Towards Maximizing the Accuracy of Human-Labeled Sensor Data

Stephanie L. Rosenthal
Carnegie Mellon University
Computer Science Department
Pittsburgh PA USA
srosenth@cs.cmu.edu

Anind K. Dey
Carnegie Mellon University
Human-Computer Interaction Institute
Pittsburgh PA USA
anind@cs.cmu.edu

ABSTRACT
We present two studies that evaluate the accuracy of human responses to an intelligent agent's data classification questions. Prior work has shown that agents can elicit accurate human responses, but the applications vary widely in the data features and prediction information they provide to the labelers when asking for help. In an initial analysis of this work, we found the five most popular features, namely uncertainty, amount and level of context, prediction of an answer, and request for user feedback. We propose that there is a set of these data features and prediction information that maximizes the accuracy of labeler responses. In our first study, we compare accuracy of users of an activity recognizer labeling their own data across the dimensions. In the second study, participants were asked to classify a stranger's emails into folders and strangers' work activities by interruptibility. We compared the accuracy of the responses to the users' self-reports across the same five dimensions. We found very similar combinations of information (for users and strangers) that led to very accurate responses as well as more feedback that the agents could use to refine their predictions. We use these results for insight into the information that help labelers the most.

Author Keywords
Labeling Sensor Data, Active Learning

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors, Experimentation

INTRODUCTION
When collecting data from users for machine learning-based applications, learning agents must acquire labels to train accurate supervised models. As many labels are hard to sense accurately and implicitly, users themselves often carry the burden of labeling their own data using a diary at the end of the day [5] or with feedback throughout the day [15]. Both of these data collection methods can be prone to inaccuracies if users forget the context of their activity, do not understand which data the system is requesting a label for, or have limited time to label data.

IUI'10, February 7–10, 2010, Hong Kong, China.

While active learning provides support for prediction with a limited number of data labels, it still requires that users provide accurate labels and be available when a label is needed [18]. New techniques, like *proactive learning,* take active learning a step further and account for human inaccuracy by creating user models and determining from among a set of users, who to ask based on the need for accuracy [9][6]. Crowd-sourcing the data to websites like Amazon.com's Mechanical Turk or with games like GWAP [30] has also become a popular option for acquiring labels from people other than the actual users or data creators, but can require several people to label the same data in order to ensure accuracy [30][31].

While these techniques address the problem of labeling the data assuming that human inaccuracy is inevitable, we are interested in determining how the learning agent itself can affect the accuracy of the responses it receives. Concretely, the agent can vary the information it provides to labelers about the data as it asks questions, to maximize the accuracy of the labels it collects and the feedback it receives to refine it's predictions. After analyzing previous agents that request classification labels, we focus on a set of information that has been commonly provided to labelers: 1) *varying number of contextual features* of the data point 2) *high/low-level* explanation of those features 3) classification *prediction* 4) *uncertainty* in the prediction and 5) *user feedback* to weigh features used in classification.

In this work, we present a set of studies that explore the impact of agents providing subsets of the above information to labelers with the goal of maximizing the accuracy of the labelers' responses and encouraging feedback that would help a learning agent. We explore and identify the best subset for *people labeling strangers*' email and interruptibility data, in addition to *users labeling their own* physical activity data. Interestingly, these two subsets were nearly identical.

Our contributions are three-fold. First, we contribute a method for determining the combination of information that maximizes the accuracy of human responses by first testing all combinations and then validating the best combination against a combination suggested by experts. Second, we

contribute our best combinations for the types of tasks we tested and find that the combinations were nearly identical. Finally, we contribute an understanding of the value of different types of information and reasons for their impact so the information can be used consistently across applications for improving label accuracy.

RELATED WORK

While several different sets of guidelines have been proposed for agents' label-gathering interactions with humans (*e.g.*, [4], [10], [14] [25]), there are seven main types of proposed information: different amounts of context, human-understandable context, uncertainty, predictions, user control and feedback, action disclosure, and social interaction. Our work draws from areas of data collection and corrective feedback in determining which dimensions to focus on and how to implement them. Table 1 outlines these dimensions and previous work that includes different subsets of them. We find that the first five types of information, operationalized and described below, are the most commonly used in current applications. While there is continuing work in understanding socially appropriate times to request labels [15] and in relaying to a user how their action will affect the application [23], we focus on only the most commonly used types of information in this work.

Different Amounts of Context

Many of the guidelines suggest that applications should provide labelers with some *contextual information* about the features of the data to be labeled. However, we found that applications interpret this principle differently. When *BusyBody* asks users to estimate their own interruptibility, it does not explain what it thinks the user is doing [15]. Hoffman *et al*. request help from Wikipedia users to fill in missing summary data as the users are reading an article [13]. When users are asked if the text they are reading in the article *belongs* in the summary, important keywords are not provided in the text. When reading the summary, users *are* provided with excerpts that could be added to make the summary more complete. In studies of interruptibility, it has been shown that people make judgments with relatively small amounts of context (15 seconds) and extra context (30 seconds) does not improve accuracy [12]. We define *sufficient context* such that if fewer features are provided, labeling the data is difficult. *Extra context* includes more features than necessary. In the interruptibility work, 15 seconds of video is sufficient, while 30 seconds is extra.

Levels of Context

Recently, researchers demonstrated that labelers' accuracy can depend on the *level of contextual information* they are provided. When users understand and use their own rules for classification, they are better at making those classifications compared to classifying based on the computers' rules [28][29]. This finding is supported by work in feedback in information retrieval applications [21][22] which mask the *low-level* sensor-level features that computers use and collect (*i.e.,* individual keywords in documents or accelerometer data) and allow users to search for information using *high-level* meaning attributed to the low-level data (*i.e.,* summaries of documents or physical motion inferred from accelerometers). However, because it is often difficult to generate the high-level explanation of context, many applications provide only the low-level raw data, like pictures, to labelers instead of a summary with the assumption that they can find their own meaning [30][31].

Prediction and User Feedback

Other work has focused on making the classification task easier for labelers by providing a classification *prediction*. Here, the user only has to confirm an answer *vs.* generate one, simplifying their work (*e.g.,* [28]). An interface may automatically fill in fields in a form or provide a prediction for which folder to sort a piece of email into (*e.g.,* [8],[11]). Users could also provide *corrective feedback* for incorrect predictions to improve later classification [8]. In the active learning community, Raghavan asked people to label text documents as news articles, sports, etc., and also asked them to pick words (features of the classifier) that should have high weight for each class [20]. Participants knew the article words they were looking for and could identify them easily. The classifier could correctly weight the important features faster than asking for article labels alone, because people had narrowed down the important features. This same method can easily be used for email classification and the other domains. For example, in CueTIP, users see their handwriting and the word prediction and can make corrections [26]. The OOPS toolkit helps users "discover" if the learner's prediction is incorrect and then provides a set of interaction techniques for the user to correct it [16]. Scaffidi allows users to provide feedback by creating rules for the classifier to make better predictions of phone numbers and other personal information [24]. By asking for feedback and providing predictions, an agent can correct errors and use feedback to improve its learning.

Work	Uncertainty	Prediction	Amount of Context	High/Low Context	User Feedback	Action Disclosure	Social Interaction
Horvitz [14]	X	X	X		X		X
Bellotti and Edwards [4]		X	X		X		
Erickson and Kellogg [10]	X		X	X	X	X	
Hoffman et. al [13]	X	X	X	X			
Mankoff et. al [16]		X	X		X		
Cutlotta et. al [8]	X	X		X			

Table 1. Applications provide users with different types of information to help in labeling data. The most popular are Uncertainty, Prediction, Amount and Level of context, and User Feedback.

Uncertainty

Finally, many agents calculate a prediction probability in order to decide whether it should ask for help. Studies of 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.,* [3][17]), even if the learner does not provide the exact uncertainty value [1]. We compare the accuracy of responses from labelers who receive an indication of uncertainty *vs.* those that do not.

While each guideline has been shown to affect usability or accuracy of responses, they are also already commonly calculated in the machine learning process. The learner uses the different weights calculated from *user feedback* on *its contextual features* to make *predictions*, must know *its uncertainty* to determine whether to ask for assistance, and must be able to explain those features to human helpers. Because the learner already calculates these parameters, it should require only minimal additional computation to generate the questions using them compared to high benefit of receiving more accurate responses.

We are interested in determining which *combination* of this popular information maximizes the accuracy of responses to an agent's questions. Additionally, with the recent popularity of crowd-sourcing data labels to people who have not witnessed the generation of the data (*e.g.,* [12], [30], [31], [1]), we are interested in the best way to elicit responses from these strangers as well. Next, we present a set of studies to understand how the content of an agent's questions affect the accuracy of labelers' responses.

STUDY DESIGN

In order to investigate the impact of the varying the content of an agent's questions along the five dimensions presented above, we designed a set of studies that compared the accuracy of labelers' responses based on the information a learning agent provides. To understand information needs for both users labeling their own data and people labeling strangers' data, we developed real tasks for labeler populations – physical activity recognition for users and email sorting and interruptibility estimation for strangers.

Tasks and Materials

Subjects were told that they were testing new technologies that learn by asking questions. They were to complete a primary task, and the application would classify their actions. The application would interrupt their task to ask them for help if it could not confidently label the data itself. They were informed that they could answer the questions if they had time, and that the application would continue to learn whether or not they answered. Participants were told that they would be given a second similar "performance" task that the application could help them complete more quickly if they helped it learn first. They were also reminded that answering was not their primary task and doing so may slow the completion of that task. In this way, we model tradeoffs of time versus improved performance that labelers might consider in real applications.

Figure 1: The agent interrupted a subjects' task to ask which activity there were performing.

User Labels – Physical Activity Coach

The sensors on mobile applications are often hidden and their data is hard to explain, but they capture activities that their users are aware of, such as exercise patterns [7]. In this task, a physical activity coach performs an activity recognition task using sensors from a mobile device to identify exercises the user performs. An application like this one may record users' activities for doctors to analyze physical activity levels, and thus users have an interest in answering its questions to ensure it correctly identifies their activities. We test our physical activity coach's questions to show that users can accurately label their own physical data and can provide feedback to improve its predictions.

Subjects were told they were testing a new physical activity coach on a handheld device that could detect the different activities they performed (Figure 1). The subjects' primary task was to perform each of the 12 physical activities from a list provided (Table 2). Subjects were given all equipment required to complete the activities, including a soccer ball, tennis balls, rackets, step stools, and golf clubs.

They were required to carry a Nokia 770 Internet Tablet that would recognize their activities and beep when it had questions. They were to respond to questions on the tablet using a stylus on a virtual keyboard. We randomly pre-selected 8 out of the 12 activities to ask participants about. Questions were sent from the experimenter's computer, 10-20 seconds after each activity was initiated. Subjects had 12 minutes to complete as many activities as possible, while answering the agent's questions when they had time.

Activity	Description
Walk	Walk around the room once
Soccer	Dribble a soccer ball around the room once
Steps	Step up and down off a stool 10 times
Tennis	Bounce a tennis ball on a racket 10 times
Golf	Putt golf balls on a mini course 5 times
Hula Hoop	Use a hula hoop 10 times
Read	Sit and read 2 pages of a travel book
Toss Ball	Throw a ball in the air 10 times
Bounce Ball	Bounce a ball on the ground 10 times
Jump	Jump up and down 20 times
Jumping Jacks	Do 10 jumping jacks
Push Objects	Push 5 chairs from table to the wall

Table 2. The participants were told that the Physical Activities Coach could detect these tasks.

Strangers' Labels – Email Sorting

While most users do not think of their desktop computers as learning from their actions, word processors learn to spell-check new words and email applications learn which emails are spam and which are not. Because these labels are subjective, the user carries the burden of having to label their own data and may make mistakes. In this task, subjects take notes on a stranger's non-personal emails. The email sorter tries to classify emails in the inbox into a folder. If it is uncertain, it requests help from the subject.

The participants' primary email task is to read provided emails about an upcoming academic conference and consolidate all the changes that need to be made to the conference schedule and website [27]. They were given a spreadsheet with information about conference speakers, sessions, and talks, and asked to make changes to it based on change requests in the email, in 12 minutes. The emails and task were modified from the RADAR dataset [27]. The emails in the data set were labeled with a folder name, which was removed to test the participants. Additionally, we added high-level summaries of the emails and low-level keywords for the agent to use to ask for help.

They were given an email application with the emails and were told that the classifier had sorted most emails into folders based on the type of changes that needed to be made (schedule or website). The email interface was built with Adobe Flex and presented on a 15" Apple MacBook Pro. The participants should try to sort the "Unsorted" emails and answer the questions that popped up automatically when the participant read an email while updating the spreadsheet with the relevant information (primary task).

Stranger's Labels – Interruptibility

With crowd-sourcing technologies widely available today, we conducted an additional task on Amazon.com's Mechanical Turk, an actual system that is often used to pair label requestors with people willing to label data. The labelers on this website have never seen the applications that collect the data; they only fill out forms online for a small fee. These labelers are a perfect example of strangers who are willing to answer short questions like our agents'.

The problem of recognizing when someone is interruptible has been widely studied in the literature (*e.g.*, [12][15]). Specifically it has been shown that strangers are fairly accurate at rating someone else's interruptibility. We recruited subjects from Amazon.com's Mechanical Turk to estimate the interruptibility of office workers from video data previously collected. When the interruptibility video data was collected, office workers made the ratings without specifying who the interrupter was. Our dataset included 586 45-second videos from 5 offices at a university that had been labeled with an interruptibility value from 1 (Highly Interruptible) to 5 (Highly Non-Interruptible) by the five office workers themselves. Twelve videos were selected from the data set and put on the Mechanical Turk website, two randomly chosen from each interruptibility level plus two more from randomly chosen levels. Participants on Mechanical Turk were asked to rate the person in each of the 12 videos on the same 1-5 scale (Figure 2).

Figure 2: The agent asked participants to judge whether people were interruptable in their offices.

Varying Agent-Provided Information

To understand what information an agent should provide to maximize the accuracy of labelers' responses in each of these tasks, we vary the information the agent provides across the five dimensions presented above, namely providing *uncertainty, different amounts of context, high/low-level context, predictions,* and requesting *user feedback*. We examine all dimensions at once to find dependencies and correlations between them for a 2x3x2x2x2 design. Table 3 describes each dimension, the possible values, and an example on how each was used for each task. Along the dimensions, our content serves as exemplars for the definitions above so our results can be easily generalized to other similar applications and tasks.

Uncertainty

Along this dimension, we varied whether the agent told subjects it was uncertain about which classification to make. Half of the participants were told by the agent that it *"Cannot determine the activity."* while the other half were given no uncertainty information.

Amount of Context

Participants received one of three conditions (split evenly across participants): no context, sufficient context , and extra context. Participants in the sufficient context condition received enough features to identify the label accurately using only that information. On average this was about two pieces of information, and subjects read statements like *"Your feet are leaving the ground."* or *"This email is from X and is about their contact information*". Participants who received extra context received redundant information and saw statements like *"Your feet are leaving the ground together repeatedly"* or *"This email is from X and about incorrect spelling of their name on the website"*. In the Interruptibility validation, participants were given 15 seconds of video for sufficient context or 30 seconds for extra context; previous work found that people make interruptibility judgments in 15 seconds.

High/Low-Level Context

We also vary the context in terms of the feature level information that is provided. Subjects in the low-level context condition receive information about sensor readings

Dimension	Description	Activity Recognition Example	Email Sorting Example	Interruptibility Estimation Example
Uncertainty	Notify labeler that it is uncertain of the label	"Cannot determine your activity."	"Cannot confidently make a prediction."	"Cannot determine if the person is interruptible."
Amount of Context	Provide varying amounts of contextual information (none, sufficient, extra)	**Sufficient**: "Your feet are leaving the ground." **Extra:** "Your feet are leaving the ground together and repeatedly."	**Sufficient**: "The email has keywords A and B." **Extra:** "The email has keywords A, B, C, and D."	**Sufficient**: (15 seconds of video) **Extra**: (30 seconds of video)
High/Low-Level Context	Give either low (sensor) level context or high (activity) level context	**Low**: "Shaking motion detected." **High**: "Your feet are leaving the ground."	**Low:** "The email has keywords A and B." **High**: "The email is best summarized by F and G."	**Low**: (raw video) **High**: "The door open and two people in the office."
Question	Ask for a label	"What activity are you doing?"	"Where does this email belong?"	"How interruptible is this person?"
Prediction	Share the expected label for the data	"Prediction: Jumping."	Prediction: "Sessions Changes."	"Prediction: 4"
User Feedback	Ask labeler to describe the important features	"How can this action be detected in the future?"	"Why is this folder correct?"	"How did you make that determination?"

Table 3. Scenario Content Dimensions, Description, and Example Sentences for each task

on the activity recognizer, keywords in an email, and raw interruptibility video footage, to help them make their classifications. For instance if someone was jumping, the sensor might read "shaking" - we do not expect users to interpret exact numerical sensor readings or graphs. With high-level context, participants received explanations such as email summaries or body motions like "your feet are leaving the floor" that correspond to sensor readings. Note that if subjects are in the no context condition from the previous dimension, this dimension is not used.

Prediction

Along this dimension, we varied whether users received a prediction from the agent. Half of the users received a correct prediction from the agent (*e.g., "Prediction: Jumping."*) and half did not receive any prediction. Because the agent always gave a correct prediction in our work, we can measure how often a human trusts the agent's prediction but cannot measure the impact of incorrect predictions.

User Feedback

After subjects gave a response to the agent's question, half of them received a follow-up question to describe their actions in the activity or reasons for classifying the email or interruption level so it could be more easily identified in the future: "*How can this activity be detected in the future?*". We use the quality of this supplemental information as a secondary quantitative measure to compare our conditions when users provide equivalent numbers of correct answers.

Putting it Together

In order to generate questions, we read down this table, top to bottom – provide 1) uncertainty 2) prescribed amount of 3) high/low level context 4) ask question 5) prediction and 6) request for user feedback. For example, when the activity recognizer combines all dimensions above, it might ask a user the following:

> **Activity Recognizer**: "*Cannot determine your activity. Your feet are leaving the ground together repeatedly. What activity are you doing? Prediction: Jumping*."
> **Human**: *Answers*
> **Activity Recognizer Follow Up**: "*How can this action be detected in the future*?"

Each sentence in this interaction is based on one of the dimensions above. Based on the agent'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 agent cannot provide high-level information, cannot ask for user feedback and can only provide sufficient context, it might instead ask:

> **Activity Recognizer**: "*Cannot determine your activity. Shaking motion is detected. What activity are you performing? Prediction: Jumping*."
> **Human**: *Answers, with no follow up*

Each participant experienced one of the 36 possible types of questions from the 2x3x2x2x2 design space for each task, removing 12 conditions for high/low-level context and the prediction (which is another form of context) when a participant receives the "no context" condition.

METHOD

The study was conducted in phases. For each population (users and strangers), we conducted an *initial* test (activity recognition and email sorting), varying the information that labelers received in all combinations. At the same time, HCI researchers who worked on applications similar to our tasks were brought together to come to a consensus on which information they thought would be best for each of our tasks – which we call the *community advice*. Then, to *validate* our results, a new set of participants completed the same task twice – once with our best combination from the initial test and once with the community advice. We validate against community advice instead of against questions without context for two reasons. We believe, and

will show in our results, that asking for help with no context is confusing and would lead to poor accuracy. Also, it does not reflect what HCI researchers have done in the past when building systems that ask for help. Our community advice reflects a more realistic metric for response accuracy improvement to compare our results against.

After completing this for both users and strangers, we conducted the third interruptibility test to further validate the combination for a larger set of strangers. We chose interruptibility because it is well known in the community and we can draw the following three parallels to the email task. Fogarty *et al.* tested their classifier on strangers, so there was a well-established baseline of accuracy to test against [12]. The raw low-level data (words/video) are human-understandable compared to the accelerometer data in the activity recognition domain. In email, participants could draw context from the emails just as interruptibility participants could use raw video. This is both a feature and a flaw in supplying raw data as there is no context lost in the explanation. Finally, participants in both tasks were tested on their ability to classify against subjective labels. For these reasons, we used the interruptibility task to validate the best combination from the email study.

Procedure

To ensure that all participants in all conditions of all tasks received the same information for the same dimensions, the questions were generated before the study began. We measure the proportion of correct answers labelers provide as well as the quality of their feedback when requested.

Initial Tests

For the *initial* tests, participants were assigned the order of tasks at random, but evenly. Participants were given an explanation of the study and signed a consent form on arrival. Subjects were given 12 minutes to complete their primary task and were told they would receive a "performance" task based on both the completion of the first one and on their responses to the agent.. After completing the task, participants were given a survey about their experiences with the questions. Then, they were given the second primary task with the same instructions and given a second survey. Upon completion, participants were told there was not time to complete the "performance" tasks and were dismissed after being paid $10.

Validation

Before the validation experiments were run, we sought advice from 3 HCI community members who work on projects similar to our email and activity recognition tasks about which information they believe each agent should use when asking for help. 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 explained each type of information and how they could be combined together. Each group (email and activity) met separately to discuss the information and then reported their internal consensus about which combination they thought would elicit the most correct answers. In the end, we had two combinations of information – one for the email sorting task with strangers and one for the activity recognition task with users.

We then analyzed the results from our initial tests to identify our best combinations of the information – one for the email sorting task and one for activity recognition. In both validation tests, participants were randomly but evenly assigned the order they would receive the two types of questions – our best combination *vs.* the community advice. Subjects performed the same procedure as the initial tests, except that they received the same task twice with different combinations of information. When they completed the second survey, they were paid $10 and dismissed.

For reasons given above, we apply the same combinations (ours and community advice) from the email task to the final interruptibility validation. Participants were randomly assigned to receive only one of either our best combination or the community advice and received 12 videos to label. All participants were paid $5 for completing the study.

We validate that our combinations are at least as good as, if not better than, the community advice for each task based on the proportion of correct answers and user opinions.

Participants

Participants, except for the interruptibility task, consisted of 60 Pittsburgh residents ages 18-61 with a variety of occupations including students, bartenders, teachers, and salesmen. 37 subjects completed both initial tests, acting as users in the activity recognition task and as strangers in the email-sorting task. Then, 11 additional subjects completed a validation of the activity recognizer task and 12 more subjects completed the email sorting validation. Only a few participants (15%) had experience with technology that learns, and all spoke fluent English. 180 participants in the Interruptibility task were recruited anonymously on Mechanical Turk, but were only allowed to complete the task once by comparing usernames.

We use this experimental method to find the combination of information that maximizes the accuracy of responses and the amount of feedback from users to label their own data and from others who label strangers' data. We will next describe the measures used to find the best combinations.

Measures

Because a machine-learning agent would benefit more from correct answers rather than incorrect ones, we assessed the user responses primarily based on correctness, but also on the quality of user feedback when available. We also gave subjects surveys about their opinions of the applications, including whether they found them to be annoying.

Users' responses were classified as *correct* answers if their last answer (some users changed their minds) was correct and *incorrect* otherwise. For example, if a subject disagreed with the prediction, but gave an equally correct reference, it was classified as correct. Synonyms were determined to be correct as long as they were not too vague. For example,

"putting" was considered synonymous with the activity "golfing", but "swinging arms" was not because it was not an accepted name for the activity (listed on instructions of activities to perform). While we do not expect participants to give exact answers, we also do not expect them to give completely incorrect or opposite values.

If users can provide accurate labels for their data, their ability to give quality, or helpful, *feedback* is of particular interest to possibly speed learning [20]. If users received a request for feedback, their response was coded based on how many features about the data were provided. A value of 0 was given to a response that provided no additional information (*e.g., "I don't know"*). For every piece of valid information, the value increased by 1. For example, "I'm doing jumping jacks if *my arms move up and down* and *my legs go in and out*", would be given a value of 2.

After completing the task, participants were given questionnaires on their *subjective* experiences with each technology. They were asked about whether they thought the application's questions were annoying and whether they found each dimension particularly useful. Responses were coded as either "Yes" or "No". Participants were also asked whether it was easy or hard to answer the questions on a Likert scale from 1 (very easy) to 5 (very hard).

RESULTS

We analyze the results of the activity recognizer initial and validation task to determine a best combination of dimensions for users labeling their own data. Then we analyze the results of the email sorting initial and validation tasks to determine the best combination of dimensions for people labeling strangers' data. We use the results of the interruptibility validation to extend the results beyond a single task and for a broader set of strangers.

Analysis

Initial Test Models - A McNemar test with the Chi-Square statistic was used to analyze the significance of the categorical response (correctness) against the categorical independent variables (our 5 dimensions) for each task. T-tests and ANOVAs were used to analyze the significance of the secondary continuous response (quality of feedback) against the independent variables. Based on the results, we define a combination of information that agents should use to ask questions of labelers to maximize label accuracy.

Validation Models - We conducted a within-subject study to validate that our guidelines result in more correct answers compared to the community input. We used T-tests to analyze the significance of the categorical response (correctness) against the two types of questions (our guidelines and the community advice).

User Labels – Physical Activity Coach

Initial Test Results

We collected 119 responses from participants, including 8 for which participants (6 of them) said they were too busy to respond. When we analyzed the remaining 111 responses for the effects of the individual dimensions on the proportion of correct labels users provided for their own data, we found that subjects were correct *nearly* 100% of the time and there was *no effect* of any of the dimensions or their combinations. However, we found that when an agent requests user feedback, subjects were able to provide on average of .81 pieces of quality feedback compared to almost 0 pieces without being asked (some subjects provided feedback without prompting). We include user feedback in our best combination, as this is a statistically significant difference ($F[6,112] = 8.87$, $p< 0.001$).

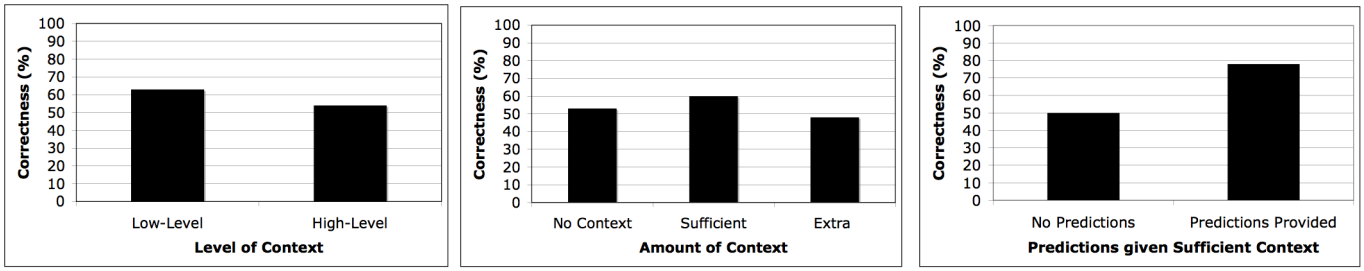
We then used the McNemar test on the amount of feedback with all five dimensions as independent variables to analyze the significance. We find that subjects who received sufficient context provide a significantly larger amount of quality feedback (.77 pieces) compared to those provided either no context (.30 pieces) or extra context (.31 pieces) ($F[2,2]=5.38$, $p<.002$). Additionally, subjects who received low-level context provided statistically significantly greater amount of feedback (.58 pieces) compared to high-level context (.34 pieces) ($F[1,1] = 3.33$ $p<0.05$). There were no significant effects and no combined effects for providing predictions or uncertainty so we use qualitative results to understand the impact of those dimensions.

We find that 25% of subjects who did not receive predictions reported it hard or very hard to answer the questions. Additionally, 0% of subjects with predictions reported the task difficult and 83% thought the questions were useful. There were no effects of uncertainty on the qualitative data so we do not include it in our best combination. Based on these results, we determine that the best combination for a user labeling their own data is the following: ***no uncertainty, do provide sufficient low-level context, predictions, and request user feedback.***

Validation Results

We validate our best combination against the HCI community advice, which varies from our combination in two ways (with differences shown in bold): *do not explain uncertainty, but* ***provide high-level*** *and* ***extra context****, predictions, and request user feedback*.

We collected 113 responses from participants including 11 non-responses. Four participants were too busy to respond at least once. We found that for both conditions, subjects gave correct responses 100% of the time and there were no statistically significant effects on feedback quality, so we use the qualitative results to differentiate the conditions. Subjects found that our dimensions were useful but only 30% realized they were receiving contextual information. Subjects did not prefer either system and could not identify which one learned more, but 70% of participants thought the system using our guidelines was smarter. Users that believe a computer is smarter will respond with more sophistication than to one they think is not as smart [19]. So, we conclude that our combination is at least as good as, if not better than, other combinations of the information.



(a) Subjects' Correctness by Level of Context (b) Subjects' Correctness by Amount of Context (c) Subjects' Correctness by Prediction

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

Strangers' Labels – Email Sorting

Initial Results

We collected 153 responses from participants including 13 non-responses. Four participants answered that they were busy at least once. We first analyzed the effects of each individual dimension on the proportion of correct answers. Subjects answered a statistically significant larger proportion of questions correctly when given low-level context (63%) versus high-level context (54%) ($\chi^2[2,2] = 10.57$, $p<.01$) (Figure 3a). 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 ($\chi^2[4,2] = 11.04$, $p<.01$) (Figure 3b). No other single dimension was significant.

To understand how the other three dimensions affected user performance, we analyzed the effects of pairing them with the significant dimension and each other. Subjects provided statistically significantly more 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%) ($\chi^2[2,2] = 7.72$ $p<.01$) (Figure 3c). We found that if we provide sufficient context, providing uncertainty increases the proportion of correct answers significantly from 46% to 70% ($\chi^2[4,4] = 11.56$ $p<.01$). There is a significant paired effect of prediction with uncertainty ($\chi^2[2,2] = 8.70$ $p<.01$). Finally, we found that requesting user feedback 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 each dimension was. We found that 50% of subjects thought the questions were useful to them during their task while 41% found answering them annoying. 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 received predictions found them useful. 78% and 71% of subjects who were asked for feedback and who received uncertainty respectively, found it useful. We conclude that the agent should use the following combination when asking strangers questions: ***provide uncertainty, sufficient low-level context, predictions, and request user feedback***.

Validation Results

HCI researchers that work in the email domain came to the following consensus on our dimensions (with differences shown in bold): *provide uncertainty, low-level* ***extra context****, predictions, and* ***do not request user feedback***.

We collected 301 responses including 4 non-responses. Three participants refused to respond at least once. We found a significant effect of the combination on the proportion of correct responses ($t[2,250] = 2.48$, $p<.01$). Subjects who received our combination were 100% correct, while those who received the community advice were 94% correct. A majority (8/11) people preferred the community advice but (7/11) people thought our agent was learning more. When we analyze the dimensions that differed between combinations, more people preferred our context (58% *vs*. 40%) and predictions (63% *vs*. 40%).

Interruptibility Results

Participants in this study were required to answer all 12 questions. Half of our 180 participants estimated the interruptibility for 12 videos with the best combination from the email task and half received the email community's advice. We analyzed the average mean squared error (MSE) of each participant's estimation compared to the true interruptibility across the videos and performed a between-subjects ANOVA analysis to compare the error between the combinations. We removed 16 of the 180 subjects that had MSE results that were more than 3 times the median of the entire data set (average MSE=1.37, outliers > 4.11). Subjects who received our combination had a statistically significant lower average MSE (mean 1.17, std. dev. 0.62) than those who received the community advice (mean 1.42, std. dev. 0.92) ($F[1,164] = 6.02$, $p < 0.01$). Subjects who received our combination were correct or off by one level of interruptibility 85% of the time, while subjects that received the community advice were correct 80% of the time. Both of these are better than the previously published interruptibility result, reporting a 65% off-by-one accuracy with only sufficient context.

DISCUSSION

Our results show that we were able to find a combination of information for agents to provide strangers that maximizes labeling accuracy. Additionally, our users were aware of their activity and had no trouble labeling it, so we can find a best combination that maximizes accuracy *and* amount of

feedback users give. Next, we discuss differences between users and strangers in each dimension to explain our results.

Benefits of Providing Information to Labelers

We found that each of the five dimensions – amounts of context, level of context, uncertainty, prediction, and requests for user feedback – had a positive effect on the labelers as they were performing their primary task. First, labelers used the context and prediction to match the agent's focus. For example, many labelers used the key words and summaries of the emails when deciding on a label instead of reading the entire email. As a result, the questions did not take as long for labelers to answer compared when they had to pick out the important context themselves. Additionally, labelers checked to confirm their label was consistent with the given context.

Although labelers were frequently interrupted with questions in their 12-minute tasks, they almost always answered when it was prefaced with uncertainty. For example, in the activity recognizer task, one participant who was interrupted only seconds after starting a task said, "It's interrupting me again! Oh, well, I guess it must be hard to distinguish between these [activities]." This shows that users excused the interruption when they felt they could help the agent. However, when we asked participants whether they valued uncertainty, they did not remember if they had received that information and therefore reported it as being not useful. We believe labelers underestimated the usefulness of uncertainty for the usability of the questions.

Finally, when labelers were asked to provide feedback about the label, they sometimes changed their labels to the correct answer when they thought about a reason for the label. While it may be difficult for a system to incorporate such freeform feedback, we find that the agent will benefit from increased response accuracy *just by asking* the question and irrespective of using the response. To make it easier to use such feedback, the agent could ask a multiple-choice question. Overall, we found each piece of information was useful for both user and stranger labelers.

Limitations of These Five Dimensions

While the five dimensions we chose were able to help the labelers focus their answers, they did not provide any new information for users to use in the physical activity task. For example, when users were given context, they already knew what they were doing. As a result, we found that varying the dimensions had little impact on the user's accuracy. We found that users pulled out the tablet and started typing often without even reading the question. We do not, however, believe this is universally true for all user tasks – users may misfile their emails in folders.

Additionally, users had a lot of trouble giving feedback about their physical activities. While they knew that they were golfing and not playing soccer, they were not able to provide much information about what actions constituted the activity like swinging arms or kicking their leg. Often, users thought for a long time about what to write the first time they saw the request for feedback, because they do not usually think about what constitutes a physical activity. Participants had less trouble expressing their feedback in the email and interruptibility tasks, because they had to develop their classification rules while performing the task. While the feedback is useful to a machine learning application, it may be too difficult for labelers to provide if they are not consciously making the classification.

Similar Combinations of Information

We observe that the best combinations for both users and strangers were nearly the same – only differing by uncertainty. We had assumed that because strangers did not know the context the data was drawn from, the agent would need to provide extra context to maximize accuracy compared to users. However, because strangers had some existing domain knowledge about sorting email and determining if someone is interruptible before the study, they did not need as much context to be accurate. We also found that just as users did not require high-level context about their own activities, the strangers did not require high-level context in the email or interruptibility tasks because the raw data (email keywords and video clips) were already human-understandable. This is significant because it reduces computation time for constructing questions, and eliminates the need to translate low-level sensor data into high-level context, allowing more time for processing data.

The only difference between the two combinations is *uncertainty*. Uncertainty offers no help to the labeler but indicates that the classification is hard. Users were aware of the difficulty of activity recognition without the acknowledgement from the system, reporting that they were impressed that a mobile device was able to recognize their activities. Receiving uncertainty did not change the users' opinion and there were no significant changes in accuracy as a result. However, strangers saw human-understandable data and assumed the classification was, in fact, easy. When strangers received uncertainty, we believe they recognized the difficulty of the task and tried harder, resulting in higher accuracy responses. In general, labelers that realize the classification is hard do not require *uncertainty* information.

Accuracy of the Agent

In this work, we wizard-of-oz'd the agents' questions to ensure they were timed correctly and included accurate information. The context that the agents provided *did* accurately represent the data and the high-level context appropriately summarized the sensors. As a result, the labelers could trust and use this data to their advantage when responding. In actual implementations, agents may not always be able to extract this information accurately. It is unclear how labelers would react to incorrect context.

Additionally, the questions were asked in the middle of activities while users were performing them. Because users knew which activity they were currently performing, the agent's information did not affect the user accuracy. If the questions were mistimed or delayed, it is unclear how this would affect the accuracy of users' responses.

While all of the predictions that were provided were the correct, the labelers often did not trust the predictions. This could be because labelers were told the agent asked when it was not confident in its prediction. While the predictions were shown to increase accuracy, we do not believe this is due to their correctness. Rather, they helped labelers narrow down the labels from which they decided on their own.

CONCLUSION

Researchers often instrument an interface or environment with sensors to collect data for learning but it can be difficult to label that data accurately. To automate the process of collecting the most *accurate* labels possible, we use an agent to ask questions. The contribution of this work is three-fold. First, we contribute a two-step method to test combinations of information – an initial step and a validation. While users who label their own data were typically very accurate at labeling their physical activities, we contribute a combination of information that maximizes accuracy *and* the quality of feedback the user provides. Additionally, we found a combination of information that maximizes accuracy of people labeling strangers' data. These more accurate labels and feedback can improve learning. Finally, we observed that the 2 combinations were nearly the same. We believe these validated combinations are applicable far beyond our 3 tasks and could be used today to collect more accurate labels when labelers have domain knowledge about the data they are working with.

This work focuses on a specific set of dimensions for classification problems. Other dimensions may also impact how labelers answer questions and need to be validated using our approach. We would also like to see how well our results apply to other domains and tasks where users and strangers have less domain knowledge about the collected data. Future work is needed to test these questions in long-term data collections and active learning applications and to understand the usability of proactively asking for help.

REFERENCES

1. Amazon.com Mechanical Turk – Artificial, Artificial Intelligence, https://www.mturk.com/mturk/, 2009
2. S. Antifakos, A. Schwaninger, and B. Schiele. Evaluating the Effects of Displaying Uncertainty in Context-aware Applications. *Proc. UbiComp 2004*, 54–69, 2004.
3. S. Banbury, *et al*. Being certain about uncertainty: How the Representation of System Reliability Affects Pilot Decision Making. Proc. *HFES, Aerospace Systems*, 36–39(4), 1998.
4. V. Bellotti and K. Edwards. Intelligibility and accountability: human considerations in context-aware systems. *Human.-Computer Interaction* 16(2), 193-212, 2001.
5. S. Carter, J. Mankoff. When participants do the capturing: the role of media in diary studies. *Proc. CHI 2005*, 899-908, 2005.
6. D. Cohn, L. Atlas and R. Ladner. Improving Generalization with Active Learning. *Machine Learning*. 0885-6125:15(2), 201-221, 1994.
7. S. Consolvo, *et al*. Activity Sensing in the Wild: a Field Trial of Ubifit Garden. *Proc. CHI* 2008, 1797-1806, 2008.
8. A. Culotta, T. Kristjansson, A. McCallum, P. Viola, Corrective Feedback and Persistent Learning for Information Extraction. *Artificial Intelligence*, 170(14-15), 1101-1122, 2006.
9. P. Donmez and J. G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. *Proc of CIKM '08*, 619-628, 2008.
10. T. Erickson and W. A. Kellogg. Social translucence: an approach to designing systems that support social processes. *ACM ToCHI* 7(1) 59-83, 2001.
11. A. Faulring, *et al*. Successful User Interfaces for RADAR. *CHI 2008 Workshop on Usable Artificial Intelligence*, 2008.
12. J. Fogarty, *et. al*. Predicting human interruptibility with sensors. *ACM ToCHI* 12(1), pp. 119-146, 2005.
13. R. Hoffmann, *et al*. Amplifying community content creation with mixed initiative information extraction. *Proc. CHI '09*, 1849-1858, 2009.
14. E. Horvitz. Principles of mixed-initiative user interfaces. *Proc. of CHI '99,* 159-166, 1999.
15. E. Horvitz, P. Koch, J. Apacible. BusyBody: creating and fielding personalized models of the cost of interruption. *Proc. CSCW 2004,* 507-510, 2004.
16. J. Mankoff, G. Abowd, S.E. Hudson, OOPS: a toolkit supporting mediation techniques for resolving ambiguity in recognition-based interfaces. *Computers & Graphics*, 24(6), 819-834, 2000.
17. S. Mcnee, *et al*. Confidence Displays and Training in Recommender Systems. *Proc. INTERACT*, 176–183, 2003.
18. T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
19. J. Pearson, J., *et. al.* Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. *Proc. CHI '06,* 1177-1180, 2006.
20. H. Raghavan, O. Madani, R. Jones. Active Learning with Feedback on Features and Instances. Journal of Machine Learning Research 7, 1655-1686, 2006.
21. Y. Rui, *et al*. Relevance Feedback: a Power Tool for Interactive Content-based Image Retrieval. *IEEE Trans. on Circuits/Systems for Video Technology*, 8(5): 644–655, 1998.
22. G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4): 288–297, 1990.
23. T. Salvador and K. Anderson. Practical Considerations of Context for Context Based Systems: An Example from an Ethnographic Case Study of a Man Diagnosed with Early Onset Alzheimer's Disease. *UbiComp 2003*, 243–255, 2003.
24. C. Scaffidi. Topes: Enabling End-User Programmers to Validate and Reformat Data. Technical Report CMU-ISR-09-105, 2009.
25. N. Shadbolt, A. Burton. Knowledge Elicitation. Evaluation of Human Work: Practical Ergonomics Methods, 321-345, 1990.
26. M. Shilman, D. S. Tan, P. Simard, CueTIP: a mixed-initiative interface for correcting handwriting errors. *Proc. UIST '06*. 323-332, 2006.
27. A. Steinfeld, *et al*. 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.
28. S. Stumpf, *et al*. Predicting User Tasks: I know What You're Doing. *AAAI 2005 Workshop*, 2005.
29. S. Stumpf, *et al*. Toward Harnessing User Feedback for Machine Learning. *Proc. IUI 2007*, 82 – 91, 2007.
30. L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. ACM Conference on Human Factors in Computing Systems. *Proc. CHI 2004*, 319-326, 2004.
31. L. von Ahn, *et al*. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 1465-1468, 2008.