# TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing

V. K. Chaithanya Manam
Purdue University
West Lafayette, Indiana, USA
vmanam@purdue.edu

Dwarakanath Jampani
Purdue University
West Lafayette, Indiana, USA
djampani@purdue.edu

Mariam Zaim
Purdue University
West Lafayette, Indiana, USA
mariam@zaeem.org

Meng-Han Wu
Purdue University
West Lafayette, Indiana, USA
wu784@purdue.edu

Alexander J. Quinn
Purdue University
West Lafayette, Indiana, USA
aq@purdue.edu

## ABSTRACT

Developing instructions for microtask crowd workers requires time to ensure consistent interpretations by crowd workers. Even with substantial effort, workers may still misinterpret the instructions due to ambiguous language and structure in the task design. Prior work demonstrated methods for facilitating iterative improvement with help from the requester. However, any participation by the requester reduces the time saved by delegating the work—and hence the utility of using crowdsourcing. We present TaskMate, a system for facilitating worker-led refinement of task instructions with minimal involvement by the requester. Small teams of workers search for ambiguities and vote on the interpretation they believe the requester intended. This paper describes the workflow, our implementation, and our preliminary evaluation.

## CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; *Human computer interaction (HCI)*.

## KEYWORDS

Crowdsourcing, task instructions, ambiguities, workflow

**ACM Reference Format:**
V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308560.3317081

## INTRODUCTION

Crowdsourcing platforms offer the opportunity to complete digital work with quality levels which are only achievable with human power, but with the on-demand availability and scale of cloud computing. An oft-cited challenge for workers is ambiguous instructions, which may lead to seemingly wrong results for the requesters and frustration for workers [11]. Ultimately, this squanders the time and efforts of both the requester and worker.

Instructions for crowdsourcing tasks are generated in a process that is similar but typically less rigorous than the process used by expert annotators in behavioral sciences [15, 18]. In crowdsourcing, experts independently examine a sample of data, generate instructions specially around possible ambiguities identified in the sample data,and finally discuss and improve the instructions based on the feedback from others [12]. Writing comprehensive instructions that cover all the variations of the data in a dataset would require close examination of all the data which is typically impractical in the crowdsourcing settings.

Additionally, there exist several other methods to improve the quality of the work in crowdsourcing platforms while posting a task to multiple workers and later collecting their responses [9] via voting, ranking, or clustering. However, if the instructions are unclear, ambiguous, or do not provide enough information about the task, these existing mechanisms can lead to the rejection of conscientious efforts of workers, reducing efficiency of the system.

Ambiguities typically arise while to a novice requester, or a set of requesters try to transpose an idea following their own mental models into system-executable instructions, they may fail to explain their task clearly because of lack of time or skill. Classifying the ambiguities in a systematic way helps researchers study and resolve them efficiently. WingIt [16] presented a classification of ambiguities in the task into three major categories: input, process and output. Input ambiguity consists of form parameters that the requester supplies. Process ambiguity contains the procedural information about how to do the task, as well as any context provided by the requester. Output ambiguity contains the form fields, associated labels, and any text the worker is supposed to provide the requester.

Existing research on improving the quality of instructions in crowdsourcing can be classified into two major categories: reactive mechanisms [16] amd proactive mechanisms [3, 6, 7]. In a reactive mechanism, the requester will post all the tasks and when ever a worker identifies a problem, he/she will ask the requester and get clarification. All the clarifications provided by the requester are added to the original instructions while other workers are working on them. In reactive mechanism, the requester should be available and reply to the workers quickly while workers are working on the tasks until the task is completed by all workers. However, this might not be practical for busy requesters. In any proactive mechanism,


*WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA*

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317081

the requester posts a few instances of tasks and requests workers to provide feedback. Based on the feedback received from the workers, the requester improves the instructions. This is an iterative process and continues until requester gets correct results. Modify ing instructions through iterative processes based on the workers' feedback consumes lot of time for the requester. In this paper, we propose a workflow that delegates the task refinement to workers and thereby reducing the requester time.

This paper presents TaskMate, a workflow for engaging crowd-based workers to improve the quality of instructions for a given task. TaskMate allows workers to collaborate with each other to solve a task, therefore minimizing the effort and time that the requester needs to invest. To achieve this goal, we propose a workflow that divides the overall problem into small, manageable, and verifiable steps.

The key contributions of this paper are as follows:

- A novel workflow that will help busy requesters to create a high quality instructions in short time by recruiting a set of reliable crowd workers for a given task.
- TaskMate, a system that implements this workflow for information search tasks, the most common task type on Mechanical Turk [10].

## RELATED WORK

Existing research on crowdsourcing emphasizes the impact of task instructions and design on the quality of the outcome. Gadiraju et al. identified the importance of task instructions in crowdsourcing and enhanced task clarity by modeling a solution with predictive features which would help enhance task clarity [5]. Wu et al. studied in detail how the quality of instructions will affect the quality of crowd work [19].

### Task authoring interface enhancements

Fantasktic [8] is a system designed to help novice requesters design better tasks through three task design techniques: a guided task specification interface, a preview interface, and a worker tutorial. A guided task specification interface provides guidelines and recommendations to the requester when he/she is creating a task. A preview interface showcases the task as will be seen by a worker to the requester creating the task. A worker tutorial is generated automatically based on the sample answers provided by the requester. The quality of responses from workers showed significant improvement when the requester created a task following the guidelines of the guided task interface. However, task previews and worker tutorials did not have any impact on the workers' response. Guided task specification does not address the issue of ambiguities in the instruction. Moreover, it is not feasible to create guided interfaces for all of the different kinds of tasks on crowd platforms.

### Worker-requester collaboration

Some recent approaches have demonstrated how workers and requesters can work together to efficiently improve an imperfect task design. Daemo [7] allows a requester to post few instances of the tasks to the workers and get their feedback. The requester can then improve the quality of his/her task based on workers' feedback. Accuracy of results were observed to be significantly higher for tasks that incorporated workers' feedback than for tasks that were originally posted.

Wingit is a reactive system in which a worker can edit the instructions or ask a question with a guessed answer to disambiguate their question when they encounter an ambiguity. The requester can then either trust the worker's guess and allow them to continue working on the task based on his/her guess—or the requester can review the worker's guess and reply with an acknowledgement containing the correct answer. In their evaluation, the workers demonstrated low proficiency identifying the ambiguities in the task instructions.

Structured labeling [13] is a tool that allows requesters to produce a clustered dataset where each cluster is labeled by a single person. Structured labeling enables requesters to independently learn about their task rather than through the worker-feedback.

Revolt [4] is a collaborative system developed to deal with ambiguous instructions in image-labeling tasks. Revolt produces structured labels created by the crowd. It allows multiple workers to label the images with specified instructions. In case of a conflict, workers relabel the images based on a description provided by other workers. The Revolt system is designed to improve the accuracy of results produced by workers only when there is a conflict. However, the Revolt system does not change the quality of instructions provided by the requester. Sprout [3] demonstrated a workflow that allows a requester to write minimum instructions for the task and then take the help of the crowd to create clear and detailed instructions. Taking feedback from workers and then improving the instructions for a task is extremely exhaustive in terms of both time and money for the requester.

With TaskMate, we show how workers acting without participation from the requester can identify and resolve ambiguities in the instructions, resulting in reduced response time while increasing overall operation efficiency.

### Comparing TaskMate with WingIt

WingIt proposed Q&A and Edit methods for workers to resolve ambiguities in the instructions. The requester can either trust a single worker's judgment on the ambiguities or requester is available until the task is completed, to reply back to the workers' questions quickly when there is any ambiguity in the instructions. Trusting the judgement of a single worker may not always turn out to be the correct option. Additionally, the requester may not be available to reply to workers' questions in an on-demand basis. To overcome these problems, TaskMate employs multiple workers' judgments to resolve ambiguities in the instructions while the requester can choose to stay unavailable to the workers at any time during the process. By using WingIt, it is possible to post all the instances of the task and ambiguities are resolved while workers are working on the task. However, by using TaskMate, the requester need to post few instances of the task for improving the quality instructions. WingIt can resolve instance specific ambiguities but TaskMate may not.

## TASKMATE

In this section, we present TaskMate, our system for improving the quality of instructions. The design decisions for TaskMate are based on previous work and the authors' hefty experience running crowdsourcing tasks.

TaskMate aims to identify ambiguities in the task, resolve them via majority voting, and generate clear instructions which clarifies all the ambiguities that are identified. TaskMate allows users to describe the task that they want to post on a crowdsourcing platform containing basic information. Users spending less than five minutes can receive high quality instructions within a short amount of time. We added an input decomposition at the beginning to identify all the list of possible ambiguities/problems following Soylent [2] and PlateMate [17]. The TaskMate framework consists of the following five stages: 1) Identify 2) Resolve 3) Merge 4) Verify 5) Select. The overall framework is composed of a multitude of iterative and parallel processes, as defined by Little et al. [14]. Figure 1 shows pseudo-code for the entire TaskMate workflow and Figure 2 explains the workflow with an example task.

## Identify

The goal of this stage is to detect all the problems with the task instructions. Workers are presented with the original task instruction, asked to work on it, and then submit the results. When a worker is working on the task and finds any problem with the task, he/she will describe the problem in the form of question along with a set of possible answers. From the set of possible answers, the worker will choose one answer and complete the task. In the example task shown in Figure 2, it is not clear whether the requester is looking for top 3 computer science programs with in USA or Global or Europe. The worker will ask a question "Which ranking are you looking for?" with a list of possible answers (USA/Global/Europe). By the end of this stage, we will have a list of all problems with the given instructions along with a list of possible solutions for each problem.

## Resolve

The goal of this stage is to guess correct solutions for each problem identified in the previous stage. Workers are presented with the original task ("List the top 3 Computer Science programs."), list of problems and the corresponding solutions to each problem. Workers were asked to guess what requester might be looking for from the list of possible solutions for each problem. Based on majority voting, we will get the correct solution for each problem. By the end of this stage, we will have a list of problems and solutions to each problem. In the example task, workers will vote for the two problems ("Which rankings are you looking for?", "Which level of college ranking are you looking at?"), based on majority voting, we will select the correct solution for each problem.

## Merge

The goal of this stage is to incorporate all the problems identified in the original task and create a new task instruction. We can also have the task along with Q&A as given in Q&A method in WingIt [16]. However, the task instructions along with Q&A increases the length of the task and the workers need to spend more time reading and understanding the task. Previous studies [19] have shown that uptake rate will be less for lengthy tasks. So, to make tasks more concise or comprehensive, we ask the workers to merge the Q&A's in the task and create a new task instructions (e.g.,"List top 3 Computer Science Undergraduate programs in the world."). In order to get better instructions, we will ask more than one worker to combine original task and Q&A's. By the end of this stage, we will have a set of new task instructions ({"List top 3 Computer Science Undergraduate programs in the world.", "List the top 3 Computer Science business programs."}).

## Verify

The goal of this stage is to verify whether all the ambiguities are correctly incorporated in to the new task instructions without changing the meaning of original task instruction. Workers are presented with original task, list of ambiguities along with solutions, and new task instructions that we got at the end of Merge stage. Workers were asked to verify either all the ambiguities are incorporated correctly or some of them are missing for each of the new task instructions. In the example task, workers will approve "List top 3 Computer Science Undergraduate programs in the world" and reject "List the top 3 Computer Science business programs". By the end of this stage, we will have a set of valid task instructions that incorporated all the problems in the original task instructions.

## Select

The goal of this stage is to select a single task instruction from a set of task instructions that are clear and concise. Workers are presented with all the verified new task instructions and were asked to select one of them. Based on majority voting, TaskMate will finalize one task instruction ("List top 3 Computer Science undergraduate programs in the world."). This finalized instruction will be given to the requester, and he/she can use this instruction to post all the instances of the tasks on crowdsourcing platform.

## System Usage

Our system is intended for requesters who aim to post a task on a crowdsourcing platform. With TaskMate, we hope to decrease the number of iterations that a requester has to go through in order to generate a clear set of instructions. Requester submits the task instructions to TaskMate, all the stages are triggered and posted to a crowdsourcing platform automatically. The requester can monitor the real-time progress of each of these stages. Once all five stages of TaskMate are complete, the 'new and improved' task is reported back to the user.

# STUDY DESIGN

We conducted experiments on Amazon's Mechanical Turk (AMT) to investigate how TaskMate improves the quality of instructions provided through the system. Here, we describe the tasks, participant recruitment, and payment associated with our study.

To test TaskMate, we created tasks based on the taxonomy of ambiguities proposed in WingIt [16] and described them in Table 1. In our study, we planted ambiguities in the thirty tasks that comprise of all the different types of ambiguities. We are interested in studying the effect of each stage in TaskMate and we do not differentiate workers based on gender, geographic location, prior experience, and other personal characteristics. We randomly assigned three workers to work on each stage. Workers who worked on any stage are prevented from working on other stages for a given task.

We paid $0.40 for identify, $0.10 for resolve, $0.50 for merge, $0.10 for verify, and $0.15 for select per HIT. A total of 192 distinct

```
def improve_instructions(draft):
   # DETECT issues (using AMT)
   issues_qa = identify(draft, num_workers)
   # Each worker produces ≥ 0, Q → A+ pairs,
   # with ≥ 1 possible answers for each question.
   # issues_qa is like:
   #   [
   #     { "What size?" : ["big", "small"]},
   #     { "What color?": ["red", "green"],
   #       "How big?" : [">1 mile diameter", "small", "tiny"]}, ...
   #   ]

   # VOTE on resolution of issues (using AMT)
   changes = resolve(draft, issues_qa, num_workers)
   # changes is like:
   # { "What size?": "big", "What color?": "red", ... }

   # INTEGRATE changes
   # Several workers do this, resulting in multiple
   # candidate versions
   candidates = merge(draft, changes, num_workers)
   # Candidates is like:
   # set(["Find the big, green ball.",
   #     "Get the big green (?) ball(s).",
   #     "Get the bright, candy-colored ball that is small.",
   #     "Get skldfjsdflkjl.l;llllll;;", ...)

   # CULL wrong candidates
   correct_candidates = verify(draft, changes, candidates, num_workers)
   # correct_candidates is like:
   # set(["Find the big, green ball.",
   #     "Get the big green (?) ball(s).",
   #     "Get the bright, candy-colored ball that is not small.", ...)

   # SELECT best candidate
   winner = select(draft, changes, correct_candidates, num_workers)

   return winner
```

**Figure 1: Pseudocode for the TaskMate workflow.**

workers have participated in our study. Our experimental tasks, problems associated with each task, and the improved task are given in Table 2.

## DISCUSSION & FUTURE WORK

Our study shows that workers are not able to detect ambiguities in the task significantly. This can be improved by training workers on identifying each of these ambiguities. In TaskMate, we observed that the task completion time is in the order of minutes. This can be due to the worker-demographics, worker-availability, the relative attractiveness of work, and the amount we pay. In order to reduce the completion time, we can either increase the amount per task or use retainer model [1]. We have not used any Natural Language Processing (NLP) techniques to validate the task provided by the

| Input | Process | Output |
|---|---|---|
| Entity (IE) | Steps (PS) | Entity (OE) |
| Syntax (IS) | Words (PO) | Exception (OX) |
| Wrong (IW) | Wrong (PW) | Units (OU) |
| Units (IU) | | Format (OF) |
| | | Precision (OP) |

**Table 1: Classification of ambiguity in task instructions [16].**

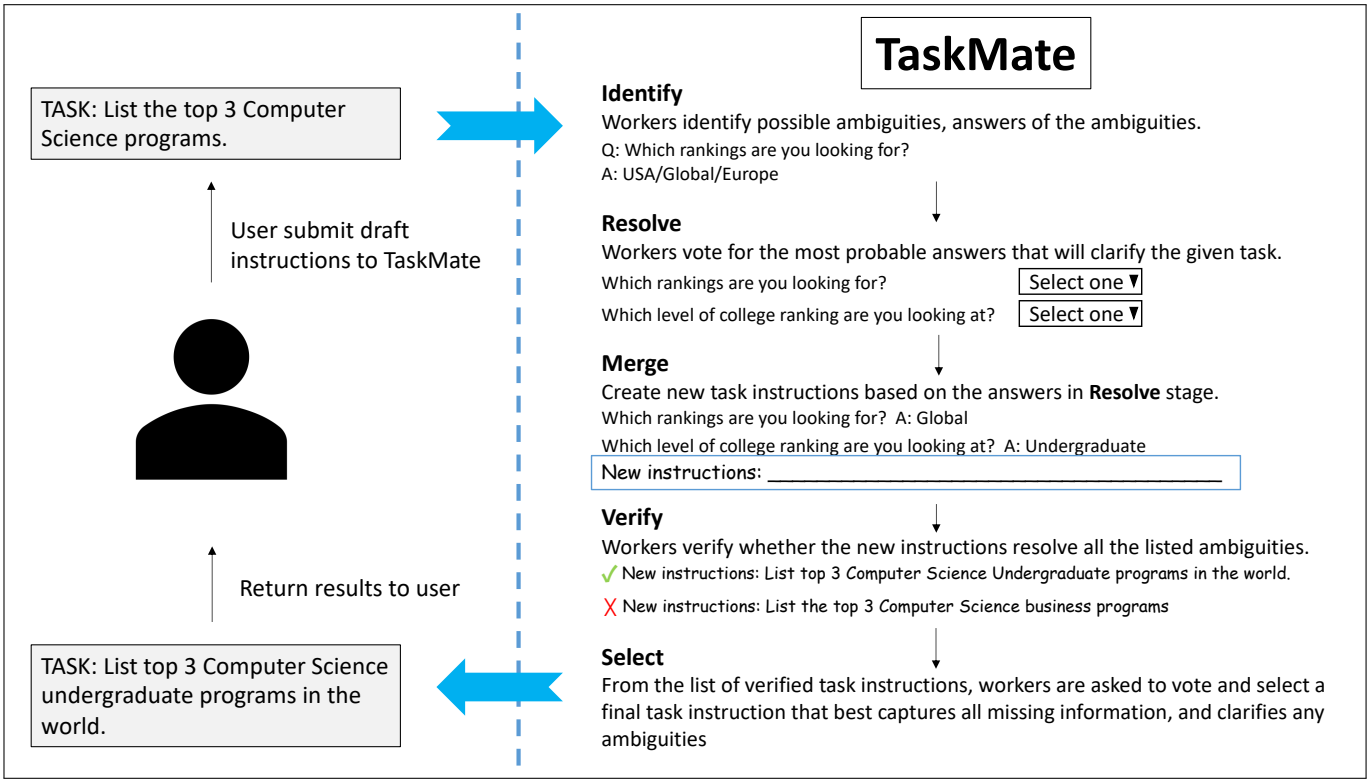


**Figure 2: TaskMate workflow with an example task**

requester. For future work, we could use NLP initially to identify some of the ambiguities by using word sense disambiguation. Furthermore, we can correct grammatical and syntactical errors in the original instructions automatically.

In the current system, workers need to identify all the ambiguities in the first stage itself. In the future, we can extend TaskMate as a cyclic process by asking the worker at each stage where any new ambiguities are found or not. If any new ambiguities are found, we can move back to the first stage and continue until all ambiguities are resolved. We've observed that workers are able to identify some of the problems in the task and improve the instructions based on the identified problems. We've also found that workers were able to identity ambiguities in the task and the list of possible solutions but they were not able to find the correct one that the requester wants in resolve stage. So, in the future, we can use the requester in resolve stage in order to achieve better results.

## CONCLUSION

In this paper, we present TaskMate: a mechanism to improve the quality of instructions in crowdsourcing. TaskMate consists of five stages: Identify, Resolve, Merge, Verify, and Select. In the Identify stage, each worker will identify the problems with the task instructions and a set of possible solutions by working on the task. In Resolve stage, all the ambiguities identified in the previous stage are resolved by choosing the best possible solution to the resolve the ambiguity. In Merge stage, workers will write a new task by merging the original instruction along with the list of ambiguities and clarifications. In Verify stage, workers will verify whether the new merged task instruction has incorporated all the ambiguities associated with the set of instructions generated from previous stages as output of the original instruction. In the final stage, Select, workers will vote for the best instruction which does not contain any form of vagueness or lack of clarity. Our results show that workers were able to come up with instructions that are clear and precise.

## REFERENCES


[1] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *UIST '11: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 33–42.

[2] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *UIST '10: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 313–322.

[3] Jonathan Bragg and Daniel S Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *UIST '18: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, ACM, New York, NY, USA, 165–176.

[4] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *CHI '17: Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2334–2346.

[5] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *HT '17: Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 5–14.

[6] Snehalkumar Gaikwad, Nalin Chhibber, Vibhor Sehgal, Alipta Ballav, Catherine Mullings, Ahmed Nasser, Angela Richmond-Fuller, Aaron Gilbee, Dilrukshi

Gamage, Mark Whiting, et al. 2017. Prototype Tasks: Improving Crowdsourcing Results through Rapid, Iterative Task Design. *arXiv preprint* arXiv:1707.05645 (2017).

[7] Snehalkumar (Neil) S. Gaikwad, Mark E. Whiting, et al. 2017. The Daemo Crowdsourcing Marketplace. In *CSCW '17: Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, New York, NY, USA, 1–4.

[8] Philipp Gutheim and Bjorn Hartmann. 2012. *Fantasktic: Improving Quality of Results for Novice Crowdsourcing Users*. Master's thesis. EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-112.html [Online; accessed September 1, 2018].

[9] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *HCOMP '10: Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA, 64–67.

[10] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829–840.

[11] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd Work. In *CSCW '13: Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 1301–1318.

[12] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *CHI '14: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3075–3084.

[13] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *CHI '14: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 3075–3084.

[14] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 68–76.

[15] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. *CAM Journal* 10, 2 (1998), 31–36.

[16] VK Chaithanya Manam and Alexander J Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions.. In *HCOMP '18: Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Palo Alto, CA, USA, 108–116.

[17] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: Crowdsourcing Nutritional Analysis from Food Photographs. In *UIST '11: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1–12.

[18] Cynthia Weston, Terry Gandell, Jacinthe Beauchamp, Lynn McAlpine, Carol Wiseman, and Cathy Beauchamp. 2001. Analyzing interview data: The development and evolution of a coding system. *Qualitative sociology* 24, 3 (2001), 381–400.

[19] Meng-Han Wu and Alexander J. Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *HCOMP '17: Proceedings of the 5th AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Palo Alto, CA, USA, 206–215.

| # | Original Task | Instance | Problem and solution | Improved Task (TaskMate) |
|---|---|---|---|---|
| 1 | Find the URL for the following professor in Electrical and Computer Engineering at Purdue | Raghunathan | Q: Which one? (IE)<br>A: Most senior professor; prefer Computer Engineering: Vijay Raghunathan<br>Q: Which URL? (OE)<br>A: Most specific home page | Find the personal URL for Vijay Raghunathan, a professor in Electrical and Computer Engineering at Purdue, on the universities information page. |
| 2 | Find the address of a movie theater close to the following airport in the city below | Chicago | Q: Which one? (IE)<br>A: Largest in the area<br>Q: How close? (OP)<br>A: Closest based on driving time. | Find the address of a closest movie theater to the airport Chicago O'Hare. |
| 3 | When was the following movie released? | Harry Potter | Q: Which one in the series? (IE)<br>A: Most recent<br>Q: Date or just year? (OF)<br>A: Date | What year was Harry Potter and The Sorcerer's Stone released in movie theaters in the United States? |
| 4 | Find the weight of the any smartphone with the following brand that can be purchased on BestBuy.com for $30000 to $49900. | Moto G | Q: Which model? (IE)<br>A: Most recent<br>Q: Do you mean "$300.00 to $499.00"? (PS)<br>A: Yes.<br>Q: Weight in g or oz? (OU)<br>A: g | Go to BestBuy.com and do a search for Moto G brand smartphone (cell phone only) from the current year that is priced between $300-$499. From the detail listing page, copy the weight that is given. |
| 5 | Find the cost of attending this university for undergraduates. | MIT | Q: domestic or out-of-state? (OE)<br>A: domestic<br>Q: Total or individual components (tuition, fees, books, etc.)? (OP)<br>A: Sum of tuition and fees only | Find the cost, including the tuition and fees, of attending the School of Architecture and Planning at MIT without a full scholarship or in-state residency discounts while living off campus. |
| 6 | Find the i-10 index and h-index for the following professors. Go to Microsoft Scholar and search for the person's name. | Michael Bernstein (Stanford) | Q: What are i-10 and h-index? (PO)<br>A: Use Google Scholar<br>Q: All or last 5 years? (OE)<br>A: All | The i-10 index refers to the number of papers with 10 or more citations. The h-index is an author-level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar. Search Google Scholar for the i10 index and h-index since 2014 for the following professors: Michael Bernstein (Stanford). |
| 7 | Use search engine to find the home page URL for the following organization | WMT | Q: Are these stock market names of the companies Walmart (WMT)? (IE)<br>A: Yes | Use search engine Google to find the home page URL for the following organization: Walmart (WMT). |
| 8 | Go to the Google Store (https://store.google) and search for the device below. Enter the price of the least expensive option. | Samsung Pixel | Q: Did you mean "Google Pixel"? (IW)<br>A: Yes.<br>Q: What model? (IE)<br>A: Google Pixel 3<br>Q: Do you mean "store.google.com"? (PW)<br>A: Yes. | Please, enter the Google Store website (https://store.google.com) and look for the price of the following device: Google Pixel 3. Write down the cheapest option you can find. |

| # | Original Task | Instance | Problem and solution | Improved task (TaskMate) |
|---|---|---|---|---|
| 9 | Find the maximum extendable ROM card size for the following devices | Moto G4 Plus | Q: Do you mean extendable memory card? (IW)<br>A: Yes<br>Q: GB, MB, inches, cm, mm? (OU)<br>A: GB | Find the maximum extendable memory card size for the Moto G4 Plus. |
| 10 | Find a city that has an average temperture between 20 to 30 during May in the following state | Indiana | Q: Do you mean temperature? (IS)<br>A: Yes<br>Q: Celsius or Fahrenheit? (IU)<br>A: Fahrenheit<br>Q: What if no city in that state fits the criteria? (OX)<br>A: Enter "none" | Please search in google and find and list all the cities in Indiana state, USA where More then 5000 residents reside and average temperature between 20 to 30 during May. Temperature must be in Fahrenheit. |
| 11 | Search for flight information from Chicgo to the following city in December with maximum price 700 | Beijing, China | Q: Q: Do you mean Chicago? (IS)<br>A: Yes<br>Q: Cost in US dollars or Euros (IU)<br>A: US dollars | Use Expedia to find the least expensive round-trip flight from Chicago, IL to Beijing, China during December. Include the name of the airline and the dates for each flight (maximum price $700 USD) |
| 12 | Find the URL for graduate program admission process in the following universities | MIT | Q: Where can I find it? What should I search for? (PS)<br>A: Search in the admission page | Find the URL for the graduate Political Science admissions page at MIT for the year 2019. |
| 13 | Find three mobile phones from below manufacture that cost less than $300 | Apple | Q: Where can I find it? What should I search for? (PS)<br>A: Search in amazon or any shopping website | In US dollars, find 3 mobile phones (old or new) from Apple that cost less than $300. |
| 14 | Find the impact factor for the following journals | Transaction on Human Computation | Q: What is impact factor? (PO)<br>A: Search online and find out | Find the impact factor of the academic journal "Transaction on Human Computation" on students, based on the number of times the academic journal has been cited, using journal citation reports, given that an average article is cited 100 times, as measured by Thomson Reuters. |
| 15 | Go to amazon.com and search for the best price of the following car, base model, year 2017 | Honda CRV | Q: Old or New? (IE)<br>A: New.<br>Q: Do you mean cars.com or some car website instead of amazon.com? (PW)<br>A: Yes | Go to Amazon.com and search for a GPS system for a 2017 Honda CRV. Find the best price. |
| 16 | Go to cars.com and find the best price for the following mobile phones | iPhone 6s | Q: Old or New? (IE)<br>A: New.<br>Q: Do you mean Amazon.com or some shopping website instead of Cars.com? (PW)<br>A: Yes | Go to the Apple website and find the best price for the iPhone 6s. |

| # | Original Task | Instance | Problem and solution | Improved task (TaskMate) |
|---|---|---|---|---|
| 17 | Go to bestbuy.com and then search for the following key words. Send the top 3 links that you get | Buy a car | Q: Do you mean google.com instead of bestbuy.com? (PW)<br>A: Yes | Go to bestbuy.com. If it asks for a country, use USA. Search for the following keyword: Buy a car. Enter the first 3 links that you get. |
| 18 | Find the URL and the fee for undergraduate engineering program at the following Universities | University of Texas, Austin | Q: Is it for domestic or international? (OE)<br>A: International<br>Q: Is it for Semester or Yearly or 4 years fee? (OE)<br>A: Yearly<br>Q: Total or individual components (tuition, fees, books, etc.)? (OP)<br>A: tuition+fees only | Please find the URL and the total tuition cost for the all undergraduate engineering degree programs at the college of engineering at the University of Texas' Austin location only. |
| 19 | Find the city fuel efficiency of the following Indian Car | Mahindra XUV | Q: Manual or Automatic? (IE)<br>A: Manual<br>Q: Is mileage in MPG or Kilometers per liter? (OU)<br>A: Kilometers per liter | Find the city fuel efficiency in kmpl of the following Indian Car: Mahindra XUV (petrol variant). |
| 20 | Find the average temperature of the following city in Decmber | Chicago, USA | Q: Do you mean December? (IS)<br>A: Yes<br>Q: Is temperature in Celsius or Fahrenheit. (OU)<br>A: Fahrenheit | Find the average high and low temperatures in Chicago, USA during the month of December. Also provide precipitation averages if listed. |
| 21 | Find the fuel tank capacity of the following vehicle | Chevrolet Cruze | Q: Gallon or liters? (OU)<br>A: Gallons | Find the fuel tank capacity in gallons of a Chevrolet Cruze produced after 2015 with a link to where you found the information. |
| 22 | Find the date when the 1st model of following phone was launched | Google Nexus | Q: Date format? (OF)<br>A: MM/DD/YYYY | Find the date when the 1st model of the Nexus One was first available for purchase. |
| 23 | Find the launch date for the next model of the following mobile phone | Nexus Pixel | Q: What if new phone is not announced? (OX)<br>A: Write ""Not announced""<br>Q: Date format? (OF)<br>A: MM/DD/YYYY | Find the launch date for Google Pixel the next model. |
| 24 | Find the recipe of the following food | Fried Rice | Q: Which fried Rice? (IE)<br>A: Chicken Fried Rice<br>Q: URL or entire recipe? (IW)<br>A: URL<br>Q: How do i find the recipe? Do I have to search online or ask someone and enter it here (PS)<br>A: Search online | Use google and find one recipe for Vegetable Fried Rice using boiled white rice (It should only take 10 minutes to cook) and that can be used in a fast food restaurant business. |
| 25 | Find three URLs for the best used cars that are priced maximum 7000 dollars with in 100 miles from ORD Chicago | Honda Accord 2007 | Q: Which model? Base model or high end. (IE)<br>A: Base model<br>Q: What does ""best"" mean? (PO)<br>A: Lowest Price | Find three URLs for the best used car below that are priced maximum 7000 dollars that allows loan payment within 100 miles from ORD Chicago: Honda Accord 2007. |

| # | Original Task | Instance | Problem and solution | Improved task (TaskMate) |
|---|---|---|---|---|
| 26 | Find the Official twitter account for the following celebrities | Kristen Stewart | Q: What if no account matches given criteria? (OX)<br>A: enter "none" | Find the official, verified twitter account for Kristen Stewart, who is a famous worldwide celebrity. Look for the verified check mark that lets you know you found the official account for Kristen Stewart. When you have found the official account, copy and paste the URL of the twitter account into the field below. |
| 27 | Find the price of the following car in Chicago | Honda Accrd | Q: Coupe or Sedan? (IE)<br>A: Sedan<br>Q: New or Used? (IE)<br>A: New<br>Q: How to find the price? Should I search online or call the car company? (PS)<br>A: Search online<br>Q: Do you mean Honda Accord? (IS)<br>A: Yes | Find the average price of the car below: New 2019 Honda Accord, Chicago IL. |
| 28 | Find the lowest price for the device in the US or Canada. | iPhone 7TT | Q: Did you mean "iPhone 7 Plus"? (IW)<br>A: Yes.<br>Q: How can I search by country? (PS)<br>A: Search amazon.com and amazon.ca. Then take minimum of the two.<br>Q: USD or CAN? (OU)<br>A: USD | Find the lowest price for the new device iPhone 7 plus in the US or Canada the payment method can be made through debit. |
| 29 | Find names and URLs of the professors who's area of research is HCI in the following Universities | CMU | Q: How do I find the names? Via online search or contacting someone in the university? (PS)<br>A: Search online<br>Q: Do you need one or more than one Prof. name? (OE)<br>A: only one | Look for the professor personal website URLs in charge of research about Human-Computer interaction in the Carnegie Mellon University. One professor per task. |
| 30 | Find the store name and the address near 465 Northwestern Ave, West Lafayette, IN 47907 where we can buy the following plants | Spider Plant | Q: Near means ? Is it 3 miles or 30 miles? (IE)<br>A: Within 10 miles radius<br>Q: What if no account that matches given criteria? (OX)<br>A: enter "none" | Find the closest store with plants to 465 Northwestern Ave, West Lafayette, IN 47907. |

**Table 2: Experiment data**