A Capability Requirements Approach for Predicting Worker Performance in Crowdsourcing

Umair ul Hassan
Digital Enterprise Research Institute
National University of Ireland
Galway, Ireland
umair.ul.hassan@deri.org

Edward Curry
Digital Enterprise Research Institute
National University of Ireland
Galway, Ireland
ed.curry@deri.org

***Abstract*—Assigning heterogeneous tasks to workers is an important challenge of crowdsourcing platforms. Current approaches to task assignment have primarily focused on content-based approaches, qualifications, or work history. We propose an alternative and complementary approach that focuses on what capabilities workers employ to perform tasks. First, we model various tasks according to the human capabilities required to perform them. Second, we capture the capability traces of the crowd workers performance on existing tasks. Third, we predict performance of workers on new tasks to make task routing decisions, with the help of capability traces. We evaluate the effectiveness of our approach on three different tasks including fact verification, image comparison, and information extraction. The results demonstrate that we can predict worker's performance based on worker capabilities. We also highlight limitations and extensions of the proposed approach.**

Keywords*—*microtask, taxonomy, crowdsourcing, performance

I. Introduction

Crowdsourcing has emerged as a powerful paradigm for solving complex problems at large scale with the help of people [1]. Crowdsourcing has been fuelled by the rapid development in web technologies that facilitate contributions from millions of online users [2]. Practitioners in the crowdsourcing community have grouped web-based platforms into four groups[1], as shown in Table I. *Microtask* platforms are used for outsourcing small tasks that can be completed within minutes by an average online user. *Macrotask* platforms are designed to manage large projects requiring weeks or months of effort, possibly with multiple skilled people. *Crowdfunding* platforms allow people to gather money from the crowd, for a specific project or cause, generally against some kind of recognition for the contribution. *Contest* platforms allow organizations to solicit best solutions, designs, or ideas by offering rewards for winning entries. The scope of this paper is limited to the microtask platforms, therefore the terms microtask and task will be used interchangeably. The methods described here may also be applicable to macrotask platforms.

TABLE I. Four groups of crowdsourcing platforms

Group	Platform
Microtask	MTurk, Crowd Flower, Click Worker, Mobileworks
Macrotask	Innocentive, Quriky, Apache Foundation
Crowdfunding	Kickstaters, Indiegogo, Seedups, crowdrise
Contest	99designs, crowdSPRING, Kaggle, innocentive

Microtask platforms like Amazon Mechanical Turk[2] and CrowdFlower[3] enable access to large numbers of readily available online workers. On one side, this enables researchers to quickly collect large amounts to human labeled datasets at low costs. For example, human judgements of relevance between search engine queries and results [2]. On the other side, software programmers can outsource simple data processing tasks such as verification of matching between two database entities [2]. The growing number of crowdsourcing applications and platforms poses challenges in terms of interoperability, workforce management, and quality assurance.

Generally crowd workers tend to select large numbers of microtasks, in order to maximize their earnings [3]. This behavior can result in low quality of work due to the mismatch between worker capabilities and task requirements. Existing methods for quality control either user ground truth data or expert review. These methods generally occur after the completion of tasks, therefore requiring re-submission of low quality tasks. After-the-fact quality control is expensive in terms of both time and cost. The problem is especially difficult for subjective or open ended tasks. Microtask platforms also provide qualifications for filtering workers, such as Amazon Mechanical Turk. Qualification is based on metrics such as percentage of rewarded tasks and number of completed tasks. Nonetheless, there can be high variability in the quality of work due to the differences of skills, knowledge, abilities, and other characteristics of the crowd workers [4].

Recently there has been an increase in the efforts directed towards investigating better ways of assigning tasks to appropriate workers. The objective of *task routing* is to improve the quality through assignment, before a worker performs the task. A variety of approaches have been proposed to address *task routing* problem in crowdsourcing. An approaches matches social network profile of worker with the task contents for making routing decisions [5]. Interaction tracing approaches aim to predict the future performance of a worker by capturing

[1] http://dailycrowdsource.com/taxonomy

[2] http://www.mturk.com

[3] http://www.crowdflower.com

978-1-936968-92-3 © 2013 ICST
DOI 10.4108/icst.collaboratecom.2013.254181

log of events within the task interface [6], [7]. Although appropriate for their particular contexts these approaches are difficult to generalize for environments with heterogeneous tasks and workers. In this work, we aim to develop a general theory of human task performance in heterogeneous crowdsourcing environments. Specifically, we are interested in heterogeneity due to the knowledge, skills, or abilities required from crowd workers.

We present a capability-centric approach for analyzing and predicting human performance on heterogeneous tasks in crowdsourcing platforms. Workers employ specific capabilities to perform tasks in a particular context, which may be useful for predicting worker performance in another context. For example, if a worker has performed well on the task of tagging images then she might also perform well on the task of image comparison, as compared to the task of text translation. Conversely, she may require less training for visual skills as compared to linguistic skills. Using this information about a worker's capabilities the crowdsourcing platform can make intelligent routing decisions.

In this paper, we examine whether the worker's capabilities, as they are been employed, can be used to predict the quality of worker's responses on different tasks in the future. The specific contributions of this paper are as follows:

- A capability requirements approach for modeling heterogeneous tasks in a crowdsourcing platform. A taxonomy of worker capabilities for microtasks, based on the capability requirements analysis of two popular crowdsourcing platforms.
- A probabilistic model for profiling worker capabilities and for predicting worker performance on new tasks. The model is further utilized for making task assignment decisions.
- Evaluation of the proposed approach for three different task contexts. The results demonstrate the applicability of proposed approach for predicting worker performance on heterogeneous tasks.

The rest of this paper is organized as follows. Section II provides an research background for task analysis and worker modeling. Section III describes the taxonomic approach for associating worker capabilities with microtasks. Section IV details the probabilistic model for profiling workers capabilities based on their performance of tasks. Section V reports the results of the experiment performed for the evaluation of proposed approach. Section VI provides an overview of related research work. Section VII concludes the paper with the summary of results, limitations, and future work.

II. Background

Researchers in behavioral sciences have been actively working on human task performance for decades. They have extensively defined and categorized several human capabilities through empirical and theoretical methods [8]. For instance, Fleishman studied the effects of cognitive, psycho-motor, and physical abilities of pilots on their task performance. The objective of these categorizations was to support researchers and practitioners in developing systems and theories, that utilize various conditions of training, performance, and transfer of learning. We share the same vision in the context of crowdsourcing research. We endeavor to apply the vast body of knowledge from behavioral sciences to the study of human performance in crowdsourcing platforms, specifically for microtask platforms. To the best of our knowledge, this is the first attempt to bring these fields together.

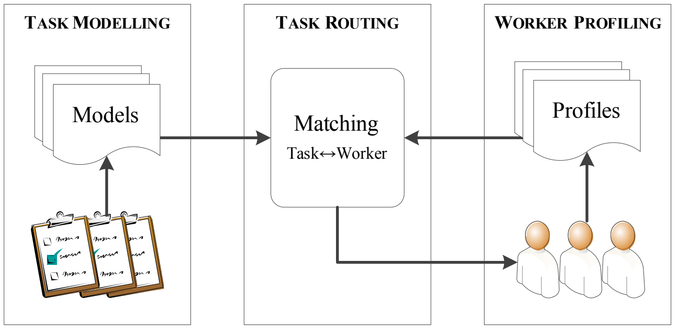


Figure 1. Overview of the *task routing* process in a heterogeneous crowdsourcing platform. Multiple types of task and workers are represented with multiple models and profiles, respectively.

A. Task Routing

Routing of tasks to appropriate workers based on their suitability is a well known problem in business process management, also know as the *personnel selection* problem in organizational psychology. Comparatively the task routing problem has been studied less in the crowdsourcing and human computation research [9]–[11]. Nonetheless, there are three major aspects of task routing in crowdsourcing plafrorms, as shown in Fig. 1.

1) Task modeling: This is the process of identifying characteristics of a task that can help in determining the best workers for a task. A task can be modeled either in terms of the required actions, activities, and operations [12] or the human abilities needed to perform the task [13]. For instance, an audio transcription task involves two activities; listening to audio and writing the text. The same task requires oral comprehension and written expression capabilities from the workers. In this paper, we follow the latter approach where a task in modeled in terms of the capability requirements.

2) Worker profiling: This process is concerned with identifying and gathering information about worker's characteristics to support routing decisions. Worker profile information can be generated by observing worker's performance on test tasks [14], observing worker's interaction patterns [6], or retrieving worker's information from external documents[15]. In this paper, we follow a task performance based approach for profiling a worker's capabilities.

3) Task routing: This process is designed to decide which task is appropriate for which worker, whilst considering various constraints. The actual choice of routing algorithm is dictated by the specific of task models and worker profiles. Techniques such as graph matching, fuzzy matching, semantic similarity, and machine learning have be used for routing [16]. In this

TABLE II. COMMON TASKS FOUND IN TWO POPULAR PLATFORMS: AMAZON MECHANICAL TURK (AMT) AND CROWDFLOWER (CFL)

Task	**Description**	**AMT**	**CFL**
Translation	Translation of content form one language to another language	✓	✓
Transcription	Conversion of audio content in audio or video files to text	✓	✓
Digitization	Conversion of textual content in images or videos to text	✓	✓
Fact Verification	Verification of given data or information	✓	✓
Content Creation	Creation of textual, audio, or video content	✓	✓
Categorization	Classification of items into different categories	✓	✓
Item Comparison	Comparison on images, audio, or video content	✓	✓
Content Tagging	Tagging of textual, audio, or video content with keywords	✓	✓
Ranking	Ranking items according to given criteria	✓	✓
Web Research	Information collection from Web pages or search engines	✓	✓
Information Extraction	Extraction of specific data or information from content	✓	✓

paper, we propose a probabilistic approach for predicting performance of workers on future tasks and making routing decisions.

B. Worker Capability

Since the tasks are modeled according to the human capabilities, an appropriate definition of capabilities is needed. We base the definitions according to the previous research work on Occupational Information Systems [17] and Human Capability Frameworks [18].

Capability is defined as the ability of humans to do things in terms of both the capacity and the opportunity. There are four types of capabilities described in the literature [17], [18].

1) *Knowledge* is the collection of discrete but related facts and information about a particular domain that is acquired through formal education or accumulated through experience. Alternatively, it is the body of information applied directly to the performance of a specific task.
2) *Skill* is the proficiency needed to perform a task that is developed through practice.
3) *Ability* is the capacity to engage in a specific behavior and it is considered to be stable over time.
4) *Other characteristics* can include miscellaneous factors such as motivation, attitude, and social relations.

Opportunity refers to the option available to a person to use his/her capabilities for performing a task, possibly in return of a reward. *Reward* is the benefit gained from the use of capabilities, while performing tasks, when given the opportunity [18]. *Matching* covers the process of comparing capabilities and opportunities for the finding the suitability between worker and task [18].

III. CAPABILITIES TAXONOMY FOR MICROTASKS

This section describes the capability requirements approach for modeling tasks in microtask platforms. The capability requirements approach focuses on what the human workers are able to do. Therefore, a task is described in terms of the capabilities employed by a worker whilst performing it. We identified common tasks found in two microtask platforms, to model the capability requirements of tasks. Table II highlights the common tasks along with their description. Each of these tasks is then compared and contrasted with the a list of human capabilities.

The list of possible human capabilities can be extensive including domain knowledge, skills, and abilities. Therefore, a taxonomy of human capabilities that are relevant the crowdsourcing and microtask platforms is needed. The general objective of such taxonomy is to help in standardizing the methods of studying human performance and in generalizing the methods to new tasks. Specifically, we are interested in the use of capabilities taxonomy for establishing similarities between heterogeneous tasks. Such taxonomy can further help is defining groups of tasks with similar capability requirements and finding the most suitable match of people to tasks.

We adapt a well defined taxonomy of basic human abilities, developed by Fleishman et al. [13], for the capabilities needed for common tasks in microtask platforms. Fleishman's taxonomy groups 52 relatively enduring human abilities into three broad areas; cognitive, psycho-motor, and physical. We have identified the 8 abilities that are relevant to the microtask platforms, by reviewing descriptions of common tasks. The following list provides definitions of the identified abilities:

- *Comprehension* (C): The ability to understand the meaning or importance of something
- *Bilingualism* (B): The ability to speak and understand two languages
- *Writing* (W): The ability or capacity to write text in a given language
- *Comparison* (M): The ability or capacity to compare things based on some criteria
- *Judgment* (J): The act or process of judging; the formation of an opinion after consideration
- *Perception* (P): The ability or capacity to perceive items visually or phonetically
- *Identification* (I): The process of recognizing something

TABLE III. THE MAPPING OF CAPABILITIES TAXONOMY AND COMMON MICROTASKS IN TERMS OF CAPABILITY REQUIREMENTS

Microtask	**C**	**B**	**W**	**M**	**J**	**P**	**I**	**R**
Translation	✓	✓	✓					
Trannscription	✓		✓			✓		
Digitization	✓		✓					
Fact Verification	✓				✓		✓	
Content Creation			✓					
Categorization				✓	✓			✓
Item Comparison				✓		✓	✓	
Content Tagging					✓	✓	✓	
Ranking				✓	✓			✓
Web Research	✓						✓	✓
Info. Extraction	✓						✓	✓

- *Reasoning* (R): The ability to draw conclusions from facts, evidence, relationships, etc.

Table III shows the mapping matrix of identified human abilities versus the common tasks. The capabilities defined in our taxonomy are either cognitive or perceptual, as most of the microtasks in current crowdsourcing platforms involve these kind of abilities. A detailed comparison of our taxonomy and Fleishman's taxonomy is provided in Table V for reference.

IV. CAPABILITY TRACING

Thus far we have described common tasks in microtask platforms, as well as the human capabilities required to perform those tasks. Given the capability requirements of tasks, there is a need to model the performance of tasks by worker as they employ the capabilities. Analysis of the relationship between the capabilities and the outcome of task can help measure and predict the performance of workers across heterogeneous tasks. This section describes a probabilistic approach for modeling capabilities of a worker as they as employ for performing tasks. The approach is inspired by the technique called *knowledge tracing* [19]. Knowledge tracing is used in tutoring systems to estimate the probability of a student knowing a skill given the observations of her attempting to utilize the skill during test tasks. First, we discuss the *capability tracing* model and its parameters. Then, we demonstrate its use for predicting performance of worker based on the capabilities associated with tasks.

Similar to knowledge tracing, we model the capability as a latent binary variable. A worker's overall capabilities is assumed to be a set of these latent binary variables. Each latent variable is updated based on the correctness of the observed evidence of a capability employed by the worker. The evidence is based on the test tasks associated with the capability in question. The model further assumes that the observation of evidence is also a binary variable, which indicates whether the worker's response to a task is correct or incorrect.

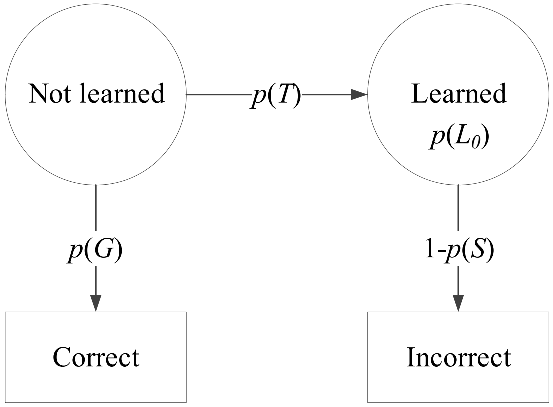


Figure 2. The probabilistic network model of a worker learning to apply a capability while performing tasks. The two states of binary variable for the learned capability of worker are represented with circles. The two possible observations of capability being applied correctly are shown as boxes.

Fig. 2 shows the probabilistic network representation of the two states of worker's capability, as well as their relationships with the two states of a worker's observed responses. The binary variable representing a capability of worker can transition from ”Not learned” state to ”Learned” state, however vice versa is not possible. In the ”Not learned” state the worker can correctly apply the capability by *guess*. The worker can incorrectly apply the capability even when in the ”Learned” state, which is called the *slip*. There are four parameters in the capability tracing model for each capability of a worker, as described below:

- $p(L_0)$: The initial probability of a worker being capable of successfully employing the capability.
- $p(T)$: The probability of worker moving from not learned to learned state when given the opportunity to employ the capability.
- $p(G)$: The probability of guess i.e. the evidence of task performance in correct when the worker have not learned to employ the capability.
- $p(S)$: The probability of slip i.e. the evidence of task performance is incorrect even when the worker has learned to employ the capability.

Given the parameters of model for all capabilities, the worker's profile of capabilities is updated after each observation of worker performing a task. If a worker employs the capability correctly then $p(L_n|O_n^+)$, the conditional probability of the worker learning to employ the capability, is updated as follows

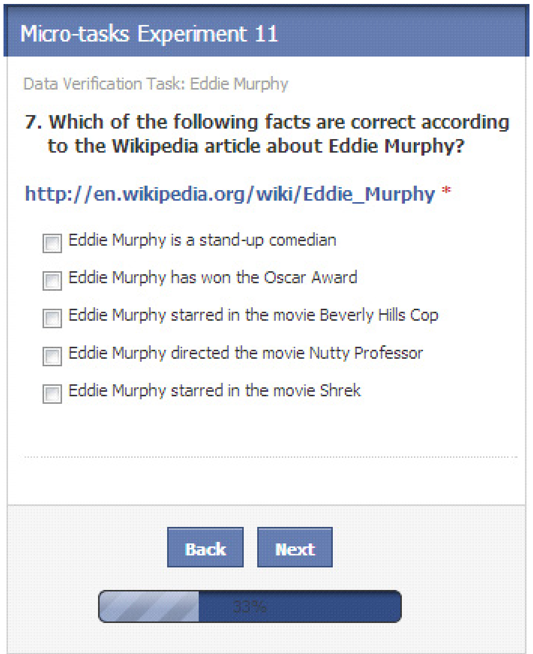
$$\frac{p(L_{n-1}) \times (1 - p(S))}{p(L_{n-1}) \times (1 - p(S)) + (1 - p(L_{n-1})) \times p(G)}$$

If the worker employs the capability incorrectly then $p(L_n|O_n^-)$, the conditional probability, is updated as follows

$$\frac{p(L_{n-1}) \times p(S)}{p(L_{n-1}) \times p(S) + (1 - p(L_{n-1})) \times (1 - p(G))}$$

Using the above two conditional probabilities, the probability of a worker learning to employ the capability is updated using the following equation:

$$p(L_n) = p(L_n|O_n) + (1 - p(L_n|O_n) \times p(T) \qquad (1)$$



Micro-tasks Experiment 11

Data Verification Task: Eddie Murphy

7. Which of the following facts are correct according to the Wikipedia article about Eddie Murphy?

http://en.wikipedia.org/wiki/Eddie_Murphy *

□ Eddie Murphy is a stand-up comedian

□ Eddie Murphy has won the Oscar Award

□ Eddie Murphy starred in the movie Beverly Hills Cop

□ Eddie Murphy directed the movie Nutty Professor

□ Eddie Murphy starred in the movie Shrek

Figure 3. An example of the Fact Verification task that requires workers to verify some facts about actor Eddie Murphy according to his Wikipedia article.

Note that $p(L_n) = p(L_{n-1})$ for a capability that is not employed for the observed task. Given $p(L_n)$ for the capability of worker, the future performance of worker for a new task is predicted using the following equation:

$$p(correct_{n+1}) = p(L_n) \times (1 - p(S)) + (1 - p(L_n)) \times p(G) \quad (2)$$

Since a task can require more that one capabilities then the performance of worker is predicted based on the set of capabilities C, as follows:

$$p(correct_{n+1}) = \frac{1}{|C|} \sum_{c \in C} p(correct_{n+1})^c \quad (3)$$

V. Experiment

A. Experiment Settings

In order to evaluate the capability tracing model, we created an ad-hoc collection of tasks. The collection is based on 10 Wikipedia articles about popular actors and actresses. The article were manually selected to include artists from US and India, to cover the knowledge of international films. The task collection consists of three different types of tasks, as described below:

- *Fact Verification Task* involves verification of 5 candidate facts about an artist by comparing them with the corresponding Wikipedia article, as shown in Fig. 3. The candidate facts were generated from each Wikipedia article, which included both correct and incorrect facts.
- *Image Comparison Task* requires workers to compare the image of an artist on corresponding Wikipedia article with 3 candidate images, to judge if the person is same. The candidate images were collected from the Web to ensure similarity with Wikipedia image.
- *Information Extraction Task* asks workers to extract certain entities from Wikipedia article i.e. the list of all cities where an artist has lived.

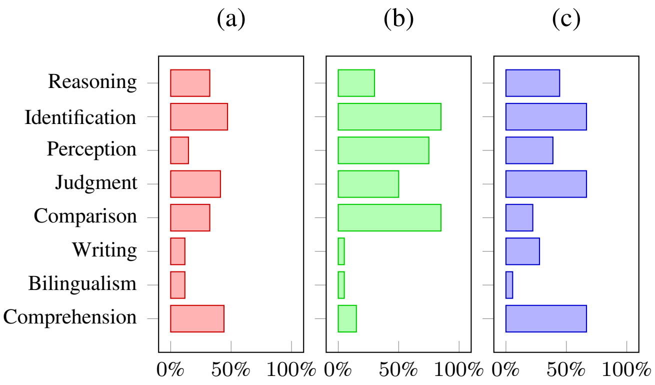


Figure 4. The distribution of crowd opinion about capabilities required for performing three types of microtasks: fact verification task (a), image comparison task (b), and information extraction task (c). In this case the crowd include workers who completed all three types of tasks.

The ground truth for each set of tasks was created manually by expert editors. The ground truth was used both for training the capability tracing model and evaluation of correctness of worker responses.

Crowdsourcing was performed through a purpose built web application. Crowd workers were hired by asking students to participate through a university wide email and by asking workers on a microtask platform[4] through a proxy task. A total of 34 people participated in the experiment. The workers were asked to complete all three types of tasks. Although all workers completed Fact Verification tasks, only 20 workers complete both Fact Verification and Image Comparison tasks, and 17 workers completed all three types of tasks. The workers were further asked to specify if they were knowledgeable about films from US and/or India. Out of all workers, 32% were knowledgeable about films from both US and India, 41% about films from US only, and 24% about films from India only.

In the following, we present two parts of the experiment focused on solicitation of capability requirements of tasks and evaluation of the capability tracing model. The underlying objective of both parts is to demonstrate the utility of proposed capability-centric approach for modeling tasks and worker performance in crowdsourcing platforms.

B. Task Capabilities

Given that the capability requirements approach maps each task with the appropriate human capabilities, we asked workers to choose the capabilities from the capabilities taxonomy that are important for each task. Fig. 4 shows the distribution of crowd opinion about the capability requirements of tasks, among the set the workers who completed all the three types of

[4] http://www.shorttask.com

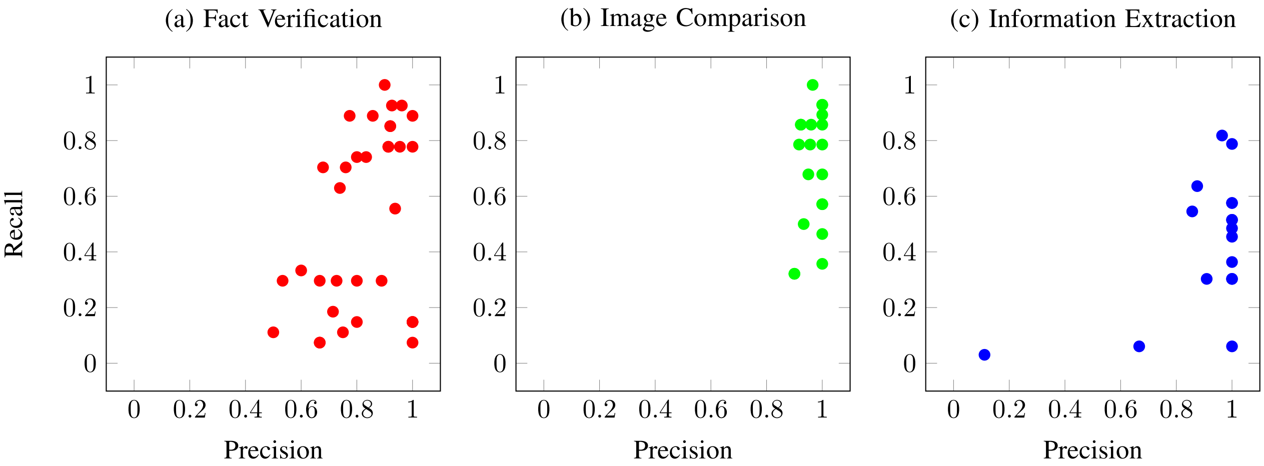


Figure 5. The performance of crowd workers for three types of tasks. Plot (a) shows 34 workers for *fact verification* task, plot (b) shows 20 workers for *image comparison* task, and plot (c) shows 17 workers for *information extraction* task.

tasks. As can be seen, more than majority of workers believed that *identification* capability is essential for all three types of tasks. Majority of workers also agreed that the *judgment* and *comprehension* capabilities are important for the Fact Verification task. In the case of the Image Comparison task most workers specified that *comparison* and *perception* are important as well. There is general consensus between workers that *comprehension*, *judgment* and *reasoning* capabilities are also useful for Information Extraction tasks. We selected top-3 capabilities for each type of task for building capability tracing models.

C. Performance Prediction

In this section, we present the results of experiments designed to illustrate the utility of capability tracing model. In order to calculate the correctness of a worker's response to a task (i.e. selected facts, images, and/or cities) we compared it with the ground truth, assuming that the ground truth provides the correct response of each task. Considering that the true positive (tp) are the cases when both worker and ground truth select the item, true negative (tn) are the cases when both do not select the item, false positive (fp) are the cases when worker selects the item which is not selected by ground truth, and false negative (fn) the worker fails to select the item which is selected by ground truth. Subsequently, we calculate three evaluation metrics: precision as $P = tp/(tp+fp)$, recall as $R = tp/(tp+fn)$, and accuracy as $A = (tp + tn)/(tp + tn + fp + fn)$.

Fig. 5 shows the distribution of workers in terms of precision and recall, for three types of tasks. As can be seen, workers are generally precise in there responses but there is variation across complete range of recall. The best workers perform with both precision and recall above 0.8. Interestingly, no worker achieved the highest recall for the information extraction task, which highlights the difference between workers and ground truth in terms of the entities extracted from the Wikipedia articles. Nevertheless, these distributions emphasize that in order to achieve high accuracy tasks should be assigned to workers that lie in the top-right quadrant of the plots.

The second aim of the experiment was to compare the quality of prediction of capability tracing, in terms of the predicted correctness of worker responses to the tasks. Following routing strategies were compared:

- *Accuracy* (AC) is the baseline approach that considers the previous accuracy of worker's responses as the indicator of predicted correctness of future tasks. Therefore in this case $p(correct_{n+1} = A$, where A is worker accuracy on observed tasks.
- *Capability Tracing* (CT) uses the capability tracing model to predicting performance of workers. The correctness of worker response to new tasks is calculated according to Equation 3.

We used Bayesian knowledge tracing tool available at PSLC DataShop[5] for building capability tracing models. The model was fitted to training data using the EM algorithm. Furthermore, the values of model parameters were set to default i.e. $p(L_0) = 0.5, p(T) = 0.4, p(S) = 0.2, p(G) = 0.2$. We compare the predicted correctness of worker response for new tasks with the actual correctness according to ground truth.

Table IV shows the comparison of baseline and capability tracing, where the workers' capabilities are observed on one type of task and their performance is predicted on another type of task. For instance, the capability tracing model was trained using worker's observed data on Fact Verification task. Then the Image Comparison tasks were predicted using $p(correct_{n+1})$ i.e. probability of a worker correctly applying a capability on new task. Results show that the capability tracing approach is comparable to the baseline approach in general and achieves better accuracy of prediction between similar tasks. The Fact Verification and Information Extraction tasks have similar capabilities requirements, therefore capability tracing can better predict the performance of workers between them. The drop in prediction quality of capability tracing for Image Comparison task can be attributed to the little variation is the performance of workers on this task.

[5] https://pslcdatashop.web.cmu.edu/

TABLE IV. COMPARISON OF CORRECT PREDICTIONS MADE BY BASELINE (AC) AND CAPABILITY TRACING (CT) ON NEW TASKS. THE PERCENTAGE OF CORRECT PREDICTIONS ARE AVERAGED OVER 17 WORKERS. FOR THE SAME TYPE OF TASK 50% WERE USED AS OBSERVATION TASKS AND 50% AS NEW TASKS

Observation Tasks	**New Tasks**	**AC**	**CT**
Fact Verification	Fact Verification	63%	65%
Image Comparison	Image Comparison	66%	61%
Information Extraction	Information Extraction	48%	48%
Fact Verification	Image Comparison	70%	70%
Fact Verification	Information Extraction	51%	53%
Image Comparison	Information Extraction	53%	45%

VI. RELATED WORK

In this section we review the related literature in crowdsourcing and human computation research along two areas: user & performance modeling and task routing.

A. User & Performance Modelling

Karam et al. [20] have defined a meta-data model to represent various aspects of users in human computation applications, such as social profiles, activity history, actions, etc. Rzeszotarski and Kittur [6] defined users models using the logs of user interaction events such as clicks, keystrokes, focus time, etc. Gomez and Laidlaw [7] developed a framework to support user interface design decision based on interaction histories from crowds. In its current state, most of the user modeling research in crowdsourcing is limited to either representation models or event logging frameworks. This paper presents a complementary approach that models a worker in terms of human capabilities employed while performing crowd sourced tasks.

Moris et al. [21] have studied the effects of priming on task performance in microtask platforms. The study found that by using primes like images and music the performance of crowd workers can be improved in the short term. Le et at. [22] investigated the effects of initial worker training for improving the quality of microtasks, specifically for relevance judgement tasks. Rogstadius et al. [23] examined the relationship between worker motivation and task performance in crowdsourcing markets. By comparison, this work categorizes tasks in terms of human capabilities while considering worker performance on heterogeneous tasks.

B. Task Routing

Matching between tasks and workers has been an active area of research among crowdsourcing and human computation. Law et al. [24] studied the effects of self-rated expertise, interests, confidence and understanding, on pull based task selection by crowd workers. The study remained limited to relevance judgment tasks only. Zhang et al. [25] proposed peer routing, a rules-based method of incentive calculation that encouraged people to jointly contribute the task solution and make the routing decisions. Peer routing relies on worker's assessment of other workers, as opposed to worker's specific capabilities discussed in this paper. Ho and Vaughan [26] formalized the task assignment problem in online setting for heterogeneous tasks in crowdsourcing markets. They also proposed an online algorithm that is competitive to the offline version of assignment algorithm. Our approach is complementary to these assignment algorithms as this worker focuses on the actual human capabilities for microtasks.

The task routing problem has also been studied in the context of online communities websites. Zhou et al. [27] combined three approaches to profiling users, based on information available in online question answering systems, for actively pushing question to appropriate users. The value of intelligent task routing in community maintained knowledge system has been demonstrated in recent studies [28]. We attempt to study the capabilities for human task performance, instead of explaining the human behaviour in a specific kind of crowdsourcing system.

VII. CONCLUSION

First this paper presents a taxonomy of worker capabilities in crowdsourcing platforms for microtasks. The taxonomy is based on well-established work on human abilities and performance in behavioural sciences research. We have provided a mapping of capabilities taxonomy with commonly found microtasks in existing online marketplaces. We also compare the capabilities taxonomy with the Fleishmans taxonomy of human abilities. The taxonomy is first step towards an effective categorization of microtasks that can be used to reason about performance and learning of workers in crowdsourcing platforms.

Second this paper introduces and evaluates capability tracing, a technique for measuring latent capabilities or workers to make inferences about their performance on heterogeneous tasks. We demonstrate a generalized approach to worker performance analysis, by analysing microtasks according to the capabilities required from workers. We use Bayesian networks to model different states of worker capabilities. Using this probabilistic model we predict worker performance of new tasks. We evaluate the capability tracing approach with the help of three different microtasks. The results show that the proposed approach is comparable with the baseline approach in terms accuracy.

Future work on worker capabilities and performance modelling is to be done in several areas of microtask platforms. So far, we have focus on task performance from a human ability point of view. Knowledge and skill constitute another type of human capabilities that are relatively unstable overtime as compared to basic human abilities. As a next step we would like to study these capabilities in specific applications of crowdsourcing. There is a large body of knowledge in cognitive science and educational data mining fields that may provide the foundational theory required for this purpose.

Finally, promoting learning through active feedback and practice can help improve the quality of worker output and serve as a motivating factor. The research work on intelligent tutoring systems can provide starting point in terms of theory and practice. Furthermore, developing standard tests for capabilities can help generalize worker performance across multiple domains and applications.

Acknowledgements

The work presented in this paper is funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). We thank all volunteers who participated in the study and staff who provided helpful comments on this document.

References

[1] J. Surowiecki, *The wisdom of crowds*. Random House Digital, Inc., 2005.

[2] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Communications of the ACM*, vol. 54, no. 4, pp. 86–96, 2011.

[3] L. B. Chilton, J. J. Horton, R. C. Miller, and S. Azenkot, "Task search in a human computation market," in *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2010, pp. 1–9.

[4] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 286–295.

[5] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Pick-a-crowd: tell me what you like, and i'll tell you what to do," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 367–374.

[6] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: using implicit behavioral measures to predict task performance," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 13–22.

[7] S. Gomez and D. Laidlaw, "Modeling task performance for a crowd of users from interaction histories," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 2012, pp. 2465–2468.

[8] E. A. Fleishman, "Systems for describing human tasks." *American Psychologist*, vol. 37, no. 7, pp. 821–834, 1982.

[9] E. Law and L. v. Ahn, "Human computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121, 2011.

[10] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 1301–1318.

[11] U. Ul Hassan, S. O'Riain, and E. Curry, "Slua: Towards semantic linking of users with actions in crowdsourcing," in *Proceedings of 1st International Workshop on Crowdsourcing the Semantic Web*, Sydney, Australia, 2013.

[12] L. G. Militello and R. J. Hutton, "Applied cognitive task analysis (acta): A practitioner's toolkit for understanding cognitive task demands," *Ergonomics*, vol. 41, no. 11, pp. 1618–1641, 1998.

[13] E. A. Fleishman, "Toward a taxonomy of human performance." *American Psychologist*, vol. 30, no. 12, pp. 1127–1149, 1975.

[14] U. Ul Hassan, S. O'Riain, and E. Curry, "Effects of Expertise Assessment on the Quality of Task Routing in Human Computation," in *Proceedings of the 2nd International Workshop on Social Media for Crowdsourcing and Human Computation*, Paris, France, 2013.

[15] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise retrieval," *Foundations and Trends in Information Retrieval*, vol. 6, no. 2–3, pp. 127–256, 2012.

[16] U. Ul Hassan, S. O'Riain, and E. Curry, "Towards Expertise Modelling for Routing Data Cleaning Tasks within a Community of Knowledge Workers," in *Proceedings of the 17th International Conference on Information Quality*, Paris, France, 2012.

[17] N. G. Peterson, M. D. Mumford, W. C. Borman, P. Jeanneret, and E. A. Fleishman, *An occupational information system for the 21st century: The development of O* NET*. American Psychological Association, 1999.

[18] R. Tipples, "The human capability framework-an important and useful framework for understanding the labour market?" *New Zealand Journal of Employment Relations*, vol. 29, pp. 3–20, 2004.

[19] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1994.

[20] R. Karam, P. Fraternali, A. Bozzon, and L. Galli, "Modeling end-users as contributors in human computation applications," in *Model and Data Engineering*. Springer, 2012, pp. 3–15.

[21] R. R. Morris, M. Dontcheva, and E. M. Gerber, "Priming for better performance in microtask crowdsourcing environments," *IEEE Internet Computing*, vol. 16, no. 5, pp. 13–19, 2012.

[22] J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution," in *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 2010, pp. 21–26.

[23] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets." in *ICWSM*, 2011.

[24] E. Law, P. N. Bennett, and E. Horvitz, "The effects of choice in routing relevance judgments," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 1127–1128.

[25] H. Zhang, E. Horvitz, Y. Chen, and D. C. Parkes, "Task routing for prediction tasks," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 889–896.

[26] C.-J. Ho and J. W. Vaughan, "Online task assignment in crowdsourcing markets." in *AAAI Conference on Artificial Intelligence*, 2012.

[27] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, "Routing questions to the right users in online communities," in *Proceedings of the 25th IEEE International Conference on Data Engineering*. IEEE, 2009, pp. 700–711.

[28] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl, "Suggestbot: using intelligent task routing to help people find work in wikipedia," in *Proceedings of the 12th International Conference on Intelligent User Interfaces*. ACM, 2007, pp. 32–41.

TABLE V. The mapping between Microtask Taxonomy and Fleishman’s Taxonomy of human abilities.

Human Abilities	Comprehension	Bilingualism	Writing	Comparison	Judgment	Perception	Identification	Reasoning
Oral Comprehension	✓							
Written Comprehension	✓							
Oral Expression								
Written Expression			✓					
Mathematical reasoning								✓
Number facility								
Fluency of Ideas								
Originality								
Problem sensitivity								
Deductive reasoning								✓
Inductive reasoning								✓
Information ordering				✓				
Category flexibility				✓				
Flexibility of Closure								
Speed of Closure								
Memorization							✓	
Perceptual speed						✓		
Visualization								
Time sharing								
Selective attention								
Control precision								
Multiple coordination								
Response orientation								
Rate control								
Reaction time								
Arm-hand steadiness								
Manual dexterity								
Finger dexterity								
Wrist-finger speed								
Speed of limb movement								
Static strength								
Explosive strength								
Dynamic strength								
Trunk strength								
Extent flexibility								
Dynamic flexibility								
Gross body coordination								
Gross body equilibrium								
Stamina								
Near vision								
Far vision								
Visual Color discrimination						✓		
Night vision								
Peripheral vision								
Depth perception						✓		
Glare sensitivity								
Hearing Sensitivity						✓		
Auditory attention						✓		
Sound localization								
Speech Recognition						✓		
Speech clarity								