Human-Machine Hybrid Computation

Position paper for *CHI 2011 Workshop On Crowdsourcing And Human Computation*

Alexander J. Quinn
Benjamin B. Bederson
Human-Computer Interaction Lab
Computer Science Department
University of Maryland
{aq,bederson}@cs.umd.edu

INTRODUCTION

A common thread among fields such natural language processing, computer vision, and artificial intelligence is that all seek to automate tasks which humans do naturally. Examples include listening to human speech, seeing and recognizing objects, reading text, and understanding meaning. If they could be automated by computers, then they could be done on demand, and much more quickly and cheaply than human workers can do. However, for many problems, the quality provided by such algorithms is still too low to be used in applications that demand high accuracy and reliability. The other end of the spectrum is traditional human labor, in which workers can be hired to do the same types of tasks. Their accuracy is usually much higher than that of computers, but human workers take more time and require more money to do the same amount of work.

This forces the developer of a specific solution to make a choice:

a) Use all human effort and get good results with a high cost of time and money.

b) Use fully automated methods and get less accurate results quickly and cheaply.

Platforms such as Amazon Mechanical Turk (AMT) allow the developer to bridge this gap somewhat. AMT can be used to perform the same types of tasks on demand, and often more quickly and cheaply than would be possible if one had to hire new workers. However, issues such as cheating and varying abilities among the workers mean that the quality tends to be less reliable than what traditional workers would provide. This can be remedied somewhat by requesting multiple judgments on each task and/or using the workers in different roles that are coordinated to yield high quality final results. However, this in turn increases the cost.

The goal of our research is to expand the options for bridging this gap (Figure 1). Our primary strategy is to develop hybrid solutions that tightly integrate human computation with machine resources. We are developing methods and tools for doing this in ways that will be applicable across a wide range of problem domains. Below, we discuss two specific ideas we have explored.

CROWDFLOW: BLENDED COMPUTATION

We have been developing CrowdFlow, a general framework for blending human efforts with machine capabilities, especially machine learning classifiers [4]. The user of the system will specify the desired speed-cost-quality tradeoff. The system will then allocate tasks to humans and machines in a way that will fulfill the user's specification.

For example, suppose a movie web site operator had 100,000 movie reviews written by site visitors and they wanted to classify them as positive or negative (perhaps so they could show only the positive ones in order to sell more movies). Asking humans to read 100,000 reviews would be too slow and expensive. Computers can do the task, but not accurately enough for this operator's needs. The operator could set a budget of, say $1000, and have the system provide the best results possible for that budget. Alternatively, the operator could set a desired accuracy, say 90%, and the system would blend human and machine effort automatically to achieve that target.

A key part of the idea is that the humans and machines will benefit from one another because each will be doing the same types of tasks. When a human does a task, the result will be used to train the classifier, thus helping to boost the quality of the machine. Machines can also help make

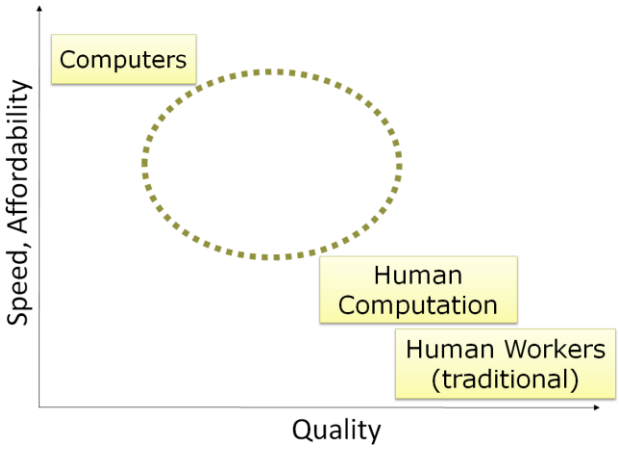

Figure 1. Hybrid services help to fill in the gap between all-human solutions (good, but slow and expensive) and all-machine solutions (fast and cheap, but poor quality).

humans more productive (and thus cheaper and faster) by providing a first cut answer to the human. If the answer is correct, the human can simply review and accept the answer, without having to enter any detailed information into the interface. If the answer is partially correct, the human can fix it. If the answer is completely wrong, the human can replace it and enter a good answer. This flow of information is illustrated in Figure 2.

Last year, we explored a possible implementation strategy by creating a prototype system in the form of a Python programming module. We could connect an arbitrary classifier via a Python wrapper. The programmer using the module would subclass our `Machine` class, providing methods such as `train(task,answer)` and `evaluate(task)`.

This version was a simplified version of the CrowdFlow model, but it was useful for understanding what will be needed to make it work. For financial constraints, the algorithm simply posted as many HITs as possible given the budget and assigned the rest to the machine. For time constraints, the code would post many HITs and let them run until the time ran out, leaving the rest to the machine. Accuracy constraints are more difficult because the algorithm does not know a priori how accurate the humans will be, and thus it does not know how many tasks should be done by humans and how many by the machine in order to achieve a given combined accuracy of the output (i.e. need exactly 85% accurate final results as cheaply and quickly). To test using an accuracy constraint, our code required that the user provide an estimate of expected human accuracy. In the future, we hope to have the code automatically gauge workers' accuracy.

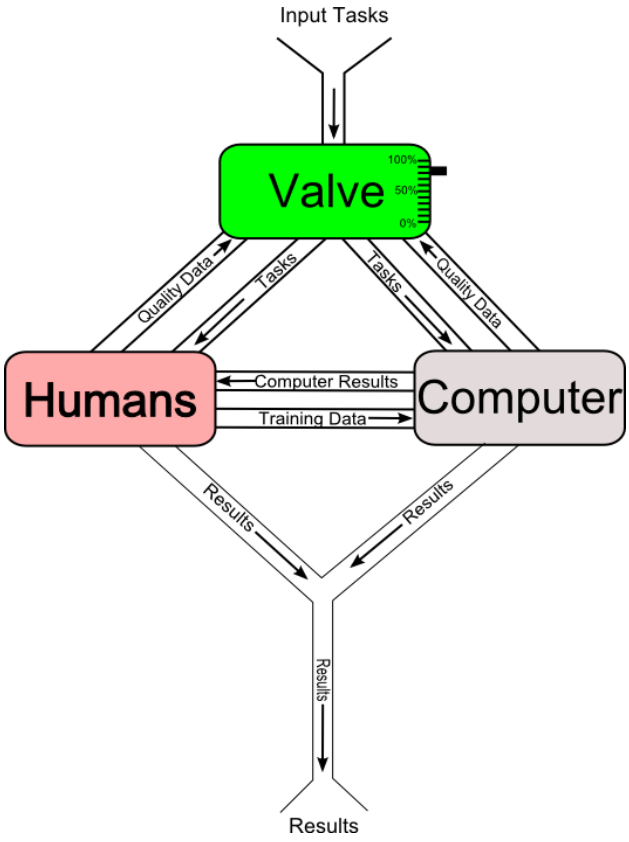

Figure 2. In CrowdFlow humans will make machines more accurate by providing training data, while machines make humans more productive by providing imperfect first cut results.

We used the prototype to evaluate the basic idea of CrowdFlow on two problem domains: sentiment analysis of movie reviews and detection of human figures in photographs. Details are in [4]. We learned that making CrowdFlow work in more realistic scenarios will entail several challenges. First, we must be able to estimate, within some confidence interval, the accuracy of the human workers, *even if there is no ground truth*. This is doubly important when you consider that without humans judgment, it is impossible to estimate the machine accuracy. Also, the system must keep updating the accuracy estimates and automatically adjust the allocation of tasks as needed.

Recently, we have been working with economics professor Ginger Jin and business professor Siva Viswanathan to run experiments to better understand the labor and economic issues that affect price and worker performance so that ultimately, we can build the models needed to be able to estimate, predict, and hopefully even manipulate the level of accuracy in the results we receive from the workers.

PARATRANS: COMPLEMENTARY ROLES

Another approach is to decompose a complex task into fine-grained sub-tasks that are either well-suited to the human workers or well-suited to the computer. One can build a continuous process that calls on Mechanical Turk workers and computer resources, as needed.

Working with linguistics professor Philip Resnik, we built a test apparatus for translating text from Chinese into English using a combination of a machine translation engine and workers on AMT. The process, which is described in [6] works as follows. The task interfaces used with AMT are shown in Figure 3.

1. Google Translate, a machine translation service, translates the source Chinese text into English.
2. Workers on AMT read the English translation and highlight the words or phrases that make it awkward or difficult to understand. Portions of the sentence that are already acceptable are left as is.
3. Our server projects those problematic phrases back to Chinese using detailed word alignment data provided by the Google Translate Research API, yielding a set of Chinese phrases.
4. Workers on AMT who speak Chinese read the Chinese phrases and provide paraphrases.

5. Our server substitutes these Chinese paraphrases into the original source sentence. The algorithm takes the cross product of all possible substitutions, including the original text, to generate a set of revised versions of the original Chiense sentence. (Some practical limits are imposed to deal with combinatorial explosion in cases where there were many non-overlapping possible substitutions.)
6. Google Translate translates the new sentences into English yielding a set of candidate translations.
7. Either a machine or a computer can be used for the last step of identifying the best candidate. In our pilot study, we asked humans to rate the candidates based on fluency (correctness of language). Since most machine translation engines normally create candidates and use heuristic-based algorithms to choose the best one, it would be natural to use such algorithms in a more realistic deployment.

The result is a new English translation that tends to be better than the initial translation provided by Google Translate but not quite as good as what a professional translator would have provided. It benefits from the assistance of humans in two crucial steps: identifying problematic sections and providing paraphrases. The rest of the work is done by machines.

The output of this human-machine hybrid process was generally better than what Google Translate provided initially. Our evaluation using the standard NIST '08 data set with oracle evaluation measured an improvement of 2.46 points using the BLEU metric, and 5.46 using the TERp metric [6]. Although the results were still not as good as what a professional bilingual translator could produce, the human tasks do not require bilingual translation abilities, and thus can draw from a much larger pool of workers. This means the work can probably be done more cheaply and more quickly than if a professional translator were employed.

CONCLUSION AND FUTURE WORK

These are only two of many possible compromise solutions that use a combination of different computational resources to achieve different points in the speed-cost-quality tradeoff space. We are actively exploring others, which we will be sharing as results become available. The ultimate goal is to be able to accurately predict—and thus direct—the speed, cost, and quality of the output, in order to give solution developers better options than are currently available.

(a)

Example	The latest research visit Jupiter was the Pluto-bound New Horizons spacecraft in chicken February 2007.
1.	River West will root for the bar 30 years of rural health personnel Awards
2.	They is not Mao Zedong's I, they have blood and meat, solid indeed in the life in me have this one community will be the human relations group of bodies, they from the Shan Xia Township Dao laid measures of unemployment pro-body sense of Shou's history, is as two era of social phenomena.
3.	(Reporter Wu Zhiqiang Wang Xiang Jiang)
4.	(International) Turkish ruling party won the Council election
5.	Strapless slippers

(b)

	Original	Revised
Example A	我需要 有人帮忙。	我需要 帮忙的人。
Example B	小萌和小琴度假去了欧洲。	小萌和 小琴放假时去了 欧洲。
1.	两国不论在国际事务,还是在各自国家建设中,都相互信任,相互支持,相互帮助,密切配合。	两国 论在国际事务,还是在各自国家建设中,都相互信任,相互支持,相互帮助,密切配合。
2.	我又开始梦游了,看到了家里的老院子,以及年轻的姐夫在院子里忙碌的身影。	我又开始梦游了,看到了家里的老院子,以及年轻的姐夫 院子里忙碌的身影。
3.	于是,也就慢慢养成了一些习惯,比如告诉自个儿,对自己狠一点,对别人好一点;	于是,也就慢慢养成了一些习惯,比如告诉自个儿,对自己狠一点, 别人好一点;

Figure 3. (a) Task UI for highlighting problematic spans of a sentence. (b) Task UI for providing paraphrases of the corresponding source language text of problematic spans.

ATTENDANCE

Alexander J. Quinn is a 5th year graduate student pursing a PhD in computer science at the University of Maryland under the direction of Professor Benjamin B. Bederson. His prior research has covered several other areas of human-computer interaction, including readability in digital libraries [5], mobile applications for story authoring [3], visualization of temporal data [7], and novice programming [1]. He is the first author of a paper at CHI 2011 about human computation [2] and is currently preparing his thesis proposal, which is about a way of applying CrowdFlow to complex analytical problems. Therefore, attending this workshop would be a valuable opportunity to learn more about the field while contributing to the ongoing conversation about how to utilize human computation.

REFERENCES

1. Quinn, A.. An Interrogative Approach to Novice Programming. In *Proceedings of HCC 2002*.
2. Quinn, A. & Bederson, B.B. Human Computation: A Survey and Taxonomy of a Growing Field, Proceedings of *CHI 2011* (in press).
3. Druin, A., Bederson, B. B., Quinn, A. J.. Designing Intergenerational Mobile Storytelling. In *Proceedings of IDC 2009.*
4. Quinn, A., Bederson, B., Yeh, T., Lin, J. CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility. Technical Report HCIL-2010-09, University of Maryland.
5. Quinn, A. J., Hu, C., Arisaka, T., Rose, A., and Bederson, B. B. Readability of scanned books in digital libraries. *Proceedings of CHI 2008*.
6. Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A., & Bederson, B.B. Improving Translation via Targeted Paraphrasing, Proceedings of Conference on Empiracal Methods in Natural Language Processing. In *Proceedings of EMNLP 2010*.
7. Wang, T. D., Plaisant, C., Quinn, A. J., Stanchak, R., Murphy, S., and Shneiderman, B. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceedings of CHI 2008*.