Using Interaction to Improve Intelligence: How Intelligent Systems Should Ask Users for Input

Anind K. Dey, Stephanie Rosenthal and Manuela Veloso
School of Computer Science
Carnegie Mellon University
{anind, srosenth, mmv}@cs.cmu.edu

Abstract

Intelligent systems will often need to collect input from users, to provide labels for training data or to correct mistakes the system makes. One interesting avenue of research is how to formulate the questions an intelligent system asks a user, in order to obtain the most accurate responses. In this paper, we study the impact of varying 5 dimensions of questions on response accuracy: indicating uncertainty, amount of context, level of context, suggesting an answer and asking for supplemental information. In a study of an email sorting task, we show that there is a combination that results in higher levels of accuracy than other combinations and validate this combination in a comparison to questions that a panel of HCI and email experts chose. The contributions of the paper are the approach to determine the best combination of dimensions, the validated combination, and a demonstration of how this type of question interaction can improve intelligent systems.

1 Introduction

There has been a lot of research on active learning systems, in which intelligent systems choose data that they are most uncertain about and need labels for, and ask users to provide those labels. Active learning is commonly used when there is lots of available data, but labeling this data is very expensive. Using active learning can reduce the number of examples that the intelligent system needs to produce an effective learning system. While much of the work in active learning has focused on selecting the appropriate examples to ask the user about, there has been considerably less attention paid to the manner in which these questions are asked.

In particular, we are interested in understanding whether we can determine the content of questions, in order to improve the accuracy of user responses. Correct labels are essential for drawing correct conclusions from observed data. Inaccurately labeled data results in poorer accuracy of a machine learner, which may believe it predicts labels correctly, but really does not. For example, a physical activity coach may predict that a user is lazy and not moving when he is really running. If the coach reports the activity data to the user's doctor, the doctor would have incorrect information with which to diagnose or may draw incorrect conclusions about the cause of an illness.

We are interested in how a computer can help elicit labels from users for data to automate the process of labeling data, as is often done with experience sampling [Larson and Csikszentmihalyi, 1983]. More specifically, we want to maximize the proportion of correct labels users provide so that data is as *accurate* as possible (*e.g.*, [Hoffmann, *et. al.*, 2009]). Knowledge elicitation techniques such as interviews and scenario simulation can be used to acquire very accurate labels from people, and even obtain supplemental information about the data, like rules or generalizations for how classify similar observations [Shadbolt and Burton, 1990]. These techniques have been traditionally performed by people, but recent work has shown that a computerized knowledge elicitor (KE) can collect labels from users when it explains scenarios about what it wants labeled [Stumpf *et al.*, 2007]. However, this work does not focus on maximizing the accuracy of the labels.

The focus of our work is to understand how varying a knowledge elicitor's questions affects the accuracy of users' responses. We vary the elicitor's questions across five dimensions taken from the HCI literature and incorporate both dimensions for contextual information and usability techniques: explaining that the active learner is uncertain, the amount of identifying context to provide, the level of that context, whether the learner predicts an answer to the question, and whether it asks for additional input. We apply this approach to an email sorting task and using the results of the study, we contribute a set of guidelines for how a knowledge elicitor should formulate questions to optimize the accuracy of users' responses. We validate those results against HCI community advice about how to ask questions and show that our guidelines are better than the community advice. Next, we discuss related work on computerized knowledge elicitation.

2 Related Work

The interpretation and understanding of language has been of interest to the social psychology and human-computer interaction (HCI) research communities for years. Differences in understanding between a researcher and respondents can affect the validity and reliability of surveys and questionnaires and the usability of interfaces [Presser, 2004]. The social psychology and HCI communities have developed guidelines on how to write survey questions and techniques like focus groups, interviews, cognitive walkthrough, and pretests to help researchers iterate on and improve their surveys and user interfaces [Jeffries *et al.*,

1991].

We focus on the active elicitation of classifications from users for learning, although other techniques like implicit learning and critiquing have also been shown to be effective for learning [Steinfeld, *et. al.,* 2006; Gajos and Weld, 2005]. The goal of our work is to present guidelines for designing understandable questions that intelligent systems can ask users to improve their learning and reasoning abilities. We draw the content of these questions from data that an intelligent system already collects and reasons about, namely uncertainty, context, prediction, and feature selection. These dimensions have been examined in previous HCI work on personalized interfaces, but their effects on how humans respond have not been widely studied. We now discuss why these 5 dimensions are relevant to questions asked by intelligent systems.

1) *Uncertainty*: Intelligent systems must calculate their uncertainty in order to decide when to ask for help. Studies on context-aware, expert, and recommender systems all show that providing users with the level of uncertainty in a system's predictions improves its overall usability (e.g., [Antifakos *et al*., 2004; Banbury *et al*., 1998]). Additionally, in one task where users had to remember a set of numbers given an imperfect memory aid, users showed improved recall when they were given the uncertainty/confidence information compared to the same display without the uncertainty information [Antifakos *et al*., 2004]. An intelligent system that tells the user that it is uncertain may not only improve its usability but also user accuracy in responding to questions.

2) *Context Amount* and 3) *Context Level*: In order for intelligent systems to interact with the environment, they receive sensor data from their environment or interface. If they are uncertain of how to act, they can obtain assistance from the user and then learn to associate their current context (state) with the new action. Additionally, the user may require the intelligent system's contextual information to understand what the system is referencing in its question. Studies have shown that when a human and robot share a common frame of reference in the environment, they can communicate more effectively (e.g., [Torrance, 1994; Topp *et al*., 2006; Steel, 2003]). However, it is unclear how much contextual information is needed to answer questions accurately and how much is too much when a user starts finding the explanation annoying. Similarly, it is unclear whether low-level context or sensor data is more beneficial than inferred or high-level context.

4) *Prediction*: When an intelligent system asks a user for help, it requires additional effort on the part of the user to understand what is being asked, limiting the user's productivity (*e.g*., [Shiomi *et al*., 2008]). In HCI, there has been an effort for agents to proactively provide predictions of the next action to take and warnings that errors are about to occur in order to reduce the cognitive load of humans who are assisting them (*e.g*., [Shiomi *et al*., 2008; Stumpf *et al*., 2005]). The idea is that confirming an answer is easier than generating an answer reducing the amount of work the user has to do and possibly increasing the accuracy of the response.

5) *Feature Selection*: Intelligent systems may collect a lot of sensor data, but may only process some of it for completing a certain task. If they have picked the wrong features of the data to process, this may affect the system's ability to complete the task correctly or may require that the system ask for help more often. If the system asks whether the features it selected are correct, it could confirm or change the features as necessary, possibly reducing the number of future questions. Although the extra question takes more time initially to answer, recent work in classifying email has demonstrated that a human may be more willing to give more time to provide the feedback as they are already interrupted, if it may decrease the possibility of additional interruptions [Stumpf *et al*., 2007; Stumpf *et al*., 2008].

We vary whether the intelligent system can provide different combinations of these dimensions in the questions it poses to users, and study the effects on the answers. We study all dimensions together to find dependencies and correlations between them and provide guidelines on the content of intelligent systems' questions to maximize the accuracy of the users' responses. We present a systematic analysis of these five dimensions below.

3 Method

We conducted a study to test the correctness of users' responses to an intelligent systems questions, based on variations in the content of those questions. In this case, the intelligent system was an email sorting system, attempting to classify all the email into distinct categories, and sort the email based on those categories. Email it could not sort was placed in the "Unsorted" folder and used to solicit input from users.

Thirty-seven participants (ages ranging from 18-61) were provided with a traditional email interface containing email about tasks for planning an upcoming academic conference. Participants were recruited from a popular experiment website through the university, but only half them were students. All of the participants were computer literate and had no trouble using Microsoft Excel, none of the participants had experience with writing machine learning algorithms, but half reported they helped a system learn – most from Gmail "reporting spam" and training speech recognition software. They were given the primary task of reading the email and consolidating all the changes that needed to be made to the conference schedule and website, using a provided spreadsheet. They were told that the application had already sorted most of the emails into folders based on the type of changes that needed to be made to the planning spreadsheet, and that it would ask them for help in sorting the rest, as they performed their planning task. Participants were given a limited amount of time to perform their task and were told that they could ignore these questions if they delayed their progress on the primary task. They were given

the incentive to respond to questions, by being told that they would complete the task a second time, and that any feedback they provided to the learning system would be used in their second trial, helping them complete this latter trial more quickly. By making the labeling task secondary, we allow more users to help when they can without requiring full commitment to making contributions all the time [Hoffmann *et. al.*, 2009].

The emails and task were selected and modified from the RADAR experiments dataset [Steinfeld *et al.*, 2006]. Each email in the dataset was labeled with a folder, used in this study. Additionally, we labeled each email with two "low level" keyword fields (one for sufficient and one for extra context) and two "high level" summary fields. The email interface was built with Adobe Flex. The email sorting application was Wizard-of-Oz'ed and questions from the application were automatically triggered when subjects clicked on "Unsorted" emails. The questions appeared in a pop-up, froze the rest of the interface, and would not disappear until users answered whether they were willing to answer the learner's question. Although users could read the subject line without being asked the questions, many subject lines did not provide enough information to classify the email before reading the email body. Additionally, subjects were allowed to read the text of the email while the question was up, but the rest of the email text was filler and did not provide any additional clues. This design ensured all participants were asked the same questions when they clicked on the same email.

3.1 Question Wording Dimensions

The five dimensions we use to formulate the KE's questions have been used in the previous work presented in the Related Work section, namely indicating uncertainty, providing context, high/low level context, suggesting an answer, and requesting supplemental information. We examine all dimensions at once to find dependencies and correlations between them. The content of the questions varied along these 5 dimensions for a 2x3x2x2x2 design, between subjects as follows:

1. **Indicating Uncertainty**: Whether the KE notifies a human that the active learner is uncertain of what to do or not (*e.g.,* [Antifakos *et al.*, 2004]). Half the subjects were told that the sorting system could not sort the email, while the other half were given no information about uncertainty.
2. **High/Low Level Context**: Whether the KE gives either low (*e.g.*, sensor) level context or high (*e.g.*, activity) level context (*e.g.,* [Salton and Buckley, 1990]). In the email talk, the low level context are keywords taken directly from the email while the high level context is a summary of the email. We ensured that the summary covered all keywords so that both conditions received the same context, just in different forms.
3. **Providing Context**: The amount of contextual or identifying information the KE should provide a user before asking for a label, namely none, sufficient, and extra information (*e.g.,* [Stumpf *et al.*, 2007]). In the email task, we vary the number (0, 2, or 4) of keywords or the length (0, 1, or 2 sentences) of the summary. Sufficient context is enough for a user to determine the folder without any other information, while extra context is twice as much information.
4. **Suggesting an Answer**: Whether the KE tells the human what the active learner predicts the answer to the question is (*e.g.,* [Stumpf *et al.*, 2005]). Half of the users received a correct prediction from the system about which folder the email should be sorted, and half received no prediction.
5. **Requesting Supplemental Information**: Whether the KE asks the human for additional input to generalize their response to other similar problems (*e.g.,* [Stump *et al.*, 2005]). For example, half the users were asked why they thought an email they provided a label for should be sorted in a particular folder.

Putting it Together

When the KE combines all dimensions above, it might ask a user the following question about an email being read:

> **Activity Recognizer**: "*Cannot determine how to sort this email. It is about an emergency and the author is not available Wednesday. What folder does this belong in? Prediction: Schedule Changes*."
> **Human**: *Answers*
> **Activity Recognizer Follow Up**: "*How can this folder be detected in the future*?"
> **Human**: *Answers e.g., "It contains the words "talk" "change" or "reschedule" and a day of the week.*

Each sentence in the interaction above is based on one of the dimensions above. Based on the KE's capability to provide information on the dimensions (*i.e.*, conditions of the study), the corresponding sentence can be removed or changed. For example, if the KE cannot provide high level information, cannot ask for additional information, cannot indicate uncertainty, and can only provide sufficient context, it might instead ask:

> **Activity Recognizer**: "*The author is John Doe, the subject is "Change Talk Time", and the email contains keywords "emergency" and "any day but Wednesday". What folder does this belong in? Prediction: Schedule Changes*."
> **Human**: *Answers, with no follow up*

Each participant experienced one of the 36 possible types of questions from the 2x3x2x2x2 design space for each domain, removing the conditions for high/low level context when participant receives the "no context" condition for the providing context dimension.

3.2 Measures

Because an active learner would benefit more from correct answers to questions rather than incorrect ones, we assessed

the user responses to the questions primarily based on correctness, but also on the quality of supplemental information when available. We also gave subjects surveys about their opinions of the applications asking questions including whether they found them annoying.

Correctness: Users responses were classified as *correct* answers if their last answer (some users changed their minds) was correct and incorrect otherwise [Hoffmann, *et. al.*, 2009]. For example, if a subject disagreed with the suggestion, but gave an equally correct reference, it was classified as correct.

Supplemental Information: If a user received a request for additional information, their response was coded based on how much additional information was provided. A value of 0 was given to a response that provided no additional information (*e.g., "I don't know"*). Every piece of valid information increased the value by 1.

Qualitative: After completing the task, participants were given questionnaires on their experiences with each technology. They were asked whether they thought the application's questions were annoying and whether they found each of the five dimensions particularly useful. Answers were coded as either "Yes" or "No" to each of the six questions. Participants were also asked the difficulty of answering the questions on a Likert scale from 1 (very easy) to 5 (very hard).

4 Results

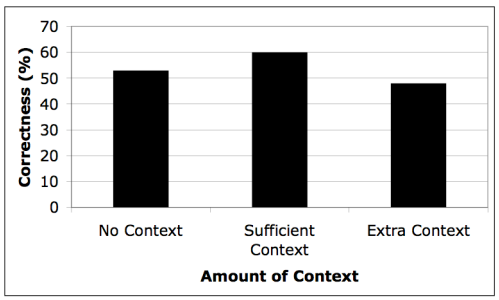
All five dimensions had an impact on the accuracy of user responses to questions. The McNemar test with the Chi-Square statistic were used to analyze the significance of the categorical response (correctness) against the categorical independent variables (our five dimensions). T-tests and One-way ANOVAs were used to analyze the significance of the secondary continuous response (quality of supplemental information) against the independent variables (our five dimensions). Based on the set of results for each domain, we define a set of guidelines that a KE should use when planning the wording of scenarios and questions to present to users. We validate those guidelines against advice we received from HCI experts on how to present and request information from users.

We analyzed the effects of each individual dimension first on the proportion of correct answers the KE received. Subjects answered a statistically significant larger proportion of questions correctly when given **low level context** (63%) versus high level context (54%)%) ($\chi^2[2,2] = 10.57$, $p<.01$). No other single dimension was significant.

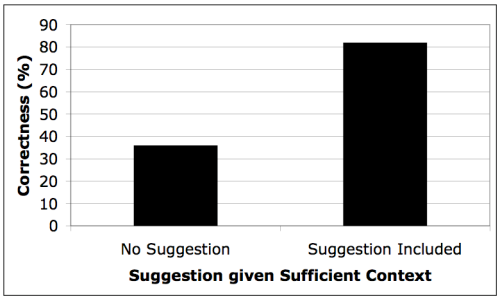
In order to understand how the remaining dimensions affected user performance, we then analyzed the effects of combining them with the significant dimensions and each other. Figure 3a shows the percentage of questions correctly answered for the different choices of context. We can see that there is an effect of context on accuracy. Subjects had significantly fewer correct answers when they received no context (53%) or extra context (48%) compared to subjects who received **sufficient** context (60%) and this effect is heightened when combined with the level of context ($\chi^2[4,2] = 11.04$, $p<.01$). Subjects also provided a statistically significant greater proportion of correct answers when they received a **prediction** with sufficient context (78%) compared to when they did not (50%) or when they received other amounts of context (55%) ($\chi^2[2,2] = 7.72$ $p<.01$). However, we found a significant paired effect of amount of context with predictions. When subjects received no suggestions, they provide more correct answers with extra context (71%) compared with sufficient context (35%), but when they receive suggestions, they provide more correct answers with sufficient context (90%) compared to extra context (50%) ($\chi^2[4,2] = 7.82$ $p<.05$).

We found that if we provide sufficient context, indicate **uncertainty** increases the proportion of correct answers significantly from 46% to 70% ($\chi^2[4,4] = 11.56$ $p<.01$). We found a significant paired effect of prediction with uncertainty ($\chi^2[2,2] = 8.70$ $p<.01$). Finally, we found that requesting **additional information** resulted in an increase from 30% to 90% in correct answers when paired with uncertainty but a decrease from 87% to 45% when no uncertainty information is provided ($\chi^2[2,2] = 12.21$ $p<.02$).

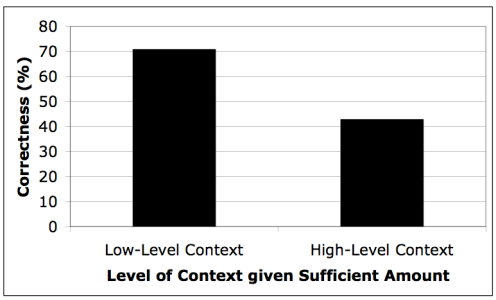
We analyzed the survey responses to understand how useful subjects felt the questions were and more specifically how useful they felt each dimension was. We found that 50% of subjects thought the email questions were useful to them when performing their task while 41% found answering the questions annoying. When we look at the perception of usefulness for each dimension, a majority of subjects who saw each dimension thought they were useful. 90% of subjects found context useful when they received at least sufficient context, and 100% of subjects who were given suggestions by the KE found them useful. Additionally, 78% of subjects who were asked to give supplemental information



(a) Subjects' Correctness by Amount of Context



(b) Subjects' Correctness by Suggestion



(c) Subjects' Correctness by Level of Context

Figure 3: Using results from our study, we developed guidelines along our five dimensions for how a KE should ask questions.

found it useful and 71% of participants who received uncertainty information found it useful.

Based on these results, we conclude that the KE should use the following guideline when asking questions: ***indicate uncertainty, provide sufficient low-level context, suggest an answer, and request supplemental information***.

5 Validation

We asked for advice from 3 members of the HCI community who conduct research in email sorting systems about how the KE should formulate questions along our dimensions. The community members understood both the technical data that could be collected from the domains and the usability requirements necessary for effective communication to users. We validate that our guidelines are at least as good as, if not better than, the community advice based on the proportion of correct answers and user opinions.

We explained each dimension to a group of HCI researchers and gave examples of how we combined the dimensions together. The researchers discussed the dimensions together and then reported their consensus about what combination they thought would elicit the most correct answers from users for the email task. The consensus reached was different then the results of our study in two ways (with differences shown in bold):

> *indicate uncertainty, provide low-level* ***extra context****, suggest an answer, and* ***do not request supplemental information***

5.1 Validation Method

We conducted a within-subject study in order to validate that our guidelines improve the proportion of correct answers that people give the KE compared to the community input. Subjects were told that they would be testing two different ways the technology learns from asking them questions. Similar to the first study, participants were told that they would complete both "learning" tasks to teach both applications first and then later they would complete performance tasks to test how well each learned from them. The order of the two conditions was randomized for each study, and the technologies were Wizard-of-Oz'ed for consistency in asking questions. Each "learning" task was 12 minutes long and the subjects were given surveys after each task. Then they were told they did not have time to complete the performance tasks. We scored participants' answers and collected qualitative measures through surveys.

5.2 Validation Results

T-tests used to analyze the significance of the categorical response (correctness) against the two types of questions (our guidelines and the community advice). We found a significant effect of the type of question on the proportion of correct responses ($t[2,250] = 2.48$, $p<.01$). Subjects who received our guidelines were correct 100% of the time, while those two received the community advice were correct 94%. A majority (8/11) people preferred the community advice but (7/11) people thought that our guidelines were learning more. When we analyze the dimensions where our guidelines and the community advice differed, more people preferred our context (58% *vs*. 40%) and suggestions (63% *vs.* 40%) to the community advice.

6 Discussion

We present two main contributions in this work, which we now discuss in detail. First, we have presented an approach to testing a knowledge elicitor's questions. This approach is thorough in its evaluation of our dimensions and can be used to investigate other dimensions and other domains. We have shown that by using our approach of testing a KE's questions, we can identify guidelines that provide users with clear enough content that they provide significantly more correct or higher quality responses than other questions (along our chosen dimensions).

Second, we have validated these guidelines against questions generated from community advice and show that for our email sorting application, our guidelines perform better and are preferred by users.

6.1 Approach

We present an approach to understanding how users' responses are affected by a knowledge elicitor's questions based on five dimensions. It requires we enumerate all possible combinations of those 5 dimensions to test the impact of each one and additionally all the dependencies between them. If we had only tested combinations of dimensions that were closely linked to the community advice, we would not have found our guidelines. While this technique ensures that we do not leave out any combination that may be the best, it can be expensive to use.

5.2 Determining the Best Guidelines

We expected to find significant effects of the dimensions in order to help us determine "best" guidelines. Some of our results show not only statistically significant but large effects of dimensions, especially pairs of dimensions. For example, when we paired the dimensions with sufficient context, we find that participants increased their proportion of correct answers by almost 30% between high- and low-level context and participants increased their proportion of correct answers from 36% to 82% when they were given suggestions. For results like these, it is easy to see what effect this difference could have on the accuracy of a machine learning system that used this labeled data. A system would have a much higher chance of finding the right boundaries to distinguish different labels if it had a lot more accurately labeled data. If these boundaries are not identified correctly, the machine learning system will produce inaccurate predictions or classifications.

Additionally, our final validation result found a statistically significant difference between the two conditions of 6%. Although this difference seems insignificant when correct classification rates are 94% and 100%, the 6% differ-

ence could drastically change a classification boundary for a learning algorithm and as a result drastically increase the error rate of the predictions. Users would much prefer a spam filter, for example, that was close to 100% accurate, rather than 94% accurate, where it was mis-classifying 6 out of every 100 emails incorrectly. Any amount of incorrectly labeled data, depending on the learning algorithm, could increase error rates so it is critical to optimize the accuracy of the users' responses as much as possible.

7 Summary

Researchers often instrument an interface or the environment with sensors to collect observational data but it can be difficult to label that data accurately. To automate the process of collecting the most *accurate* labels possible, we use a knowledge elicitor to ask users questions. This is a prime example of how HCI expertise can be used to improve the intelligence of systems. By improving the accuracy of labeled information, intelligent systems can be more accurate and provide better service to their users.

The contribution of our work presented here is two-fold. First, we present an approach for understanding how a knowledge elicitor can formulate questions to ask users about labeling data. We show that this approach successfully identifies questions that users respond well to along each dimension in order to provide correct labels.

Second, we describe a set of guidelines derived from the results of our study and validate our guidelines against HCI community advice in the relevant domains to prove that they are better, along the dimensions of response accuracy and usability. Based on additional tests we conducted in other domains (not reported here), we believe these guidelines are applicable far beyond the single email sorting task we investigated and could be used today without further validation when users have domain knowledge about the task or data they are working with.

This work focuses on a specific set of dimensions for classification problems. Additional work is needed to provide guidelines and validation for other types of questions a knowledge elicitor may ask and other dimensions that may affect how humans understand and answer questions. In addition, we would like to see how well our existing guidelines apply to other domains and tasks. Our work also does not focus on every possible domain or type of task. Additionally, future work is needed to test these guidelines in long-term data collections and active learning applications.

References

[Antifakos *et al.*, 2004] S. Antifakos, A. Schwaninger, and B. Schiele. Evaluating the Effects of Displaying Uncertainty in Context-aware Applications. *Proc. UbiComp 2004*, 54–69, 2004.

[Banbury *et al.*, 1998] S. Banbury, S. Selcon, M. Endsley, T. Gorton and K. Tatlock. Being certain about uncertainty: How the Representation of System Reliability Affects Pilot Decision Making. Proc. *HFES, Aerospace Systems*, 36–39(4), 1998.

[Gajos and Weld, 2005] K. Gajos and D. S. Weld, 2005. Preference elicitation for interface optimization. Proc. *UIST '05*, 173-182, 2005.

[Hoffmann *et. al.*, 2009] *R.* Hoffmann, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D. S. Weld. Amplifying community content creation with mixed initiative information extraction. Proc. *CHI '09*, 1849-1858, 2009.

[Jeffries *et al.*, 1991] R. Jeffries, J. R. Miller, C. Wharton, and K. Uyeda, "User interface evaluation in the real world: a comparison of four techniques," Proc. *CHI '91*, 119–124, 1991,

[Larson and Csikszentmihalyi, 1983] R. Larson and M. Csikszentmihalyi. The experience sampling method. New Directions for Methodology of Social and Behavioral Science, 15, 41-56, 1983.

[Presser, 2004] S. Presser, Methods for testing and evaluating survey questions, *Public Opinion Quarterly*, 68(1): 109–130, Mar 2004.

[Salton and Buckley, 1990] G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4): 288–297, 1990.

[Shadbolt and Burton, 1990] N. Shadbolt, A. Burton. Knowledge Elicitation. *Evaluation of Human Work:Practical Ergonomics Methods*, 321-345, 1990.

[Shiomi *et al.*, 2008] M. Shiomi, D. Sakamoto, K. Takayuki, C. T. Ishi, H. Ishiguro, and N. Hagita. A semi-autonomous communication robot: a field trial at a train station. Proc. *HRI '08*, 303–310, 2008.

[Steel, 2003] L. Steel. Evolving grounded communication for robots. *TRENDS in Cognitive Science* 7(7), 308-312, 2003.

[Stumpf *et al.*, 2005] S. Stumpf, X. Bao, A. Dragunov, and T. Dietterich. Predicting user tasks: I know what you're doing. AAAI 2005 Workshop on Human-Comprehensible Machine Learning, 2005.

[Stumpf *et al.*, 2007] S. Stumpf, V. Rajaram, L. Li, M. Burnett, and T. Dietterich. Toward Harnessing User Feedback for Machine Learning. *Proc. IUI 2007*, 82 – 91, 2007.

[Stumpf *et al.*, 2008] S. Stumpf, E. Sullivan, E. Fitzhenry, I. Oberst, W.-K. Wong, and M. Burnett, Integrating rich user feedback into intelligent user interfaces. *Proc. IUI 2008*, 50–59, 2008.

[Steinfeld *et al.*, 2006] A. Steinfeld, R. Bennett, K. Cunningham, M. Lahut, P. Quinones, D. Wexler, D. Siewiorek, P. Cohen, J. Fitzgerald and O. Hansson. The RADAR Test Methodology: Evaluating a Multi-Task Machine Learning System with Humans in the Loop. *Technical Report CMU-CS-06-125,* Carnegie Mellon University, 2006

[Topp *et al.*, 2006] E. Topp, H. Huttenrauch, H. Christensen, and K. Eklundh. Bringing together human and robotic environment representations–a pilot study. *Proc. IROS 2006*, 4946-4952, 2006.

[Torrance, 1994] M. C. Torrance, "Natural communication with robots," Ph.D. dissertation, MIT, 1994.