PARADISE-style Evaluation of a Human-Human Library Corpus

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
becky@cs.columbia.edu

Irene Alvarado
School of Engineering and Applied Science
Columbia University
ia2178@columbia.edu

Phil Crone
Simon Jerome
Columbia College
Columbia University
ptc2107@columbia.edu
sgj2111@columbia.edu

Abstract

We apply a PARADISE-style evaluation to a human-human dialogue corpus that was collected to support the design of a spoken dialogue system for library transactions. The book request dialogue task we investigate is informational in nature: a book request is considered successful if the librarian is able to identify a specific book for the patron. PARADISE assumes that user satisfaction can be modeled as a regression over task success and dialogue costs. The PARADISE model we derive includes features that characterize two types of qualitative features. The first has to do with the specificity of the communicative goals, given a request for an item. The second has to do with the number and location of overlapping turns, which can sometimes signal rapport between the speakers.

1 Introduction

The PARADISE method for evaluating task-based spoken dialogue systems (SDSs) assumes that user satisfaction can be modeled as a multivariate linear regression on measures of task success and dialogue costs (Walker, et al. 1998). Dialogue costs address efficiency, such as length of time on task, and effort, such as number of times the SDS fails to understand an utterance and re-prompts the user. It has been used to compare subjects performing the same or similar tasks across distinct SDSs (Sanders, et al. 2002). To our knowledge, it has not been applied to human-human dialogue.

For human-human task-based dialogues, we hypothesized that user satisfaction would not be predicted well by measures of success and dialogue costs alone. We expected that qualitative characteristics of human-human dialogue, such as the manner in which a dialogue goal is pursued, could counterbalance high dialogue costs. To test this hypothesis, we performed a PARADISE-like evaluation of a corpus of human-human library transaction dialogues that was originally collected to support the design of our SDS (Passonneau, et al. 2010). The communicative task we examine is to identify a specific set of books of interest from the library's holdings. This can be straightforward if the patron requests a book by catalogue number. It can be complex if the patron does not have complete bibliographic information, or if the request is non-specific. A book request is successful when the librarian identifies a specific book that addresses the patron's request.

Task success was predictive on a training set, but not on a held-out test set. Dialogue costs were less reliably predictive. Two additional factors we found to be moderate predictors pertained to the number of book requests that were non-specific in nature, and the amount and location of overlapping turns. We refer to these as qualitative features. A non-specific book request can lead to a collaborative identification of a specific book, and the costs incurred can be worth the effort. We speculate that overlapping turns during non-task-oriented subdialogue reflects positive rapport between the speakers, while the role of overlapping turns during task-oriented subdialogue is contingent on other characteristics of the task, such as whether the goal is specific or non-specific.

The three following sections discuss related work, our corpus, and our annotation procedures and reliability. We then present how we measure

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 325–331,
Portland, Oregon, June 17-18, 2011.

user satisfaction, informational task success on book requests, and various dialogue costs. This is followed by results of the application of PARADISE to the human-human corpus.

2 Related Work

It is commonly assumed that human-computer interaction should closely resemble human-human interaction. For example, the originators of social presence theory propose that media that more closely resemble face-to-face communication provide a higher degree of social presence, or awareness of the communicative partner (Short, et al. 1976), which in turn leads to communicative success. A similar idea is seen in the origins of media richness theory (Daft and Lengel 1984), which defines media with more "richness" as having more communication cues, and thus enhancing task success. A key component of this assumption is that, if computers are created with human-like qualities then people will view computers similarly to humans. We hypothesize that human-machine dialogue need not resemble human-human dialogue in all respects, thus we earlier proposed a method to investigate human-machine dialogue despite the large disparity in the spoken language processing abilities of humans versus machines (Levin and Passonneau 2006), and applied it work described in this proceedings (Gordon, et al. 2011). Here, we apply PARADISE to human-human dialogue to facilitate comparison.

Turn-taking in conversation has received a significant amount of attention. Early work examined the types of turn-taking attempts and the reasons why such attempts either succeed or fail (Beattie 1982). Recent research has focused on the acoustic, lexical, and discourse-relevant cues that indicate a transition between speakers (Beňuš 2009, Gravano and Hirschberg 2009). More recently, turn-taking has been examined in the context of multi-tasking dialogues (Yang, et al. 2011). The Loqui human-human dialogues often involve multiple tasks. We do not annotate who has the floor, but we do transcribe overlapping speech, where there may be competition for the turn (see below).

3 Loqui Human-Human Corpus

Our baseline SDS, CheckItOut, is modeled on library transactions for the Andrew Heiskell Braille and Talking Book Library of New York City, and is part of the Library of Congress. Patrons request books from librarians by telephone, and receive book orders (primarily in recorded format) by mail. Early in the project, we recorded 175 patron-librarian calls at the Heiskell Library, 82 of which we identified to be primarily about book information and book orders. These were transcribed with an XML transcription tool, and utterances were aligned with the speech signal. The total number of words is approximately 24,670, or about 300 words per dialogue. Our transcription conventions are documented on our website.[1]

To facilitate analysis of the interactive structure of many types of interaction, such as spontaneous spoken dialogue, email, and task-oriented dialogue, we previously developed Dialogue Function Unit (DFU) annotation (Hu, et al. 2009). The primary motivation was to capture information about *adjacency pairs*, sequences of communicative acts in which an initial utterance calls forth a responding one (Sacks, et al. 1974). DFUs encode links between the elements of an adjacency pair, and a restricted set of dialogue acts designed to generalize across genres of interaction. Trained annotators applied DFU annotations to all 82 dialogues.

To measure task success and dialogue costs, we developed an additional annotation process that builds on DFU annotation, as described next.

4 TSC Annotation

In our human-human corpus, each patron has a different set of goals. For most of the dialogues, at least some of the patron's goals are to request books from the librarian. Other goals include requesting an update to the patron's profile information, requesting new equipment for listening to recorded books, and so on. The three-step method developed for annotating task success, dialogue costs and qualitative features (TSC Annotation) consists of an annotation step to determine what tasks are being executed, and two tabulation steps. The 82 dialogues that had already been annotated for DFUs were then annotated for task success and dialogue costs.[2] Three annotators were trained in the annotation over the course of several one-hour sessions, each of which was devoted to a different

[1] See resources link at http://www1.ccls.columbia.edu/~Loqui/.
[2] The guidelines are at http://www1.ccls.columbia.edu/~Loqui/resources.html.

16.1.0	L	*wh- wha- do you have the author?*
		[Request-Info: author of book]
17.1.0	P	*Cesar Millan*
		[Inform: author is Cesar Millan]
18.1.0	L	*M I L A N?*
		[Request-Info: is librarian's spelling correct]
19.1.0	P	yes
20.1.0	L	<non-speaking-librarian-activity>
21.1.1	P	can you hold on just {one second}
		[Request-Action: can librarian hold]
21.1.2	L	{sure sure}
		[Confirm]
22.1.0	P	I'm back
23.1.1	L	I'm sorry I'm not seeing anything {by him}
		[Inform: Nothing by this author]
23.1.2	P	{really}
		[Request-Info: yes/no]
24.1.0	L	no
		[Disconfirm]
		BOOK REQUEST 1.1

Figure 1. Book request DTU

sample dialogue. Pairs of annotators worked on each dialogue, with one annotator reviewing the other's work. Disagreements were adjudicated, and interannotator agreement was measured on three dialogues.

4.1 Annotation

The annotation procedure starts by dividing a transcription of a dialogue into a covering sequence of communicative tasks (Dialogue Task Units, or DTUs). Each DTU encompasses a complete idea with a single goal. It ends when both speakers have collaboratively closed the topic, per the notion of *collaborative contributions to discourse* found in (Clark and Schaefer 1989). Each DTU is labeled with its type. The two types of DTUs of most relevance here are book requests (BRs; where a patron requests a book), and librarian proposals (LPs; where the librarian proposes a book for the patron). Each BR or LP is numbered. Other DTU types include Inform (e.g., patron requests the librarian to provide a synopsis of a book), and Request-Action (e.g., patron requests the librarian update the patron's profile). After the DTUs have been annotated, success and task measures are tabulated for the book requests (BR and LP): the start and end lines, the specificity of the request (a request for any book by a given author is non-specific), and whether the task was successful.

Figure 1 shows part of a *book request* DTU. The DTU in Figure 1 is unsuccessful; the librarian is unable to identify the book the patron seeks. Several DTUs might pertain to the same goal, pursued in different ways. For example, the DTU illustrated here is the second of three in which the patron tries to request a book called *The Dog Whisperer*. The dialogue contains 7 DTUs devoted to this request, which is ultimately successful.

Figure 1 also illustrates how we transcribe overlapping utterances. Each line in Figure 1 corresponds to an utterance, or in the case of overlapping speech, to a time segment consisting of an utterance with some overlap. Patron utterance 21.1.1 is transcribed as ending with overlapping speech (in curly braces) where the librarian is also speaking within the same time segment (21.1.2). This is followed by the patron's utterance 22.1.0. The next time segment (23) also has an overlap, followed by the librarian's turn 24.1.0. As a result, we can investigate the proportion of utterances in a dialogue or subdialogue with overlapping speech, and the types of segments where overlaps occur.

4.3 Interannotator Agreement

To assess interannotator agreement among the three annotators, we randomly selected dialogues from a set that had already been annotated until we identified three that had been annotated by distinct pairs of annotators. Each was then annotated by a different third annotator who had not been a member of the original pair. Interannotator agreement on DTU boundaries and labels was measured using Krippendorff's alpha (Krippendorff 1980). Alpha ranges from 0 for no agreement above chance prediction, given the rate at which each annotation value is used, to 1 or -1, for perfect agreement or disagreement.

The three dialogues had alpha values of 0.87, 0.77 and 0.66, thus all well above agreement that could have resulted from chance. The dialogue with the highest agreement had 1 book request consisting of 2 DTUs. The first DTU had a non-specific request for two books by a given author, that was later reformulated in the second DTU as a specific request--by author and titles--for the two books. The dialogue with the next highest agreement had 12 specific book requests by catalogue number, and one DTU per book request. The dialogue with the lowest agreement had 5 book requests, with one DTU per book request. Two were by catalogue number, one was by author, and one was by author and title.

5. Perceived User Satisfaction

An indirect measure of User Satisfaction for each dialogue was provided by two annotators who listened to the audio while reviewing the transcripts. The annotators completed a user satisfaction survey that was nearly identical to one used in an evaluation of CheckItOut, the SDS modeled on the library transactions; references to *the system* were replaced with *the librarian*. It contained ten questions covering the librarian's clarity, friendliness, helpfulness, and ability to communicate. The annotators rated the perceived response of the caller with regard to the survey questions. On a 1 to 5 scale where 5 was the greatest satisfaction, the range was [3.8, 4.7], thus overall, patrons were perceived to be quite satisfied.

6. Task Success

The dialogue task investigated here is informational in nature, rather than a borrowing task. That is, a book request is considered successful if the librarian is able to identify the specific book the caller is requesting, or if the librarian and patron are able to specify a book in the library's holdings that the caller wants to borrow. The actual availability of the book is not relevant. Some patrons request a specific book, and provide alternative means to identify the book, such as catalogue number versus title. Some seek unspecified books by a particular author, or books in a given genre.

We calculate task success as the ratio of successfully identified books to requested books. The total number of books requested ranged from 1 to 24. Patron-initiated book requests as well as librarian-initiated proposals are included in the tabulation. In addition, we tabulate the number of specific book requests that change in the type of information provided (RC, title, author, genre, etc.) as well as the number of book requests that change in their specificity (non-specific to specific). Finally, we tabulate how many of these changes lead to successful identifications of books.

In general, task success was extremely high. More than 90% of book requests were successful; for 78% of the dialogues, all book requests were successful. This high success rate is to be expected, given that most callers are requesting specificc books they learn about from a library newsletter, or making non-specific requests that the librarian can satisfy.

7. Dialogue Costs and Qualitative Features

Along with two measures of task success (number of successfully identified books: Successful.ID; percent of requested books that are successfully identified: Percent.Successful), we have 48 measures of dialogue costs and qualitative features. The full list appears in column 1 of the table in Appendix A. Dialogue costs consist of measures such as the total number of turns, the total number of turns in book requests, the total number of utterances, counts of interruptions and misunderstandings by either party, and so on. Qualitative features include extensive clarifications, the types of book request, and overlapping utterances.

An extensive clarification serves to clarify some misunderstanding by the caller, and generally these segments take at least ten turns.

We classify each book request into one of seven types. These are non-specific by author, non-specific by genre, specific author, specific title, specific author and title, specific set, and specific catalogue number. As shown in the Appendix, we also tabulate the total number of specific book requests per dialogue (S.Total) and the total number of non-specific requests (NS.Total).

We tabulate overlapping utterances in a variety of ways. The average number of overlapping utterances per dialogue is 13.9. A breakdown of overlapping utterances into those that occur in book requests versus other types of DTU gives a mean of 4.36 for book requests compared with 8.74 otherwise. We speculate that the difference results from the potential for overlapping utterances to impede understanding when the utterance goals are to request and share information about books. In these contexts, overlap may reflect competition for the floor. In contrast, overlapping utterances at points in the dialogue that pertain to the social dimension may be more indicative of rapport between the patron and the librarian, as a reflection of sharing the floor. We do not attempt to distinguish overlaps with positive versus negative effects. We do, however, tabulate overlapping speech in different types of DTUs, such as book request DTUs versus other DTUs.

To illustrate the role of the qualitative features, we discus one of the dialogues in our corpus that exemplifies a property of these human-human dialogues that we believe could inform SDS design: high user satisfaction can occur despite low

success rate on the communicative tasks. Dialogue 4 had the lowest task success of all dialogues (62.5%), yet perceived user satisfaction was quite high (4.7). This dialogue had a large number of book requests and librarian proposals, with a mix of requests for specific books by catalogue number, title, or author and title, along with non-specific requests for works by given authors. It also had a fairly high proportion of overlapping speech. As we discuss next, both dimensions are represented in the quantitative PARADISE models for predicting user satisfaction.

8. PARADISE Results

PARADISE predicts user satisfaction as a linear combination of task success and cost variables. Here we apply PARADISE to the Loqui library corpus, and add qualitative features to task success and dialogue costs. Six of the dialogues had no book requests, thus did not exemplify the task, namely to identify books for the patron in the library's holdings. These six were eliminated.

We split the data into independent training and test sets. From the 76 dialogues with book requests, we randomly selected 50 for deriving a regression model. These dialogues had a total of 211 book requests (mean=4.22). We reserved 26 dialogues for an independent test of how well the features from the user satisfaction model on the training set predicted user satisfaction on the test set. The test set had 73 book requests (mean=2.81).

To explore the data, we first did Analysis of Variance (ANOVA) tests on the 50 individual features as predictors of perceived user satisfaction on the training set. Certain features that are typically predictive for SDSs were also predictive here. Those that were most predictive on their own included the proportion of book requests successfully identified (Pct.Successful), and several cost measures such as total length in utterances, and the total number of interruptions and misunderstandings. However, other features that were predictive here that are not typical of human-machine dialogue were the number of utterances with overlapping speech (Simultaneous.Utterances), and the number of book requests that evolved from non-specific to specific (Change.NS.to.S).

Given the relatively small size of our corpus, and the large number of variables, we pruned the 30 features from the trained model before using them to build a regression on the test set. All analyses were done in the R Statistical Package (http://www.r-project.org/). We used the R function **`step`** to apply the Akaike Information Criterion to guide the search through the model space. The resulting model relies on 30 of the 50 variables, and has a multiple R-squared of 0.9063 (p= 0.0001342). Appendix A indicates the 30 features selected, and their p-values. For the pruned model, we selected half of the 30 features that contributed most to the best model found through the step function on the training set. The pruned model had a multiple R-squared of 0.5334 (p=0.0075). When we used the same features on the test set, the R-squared was 0.7866 (p=0.0416). However, the significance of individual features differed in training versus test. Appendix A lists the 15 features and their p-values on the training and test sets.

On the training data, the most significant features were Pct.Successful, the total number of dialogue segments pertaining to book requests (including librarian proposals; BR.request.segs), and the total number of book requests (Total.BR). The number of non specific book requests that evolved into specific requests (Change.NS.to.S) and the number of utterances per turn (Utterances.Turns) were marginally significant.

On the test data, the most significant variables were the ratio of overlapping utterances in segments that were not about book requests to book request segments (noBRLP.Overlap.per.TotalRequestSegments), the total number of non-specific book requests (NS.Total), and the number of overlapping utterances (Overlap.Utterances).

9. Conclusion

The human-human corpus examined here is an appropriate corpus to compare with human-machine dialogue, in that our SDS was modeled on the book requests in the human-human corpus. The R^2 values indicate that the regression models based on the 15 features fit the data well, yet the coefficients and probabilities are very different. In part, this is due to the large number of variables we investigated, relative to the small size of the corpus. Nevertheless, the results presented here point to a number of dimensions of human-human dialogue that contribute to user satisfaction beyond those that are typically considered when evaluating human-machine dialogue.

References

Beattie, G. W. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. Semiotica, 39 (1-2): 93-114.

Beňuš, Š. 2009. Are we 'in sync': Turn-taking in collaborative dialogues. In 10th Interspeech, pp. 2167-2170.

Clark, H. H. and E. F. Schaefer 1989. Contributing to discourse. Cognitive Science, 13 259-294.

Daft, R. L. and R. H. Lengel 1984. Information richness: A new approach to manager behavior and organization design. Research in Organizational Behavior, 6 191-233.

Gordon, J., et al. 2011. Learning to balance grounding rationales for dialogue systems. In 12th Annual SIGdial Meeting on Discourse and Dialogue (SIGdial 12).

Gravano, A. and J. Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue. In 10th Annual Meeting of SIGDIAL, pp. 253-261.

Hu, J., et al. 2009. Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. In 10th SIGDIAL on Dialogue and Discourse, pp. 357-366.

Krippendorff, K. 1980. Content Analysis: An Introduction to Its Methodology. Beverly Hills, CA: Sage Publications.

Levin, E. and R. J. Passonneau. 2006. A WOz Variant with Contrastive Conditions. In Interspeech Satelite Workshop, Dialogue on Dialogues: Multidisciplinary Evaluation of Speech-based Interactive Systems.

Passonneau, R. J., et al. 2010. Learning About Voice Search for Spoken Dialogue Systems. In 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 840-848.

Sacks, H., et al. 1974. A simplest systematics for the organization of turn-taking for conversation. Language, 50 (4): 696-735.

Sanders, G. A., et al. 2002. Effects of word error rate in the DARPA Communicator data during 2000 and 2001. International Journal of speech Technology, 7 293-309.

Short, J., et al. 1976. The social psychology of telecommunications. Chichester: John Wiley.

Walker, M. A., et al. 1998. Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. Computer Speech and Language, 12 317-348.

Yang, F., et al. 2011. An Investigation of interruptions and resumptions in multi-tasking dialogues. Computational Linguistics, 37 (1): 75-104.

Appendix A: Features

	Variable	Training Coeff.	Training p-value	Pruned Coeff	Pruned p-value	Test Coeff.	Test p-value
1	Successful.ID						
2	Pct.Successful	**0.504001**	**0.005118**	**0.356516**	**0.01219**	-0.04154	0.86744
3	Change.NS.to.S	**1.440471**	**0.023525**	**0.287376**	**0.05761**	0.10284	0.22876
4	Successful.NS.to.S	**-1.450301**	**0.048656**				
5	Change.S.to.S						
6	Successful.S.to.S						
7	BR.request.segs	-0.201228	0.119857	**-0.147057**	**0.00837**	0.02566	0.79277
8	LP.request.segs	**0.146464**	**0.073138**				
9	Total.Request.Segments						
10	Total.BR	**0.448858**	**0.001813**	**0.147945**	**0.01220**	-0.09960	0.35796
11	Segments.per.BR	**0.296577**	**0.047333**	0.123411	0.17907	-0.08707	0.59903
12	NS.Author	-0.216559	0.090830				
13	NS.Genre	-0.138867	0.249339				
14	S.Title						
15	S.AuthorTitle						
16	S.Set	**-0.953284**	**6.61e-05**				
17	S.RC	-0.158897	0.104752				
18	S.Author						
19	S.Total						
20	NS.Total			0.013265	0.75986	-0.27280	**0.00716**
21	Turns.in.BR						
22	Utterances	**-0.005613**	**0.013967**				
23	Interruptions	**0.187876**	**0.002704**	-0.050500	0.29683	-0.29078	**0.05378**
24	Misunderstandings						
25	Simultaneous.Utterances	**-0.151491**	**0.001967**	-0.008705	0.21024	0.02329	**0.04179**
26	Extensive.Clarifications	**-0.181057**	**1.76e-05**	-0.022723	0.25767	-0.08685	0.11608
27	S.U.Conventional	**0.142152**	**0.006168**				
28	S.U.Inform	**0.141891**	**0.001619**				
29	S.U.Sidebar	**0.107238**	**0.047303**				
30	S.U.BR.RC	**0.142538**	**0.006467**				
31	S.U.BR.Title	**0.245880**	**0.000415**				
32	S.U.BR.Title.and.Author	**0.136412**	**0.002581**				
33	S.U.BR.Genre						
34	S.U.LP	**0.176515**	**0.015598**				
35	S.U.R.A.	**0.171413**	**0.001459**				
36	S.U.IR.IRA	**0.166315**	**0.001994**				
37	Utterances.Turns	**-0.392267**	**0.020190**	-0.256307	0.08077	0.01731	0.95674
38	Total.Turns.BR						
39	Turns.in.BR.BR	-0.015623	0.093573				
40	BR.Utterances	**-8.875951**	**0.000603**	-1.104338	0.55174	2.59438	0.33439
41	NS.Total.per.BR	0.183761	0.177739	-0.102524	0.33547	0.31111	0.10004
42	S.U.BRLP						
43	S.U.BRLP.per.BR						
44	S.U.BRLP.per.TotalRequestSegs						
45	S.U.nonBRLP						
46	S.U.nonBRLP.per.BR						
47	S.U.nonBRLP.per.TotalRequestSegs	0.024492	0.117363	0.007839	0.33727	-0.06000	**0.00848**
48	S.nonRC						
49	S.nonRC.per.BR	-0.370227	0.064299	-0.062149	0.46085	-0.08072	0.47704
50	S.nonRC.per.TotalRequestSegs						