Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Crowdsourcing for Deployable Intelligent Systems

Walter S. Lasecki
University of Rochester, Rochester, NY 14627
wlasecki@cs.rochester.edu

Abstract

My work aims to create a scaffold for *deployable* intelligent systems using crowdsourcing. Current approaches in artificial intelligence (AI) typically focus on solving a narrow subset of problems in a given space - for example: automatic speech recognition as part of a conversational assistant, machine vision as part of a question answering service for blind people, or planning as part of a home assistive robot. This approach is necessary to scope the solution, but often results in a large number of systems that are rarely deployed in real-world setting, but instead operate in toy domains, or in situations where other parts of the problem are assumed to be solved.

The framework I have developed aims to use *the crowd* to help in two ways: (i) make it possible to use human intelligence to power parts of a system that automated approaches cannot or do not yet handle, and (ii) provide a means of enabling more effective deployable systems by people to provide reliable training data on-demand. This summary begins with a brief review of prior work, then outlines a number of different system that I have developed to demonstrate the capabilities of this framework, and concludes with future work to be completed as part of my thesis.

Background

Crowdsourcing is a form of human computation (von Ahn 2005) in which multiple workers complete small portions of a task in order to collectively arrive at a more accurate or complete answer than any individual member of the crowd would have. Human computation has been shown to be effective at solving many problems that automated system still struggle with, such as visual question answering (Bigham et al. 2010), document editing (Bernstein et al. 2010), and real-time transcription (Lasecki et al. 2012a).

Legion

Legion (Lasecki et al. 2011) is a system that allows the crowd to control existing user interfaces in real-time. Work on Legion has explored navigating robots through a maze, controlling word processing and spreadsheet software, and enabling more accurate predictive keyboards for motor-impaired users (Lasecki et al. 2011).

This work introduced a new model of crowdsourcing, called *continuous* real-time crowdsourcing, in which groups of workers simultaneously contribute to a continuous task in real-time. The system operates in a closed-loop fashion and allows workers to each get feedback from their actions and use that context to decide their next course of action. The key to making this approach work with existing interfaces is the *input mediator* which merges all input into a single control stream, making the crowd appear as a single reliable user, or *crowd agent*. We found that the most effective method for doing this was to use the cosine similarity between each worker and the crowd as a whole calculate a weight value that tracks each user's performance using the 'wisdom of the crowd'. We then select a single 'leader' to be in charge for micro-units of time (typically less than 1 second). Despite frequent shifts in leadership, the overall output is both reliable and consistent.

The crowd agent model also allows the crowd to retain more capabilities of an individual. For example, tests showed that the crowd was able to collectively remember task past information through a collective process similar to organizational memory, a phenomenon often seen on a larger scale (societies, corporate cultures, etc.)

Legion:Scribe

Legion:Scribe (Lasecki et al. 2012a) is system that provides real-time text captions of speech using non-expert workers. Professional stenographers are currently the only means of providing reliable real-time captions, but they must be trained for 2-3 years to be able to keep up with natural speaking rates that often exceed 250 words per minute, making them very expensive ($100-200/hr) and rare (thus had to schedule). Scribe allows group of 3-5 non-expert workers to keep up with live audio by merging their partial captions into a single final result that is shown to workers.

This alignment process plays the role of the input mediator, but synthesizes the inputs instead of selecting the best individual input at any given time, as Legion did. Scribe does this alignment by using an A^* search based multiple sequence alignment algorithm (Naim et al. 2013), originally inspired by those used in computational biology for genetic sequence alignment.

Legion:AR

Legion:AR (Lasecki et al. 2013a) extends the approach used in Scribe to provide consistent labels of activities being performed in a video stream using the crowd. Legion:AR advances the idea of supporting *already deployed* systems by allowing automated activity recognition systems (in our tests, an HMM-based system) to actively learn from the labels produced by the crowd. Whenever the automated system is unsure of its label for a given activity, a request is issued to Legion:AR and a label stream begins within seconds. Our goal was to reduce the training overhead involved in releasing robust activity recognition systems, and handle situations which are difficult or impossible to plan for a prior.

Chorus

Chorus (Lasecki et al. 2012b) is a system that allows the crowd to collectively act as a conversational parter. Chorus uses an incentive mechanism to encourage workers to propose and vote on potential responses to the user. To help support multi-session memory of the user even with an ever-changing workforce, Chorus includes a shared memory space which uses the crowd to extract facts that might need to be recalled later. Using this two-part system, we developed Chorus:Assist, an intelligent assistance that is able to return answers or ask followup questions about information gathering tasks. Chorus:Assist was able to consistently and correctly answer over 84% of user prompts, both primary queries and requests for details or followup information.

Chorus also holds the promise of being an unprecedented training framework for automated conversational agents, who can be used as workers just as members of the crowd can be. This allows AI systems to both learn from the conversation they observe, both the accepted content, and that which was deemed by the crowd to be not worth presenting to users. It also allows systems to propose possible responses that will be filtered out by the crowd if incorrect. This means that even still-untrained systems do not effect the user's perception of the assistant. Additionally, as the system learns, more of the responsibility can be handed automatically – allowing Chorus to smoothly scale from being entirely crowd-powered, to being entirely automated.

Chorus:View

Chorus:View uses the conversational interface pioneered in Chorus:Assist to help blind and low-vision users answer visual questions about the environment around them using streaming video. Because the interaction is continuous, the crowd maintains context and can give real-time feedback, unlike prior single-questions systems such as VizWiz (Bigham et al. 2010). In initial tests, Chorus:View shows significant improvement over VizWiz (several seconds compared to several minutes) in tasks where information had to be sought out, such as reading details on food packages.

Future Work

My work thus far has established a set of approaches to recover reliable, consistent responses from the crowd, and a framework that allows the crowd to act as a single agent and source of guidance to existing AI systems. In future work, my goal is to focus on the development of hybrid intelligence systems more seamlessly blend artificial and human intelligence to create systems that can be deployed now and train automated systems *in situ* to take over control later.

First, I will focus on creating models that take advantage of asymmetries in human and machine skillets in the input combination process. For instance, in the case of Scribe, humans tend to be very good at capturing short, predictable words, where ASR tends to be better at capturing longer, more complex words due to their unique phoneme profile. Using this, we can create a weighting scheme to favor each in their specialty. Next, I will focus on learning these in more general domains from features of the answers.

Finally, we can use the crowd to generating on-the-fly training data in ways never considered before. For example, we are developing ARchitect to investigate formalizing activity recognition tasks by extract STRIPS-style pre and post conditions from crowd labels of video with the help of multiple layers of clarifying responses and knowledge association (Lasecki et al. 2013b).

Conclusion

In my work, I have presented a model of crowdsourcing which can be used to create deployable intelligent systems in real-world domains that leverage the strengths of artificial intelligence while filling in the gaps with human intelligence. Our general framework also allows intelligent agents to learn in real-world domains without the same risk of failing at a critical task, due to the crowd's supervision. This enables systems that can be deployed today using the crowd, and scale towards fully automated in the future.

References

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proc. of the Symp. on User interface software and technology*, UIST '10, 313–322.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proc. of the Symp. on User interface software and technology*, UIST '10, 333–342.

Lasecki, W.; Murray, K.; White, S.; Miller, R. C.; and Bigham, J. P. 2011. Real-time crowd control of existing interfaces. In *Proc. of the Symp. on User interface software and technology*, UIST '11, 23–32.

Lasecki, W. S.; Miller, C. D.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. P. 2012a. Real-time captioning by groups of non-experts. In *Proc. of the Symp. on User Interface Software and Technology (UIST 2012)*.

Lasecki, W.; Kulkarni, A.; Wesley, R.; Nichols, J.; Hu, C.; Allen, J.; and Bigham, J. 2012b. Chorus: Letting the crowd speak with one voice. In *University of Rochester Technical Report*, 1–10.

Lasecki, W. S.; Song, Y. C.; Kautz, H.; and Bigham, J. P. 2013a. Real-time crowd labeling for deployable activity recognition. In *Proc. of the Conf. on Computer supported cooperative work*, CSCW 2013.

Lasecki, W.; Weingard, L.; Bigham, J.; and Ferguson, G. 2013b. Crowd formalization of action conditions. In *Student Abstracts at the AAAI Conf. on Artificial Intelligence*, In submission.

Naim, I.; Lasecki; W.S., Bigham, J.; and Gildea, D. 2013. Text alignment for real-time crowd captioning. In *North American Chapter of the Association for Computational Linguistics Conf.*, In Submission.

von Ahn, L. 2005. *Human Computation*. Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, PA.