Sources of Variability and Adaptive Tasks

Gabriel Parent, Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University
5000 Forbes av., Pittsburgh, PA, 15213
{gparent, max}@cs.cmu.edu

ABSTRACT

It is known that micro-task worker performance fluctuates. While *between*-worker variability has been studied and has been used to define filters (e.g., to filter out "bad" workers), *within*-worker variability (i.e. how each worker's performance varies over time) has received less attention. Better understanding of the sources of such variability will result in the design of better filters, and more importantly, can inspire the development of adaptive tasks. In an adaptive task, between-worker variability is reduced by adapting the type and difficulty of a job to a worker, while within-worker variability is addressed by reacting via feedback to a change in worker performance. This paper presents evidence of within-worker variability on Amazon Mechanical Turk, and defines a set of sources of variability and describes how adaptive tasks could be designed to attend to them.

Author Keywords
crowdsourcing, human computation, design principles

ACM Classification Keywords
I.m. Computing methodologies: miscellaneous.

General Terms
design, human factors, performance, reliability

CROWDSOURCING FOR SPEECH PROCESSING

Crowdsourcing platforms have been useful for gathering and processing data necessary for natural language and speech research. They have caused a boost in performance of state of the art techniques by providing orders of magnitude of more data, for a fraction of the cost. [1] have demonstrated how accurate and how inexpensive the workers' judgments can be for linguistic tasks. One reason for this is that language is an essential part of human life, and everyone develops an operational level of language usage (as opposed to specialized knowledge, such as image classification).

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Our group has used crowdsourcing to solve a critical problem for speech technology development: **speech transcription**. This consists of writing down text corresponding to an audio segment. The text can include labels such as *lip smack* and *laughs*. The need for larger and larger quantities of transcribed speech data lead to the definition of *quick transcription* guidelines. Even with these new guidelines, transcription is carried out in an average of 6 times real time, which leads to an average cost of $150/hour of speech [2]. Using a crowdsourcing platform such as Amazon Mechanical Turk (MTurk), studies have shown that the speech transcription cost could be as low as 5$/hour [3] and correspond to near-expert quality [4]. In order to achieve this, quality control via worker modeling has to be done. Our approach to speech transcription has two passes segmented in a way that decreases cognitive load and increases throughput. In our first pass of speech labeling (Figure 1), one gold-standard utterance was introduced for every 10 utterances. This allowed us to determine the average accuracy of all of the workers, and this was used to filter out poor judgments. While this approach worked well for improving the quality

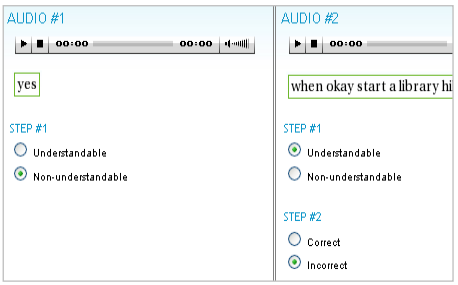

Figure 1 - First pass labeling

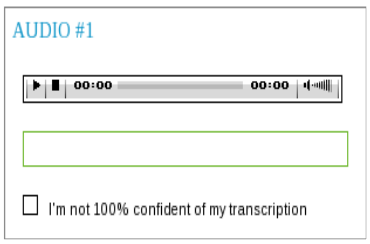

Figure 2 - Second pass transcription

of the results [4], it also introduced extraneous cost. Others have shown that unsupervised quality control, such as using inter-worker agreement, can also increase the overall quality of the results [5].

Another interesting feature for modeling the quality of a worker is self-assessment. Figure 2 shows a simple add-on to our transcription task, where the workers are asked to check a box if they are not 100% confident in their answer. Out of the 73,643 utterances transcribed, 4418 (6%) had that box checked. While the overall accuracy of the results was good for the non-checked utterances (91.4% accuracy), the utterances that were checked were generally poorly transcribed (38.1% accuracy).

FUTURE RESEARCH

Within-worker variability

Variability in the results given by workers can be divided in two categories: *between*-worker and *within*-worker. **Between-worker variability** arises from individual differences (e.g., not all workers have the same skills) while **within-worker variability** explains why the quality of a worker varying over time. Although the quality control mechanisms used so far do a good job at filtering out poor judgments, they are based on between-worker variability - in other words, most filters are based on the idea that some workers perform worse than others and thus their judgment should not be taken into account. Most research on crowdsourcing has not, to our knowledge, taken into account within-worker variability. Sources of this kind of variability include, but are not limited to:

- The learning curve
- The boredom effect [6]
- Attention
- Time of day

A learning curve is one of the elements of variability: at first, as workers learn how to do a task they need more time to complete it and their performance increases over time. [6] showed that boredom can also cause variability in individual speed and performance for workers performing repetitive computer tasks. A similar effect is observed with attention variation, which can be due to changes in attention division between the task and other activities (e.g., chatting with a friend). We found evidence in our speech transcription task that within-worker variability affects performance and throughput on MTurk.

When aggregating judgments from an increasing number of people, we found that the overall quality decreased. Gold-

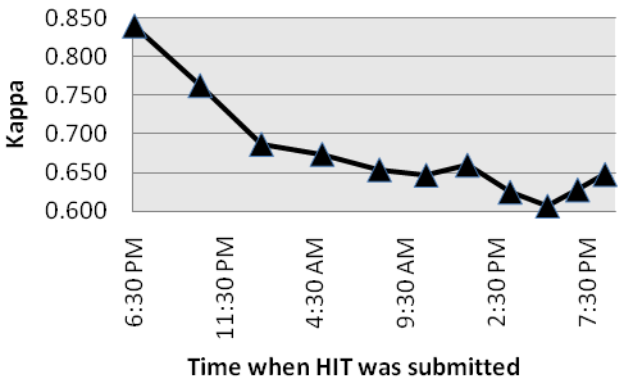

Figure 3 - Agreement with gold-standard over time

standard tasks submitted in the first few hours that the task was available on MTurk had higher agreement than rest of the job submitted (Figure 4). Also, the throughput varied over time: the time required to complete the first 20% of the data was 6 hours, while it took only 3 hours and 10 minutes to complete the last 20%. While this is evidence that within-worker variability is present, it also confounds other variables, such as the number of people working on a task at a certain time and the different pools of workers depending on the time of the day (e.g., more American workers at one time, more Indians at another). This demonstrates the need for a set of experiments that control for these variables, and provide a straightforward way of analyzing the effect of these sources of variability.

Adaptive tasks

Having a better understanding of the sources of within-worker variability will enable us to build models to predict performance and throughput over time. These models can then be used to build a *work distribution system*, which can be responsible for adapting work to each MTurk worker. There are many studies in the human-computation interaction literature that demonstrate that adaptive user interfaces improved the quality of the interaction. In a laboratory experiment, [7] showed that work-flow policies (i.e. management methods to control the way work is distributed between workers on a line of production) have an impact on between-worker and within-worker variability. By understanding this impact, it is possible to define better work distribution systems. We believe that this result applies to micro-task markets as well, and should be exploited to increase performance and throughput.

Tasks can address **within-worker variability** by adapting to the worker over time. Adapting to the learning curve could include providing more feedback while the worker is learning the task (which is implemented in CrowdFlower by

giving more gold-standard instances when a worker is starting to work on a task). The boredom effect could be reduced by detecting decreases in interest and proposing a different type of task or by proposing that the worker take a break. Similarly, a drop in the worker's attention could be addressed by using techniques to regain the worker attention, such as increasing the volume of the speech played to the worker. Another attention getter could be a pop-up window indicating to the worker the number of tasks completed which not only would get his attention but could also foster motivation. The interventions would avoid giving explicit feedback on performance which has been shown by [8] to be counterproductive in some cases.

In the case of **between-worker variability**, the instances given to the workers could be adapted as well. For example, in a speech transcription task, if a worker is known to perform well with noisy speech, he could be given more utterances with noisy speech. The difficulty of the speech utterances given to the workers could also be controlled: for example giving better workers more difficult utterances, thus decreasing between-worker variability. One could also envision ways of adapting the graphical user interface to each worker.

CONCLUSION

While much focus has been put on identifying differences between workers, less research has looked at within-worker variability. Achieving better understanding of the factors that affect the variability of the quality of a worker's judgments over time provides a more reliable confidence measure of the quality of a worker. More importantly, they can be used to design *work distribution systems*, which adapt tasks to take into account within-worker and between-worker variability, thus increasing the performance of micro-task markets.

BIOGRAPHY

Gabriel Parent has a Bachelor of Engineering in Computer Science from École Polytechnique de Montréal. He is currently enrolled as a Master of Language Technologies at Carnegie Mellon University. He organizes a monthly lunch on crowdsourcing, and incorporates crowdsourcing into his main research on adaptive spoken dialog systems.

ACKNOWLEDGMENTS

This work was funded by NSF grant IIS0914927. The opinions expressed in this paper do not necessarily reflect those of NSF.

REFERENCES

1. Snow, R., O'Connor, B., Jurafsky, D. and Ng, AY. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In Proc. of EMNLP-08. 2008.
2. Kimball, O., Kao, C.L., Arvizo, T., Makhoul, J. and Iyer, R. Quick transcription and automatic segmentation of the fisher conversational telephone speech corpus. In RT0. 2004.
3. Novotney, S. and Callison-Burch, C. Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In Proceedings of NAACL. 2010.
4. Parent, G. and Eskenazi, M. Toward better crowdsourced transcription: transcription of a year of the Let's Go bus information system data. In SLT 2010. 2010.
5. Sheng, VS., Provost, F. and Ipeirotis, PG. "Get another label? Improving data quality and data mining using multiple, noisy labelers." In Proceedings of 14th ACM SIGKDD. 2008
6. Pan, C., Shell, R. and Schleifer, M. Performance variability as an indicator of fatigue and boredom effects in a VDT data-entry task. In International Journal of Human-Computer Interaction, Vol. 6, no. 1, pp. 37-45. 1994.
7. Doerr, K. H., Freed, T., Mitchell, T. R., Schriesheim, C. A., and Zhou, X. T. Work flow policy and within-worker and between-worker variability in performance. In Journal of Applied Psychology, 89. 2004.
8. Ipeirotis, P., Get another label? Improving data quality and machine learning using multiple, noisy labelers. Presentation at the CMU crowdsourcing lunch, January 13th 2011. Pittsburgh.