to assess the inter-rater agreement of two labelers with respect to the behavior codes. The study involved a random data sample containing 1141 utterances (100 calls) from the human/male condition. The overall level of agreement was measured at 91.3%. Table 2 provides a summary of agreement for each behavior code. It shows the number of times each labeler assigned a given category and the number of times the two labelers were in agreement. These data reflect multiple codes having been broken down into their constituent parts. It should be noted that these results are based on human judgments only. It is possible to perform automated screening of various kinds to filter out obvious errors. The fact that the level of agreement is quite high is in part due to the relatively high proportion of responsive utterances, which in turn is a reflection of the success of our efforts during earlier rounds of protocol design.

Code	Labeler1	Labeler2	Agreement
AA1	1036	1065	1029
AA2	53	37	34
AA3	50	34	29
IA1	1	4	1
IA2	1	0	0
QA	0	2	0
RC	0	0	0
IN	14	0	0
DK	0	0	0
RF	0	0	0
O	59	31	18

Table 2: Agreement statistics

An analysis of the cases of disagreement revealed various probable causes. The labelers had different levels of experience with using the behavioral coding scheme. One labeler was very experienced and had been directly involved in designing the coding conventions. The other labeler had only a moderate amount of experience. In addition to misunderstandings of the coding conventions and basic mistakes, there were perceptual differences as well. Some of the observed cases of perceptual differences were (1) uncertainty whether the caller was addressing the computer or someone else in the room; (2) ambiguity in the information given—in one case, the caller gave two names in response to a last name question and it is not clear whether these are both last names or whether these are a first and a last name; (3) disagreements arising from difficulties hearing and interpreting what the caller is saying; and (4) disagreements arising from difficulty in coding subjective judgements of levels of background noise.

Human versus synthetic speech

We conducted an experiment to compare the effects of human speech versus synthetic speech on the user's behavior. In one condition, the ASQ used a recorded human voice for system prompts. In the other condition, it used a commercially available speech synthesizer, DecTalk. We considered the distribution of behavior codes associated with individual questions. In particular,

Question	AA1		AA1&AA2		AA1&AA2&AA3	
	human	synth	human	synth	human	synth
name	91.3	92.1	96.3	95.5	99.7	99.6
sex	97.3	96.9	98.5	98.4	99.7	99.4
marital status	91.9	93.7	93.1	95.6 *	99.1	99.7 *
date of birth	91.8 *	90.1	95.5	94.8	99.6	99.6
origin	97.5	99.0 *	98.1	99.4 *	98.9	99.7
race	91.2	90.1	95.5	94.0	99.3	98.7

Table 3: Percentage of responsive answers

this initial analysis examines the distribution of AA1, AA1&AA2 and AA1&AA2&AA3 classes. These classes represent answers that are concise, concise or nearly so, and responsive, respectively.

Our initial hypothesis of how the conditions might influence a user's behavior is that answers would be more constrained in the synthesized case than in the human case. That is, we expected to see a higher proportion of AA1 responses in the synthesized case. Our reasoning was that we expected that people would be more careful and conscious of their responses when it is apparent they are conversing with a computer.

Table 3 contains a summary of the results. The table shows the percentage of responses in each class (AA1, AA1&AA2, AA1&AA2&AA3) for each question. Questions with sub-questions have been combined by taking the weighted average of the results from the sub-questions. Cases where one condition shows a significant increase in responsiveness are marked by an asterisk (*). The results provide inconclusive support for our hypothesis. The synthesized condition appears to show no notable increase in the proportion of AA1 answers. Likewise, the AA1&AA2 and AA1&AA2&AA3 cases provide little in the way of conclusive evidence to suggest an effect.

There is a certain amount of variability in these results, including cases where the human condition appears more responsive than the synthesized condition. A possible factor that may contribute to this variability is the "delivery" of each question. That is, prosodically "marked" questions are likely to influence the conciseness of the user's answers and so may account for the unexpected results. This claim will require further analysis to substantiate.

Even though this analysis of effects of the human versus synthetic speech provides inconclusive support for our hypothesis, it has served to demonstrate the utility of the behavioral coding scheme for evaluating alternatives in dialogue.

Discussion

Our behavioral coding scheme provides a useful metric for performing an objective evaluation of user behavior. Such an evaluation procedure can be complemented with subjective user feedback to provide an empirically-based evaluation of a protocol. Furthermore, when incorporated into an iterative design process for refining protocols, the behavioral coding scheme provides a powerful tool for developing an effective ASQ. In particular, the behavioral coding scheme aids ASQ development in a number of respects, including:

- *Performance*. The coding scheme provides an indication of how well the system can do. System evaluation should take into account the number of adequate answers (AA1, AA2 and AA3) not just the overall number of responses.
- *Effort allocation*. The coding scheme suggests where to expend effort: which questions of the protocol would benefit most from refinement. Also, which aspects of the system (e.g., speech recognition, natural language processing or dialogue) most warrant improvement.
- *Protocol development*. The coding scheme, when used in an iterative design process, helps guide the protocol refinement and improve responsiveness.
- *Training*. The coding scheme helps identify data most suitable for training speech recognizers.
- *Validation*. The coding scheme can be used as a cross-check for validating transcriptions and so helps to eliminate transcription errors.

- *Language*. The coding scheme is usable across different languages. For instance, we made use of the same scheme when developing a prototype Census ASQ for Spanish.

The behavioral coding scheme described is intended for summarizing user responses only. It would be possible to extend the scheme to cover aspects of interviewer (or system) behavior. Although it was designed in the context of the Census task, it should be possible to adapt it easily for use in other spoken language system tasks.

Conclusion

We have presented a coding scheme for classifying the behavior of users interacting with an ASQ. This coding scheme provides a basis for evaluating the effectiveness of ASQ dialogues based on quantifying the responsiveness of users. We have made extensive use of the coding scheme in the Census task for designing the dialogue of a prototype ASQ and measuring the system's effectiveness. This paper has provided a detailed account of the behavioral coding scheme and discussed some of its uses. We reported on an experiment in which we measured the inter-rater agreement between two labelers, with respect to the behavioral coding scheme, to be 91.3%. Also, we demonstrated the utility of the coding scheme in testing the hypothesis that synthesized speech would elicit more concise responses than recorded human speech when used for system prompts. Our results provided inconclusive support for this hypothesis.

References

Cole, R.A., Novick, D.G., Fanty, M., Vermeulen, P., Sutton, S., Burnett, D., Schalkwyk, J. 1994. *A prototype voice-response questionnaire for the U.S. Census.* Proceedings of International Conference on Spoken Language Processing, 683-686. Yokohama, Japan.

Esposito, J., and Rothgeb, J. 1992. *Behavior coding manual: CATI/CAPI overlap.* Draft report. Bureau of Labor Statistics/Bureau of the Census.

Lander, T. 1994. *Behavior Code Guide*. Internal report. Center for Spoken Language Understanding, Oregon Graduate Institute.

Lloyd, P. 1992. The role of clarification requests in children's communication of route directions by telephone. *Discourse Processes*, 15: 357-374.