Fantasktic: Improving Quality of Results for Novice Crowdsourcing Users

Philipp Gutheim
Björn Hartmann

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-112
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-112.html

May 11, 2012

Acknowledgement

Many thanks to my adviser Bjorn Hartmann and to Tapan Parikh who contributed to this research through various discussions, mentoring and general support.

Fantasktic: Improving Quality of Results for Novice Crowdsourcing Users

Final Report for Master in Computer Science Degree, UC Berkeley.

Philipp Gutheim

Spring 2012

Abstract

Crowdsourcing platforms such as Amazon's Mechanical Turk and MobileWorks offer great potential for users to solve computationally difficult problems with human agents. However, the quality of crowdsourcing responses is directly tied to the task description. Creating high-quality tasks today requires significant expertise, which prevents novice users from receiving reasonable results without iterating multiple times over their description. This paper asks the following research question: How can automated task design techniques help novice users create better tasks and receive higher quality responses from the crowd? We investigate this question by introducing "Fantasktic", a system to explore how to better support end users in creating successful crowdsourcing tasks. Fantasktic introduces three major task design techniques: 1) a guided task specification interface that provides guidelines and recommendations to end users throughout the process, 2) a preview interface that presents users their task from the perspective of an agent, and 3) an automated way to generate task tutorials for agents based on sample answers provided by end users. Our evaluation investigates the impact of each of these techniques on result quality by comparing their performance with one another and against expert task specifications taken from a business which crowdsouces these tasks on MobileWorks. We tested two common crowdsourcing tasks, digitizing business cards and contact email address search on websites, with ten users who had no prior crowdsourcing experience. We generated a total of 8800 tasks based the users instructions which we submitted to a crowdsourcing platform where they were completed by 440 unique agents. We find a significant improvement for instructions based on the guided task interface which show a reduced variation of answer formats and a more frequent agreement on answers among agents. We do not find evidence for significant improvements of instructions for the task preview and the agent tutorials. Although expert tasks still perform comparably better, we show that novice users can receive higher quality results when being supported by a guided task specification interface.

Table of Contents

Introduction

Crowdsourcing platforms enable users to automate computationally difficult problems such as data verification, de-duplication, categorization and audio transcription [1,2]. Services like Amazon's Mechanical Turk and MobileWorks solve these problems through so called micro-tasks – general knowledge tasks that take several seconds to minutes – which are completed by a pool of online agents who receive monetary rewards in return. One of the challenges of crowdsourcing systems is that human agents produce results of varying quality, with unpredictable speed which can be of great importance for systems that expect similar input within a given time frame [3]. This imposes substantial challenges on users to obtain desired work products from online crowds. These challenges can be categorized in the following three areas [4,5,6]:

1. *Incentivization problem*: Online agents can be untrustworthy and may act maliciously to maximize their own payout.
2. *Human Error Problem*: Despite their intention to carry out the instructions as specified, agents may still make errors.
3. *Task specification problem*: User-generated tasks can be ambiguous in certain cases, could lack guidelines and may even be contradictory in itself.

Researchers have developed various techniques to address these challenges such as multi-worker redundancy, *i.e. majority votes [7,8], peer review mechanisms [3], and probabilistic confidence metrics [9].* Other researchers suggested mechanisms such as injection of test tasks, so called gold standards [4], and ways to increase the potential complexity of projects by introducing workflows which decompose a project into smaller, chained tasks [10,6]. In addition, Little et. al. have investigated whether a workflow that lets agents collaboratively improve their answers increases the quality of results [11]. Recently, some of these techniques have been adopted by commercial crowdsourcing platforms allowing the user to use majority vote to find the most likely answer, providing different workflows or filtering agents with experience in a specific domain [6,12].

One limitation of these techniques is their dependency on existing, well-defined task specifications. Most often, these specifications have been developed around a specific use case and evolved gradually over multiple iterations of task submission, answer evaluation and instruction refinement [1]. As a result, the task specification is predominantly based on the researcher's expertise and leverages a set of narrowly defined instructions specific to the application. In contrast, novices who use existing general-purpose interfaces of commercial crowdsourcing platforms face significant challenges designing instructions that generate reasonable results.

Anecdotally, the authors have observed novice users iterating their instructions and eventually giving up because they were not able to improve the results over several iterations. As an example, one user submitted a set of tasks which instructed agents to do a web search for a book title and asked to evaluate and retrieve as many book reviews of a certain book as possible. The initial results from the task returned only a single review per agent even though a web search would have returned significantly more. After several iterations with the task instructions the agents still returned only a single review. As a result, the user had to consult a more experienced crowdsourcing user to generate the desired list of reviews.

> *“I need to find all book reviews for this book: [Book Title]*
>
> *I need the following information:*
>
> *The URLs for all reviews that are out there*
>
> *Your classification of whether the review is a 'Thumbs Up' or 'Thumbs Down'”*

figure 1: The initial set of instructions used by novice user to retrieve book reviews

> *“In this project, we want to collect opinions about a certain book from professional book critics and blogger. Please do the following steps:*
>
> 1. *Go to* www.google.com *and search for [Book Title]*
> 2. *Click on first result.*
> 3. *[additional instructions for the search result]*
> 4. *[if condition for website content is met]*
> 5. *[if another condition for website content is met]*
> 6. *Go back to search results*
> 7. *Click on the second result and repeat steps 3 through 6.*
> 8. *Repeat steps 3 through 7 until you have clicked and evaluated a total of 50 results.”*

figure 2: The instructions after several iterations

We have developed Fantasktic to investigate how novice users can be better supported in creating successful crowdsourcing tasks. We hypothesize that *guidelines and a preview during submission helps novice end users to generate better instructions while automated tutorials allow agents to infer the unexpressed, latent, principle underlying the end user's judgment when instructions lack specification*. Fantasktic consists of three components that support the task specification process:

1. Guidelines and recommendations for the end user during the task specification process to help defining what information the instructions should contain, what action an agent should carry out in particular edge cases as well as how results should be formatted.
2. A preview of the task from the perspective of an agent, enabling an alternate review of the effectiveness of the task specification and the potential lack of information.
3. An automated way for end users to provide example answers that will generate an interactive tutorial for agents who are working on the tasks the first time.

The main contribution of this paper is to investigate the effectiveness of potential techniques to improve support for non-expert end users in crowdsourcing. The paper contributes to the existing body of literature in crowdsourcing by focusing on way to support novice end users to write better tasks instructions.

We explored this research goal by designing, implementing and evaluating a task submission system that includes an interface with guidelines and recommendations to create better instruction sets, a task preview interface showing end users their tasks from the perspective of an agent and a way to automatically generate tutorials for online agents working on an end user's tasks. We juxtapose how the proposed techniques impact the results by analyzing the similarity of answers. Our analysis distinguishes between different information (e.g. two different telephone numbers) and similar information (e.g. same telephone number but differently formatted). We evaluate different information by comparing the cosine similarity of a vector of words in a specific answer instance with the vector of all unique words for that business card given a specific instruction set. In the case of similar information, we compute the string distance for each combination of answers and compare it with the distances when we removed spaces and special character and converted all characters to lower case. We compare the techniques with one another and in relation to results from expert instructions which we extracted from a business card digitization service that crowdsources the digitization process on a regular basis.

We base this analysis on a study of ten novice end users who used Fantasktic to create business card digitization tasks. Based on the different steps within the interface, we extract four variations of instructions which we use to generate batches of 20 business card tasks totaling 440 batches or 8800 tasks. We submitted these tasks into a crowdsourcing platform were we had 440 unique workers complete one single batch of cards. In addition, we had five participants create contact email search tasks on which bases we generate additional four instruction sets. To evaluate whether our observations from the business card tasks are applicable across different task types and not unique to business cards, we analyzed the instructions from the email search tasks if they show similar pattern between instructions using one of our techniques. Similarly, looking up email addresses on websites is a common task (web research) in which agents might face challenges that are comparable to those in business card tasks such as multiple emails, submission forms only or partially relevant email addresses.

We find a significant improvement for instructions based on the guided task interface for both reduced variation of answer formats and a more frequent agreement on answers among agents. Furthermore, there is statistical evidence that the task preview or the agent tutorials improve instructions significantly. Although expert tasks still perform comparably better, we show that novice users can receive higher quality results when being supported by a guided task specification interface.

The remainder of this paper is organized as follows: We begin with a review of related work and present how Fantasktic relates to quality assurance and interface design in crowdsourcing and in respect to learner-centered design. We then discuss the architecture and implementation of Fantasktic and its three components in particular. We follow with an in depth evaluation of our study and its results. We conclude with a summary of our findings and a discussion of their inherent limitations.

Related Work

Fantasktic is related to prior work in ensuring quality in crowdsourcing tasks, simplifying task specification in crowdsourcing systems and learner-centered design.

Ensuring Quality in Crowdsourcing Tasks

Various strategies have been used to improve quality of results from crowdsourcing platforms. The ESP Game [13] introduced the notion of using multi-worker redundancy, i.e. majority votes, to correct for potential errors in crowd work and showing that sending the same question to multiple workers could be effective in eliminating both human error and malicious workers. Le et al. [4] take a different approach by demonstrating how qualifying tests, so called gold standards, can be an efficient way to preselect qualified users. Gold standard tasks are test tasks with known answers that are inserted into regular sets of tasks as a way to sample the quality of answers of individual agents [4]. Some papers proposed alternative workflows such as peer-review mechanisms that asks agents to rate a sample answer [3] or an iterative workflow which let's agents collaboratively improve answers from previous agents [11]. Similarly, Turkomatic, a system that enables users to crowdsource complex tasks, automatically decomposes a complex task into smaller, verifiable subtasks that are chained after one another [10]. Last, Get another Label [9] discusses the use of a predictive model to obtain a confidence metric on the accuracy of answers based on agreement among users.

The guided task creation and the task preview interface in Fantasktic focus on the task specification step which is prior to the task processing step in which most of the above techniques are applied. However, the effectiveness of the majority vote mechanism is dependent of the input being identically formatted. The automated tutorial technique which we evaluate shares similarities with the gold standard mechanism in the sense that tasks with known answers are presented to the agent and their answers are compared to the gold answer. However, our approach differs in several ways: Instead of inserting gold tasks across a batch of tasks without the agent's knowledge, we present gold tasks at the beginning of a batch and inform agents about it. As a result, the two techniques address different quality assurance problems: While the sampling mechanism addresses the incentivization problem, i.e. untrustworthy agents, the tutorial technique focuses on task specification and human error problem by reducing the perceived ambiguity for agents to carry out the task instructions assuming a trustworthy agent.

Simplifying Task Specification in Crowdsourcing Systems

Researchers have built innovative applications that rely on crowds. These applications usually hide the complexity of task specification from their users and leverage use case specific task designs that have been hand-crafted by researchers. For instance, Bernstein et. al. proposed Soylent, a powerful word processing tool to shorten, proof read, and edit documents with the help of crowdsourcing [3]. By embedding Soylent into Microsoft Word, the researchers were able to provide a complex service through a simple interface that would hide most of the complexity of task specification, error control and turn around times. Similarly, researchers demonstrated that crowdsourcing can be a powerful tool for visually impaired individuals by presenting a mobile application called VizWiz which lets users audio record a question to a photo taken of an object and receive an answer from a crowdsourcing platform in almost real-time [14]. VizWiz lets users record their problem or question verbally and submits these audio recordings as part of the task instructions to Amazon's Mechanical Turk where they are played by agents. More recently, Turkomatic presented a different approach by proposing a tool that used agents to assist in decomposing a complex task posed in natural language, finding that in certain circumstances the crowd could be trusted to automatically decompose complex work [10]. While these applications are successful, their approach does not generalize to situations in which users want to express new tasks themselves.

Fantasktic is designed as a general-purpose interface and hence cannot leverage task specific optimization that most applications can apply. Turkomatic presents an approach which could potentially be used for a similarly broad variety of task types as Fantasktic. However the researchers report that short, less specific instructions such as “Write a 3-sentence essay about X” or “create a list of people who are Y” returns reasonable results only in some cases.

Learner-Centered Design

Learner-centered design in human-computer interaction aims to create and evaluate learning aids using software tools [15]. Learner-centered design distinguishes itself from user-centered design by not only promoting usability but facilitating the user's understanding of the presented content which hence requires a different design process. Among various techniques, learner-centered designed software leverages two concepts that are of particular interest in respect to this paper, namely apprenticeship-based learning and case-based learning.

Apprenticeship-based learning combines learning of conceptual knowledge with process learning. It can entail techniques such as authentic task design, presenting tasks sequentially, collaboration tools for learners and scaffolding. Scaffolding is a technique to provide coaching, communicate process, and facilitate the learner's articulation in a gradually receding manner as the learner progresses and has been applied in various learning based software today [16]. Researchers have used various design principles to develop apprenticeship-based learning software which includes the suggestion of potential problems and solutions, usage of multiple, linked representations of the project, the design for use in practice and providing adaptable scaffolding [16,17].

The guided task creation interface for end users leverages some design principles for apprenticeship-based software by outlining potential challenges which might occur when processing the tasks and provides suggestions. The task preview interface enables end users to view a different representation of their task by showing a rendered task instance from the perspective of an agent in order to facilitate a review of the effectiveness of their instructions.

Case-based learning presents stories of analogous experience as a way to let learners infer the underlying principle of judgment and to promote the transfer of knowledge [18,19]. The complexity of the cases determines the level of sophistication of the system. For instance, simple “learning by example” mechanisms usually do not require a case library because the subject matter can be conveyed to the learner by a simple example [19].

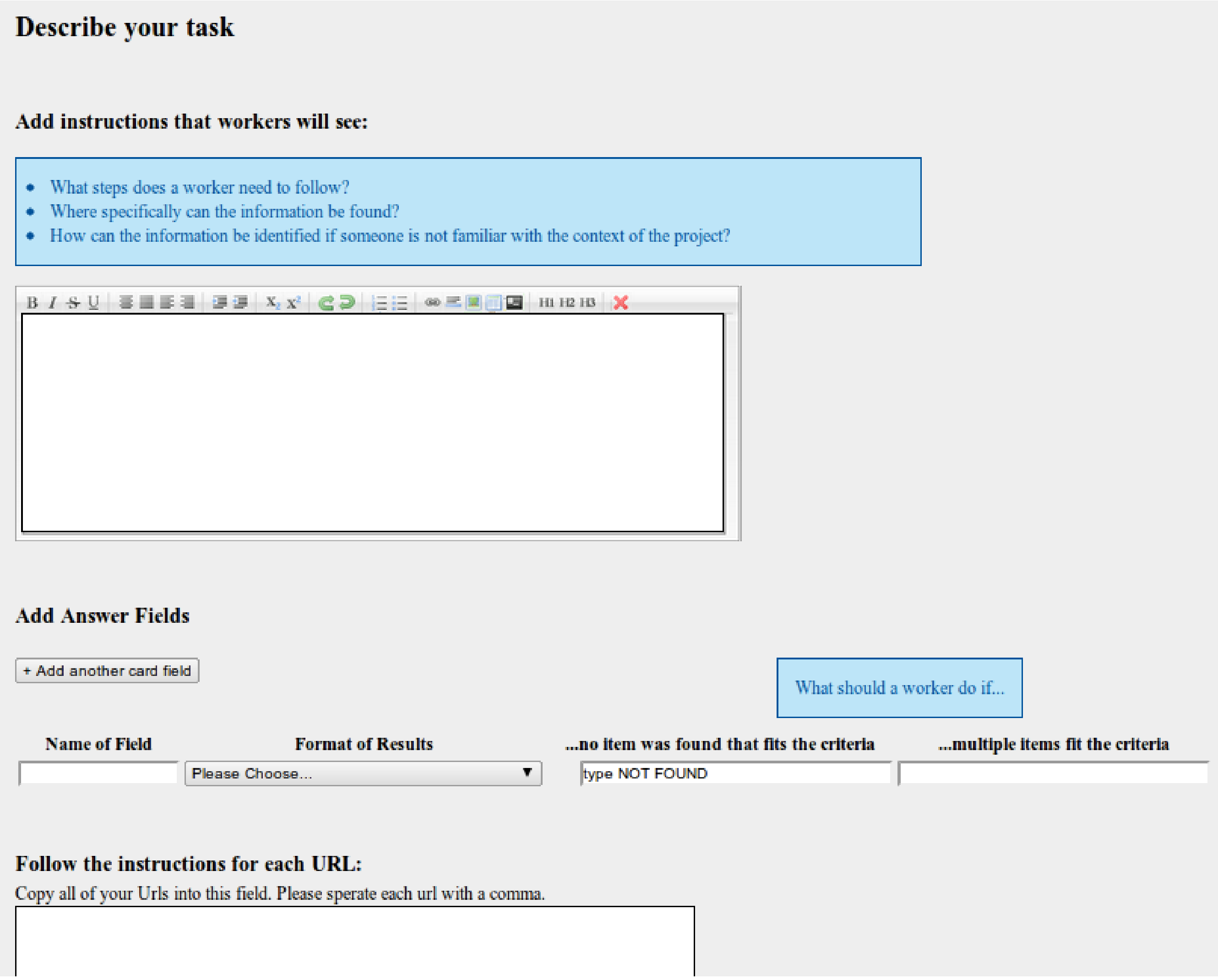
The automated tutorial mechanism makes use of the case-based learning approach by providing example answers to specific task instances. The technique presents the case in an interactive manner by rendering the task, having agents answer it on their own and upon submission informing the agent which answer matches the gold answer and which one does not. This process leverages the principle of “learning by example” to allow agents infer the judgments they need to apply in case instructions are ambiguous.

The Fantasktic Interface

Fantasktic is a system to investigate how to better support end users in creating successful crowdsourcing tasks. The system comprises three main components: a guided submission interface, a task preview interface and an automated tutorial generator. Our research is motivated by enabling novice users to receive reasonable answers from crowdsourcing platforms without having to iterate multiple times over their instructions. We hypothesize that the guidelines and preview techniques will improve the instructions by specifying formats and edge cases for the task, whereas the automated tutorial mechanism allows agents to infer the end users unexpressed specifications when a task lacks further instructions.

Guided Task Specification Interface

Fantasktic's task specification interface guides the user through the process of creating a task (see figure 3). The goal for the interface is to raise awareness towards potential problems and to provide recommendations how specific edge cases could be handled. Tasks may contain various ambiguities for agents such as cultural or contextual knowledge, particular edge cases such as unaccessible media, missing information or several information that fit the criteria in the instructions. Similarly, the format of the answer e.g. for telephone numbers (international format, national format, numbers only, etc), are relevant to reduce uncertainty for agents and to ensure same formats across tasks.



Describe your task

Add instructions that workers will see:

- What steps does a worker need to follow?
- Where specifically can the information be found?
- How can the information be identified if someone is not familiar with the context of the project?

Add Answer Fields

Name of Field	Format of Results	...no item was found that fits the criteria	...multiple items fit the criteria
	Please Choose... ▼	type NOT FOUND	

Follow the instructions for each URL:

Copy all of your Urls into this field. Please sperate each url with a comma.

figure 3: The guided task specification interface includes info boxes and additional options

To address these challenges, a notification on top informs the user about potential ways to improve the effectiveness of specifications such as subdividing a larger task into small steps and providing additional information if the task requires contextual knowledge. An additional section allows users to specify the type and format of each answer fields (a piece of information such as name or title of a person which is entered into an HTML form field). Example types are “Text: Capitalize Each Word”, “Text: EXACT copy” and “Phone Number: 5104938204” for telephone numbers (see table 1). These types are used as additional instructions to inform agents how to format the answer. In addition, the section asks the user to specify for each answer field what action an agent should carry out if the information cannot be found or if multiple information fit the criteria in the instruction. In addition, it suggests to set the default answer to “NOT FOUND” in case information are not existing.

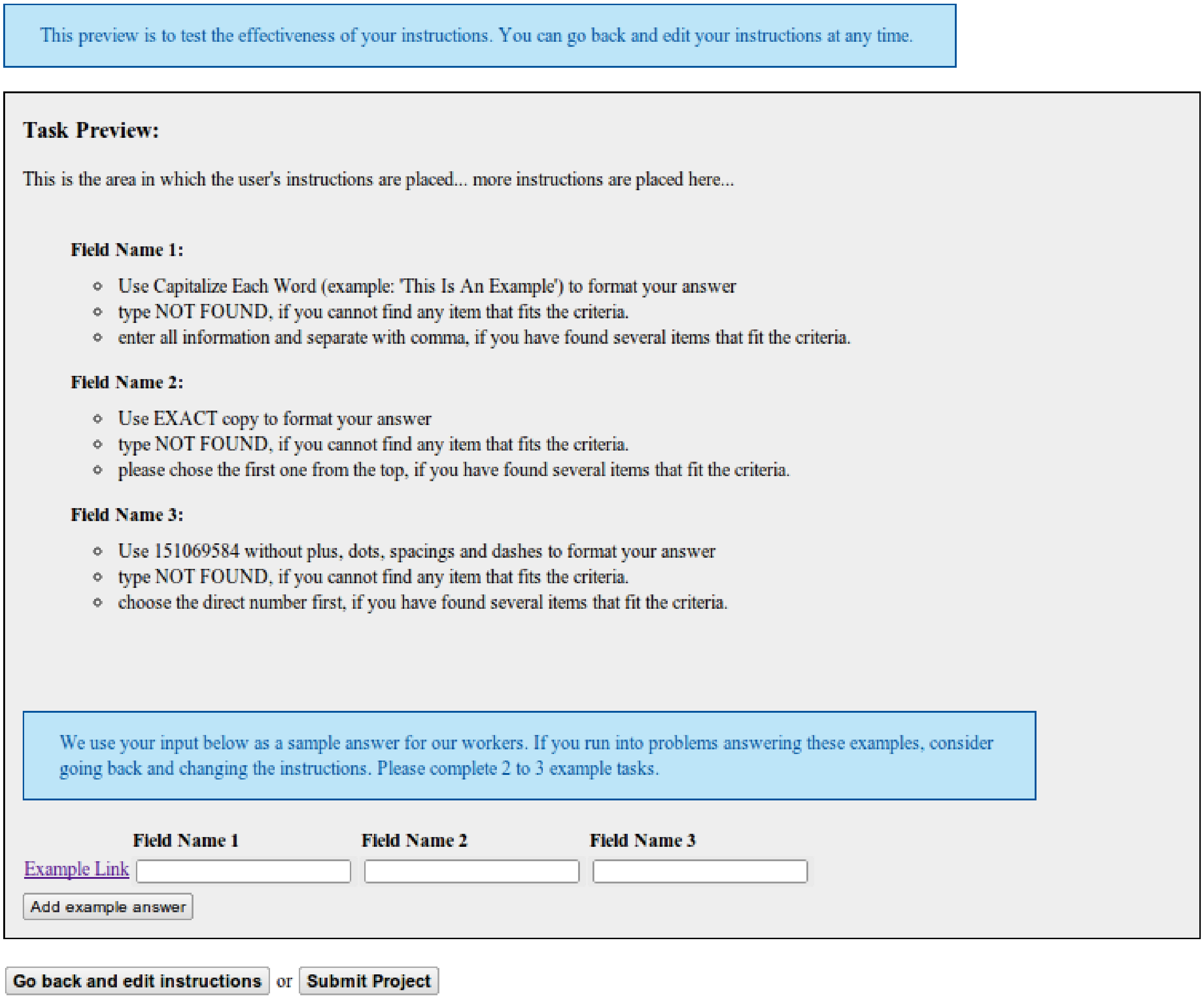
Name of Field Type	**Format of Field**
Text	Capitalize Each Word
Text	Use sentence case
Text	EXACT copy
Text	lower case
Text	UPPER CASE
Comma separated values	word1,word2,word3
Email Address	email@address.com
Phone Number	15106584724
Date	MM/DD/YYYY
Date and Time	MM/DD/YYYY HH:MM:SS
Arbitrary Number	23 or 54.25

table 1: List of all available answer formats

Task Preview

The task preview interface renders the task from the perspective of an agent, enabling users to change roles and get an alternative view of their task. The goal for the task preview is similar to the guided task specification interface focusing on potential missing or contradicting instructions. By looking at the task preview, users may be encouraged to update their instructions and improve them.

The top of the task preview contains an info box which describes the purpose of the preview screen. The interface displays the user's instructions below the info box together with instructions for type and format for the specific answer fields. At the bottom, the interface renders HTML input fields for agents to type in the answers. In addition, the interface provides example data such as URL, images or text that the end users is submitting as part of the task.



This preview is to test the effectiveness of your instructions. You can go back and edit your instructions at any time.

Task Preview:

This is the area in which the user's instructions are placed... more instructions are placed here...

Field Name 1:

- Use Capitalize Each Word (example: 'This Is An Example') to format your answer
- type NOT FOUND, if you cannot find any item that fits the criteria.
- enter all information and separate with comma, if you have found several items that fit the criteria.

Field Name 2:

- Use EXACT copy to format your answer
- type NOT FOUND, if you cannot find any item that fits the criteria.
- please chose the first one from the top, if you have found several items that fit the criteria.

Field Name 3:

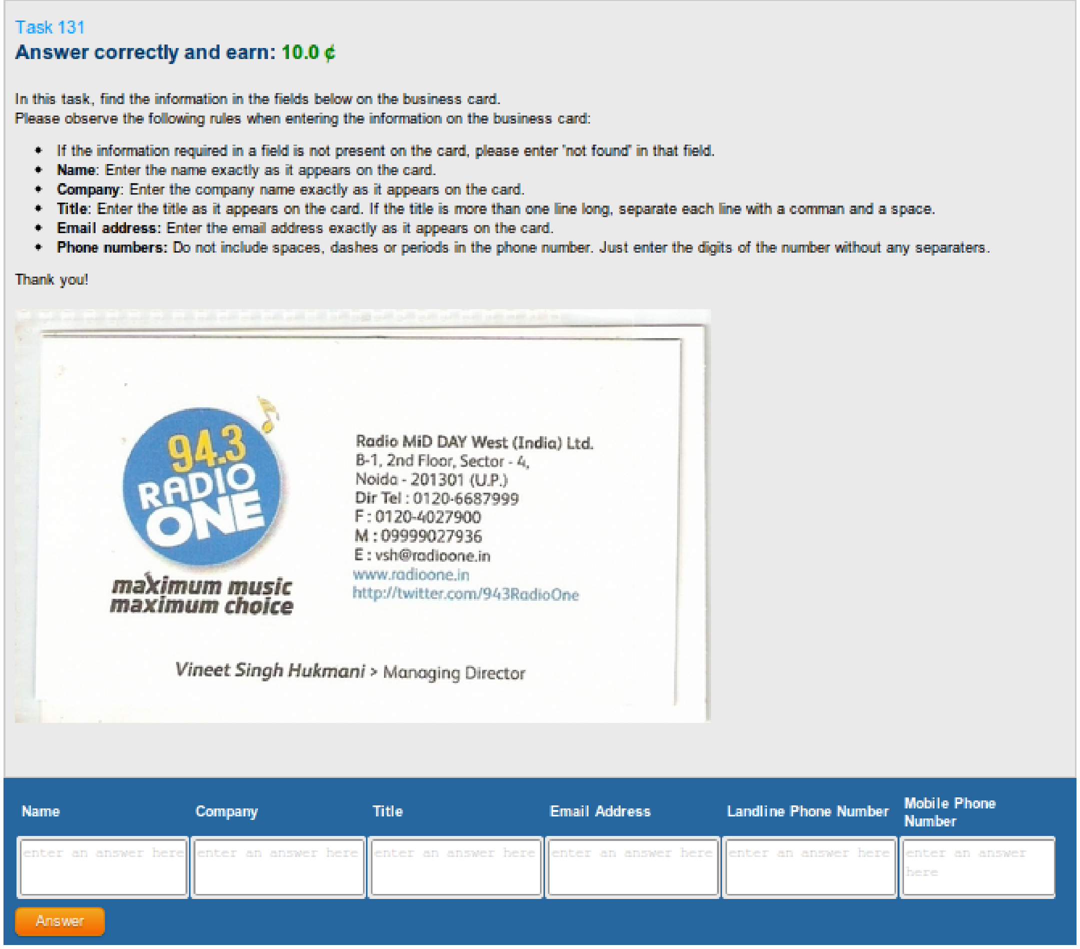
- Use 151069584 without plus, dots, spacings and dashes to format your answer
- type NOT FOUND, if you cannot find any item that fits the criteria.
- choose the direct number first, if you have found several items that fit the criteria.

We use your input below as a sample answer for our workers. If you run into problems answering these examples, consider going back and changing the instructions. Please complete 2 to 3 example tasks.

figure 4: The task preview screen renders a task from the perspective of an agent

Automated Tutorial Generator for Online Agents

The interface enables users to submit sample answers for several tasks which will be used to generate tutorials for agents. The goal of this technique is to enable agents to infer the pattern and underlying judgment principles of users in situations in which the instructions lack clarification.



Task 131

Answer correctly and earn: 10.0 ¢

In this task, find the information in the fields below on the business card.
Please observe the following rules when entering the information on the business card:

- If the information required in a field is not present on the card, please enter 'not found' in that field.
- **Name**: Enter the name exactly as it appears on the card.
- **Company**: Enter the company name exactly as it appears on the card.
- **Title**: Enter the title as it appears on the card. If the title is more than one line long, separate each line with a comman and a space.
- **Email address:** Enter the email address exactly as it appears on the card.
- **Phone numbers:** Do not include spaces, dashes or periods in the phone number. Just enter the digits of the number without any separaters.

Thank you!

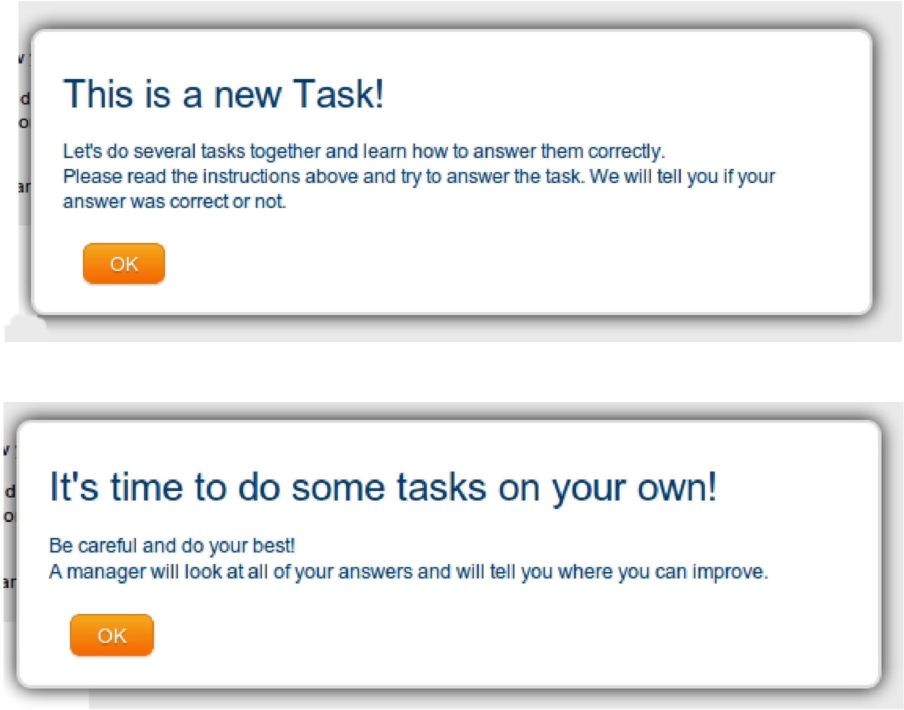
Radio MiD DAY West (India) Ltd.
B-1, 2nd Floor, Sector - 4,
Noida - 201301 (U.P.)
Dir Tel : 0120-6687999
F : 0120-4027900
M : 09999027936
E : vsh@radioone.in
www.radioone.in
http://twitter.com/943RadioOne

Vineet Singh Hukmani > Managing Director

Name	Company	Title	Email Address	Landline Phone Number	Mobile Phone Number
enter an answer here	enter an answer here	enter an answer here	enter an answer here	enter an answer here	enter an answer here

figure 5: The tasks submission interface for agents

Based on the sample answers submitted by a user, the system generates a tutorial for agents which is presented before the agent starts completing tasks. The presentation of a tutorial is similar to the presentation of a task: The instructions of the user are rendered on top, the business card or URL is provided below the instructions and the answer fields are placed at the bottom (see figure 5).



This is a new Task!

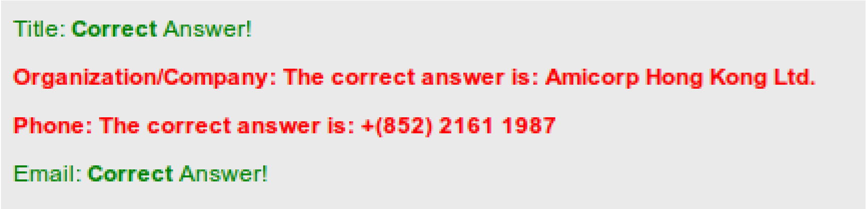
Let's do several tasks together and learn how to answer them correctly.
Please read the instructions above and try to answer the task. We will tell you if your answer was correct or not.

It's time to do some tasks on your own!

Be careful and do your best!
A manager will look at all of your answers and will tell you where you can improve.

figure 6: Info boxes that prompt the agent before starting the tutorial (above) and after completing it (below)

If an agent sees a tutorial for the first time, the system shows an info-box which purpose is to inform the agent about the upcoming tutorial tasks. Once the agent has completed a tutorial task and clicked on 'Answer', the system displays information for each answer field whether it has been answered correctly or what the correct answer should have been (see figure 7). Once the set of tutorial tasks have been completed, a non-tutorial task is presented and a prompt notifies the agent that the tutorials have been completed and that the following tasks will be 'real' tasks. As described earlier, rather than trying to filter out untrustworthy agents, this techniques aims to remove remaining ambiguities because of lacking instructions and to reduce perceived uncertainty caused by the complexity of the task.



Title: Correct Answer!

Organization/Company: The correct answer is: Amicorp Hong Kong Ltd.

Phone: The correct answer is: +(852) 2161 1987

Email: Correct Answer!

figure 7: Prompt for agents after submitting a tutorial task

Implementation

The researchers had access to a private instance of the commercial crowdsourcing platform MobileWorks to support the techniques described in the paper. The back-end service on the server side is based on the django framework and the client side to render tasks for agents as well as the submission interface for users is based on HTML, CSS and jQuery. We manually generated combinations of instructions from the user and submitted them through the MobileWorks API. We set up batches of tasks each one of them containing the same 20 business card digitization tasks but with variations of instructions. We restricted agents from completing several batches by logging their IP address and prohibiting them from accepting more tasks as soon as they completed the first batch.

Evaluation

The goal of our evaluation is to understand whether automated task design techniques can help novice users create more successful crowdsourcing tasks. We investigate this research question by presenting Fantasktic which introduces three new task design techniques: A guided task specification interface, a task preview screen and an automated mechanism to generate tutorials for agents. For the purpose of this study, we assume trustworthy agents and focus on problems that are associated with ambiguous instructions.

Study Design

To evaluate our hypothesis, we conducted a study with ten novice end users who used Fantasktic to create business cards digitization tasks and simple email web search tasks. Users completed these tasks in two different interfaces: 1) A simple HTML submission form containing a field for written instructions, a field to specify the answer boxes and a field to paste in the URLs of 20 scanned business cards (see appendix). 2) A submission form with additional options for answer fields to specify edge cases (see figure 3) and a second screen to preview the task (see figure 4).

First, we manually created a test set of 20 business cards and 20 websites. Afterwards, we recruited ten study participants who did not have any prior crowdsourcing experience and let them generate variations of 4 instructions for a business cards task using Fantasktic. For each of these instruction variations we used the same set of business cards to generate batches of 20 tasks totaling 440 batches or 8800 tasks. We submitted these tasks into a crowdsourcing platform were we had each one of the 440 unique workers complete a batch of cards. Because it is considered a personal item, business cards vary widely in the way they contain and present information about a person though they are easy to understand for study participants without the specific context of the project.

Task Submission with Novice Users

We recruited ten users: five graduate students from the School of Information and the Architecture department and five professionals who work in project management-related positions at large IT companies. The demographic of the test pool had a mixed gender (4 female and 6 male) and their age ranged between mid-twenty and mid-thirty. We introduced each user to the study via scenarios. The business card scenario was to use a crowdsourcing service to convert physical business cards collected during a conference into Microsoft Outlook contacts. We presented participants three printed examples of business cards to allow them to familiarize themselves with the task. In addition, we provided the with a printed spreadsheet which showed a table of four columns with the headlines “Name”, “Title”, “Organization” and “Cell Phone Number” and three rows which contained the name, title, organization name and cell phone number of each of the three printed business cards. We specifically instructed users that the results generated from the crowdsourcing project were required to match exactly the format and content as seen in the excel sheet in order to be imported into Outlook. We followed the same process with the email search scenario for which we asked users to collect contact email addresses from websites to add them to their newsletter. In addition, we provided a general introduction of what crowdsourcing is. We also described how the service in the study would decompose the submitted project into small tasks of a single image of a business card or URL of a website and would display the user's instructions to inform the online worker what actions had to be carried out.

We showed each user two interfaces for submitting a project into the crowdsourcing service throughout the study. First, we showed a simple interface that contained a text area to place the instructions for a worker, a section to add and name an answer field for the task, e.g."Name", "Title", as well as a section at the bottom to upload the business cards or add the URL of the websites (see appendix). We asked users to submit their business card project via this interface without additional guidance. Second, we showed users the Fantasktic interface and asked them to resubmit their project again through the new interface (see figure 3). After users completed the first step (guided task specification) in the interface, we stored their task specification in order to decompose the impact of the different components of Fantasktic during the evaluation phase. In the second step (task preview), we showed users a preview of their task from a worker perspective and asked them to provide sample answers to three business cards which would be used as examples for workers (see figure 4). We recorded changes in the instructions that were made during this step.

In addition, five of the user created instructions for the email search scenario. Our goal with this task was to make qualitative observations if the impact of Fantasktic compared to the simple interface can be replicated with different tasks and not specific to business card digitization tasks. In order to limit potential learning effects, we had participants create tasks for both business cards and email search through the baseline interface first and then introduced them to the new interface.

Batch	Participant	Condition	Business Card	Agent
1	P1	Baseline	Card 1	Agent 1
1	P1	Baseline	Card 1	Agent 2
1	P1	Baseline	Card 1	...
1	P1	Baseline	Card 1	Agent 10
1	P1	Baseline	Card 2	Agent 1
1	P1	Baseline	Card 2	Agent 10
1	P1	Baseline	...	
1	P1	Baseline	Card 20	Agent 10
2	P1	Guided Interface	Card 1	Agent 11
2	P1	Guided Interface	Card 1	Agent 20
2	P1	Guided Interface	...	
2	P1	Guided Interface	Card 20	Agent 20
3	P1	Task Preview	Card 20	Agent 30
3	P1	…		
4	P1	Agent Tutorial	Card 20	Agent 40
5	P2	Baseline	Card 1	Agent 51
6-8	P2	Agent Tutorial	Card 20	Agent 80
9-40	...			
41-44	P10	Agent Tutorial	Card 20	Agent 440

table 2: Illustration of the study setup and how tasks are assigned into batches

Submitting Study Tasks Into Crowdsourcing Service

To evaluate Fantasktic quantitatively, we submitted business card tasks into the crowdsourcing platform MobileWorks. For each user, we generated 4 batches of (the same) 20 business card tasks (see table 2) each of which had a different task specification:

1. The first batch was generated from the instructions provided through the simple interface.
2. The second batch is based on the instructions of the first step in Fantasktic, the guided task interface
3. The third batch contained the edits which were made as a result of the preview in Fantasktic.

4. The fourth batch was based on the third batch but in addition used the user's example answers to generate a tutorial for workers upfront. To workers, the tutorials looked like one of the other 22 tasks but on submission would prompt a notification whether all information were correct (a match with the user's example answers) or were incorrect and would need to be corrected to the displayed answer (see figure 7).

To ensure that no learning effects would bias the results, we recruited 10 new online worker for each batch that was submitted into the crowdsourcing platform. Each worker who we recruited completed two batches of business card tasks in the sequence of first and third batch or second and fourth batch. For the entire study, we generated 44 batches of 20 business cards, i.e. 8800 tasks. These tasks were answered by 440 unique workers who answered a single batch of 20 business cards each.

Measures

Covering every edge case for business cards is a challenge since they do not necessarily follow a standard format. This leaves sufficient space for subjective judgment of an agent completing this presumably trivial task. We define three performance measures to compare results from our study: The comparison of answer formats across the techniques, the measurement how many answers provided the same information and the benchmark of each technique compared to expert instructions.

Condition: Baseline UI	**Condition: Guided UI**	**Condition: Task Preview**	**Condition: Agent Tutorial**
Simple UI	Guided UI	Guided UI	Guided UI
		Task Preview	Task Preview
			Agent Tutorial

table 3: Illustration of the setup of study conditions

Similarity of Answers

For each business card, given a scenario and given instructions, we compute the string distance of each of the ten answers with one another and determine the median string distance. In addition, we compute the distances after converting the answers to lower case and removing special chases and spaces to better attribute what 'type' of formatting accounts for the differences.

Similarities in Information

We measure how much of the variance in answers is due to different information, e.g. different telephone numbers, rather than different formatting of the same telephone number. We run a one-way ANOVA test ($\alpha=0.05$) on our results to determine if the distribution between the conditions differs significantly.

Benchmark To Expert Instructions

We extracted the set of instructions used by a business cards digitization service and ran them as part of our study (without an initial tutorial). We compare the string distance and vector similarity of the expert results with those from the different techniques. Similarly, we run a one-way ANOVA test ($\alpha=0.05$) on the dataset to find out whether the proposed techniques show a significant improvement against the baseline towards expert task performance.

Results

We report comparisons for formatting differences and content differences and describe qualitative observations made throughout the study.

Comparing Similarity of Answers

Comparing the overall similarity of answers for business cards in the different scenarios, we find a statistically significant difference between the baseline and the guided interface condition as well as the guided interface and the expert tasks ($P < 0.01$). We do not see significance between the guided interface and the task preview ($P \approx 0.75$) as well as the guided interface and the agent tutorials ($P \approx 0.35$). These results suggest that the guided task interface efficiently reduces the variance of answers while the task preview and the agent tutorials do not improve the results significantly. For the purpose of this research, we tested our conditions in a non-independent manner by conducting each step of the study based on one another, i.e. the task preview builds on instruction sets from the guided task interface. As a result, the effectiveness of each technique can be approximated by observing the relative differences between the previous technique.

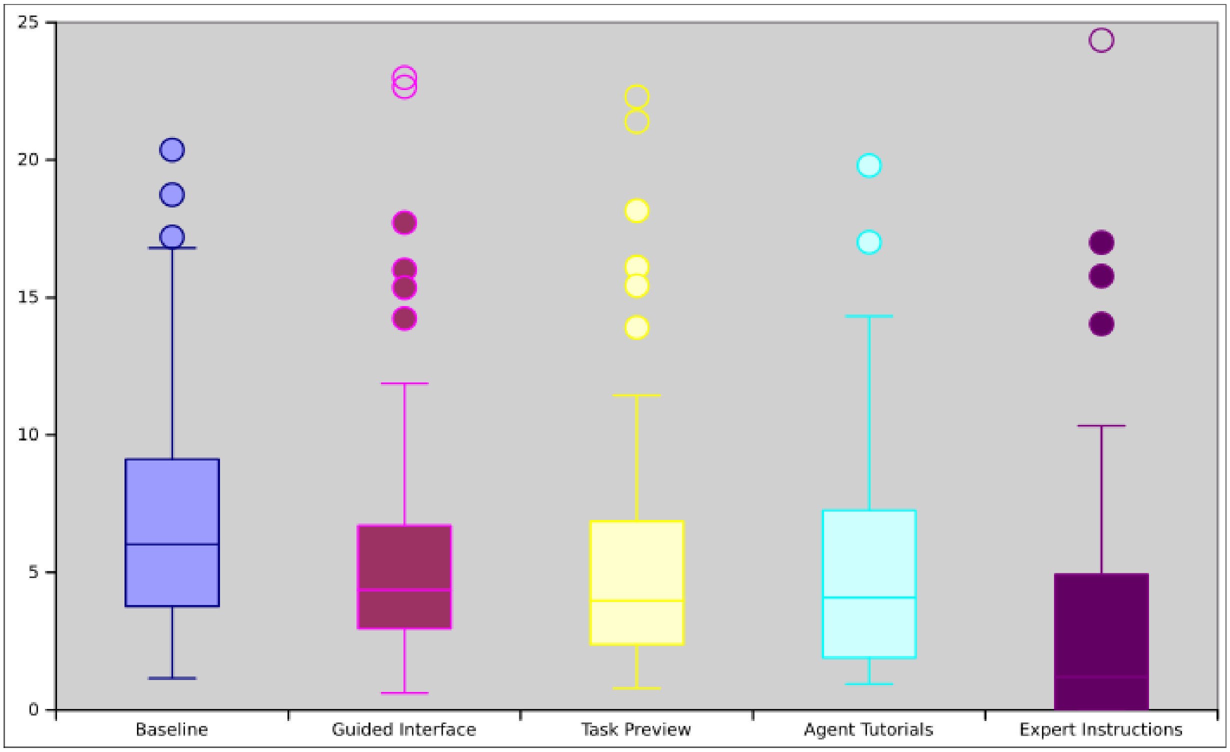


figure 8: Overall string distance of answers for the 5 scenarios

Furthermore, we did a more granular analysis of the conditions for different fields. For the name field, the graph shows a significantly smaller interquartile range across the conditions with most upper quartiles below values of 5. In addition the test shows a weak significance between the guided interface and the agent tutorials ($P < 0.03$). These observations indicate that the name field accounts for significantly less variation in answers compared to the overall variation in answers. Hence, the potential improvement for the tested conditions is limited.

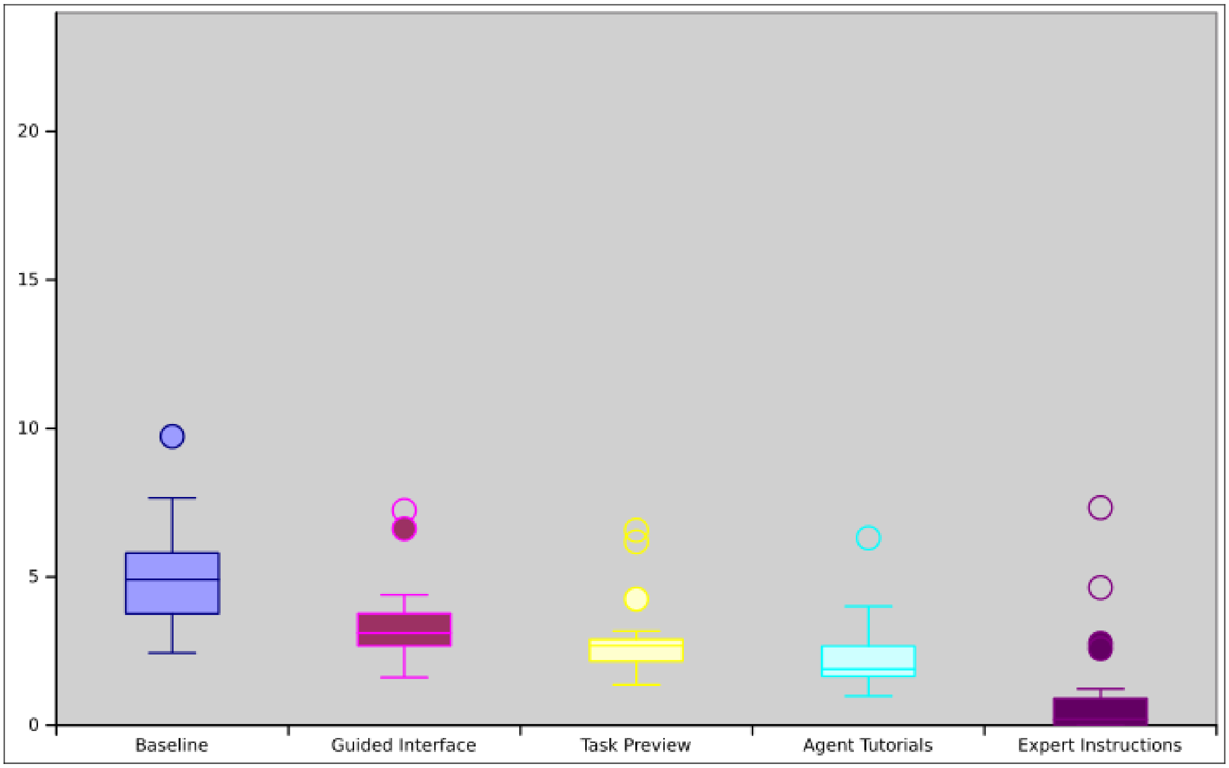


figure 9: String Distance for the 5 scenarios for the ***name field***

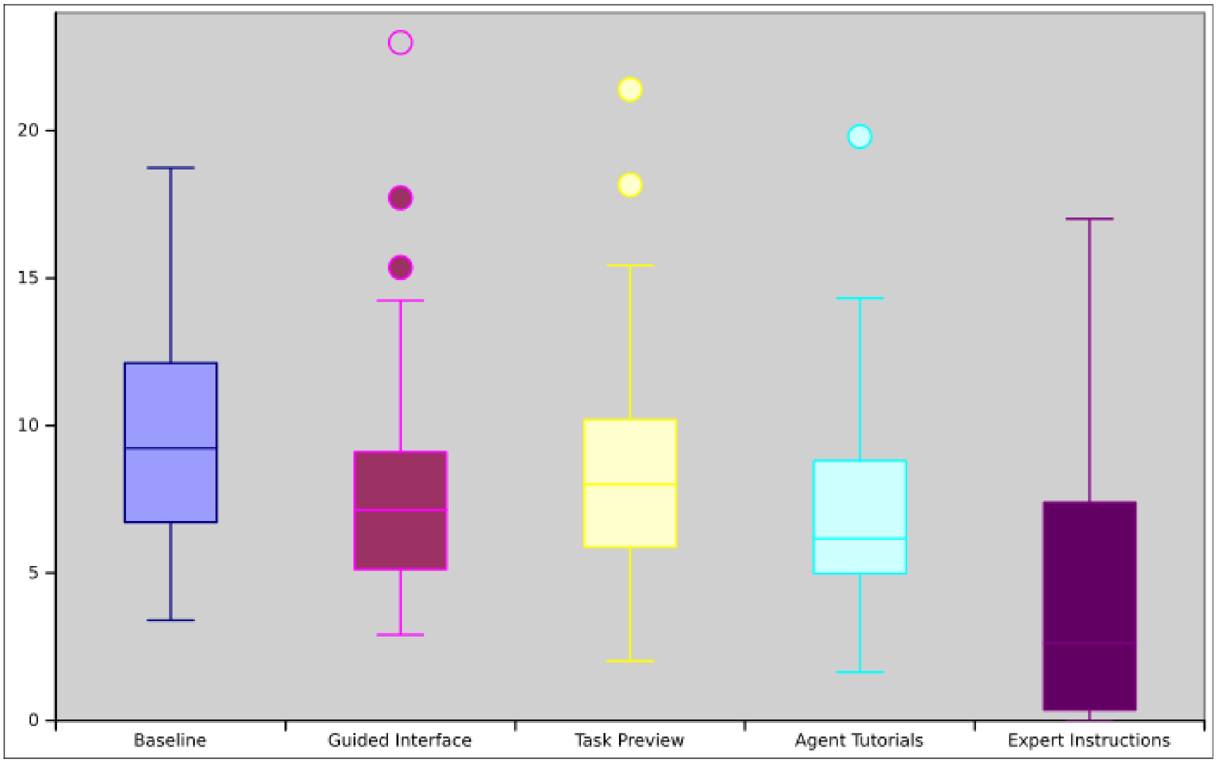


figure 10: String Distance for the 5 scenarios for the ***organization field***

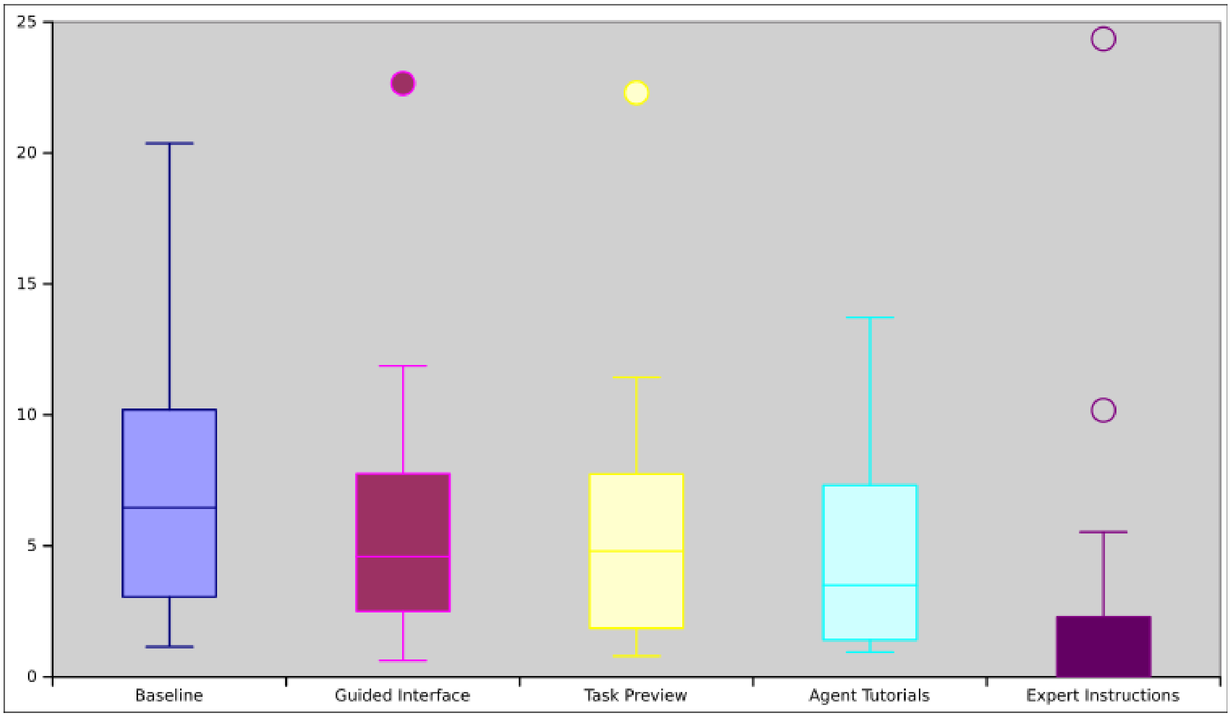


figure 11: String Distance for the 5 scenarios for the ***title field***

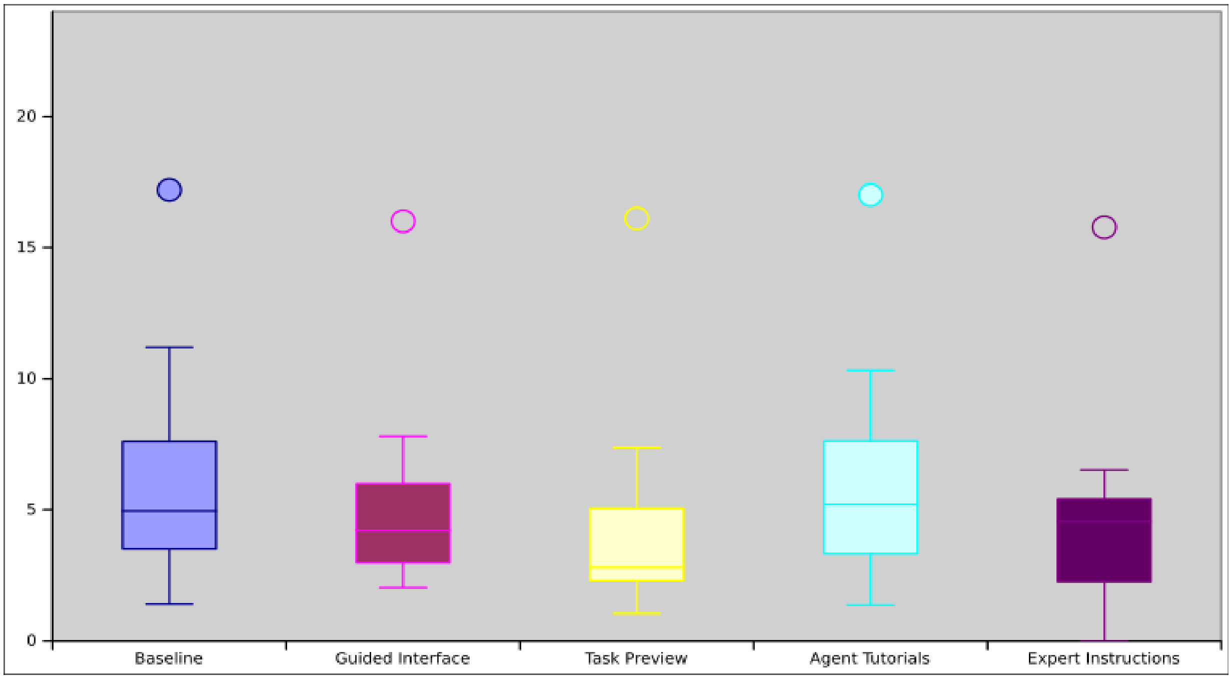


figure 12: String Distance for the 5 scenarios for the ***cell phone field***

Comparing Answer Similarity Attributed to Formatting

Besides comparing the overall string distance, we applied changes to the answers and recomputed the distance. We tested our results for 1) removing spaces in answers, 2) removing any character except alphanumeric characters, 3) converting all characters to lower case. This allows a more granular insight into what formatting descision impact the overall difference. We plotted the results in a graph of 16 bars. Each bar shows the fraction in the answers that can be attributed to differences in cases, usage of special characters, or spacing. The fraction of unattributed string distance accounts for cases in which different information has been provided as the answer, e.g. two different telephone numbers (the string distance is very high for two different information) as well as spelling errors. The numbers below the bars encode the following information: bars 1 through 4 encode the baseline scenario in the order of name field (1), title (2), organization (3), and cell phone number (4). This order continues with bars 5 through 8 which shows the results for the guided interface. Bars 9 – 12 show the results for the task preview screen and bars 13 – 16 illustrate the worker tutorial technique.

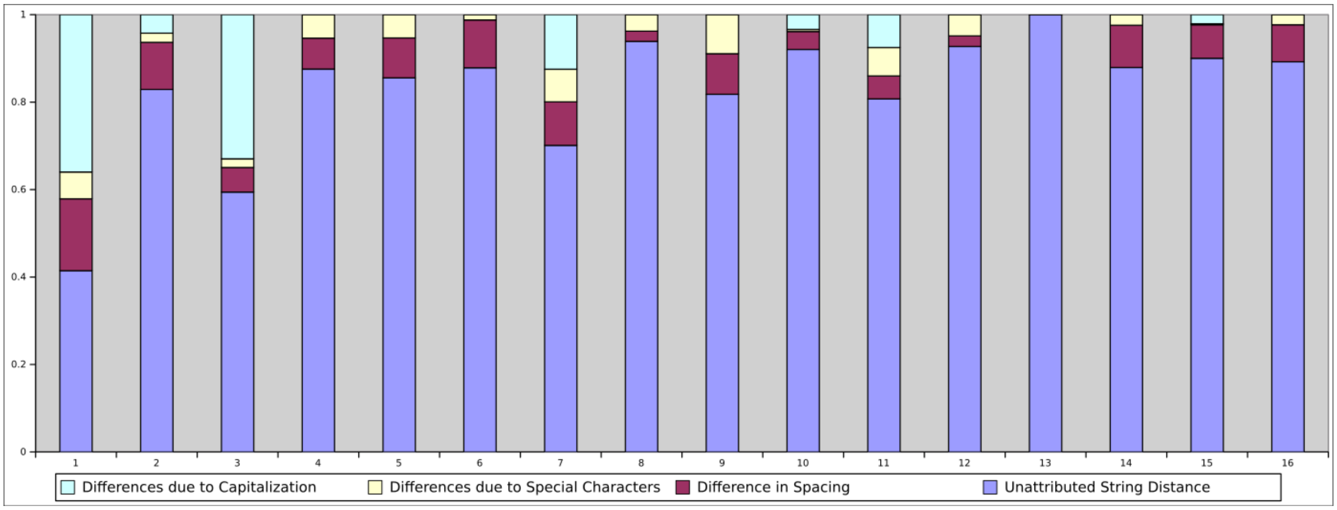


figure 13: Breakdown of differences in formats per field per condition

Answers in the baseline scenario, in particular the name and the organization, show significant differences in capitalization. For instance in case of the organization field, the results show a significant reduction in capitalization across the scenarios. While spacing accounts for an additional 10% to 20%

depending on the field, special characters represent a minor fraction. The results for the guided interface, the task preview screen and the automated tutorials show significantly less differences due to formatting. This indicates that the techniques are effectively reducing the variance in answer formats. Compared to the other scenarios, the automated tutorial technique shows a reduction in differences for special characters while spacing accounts for a similar fraction across all scenarios.

Comparing Similarities in Information

We first alter the answers by removing multiple spaces, special characters applying the changes (1) to (3) from the string distance comparison. For a specific business card, given a scenario and given instructions, we use the set of answers to create a vector of all unique words as well as vectors of words for each answer. We compute the cosine similarity of each combination of vector with one another.

Besides comparing the answer format, we extracted all words used in answers for a particular business card, given a set of instructions and interface. For each answer given for this batch, we compared its cosine similarity with the vector of all unique answers and computed the median similarity. We computed the median similarity for each batch and compared the results across the scenarios.

The ANOVA test indicates a significant improvement of the guided interface over the baseline scenario. Between the guided interface, the task preview, the agent tutorial and the expert instructions scenario, we do not find this significance. The results show that any of the techniques reduces the amount of answers in which the agent chose different information, e.g. different telephone numbers. Furthermore, it shows that the techniques achieve a performance similar to the expert instructions.

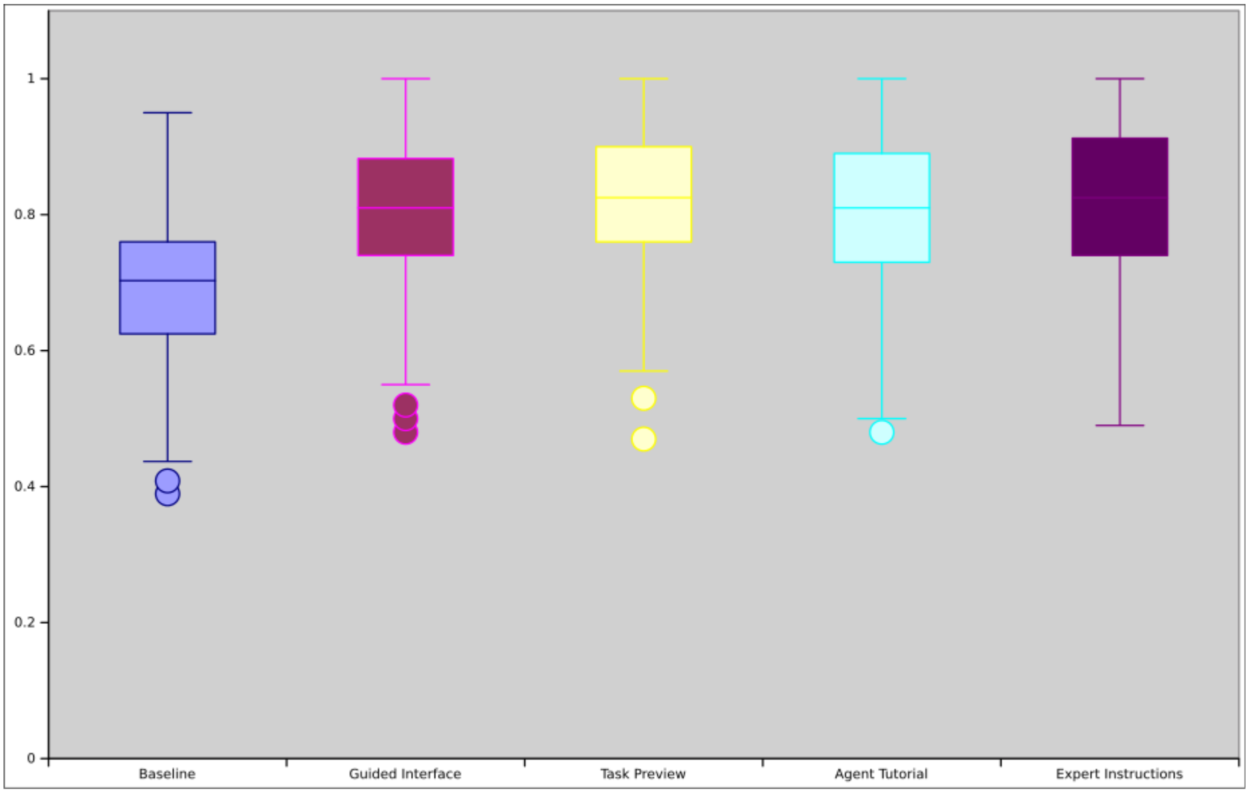


figure 14: Comparison of similarity in answers for different conditions

Qualitative Observations

In this section we share qualitative observations about the participant's behavior during the study and an analysis of the instructions that were created for the email address search task.

During the task preview we observed that participants went through two to three task examples. Seven of the ten participants recognized that their instructions were insufficient to cover an edge case they were facing. Several of the participants spoke out their observations, e.g. “This is good. This [business card] doesn't have a title” or addressed the researcher asking “I assume the handwritten phone number is the cell phone number?”. We observe that the majority of participants faced problems when providing example answers, only four returned to the instruction screen and made changes to address this problem.

We compare the example answers that participants provided for tutorials and the results that were returned for those batches. We observe that example answers strongly determine whether the tutorial will be effective or even detrimental. For instance, one participant provided examples that did not follow his own instructions: the example answer for the cell phone should have been a single telephone number with digits only, e.g. 150459425. Instead, the example answer listed all numbers displayed on the card and introduced a new formatting style, e.g. “+1 510 659 8594 / +1 220 594 5859 / +1 453 546 4523”. The tutorial presented agents two conflicting guidelines and as a result some agents adopted the guideline from the tutorial and some followed the instructions which created large amount of variation in the answers among the agents.

The set of instructions for email address web search show similar pattern for the different interfaces as the business card tasks. Instructions in the baseline condition tend to provide less information on average than instructions in the other conditions. During the task preview, participants were facing similar challenges as described above. However, only in a few cases, the participant went back and updated the instructions.

Discussion

In this section we discuss several important limitations of Fantasktic and the design of our study.

Study Limitations

Our study has several limitations that need to be addressed in future work. For the purpose of this study, we analyzed the proposed techniques in a non-independent way. The task preview displayed the revised instructions from the guided task interface and the tasks in the agent tutorial condition used the advanced instructions from the task preview. In addition, a separate study needs conduct a similar test on a larger sample size of novice users.

Conclusion

In this paper we investigated how we can design an automated task design technique that helps novice users to create successful tasks. We introduced Fantasktic as a way to test three potential techniques: a guided task specification interface, a task preview interface and an automated agent tutorial mechanism. The evaluation of our user study shows that the guided task interface has a positive impact by increasing the amount of agreement among agents for a particular information and reducing the amount of variability how the answers are formatted. Furthermore, we find that the impact of the task preview interface is not statistically significant and observe that most participants recognized unclear instructions when answering example tasks themselves though went back and made updates to the instructions only in the minority of cases. In addition, we discover that the impact of the agent tutorial mechanism is not significant either and find that its application can be effective and detrimental depending on provided example answer. Finally, the analysis of email search instructions indicate that the positive effect of the guided task interface can be applicable to a larger variety of tasks.

References

1 Huang, E., Zhang, H., Parkes, D.C., Gajos, K.Z., and Chen, Y. Toward automatic task design: a progress report. Proceedings of the ACM SIGKDD Workshop on Human Computation, ACM (2010), 77–85.

2 Quinn, A.J. and Bederson, B.B. Human computation: a survey and taxonomy of a growing field. Proceedings of CHI 2011, ACM (2011), 1403–1412.

3 Bernstein, M.S., Little, G., Miller, R.C., et al. Soylent: a word processor with a crowd inside. Proceedings of UIST 2010, ACM (2010), 313–322.

4 Le J., Edmonds, A., Hester, V., Biewald, L. (2010). Ensuring Quality In Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution. SIGIR.

5 Yuen, M-C., King. I., Leung, K-S. (2011). A Survey Of Crowdsourcing Systems. IEEE International Conference on Privacy, Secuirty, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 766-773.

6 Kulkarni A., Gutheim P., Narula P., Rolnitzky D., Parikh T., Hartmann B. (2012). MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture. IEEE Internet Computing, Sep/Oct 2012.

7 Ipeirotis, P.G., Provost, F., and Wang, J. Quality management on Amazon Mechanical Turk. Proc of ACM SIGKDD Workshop on Human Computation, (2010), 64–67.

8 Little, G., Chilton, L.B., Goldman, M., and Miller, R.C. Exploring iterative and parallel human computation processes. Proc. of the ACM SIGKDD Workshop on Human Computation, ACM (2010), 68–76.

9 Sheng, V.S., Provost, F., Ipeirotis, P.G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. KDD.

10 Kulkarni, A., Can, M., Hartmann, B. (2012) Collaboratively crowdsourcing workflows with Turkomatic. CSCW 2012.

11 Little, G., Chilton, L.B., Goldman, M., Miller, R.C. (2010). Turkit: human computation algorithms on mechanical turk. ACMUIST, pp. 57-66.

12 http://www.mturk.com, Amazon's Mechanical Turk crowdsourcing service. Visited 05/01/2012.

13 Von Ahn, L.; Dabbish, L. (2004). Labeling images with a computer game. Proceedings of the 2004 conference on Human factors in computing systems - CHI '04. pp. 319–326.

14 J. P. Bigham, et. al. (2010). VizWiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology. pp. 333-342.

15 Soloway E., Guzdial M., Hay K. E. (1994). Learner-centered design: the challenge for HCI in the 21st century. Interactions, 1:36–48.

16 Guzdial M., Kehoe C. (1998). Apprenticeship-based learning environments: a principled approach to providing software-realized scaffolding through hypermedia. Journal of Artificial Intelligence Education. 9:289–336.

17 Linn M.C., Clancy M. J. (1992). The case for case studies of programming problems. Communications. ACM, 35:121–132.

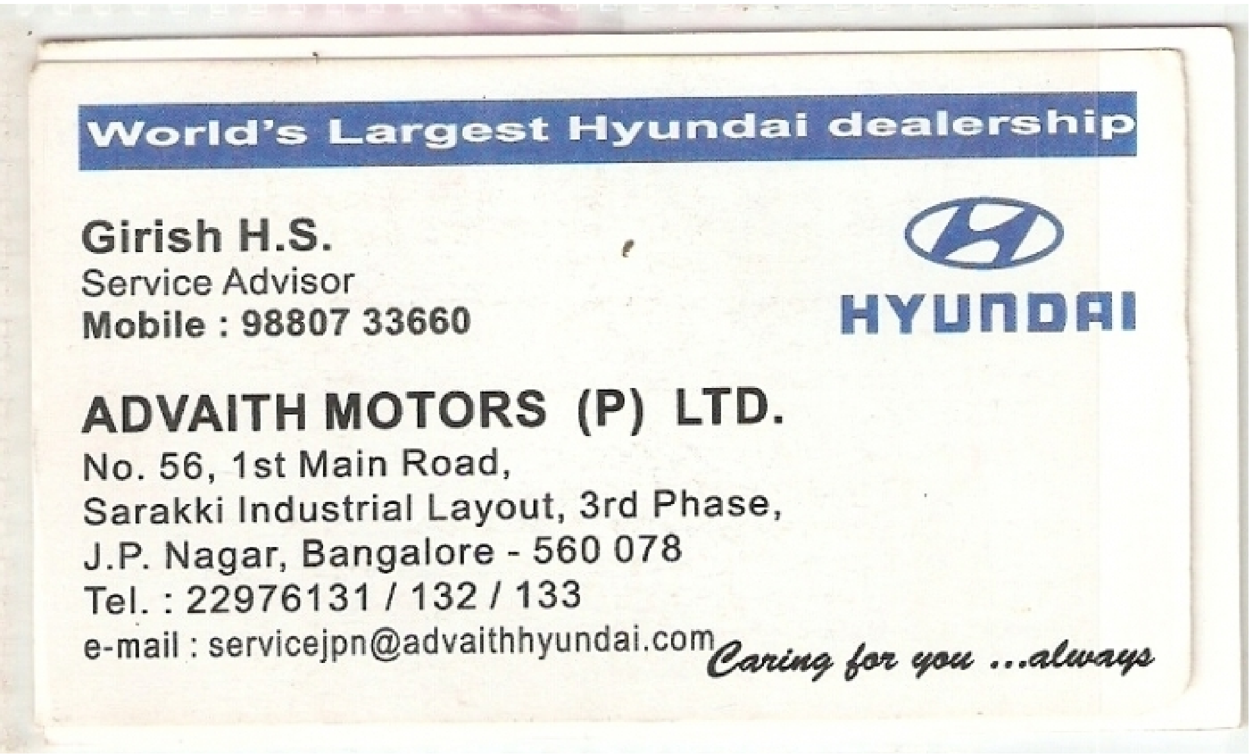
18 Kolodner J., Guzdial M (2000). Theory and practice of case-based learning aids. In David H. Jonassen and Susan M. Land, editors, Theoretical Foundations of Learning Environments. Lawrence Erlbaum Associates.

19 Lee B., Srivastava S., Kumar R., Brafman R., Klemmer S. R. (2010). Designing with Interactive Example Galleries. Proceedings in ACM Human Factors in Computing Systems (CHI).

Appendix

Example Business Cards

The following three business cards have presented to the study participants in a printed form:



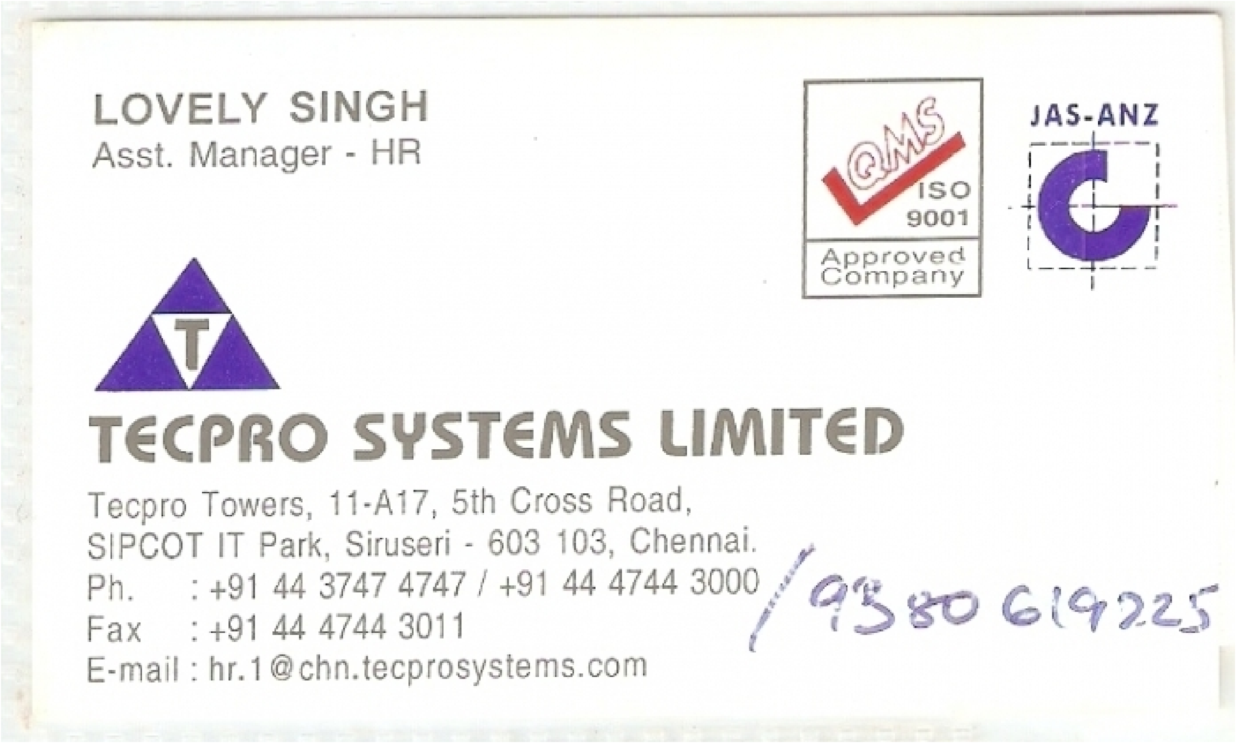
World's Largest Hyundai dealership

HYUNDAI

Girish H.S.
Service Advisor
Mobile : 98807 33660

ADVAITH MOTORS (P) LTD.
No. 56, 1st Main Road,
Sarakki Industrial Layout, 3rd Phase,
J.P. Nagar, Bangalore - 560 078
Tel. : 22976131 / 132 / 133
e-mail : servicejpn@advaithhyundai.com

Caring for you ...always



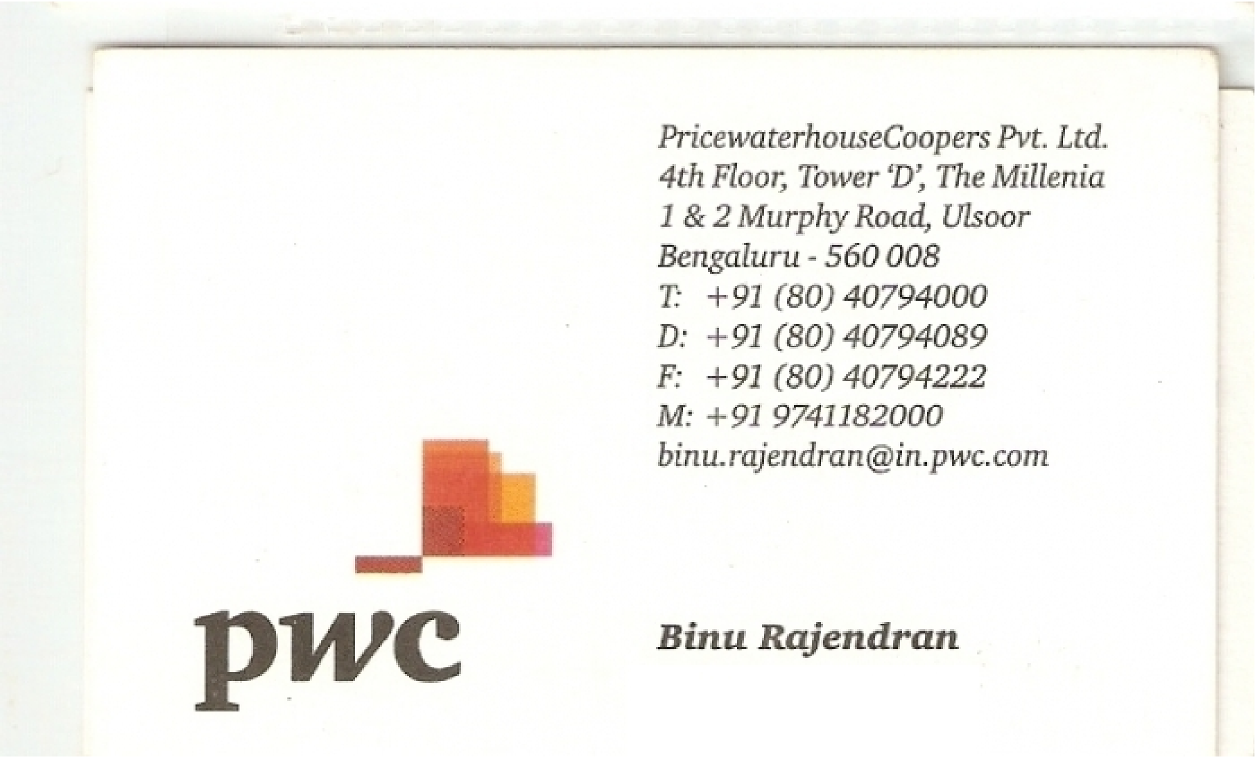
LOVELY SINGH
Asst. Manager - HR

QMS ISO 9001 Approved Company

JAS-ANZ

TECPRO SYSTEMS LIMITED

Tecpro Towers, 11-A17, 5th Cross Road,
SIPCOT IT Park, Siruseri - 603 103, Chennai.
Ph. : +91 44 3747 4747 / +91 44 4744 3000 / 9380 619225
Fax : +91 44 4744 3011
E-mail : hr.1@chn.tecprosystems.com



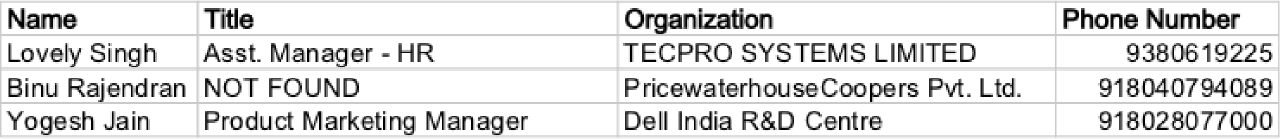
PricewaterhouseCoopers Pvt. Ltd.
4th Floor, Tower 'D', The Millenia
1 & 2 Murphy Road, Ulsoor
Bengaluru - 560 008
T: +91 (80) 40794000
D: +91 (80) 40794089
F: +91 (80) 40794222
M: +91 9741182000
binu.rajendran@in.pwc.com

pwc

Binu Rajendran

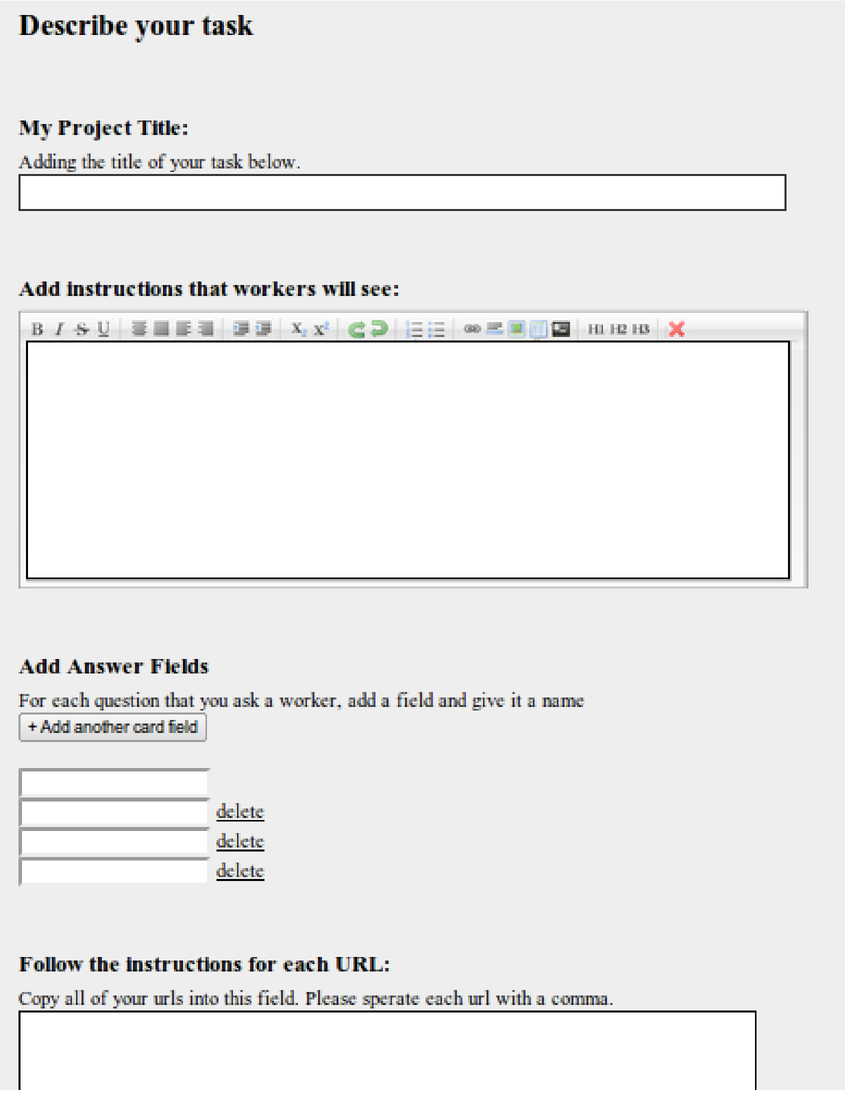
Digitized Contact Information

The figure below shows the table of digitized business card fields that has been presented to the participants in printed form.



Name	Title	Organization	Phone Number
Lovely Singh	Asst. Manager - HR	TECPRO SYSTEMS LIMITED	9380619225
Binu Rajendran	NOT FOUND	PricewaterhouseCoopers Pvt. Ltd.	918040794089
Yogesh Jain	Product Marketing Manager	Dell India R&D Centre	918028077000

The Task Specification Interface for Baseline Condition



Describe your task

My Project Title:

Adding the title of your task below.

Add instructions that workers will see:

Add Answer Fields

For each question that you ask a worker, add a field and give it a name

+ Add another card field

delete

delete

delete

Follow the instructions for each URL:

Copy all of your urls into this field. Please sperate each url with a comma.