Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Effect of Task Presentation on the Performance of Crowd Workers — A Cognitive Study

Harini Sampath, Rajeev Rajeshuni, Bipin Indurkhya
International Institute of Information Technology
Hyderabad, India

Saraschandra Karanam, Koustuv Dasgupta
Xerox Research Center
India

Abstract

We present our ongoing research on improving the task presentation using cognitively inspired features to optimize the performance of crowd workers. For the task of extracting text from scanned images, we generated three task-presentation designs by modifying two features of the task - *visual saliency of target fields* and *working memory requirements*. Our experiments conducted on Amazon Mechanical Turk (AMT) indicate that modifying visual saliency results in better performance.

Introduction

In crowdsourcing platforms, requesters post a Human Intelligence Task (HIT) online using a task template which is taken up (at some point of time) by one or more workers to solve. There is no direct interaction (physical or virtual) between the requester and the worker. Because of this asynchronous and anonymous nature of most crowd platforms, extracting high quality work from a heterogeneous set of crowd workers has been a long-known problem. Existing attempts to improve quality include aggregating responses from multiple crowd workers (Ipeirotis, Provost, and Wang 2010), filtering workers using qualification tasks (Biewald and Van Pelt 2011), and modifying the structure of the task (Kittur, Chi, and Suh 2008; Toomim 2011).

The goal of our ongoing project is to explore the effect of cognitively-inspired task designs on the performance of the crowd workers. We describe here the results from the experiments conducted on Amazon Mechanical Turk (AMT) using two cognitive features - a) visual saliency and b) working memory in the context of a form digitization task.

Template Design

We focused on the task of extracting three fields (patient name, injury, patient id) from a handwritten insurance form as shown in Fig. 1, where the user needs to read the task description, identify the target field in the form, remember its contents and type those contents into a target text box. We generated multiple cognitively-inspired templates to reduce the possibility of errors in each of these steps. We describe below briefly theories and experiments that motivated our design of the templates.

Figure 1: Sample form

Figure 2: Highlighted form

Visual Saliency

A region in an image is considered visually salient if it grabs the attention of the viewer. Saliency is a complex function of both low-level (e.g. color, contrast, orientation) and high-level (e.g. task given to the viewer) features. In the digitization task, we manipulated the visual saliency of target fields by highlighting them (Fig.2). We hypothesized that this would reduce the visual search required to identify the target fields.

Working memory

Working memory is a limited capacity system that temporarily maintains information and acts as an interface between human perception, long-term memory and action. In the digitization task, users need to hold the information from the image in their working memory before they type it in the response boxes. To evaluate the role of working memory, we placed the response boxes near the fields the users are extracting as shown in Fig.3. We hypothesized that this would reduce the need for the workers to commit all content to working memory.

Experiments on Amazon Mechanical Turk

We conducted our experiments on Amazon Mechanical Turk (AMT). The users had to extract three fields from an insurance form - Patient Name, Patient Identification Number and Patient Illness. The structure of the form was such that each

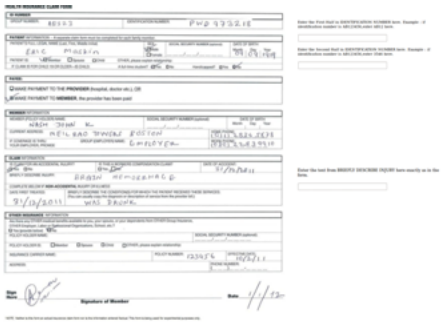

Figure 3: Working memory was manipulated by altering the location of the response boxes

of these fields had one distractor field with potential for confusion: Member Name, Group ID, and Treatment Sought respectively.

ID	Description
T1	No Modification - Baseline Template
T2	Visual Saliency - Highlight fields
T3	Working Memory - Move responses boxes

Table 1: List of Templates

Design We generated three templates by varying visual saliency and working memory as shown in Table. 1. The data was collected over a period of four days. Each day a batch of 100 HITs were posted in AMT of a particular template type. Each HIT was offered a payment of 2 cents. There were no restrictions on the number of HITs that could be completed by a participant. All the responses received were compared against the ground truth data (manually generated) and the payments were released.

Results

We used the length-normalized levenshtein distance between the response and the ground truth as the measure of accuracy. We performed a linear mixed effects analysis of the relationship between accuracy and the templates. The template was entered as a fixed effect and subjects as random effect. P-values were obtained by likelihood ratio tests of the full model with the fixed effect against the model without the effect.

For both Patient Name and Identification Number fields, T2 performed significantly better than T3 ($\chi^2(1, N = 800), p < 0.1$) and T1 ($\chi^2(1, N = 800), p < 0.05$). There was also a significant difference between T3 and T1 ($\chi^2(1, N = 800), p < 0.1$). For Patient Illness field, T2 performed significantly better than T3 ($\chi^2(1, N = 800), p < 0.05$) and $T1(\chi^2(1, N = 800), p < 0.05$). However, there was no significant difference between T3 and T1.

Discussion

Template T2, where the target fields were highlighted performed best for all the fields. This is because the confusion among the target and the distractor fields was minimum in these templates. T3 had a positive effect for both patient name and patient ID (an eight digit alphanumeric e.g. AB123456) fields, both of which do not carry a semantic content. However, patient illness, being a semantically coherent field, was not affected by this working memory modification.

Conclusion

Our results demonstrate that different designs of task presentations result in significant differences in the crowd performance. Both T2 and T3 yielded significantly better performance than T1 for patient name and patient ID fields. The performance of T2 was better than T3 for patient illness field.

For a requester unfamiliar with the cognitive nuances of task-presentation design, it is difficult to predict the performance for a particular way of task presentation. The performance also varies depending on the kind of platform, the nature of the task, the time of submission etc.

Our motivation is to build a system to recommend task templates to requesters. This system would learn how the performance of crowd workers is affected by a set of cognitive features and their interactions. The set of cognitive features would be chosen so that they would apply to a broad class of crowd-sourcing tasks. For instance, there could be other cognitive features that could affect the performance in digitization task such as number of search items, cognitive bias of crowd workers etc.,

The goal of the learning process would be to optimize for accuracy, response time, completion rate or any other significant metric for the requester. Based on this learnt model, the framework would predict the performance of each template (characterized by its feature set) in the particular crowd platform. Consequently, given a new set of template, the system can produce a ranked list of these templates based on the predicted performance, and recommend the best template for a particular task. We envision that such a framework would empower an enterprise user with a concise methodology for selection of a task template that maximizes the return of investment, by saving time and improving quality.

References

Biewald, L., and Van Pelt, C. 2011. Distributing a task to multiple workers over a network for completion while providing quality control. WO Patent 2,011,159,434.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67. ACM.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 453–456. ACM.

Toomim, M. 2011. Economic utility of interaction in crowdsourcing. In *Workshop on Crowdsourcing and Human Computation at CHI*, volume 11.