The Start of the Art: An Introduction to Crowdsourcing Technologies for Language and Cognition Studies

Robert Munro and Hal Tily
Stanford University and MIT

Introduction

More than a million workers currently login to crowdsourcing/microtasking platforms to complete short tasks for pay-per-task compensation. The platforms were originally developed to allow companies to outsource work but are now being productively used for research. On July 27th, 2011, language and cognition researchers came together for a workshop devoted to crowdsourcing technologies for language and cognition studies. While language and cognition researchers have been running some of the most varied and sophisticated crowdsourcing tasks since the earliest days of the platforms, this was the first time that researchers had come together for a workshop dedicated wholly to crowdsourcing technologies as a tool for empirical studies.

The workshop was run in conjunction with the 2011 LSA Institute at the University of Boulder and it combined presentations by researchers using crowdsourcing technologies with tutorials for those wanting to learn more about them. This paper summarizes the outcomes of the workshop. The tutorial itself is not covered here, but the participants from the tutorials were as active as the presenters in the broader discussions and so this paper draws from all participants, with thanks to everyone who attended the workshop and contributed to its success.

Discussions

Language processing was one of first large-scale uses of crowdsourcing technologies (Biewald, 2011). Shortly after *Amazon Mechanical Turk* (AMT) started to allow third parties to post tasks in 2007, a tech start-up in San Francisco, *Powerset*, began using AMT to create training data for semantic indexing and relevancy judgments for its natural language search system. Spearheaded by Biewald, these natural language evaluation and annotation tasks made *Powerset* the single biggest requester on AMT for more than a year. Innovation in crowdsourcing for language processing has moved in several directions since then. From this same start in crowdsourcing technologies for language processing, computational linguists were soon using crowdsourcing technologies for natural language processing (Snow et al., 2008) and research (Munro et al., 2010), followed by innovative work in annotation (Hseh et al., 2009), translation (Callison-Burch, 2009), transcription (Marge et al. 2010) and direct experiments (Gibson & Fedorenko, to appear; Schnoebelen & Kuperman, 2010).

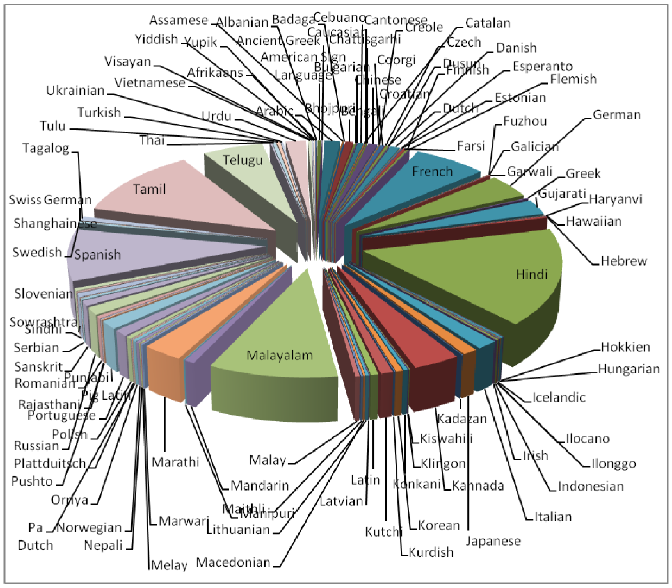


Figure 1: The languages spoken by workers on AMT (self-reported), showing 100 different languages from approximately 2000 responses (Munro and Tily, 2011)

In 2009, AMT overhauled its online interface to allow batch processing from CSV files (it previously only supported batch processing from the command line arguments.) This was a turning point for research, opening up the potential for non-programming researchers to conduct large-scale studies. While AMT is still the preferred choice of platform for researchers, many participants at the workshop were surprised to learn that it is a very small part of the overall crowdsourcing market (perhaps less than 10%). Currently, the biggest platforms are now where people are working for virtual currency inside of games.

Rather than being paid a few cents per task to working on AMT, it is just as likely that someone is being paid right now in virtual seeds within an online farming game.

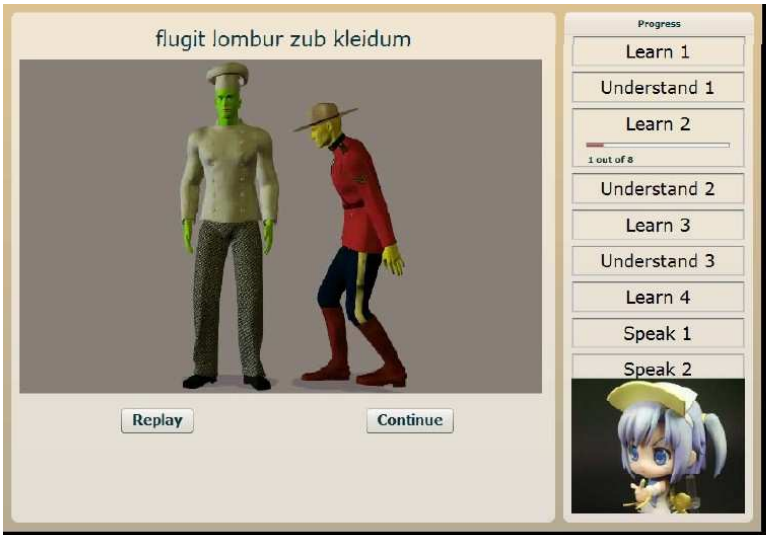

flugit lombur zub kleidum

Figure 2: Screenshot from an artificial language learning task, where the participants view an action via the video and hear/see the sentence describing that action (Jaeger et al., 2011).

Crowdsourcing/microtasking technologies are often known as 'human computing' or 'artificial artificial intelligence'. This is because the distributed online workforces are accessed much like an online computer service: data is passed out to a distributed queue, processed, and returned. It was clear from the discussions that this description does not apply for experiments accessing linguistic judgments and language performance. To be more precise, the 'computing' and both 'artificial's do not apply, as we are eliciting the actual human intelligence of the crowdsourced participants. Research has the capacity to achieve something much more exciting than fast, affordable information processing – it can give us insight into the very nature of human communications, and by extension our neurolinguistic and sociolinguistic systems (Munro and Tily, 2011). Much of the discussion in the introduction and keynote focused on the differences between experimental research and large-scale information processing, and the implications for experimental design. Large-scale crowdsourcing has consistently found that breaking tasks up into small substasks is needed to optimize accuracy, such that this strategy is now more assumed than tested (Kittur et al. 2008, Ledlie et al 2010, Munro et al. 2010, Lawson et al. 2010, Paolacci et al. 2010). This was confirmed by the professional experience of the keynote speaker (Biewald, 2011). In fact, recent work is exploring metrics to indicate where simple tasks can be embedded within more complex, dynamic workflows (Kittur et al. 2011) without even

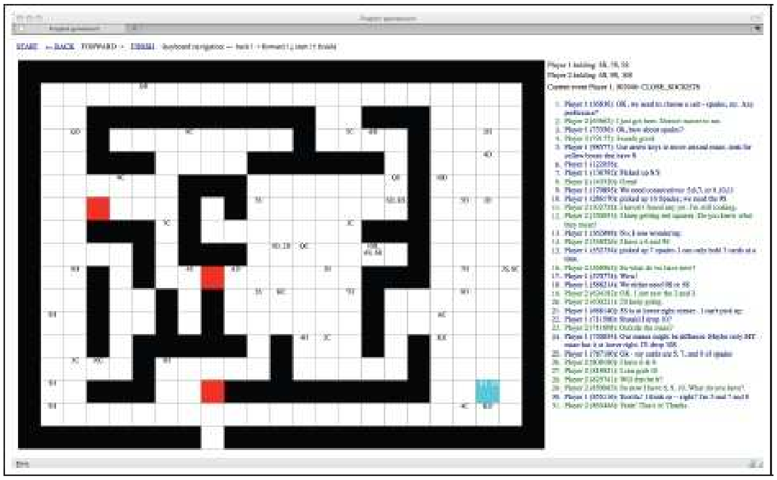


Figure 3: Screenshot from an interactive maze-like game, where participants coordinated with each other via an online chat to complete a card-collection task (Clausen & Potts, 2011).

considering the easier question of exploring where can we simply combine elements in single, larger tasks. Many of the workshop participants, and one of the presentations (de Marneffe and Potts, 2011) argued the opposite for language research, finding that workers did remain engaged for extended sets of questions/tasks, producing higher quality responses as a result. Unlike the scam click-throughs or robots that plague commercial crowdsourcing tasks, the researchers also noted the general high quality of their results. It was clear that the people undertaking the tasks were engaging with the research-focused tasks in a way that they were not engaging with commercial tasks.

Several explanations were offered for why researchers were not experiencing the amount of scammers that industry sees. Biewald suggested that the amount of scamming is a step-function, that is, there is no scamming at all until a certain volume of tasks are available, and it is simply not worth the efforts of a potential scammer to try to write programs to automatically complete a task when it is low volume (researchers rarely seek more than 100s of responses, and sometimes much less, while 100,000s are common for commercial tasks). This effectively puts researchers under the radar of this one type of scamming strategy. A second suggestion was that it would be harder to fake. While it is more difficult to automatically detect aberrant responses in the types of open-ended questions or interaction tasks that are common to linguistic experiments as there is no 'right' answer to gauge someone's performance against, the flip-side of this is that it is much harder to disguise fake responses when the response requires writing a sentence as opposed to selecting a multiple-choice question. For an interaction task, faking ‘being human’ is almost impossible, and so this might also discourage people from trying to scam these kinds of tasks. A third reason was more straightforward: linguistic experiments are fun. The motivations for why people undertake work on microtasking platforms are varied and complex (and largely limited to AMT) (Kaufmann et al., 2011). While money ranks highest for AMT, there is no majority reason and ‘fun’ is also very common. Experiments are often framed as the type of games and puzzles that people might play for free online, and it is easy to imagine that this is a motivator in itself. For people receiving virtual payment as part of a game we can assume that money is even less of a motivation. Another motivation might be that people like to contribute to science, rather than simply cutting the costs of some large business. Finally, the fact the some researchers pay above market wages will no doubt also be a good motivator for someone to pay attention when responding.

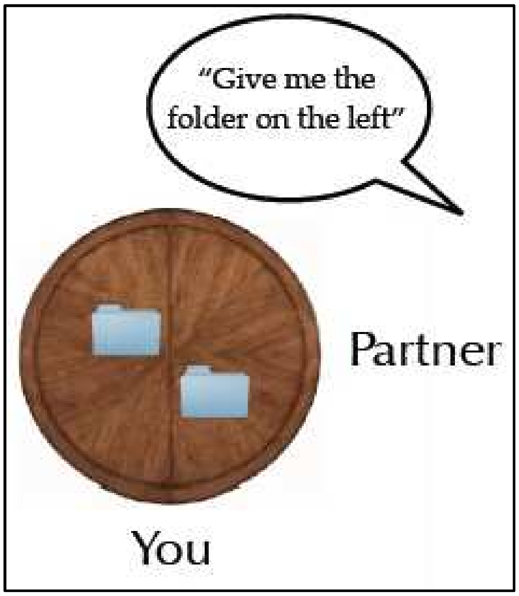
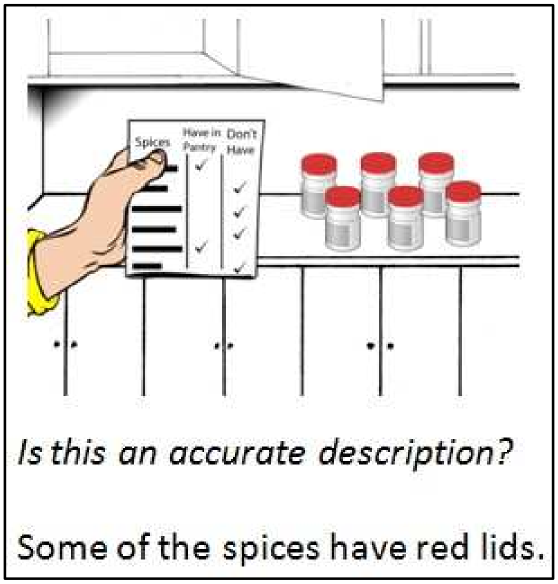


Figure 4: Screenshot from a task that exploited the ‘requester’ and ‘worker’ roles on AMT, to see whether people’s interpretation of spatial indices like ‘left’ differed according to the assumed social roles of participants (Duran & Dale, 2011).

The complexities of payment (ethics in particular) were discussed throughout the workshop. Many labs pay workers above market-wages (which are otherwise often only a few dollars an hour at best) either by choice or to meet IRB requirements. It was especially interesting to compare notes on this. Relative to the cost of hosting a lab experiment, paying higher salaries to online workers is often still a very big saving, especially in the case of shorter tasks, and if anything leads to quicker response times. The most common payment adjustment method that people used within AMT was to calculate the actual time spent through the returned metadata, and then pay the appropriate difference in wage through the built-in ‘bonus’ system. There were no dissenting voices to this approach, but participants remained concerned about how the anonymity of the worker on many platforms could still mean that it harbored an exploitative working environment. For example, the worker, even when ostensibly getting a fair wage, could still be a minor or someone coerced into giving their payment to a third party. The ability to tap online gamers, or workers from within reputable organizations, were both seen as positive future directions in this regard.



Is this an accurate description?

Some of the spices have red lids.

Figure 5: Screen shot showing an image used to elicit information about scalar implicatures in different contexts (Anand et al., 2011).

Overall, what seem to impress people the most (conference organizers included) was the great breadth of research that is now being carried out on crowdsouring platforms. The variety of linguistics within the workshop presentations was among the greatest that we have seen at *any* language or cognition workshop this year, ranging from a fine-grained distinctions in logical metonymy (Zarcone & Pado, 2011) to the interaction of human and machine topic-identification workflows (Satinoff & Boyd-Graber, 2011). The sheer inventiveness of the task designs were equally impressive, including images, sound, videos generated with artificial languages (Jaeger et al., 2011), and at the most complex full interactive games with instant-message chats (Clausen & Potts, 2011). The nature of microtasking platforms themselves was explored in a number of the presentations, including bonus payment-strategies to ensure a high retention rate of workers between tasks (Watts & Jaeger, 2011). The inherent paradigmatic biases of AMT as a experimental platform were part of many presentations, too, especially the need to model and test for any potential biases in the experimental design (Anand, Andrews & Wagers, 2011). In one interesting case, the researchers deliberately exploited the ‘requester’/’worker’ roles to simulate specific social conditions of tasks, taking advantage of the perceived power-bias for a deliberate experimental effect (Duran and Dale, 2011).

Conclusions

The sophistication of the tasks and evaluation methods that researchers are employing on crowdsourcing platforms are already an order of magnitude more sophisticated than the tasks run by commercial organizations that simply focus on throughput and 'gold' accuracy. The use of crowdsourcing platforms is also increasing at such a rate that crowdsourcing will soon become the single most common tool for empirical language and cognition studies: from discussions, it was clear that in some institutions it already has.

Despite the rapid increase in the sophistication and scale, perhaps the greatest change we are seeing is the number and nature of the researchers who are running experiments with very little overhead. Until now, a typical researcher would be about 10 years into their career before they could receive a grant to be the principal investigator for an empirical study with 100 or so participants. In this workshop, many participants learned about crowdsourcing in the morning and were able to generate experimental results by the close of day (in one case, even presenting their first analysis (Harcroft, 2011)). With any researcher now able to run experiments quickly and cheaply, *anybody* can be a principal investigator. The lowered barrier has also resulted in novel empirical research from fields like formal semantics and theoretical syntax: subfields with very little prior experimental research (experiments from both were presented in this workshop). Just as all researchers currently learn how to internally analyze language to test and generate hypotheses, it looks like an increasing number of researchers will soon be doing the same through direct experimentation. This makes for a very bright future for empirical language and cognition studies, and for crowdsourcing technologies as a whole.

References

Anand, Pranav, Caroline Andrews and Matt Wagers. (2011). Assessing the pragmatics of experiments with crowdsourcing: The case of scalar implicature. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Biewald, Lukas. (2011). Keynote. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Callison-Burch, Chris (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Clausen, David and Chris Potts. (2011). Collecting task-oriented dialogues. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Duran, Nicholas and Rick Dale. (2011). Creating illusory social connectivity in Amazon Mechanical Turk. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Gibson, Edward. and Evelina Fedorenko. (to appear). The need for quantitative methods in syntax. *Language and Cognitive Processes*.

Harcroft, David. (2011). French Semantic Role Labeling: a pilot pilot study. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Hsueh, Pei-Yun, Prem Melville and Vikas Sindhwani. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing.*

Jaeger, T. Florian, Harry Tily, Michael C. Frank, Jacqueline Gutman and Andrew Watts. (2011). A web-based (iterated) language learning paradigm with human participants. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Kaufmann, Nicolas, Thimo Schulze, Daniel Veit. (2011). More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. *Proceedings of the Seventeenth Americas Conference on Information Systems.* Detroit.

Kittur, Aniket, Boris Smus and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. *Technical Report, School of Computer Science*, Carnegie Mellon University. Pittsburgh, PA.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI '08). ACM, New York, 453-456.

Lawson, Nolan, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. (2010). Annotating large email datasets for named entity recognition with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles, CA.

Ledlie, Jonathan, Billy Odero, Einat Minkov, Imre Kiss, and Joseph Polifroni. (2010). Crowd translator: on building localized speech recognizers through micropayments. *SIGOPS Operating Systems Review* 43:4, 84-89

Marge, Matthew, Satanjeev Banerjee, and Alexander I. Rudnicky (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.*

de Marneffe, Marie-Catherine and Chris Potts. (2011). A case study in effectively crowdsourcing long tasks with novel categories. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen and Harry Tily. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings NAACL-2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Munro, Robert and Hal Tily. (2011). The Start of the Art: An Introduction to Crowdsourcing Technologies for Language and Cognition Studies. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Paolacci, Gabriele, Jesse Chandler and Panagiotis G. Ipeirotis. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5:5, 411-419.

Satinoff, Brianna, and Jordan Boyd-Graber. (2011). Trivial Classification: What features do humans use for classification? *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Schnoebelen, Tyler, and Victor Kuperman (2010). Using Amazon Mechanical Turk for linguistic research: Fast, cheap, easy, and reliable. *PSIHOLOGIJA*, 43 (4), 441-464.

Snow, Rion, Brendan O'Conner, Dan Jurafsky, and Andrew Ng. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Sprouse, Jon (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 1-13, Springer

Watts, Andrew and T. Florian Jaeger. (2011). Balancing experimental lists without sacrificing voluntary participation. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.

Zarcone , Alessandra and Sebastian Padó. (2011). A crowdsourcing study of logical metonymy. *Workshop on Crowdsourcing Technologies for Language and Cognition Studies*. Boulder, Colorado.