Original post:
http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-05/interview-crowdsourcing/

Interview: Acquiring Corpora using Crowdsourcing

Matthew Marge

SLTC Newsletter, May 2013

Crowdsourcing has become one of the hottest topics in the artificial intelligence community in recent years. Its application to speech and language processing tasks like speech transcription has been very appealing - but what about creating corpora? Can we harness the power of crowdsourcing to improve training data sets for spoken language processing applications like dialogue systems?

William Yang Wang, PhD student at the Language Technologies Institute at CMU, and his colleagues Dan Bohus, Ece Kamar, and Eric Horvitz at Microsoft Research believe we can. Published in the 2012 IEEE Workshop on Spoken Language Technology, they undertook a project to evaluate corpus collection using crowdsourcing techniques. More specifically, they collected sentences from crowdsourced workers with the purpose of building corpora for natural processing applications.

The benefits to collecting corpora using crowdsourcing techniques are numerous -- gathering data is cheap, quick to acquire, varied in nature. But this does not go without carrying risks such as quality control and workers that try to "game" the system. Wang's work explores whether this technique for data collection is a technique others building spoken language processing communities should consider. Wang discussed the project and where he sees this work heading with us.

QUESTIONS

SLTC: *What was the motivation behind collecting corpora using crowdsourcing?*

William: Well, one of the most fundamental and challenging problems in the spoken dialog research is data collection. A true story was that when I initially started my first internship at MSR, we wanted to dig directly into the problem of building a component for open-world dialog systems, but it turned out that getting an appropriate natural language data set that captures the variation in language usage associated with specific users' intentions is actually a non-trivial problem, and we also realized that many people might have encountered the same problem. Therefore, we decided to take a stab at this important problem first.

So, in the early stages, system developers typically use a deployed spoken dialog system to collect natural interaction data from users. However, even before this, data are needed for building the initial system. So this is really a chicken and egg problem. Typically what system

developers do is that they create initial grammars and prompts, either manually or based on small-scale wizard-of-Oz studies. Once this is done and a system is deployed, new data is collected and models and grammars are updated. However, there are several drawbacks with this method. First of all, the initial grammars might not generalize well to real users, and poor system performance in the initial stages can subsequently bias the users' input and the collected data. Secondly, the development lifecycle can have high costs, and refining the system's performance can take a long time. Thirdly, the systems face adoption difficulties in the early stages. Well, I mean it is simply difficult to find initial users, even graduate students, to try out these systems, because of the limited functionality and lack of robustness. Moreover, every time new functionality is added to an existing system, developers have to deal with the challenge of building or acquiring new language resources and expanding grammars. We consider using crowdsourcing to acquire natural language corpora, because of its efficiencies of collecting data and solving tasks via programmatic access to human talent.

SLTC: *Your crowdsourcing task is quite interesting in that it elicits a variety of responses from workers. You give workers some context and ask them to formulate natural language responses (you call this "structural natural language elicitation"). Why did you take this approach?*

William: In this work, we are particularly interested in the problem of crowdsourcing language that corresponds to a given structured semantic form. This is very useful, because interactive technologies with natural language input and output must capture the variation in language usage associated with specific users' intentions. For example, data-driven spoken language understanding systems rely on corpora of natural language utterances and their mapping to the corresponding semantic forms. Similarly, data-driven approaches to natural language generation use corpora that map semantic templates to multiple lexical realizations of that template. Multilingual processing tasks such as machine translation rely on the availability of target language sentences in a parallel corpus that capture multiple valid translations of a given sentence in the source language. While our work shares similarities to previous work in paraphrase generation, they only seek mappings between surface-level realizations of language without knowledge of the underlying structure of the semantics. In contrast, we focus on capturing the mapping from the structured semantic forms to lexical forms. As far as we know, this is probably the first attempt to study the use of crowdsourcing to address this structured natural language elicitation problem.

SLTC: *What were the primary methods you chose to build natural language corpora? Which method did you expect to perform best?*

William: We investigated three methods. In the *sentence-based* method, we present a corresponding natural language sentence of a given semantic, e.g. "Find a Seattle restaurant that serves Chinese food." for the frame of FindRestaurant(City=Seattle; Cuisine=Chinese). In the *scenario-based* method we adopt a story-telling scheme that

presents multiple sentences that form a scenario with a specific goal, e.g. "The goal is to find a restaurant. The city is Seattle. You want to have Chinese food." For the *list-based* method, we present a specific goal, and a set of items corresponding to the slots and values in the form of a list. For instance: Goal: Find restaurant, City: Seattle, Cuisine type: Chinese. Then for each method, we ask the crowd: "What would you say this in your own words?" While in general it is very challenging to evaluate each method with a limited number of empirical experiments, we observed that all these methods provided us accurate, natural, and relatively diverse language from the crowd. I would recommend the list-based method, simply because the seed creation process for list method requires minimum effort. In addition, the crowd workers spent less time understanding the task, because a list is typically shorter than a sentence or a paragraph.

SLTC: *How did you measure performance of the crowdsourced workers? Any good tips for quality control for those working on crowdsourcing projects?*

William: This is an excellent question. Actually, our task is quite different from traditional consensus-based crowdsourcing tasks, so accuracy is not the sole measure. For example, in speech transcription, as long as the crowd workers provide you with the correct transcript, you do not care about how many people have worked on your task. However, in our problem, it is rather different, because in addition to the semantic correctness, we also require the naturalness and the variety of the crowdsourced language. This essentially makes our task a multiobjective optimization problem in crowdsourcing. What made this evaluation more difficult is that there is no reference on the "true" distribution of natural language, so it is not easy to perform automatic real-time validation to filter out the low quality responses from sloppy workers. However, we are lucky because we used Microsoft's Universal Human Relevance System (UHRS), which is a novel vendor-based crowdsourcing platform. UHRS provides relatively high-quality results from full-time workers, and the cost is similar to that of Amazon's Mechanical Turk. In general, we find that it is essential to use repetition to encourage language variations from the crowd, and it is important to distribute the tasks in batches and launch them in different days. This is useful, because we want to avoid the scenarios where the majority of the tasks are performed by a small number of workers, which could bias the language collection. It might also be useful to remind the crowd workers right before the submission button that they are being paid, and they need to provide reasonable responses to avoid being blocked as spammers.

SLTC: *What were the greatest challenges?*

William: Crowdsourcing natural language dataset is still a relatively new research topic. While language technologies researchers have studied crowdsourcing methods for speech transcription, system evaluation, read speech acquisition, search relevance, and translation, most of them belong to consensus tasks. However, a major challenge in our work is that our task requires a certain degree of creativity, in the sense that we need the crowd to create their

own response that can be mapped to the given semantic form. This further brings up two questions: (1) how do we optimize crowdsourcing to encourage variations in the responses, while maintaining the naturalness and semantic correctness at the same time? (2) Since there is no ground truth, how do we evaluate the quality of collected natural language corpora? Our analysis suggests that the crowd can perform the task efficiently and with high accuracy, and that the proposed methods are capable of eliciting some of the natural patterns in language.

SLTC: *What were the lessons you learned from this study?*

William: As I mentioned before, crowdsourcing methods have focused largely on building consensus and accuracy (e.g. transcription). In language elicitation, the objective function is more complex; beyond accuracy (in this case semantic accuracy), we seek to elicit the natural distribution of language usage across a population of users. As such, task design and crowdsourcing controls on the worker population are important. We learned an important lesson in our first experiment: allowing the same worker to address multiple instances of the same task can lead to a lack of diversity. In the second experiment, using crowdsourcing with more controls enabled us to engage a wider population. In principle, engaging a wide population should lead to the construction of a corpus that better captures the natural distribution of language patterns, than when the corpus is authored by a single person, whether that is a crowd worker or a system designer. We seek in future study, a deeper understanding of the tradeoffs between providing enough tasks to attract workers and maintaining engagement, balancing workloads, and developing and refining these controls in a crowdsourcing platform.

Another lesson is the challenge of performing method comparisons based on data collected via crowdsourcing. As important variables (e.g. maximum number of tasks per worker, at which time of the day different tasks will be performed, etc.) cannot be controlled, it is generally challenging to ensure a balanced design for controlled experiments. We believe that repeated experiments and, ultimately, end-to-end evaluations are required to draw robust conclusions. Future work is needed to investigate the performance of deployed spoken dialog systems with language corpora elicited with different methods.

SLTC: *What kind of impact do you foresee this work having on crowdsourcing?*

William: Successful implementations of our methods may enable a number of natural language processing tasks. For instance, for data-driven natural language generation, the methods are sufficient to collect the required data, as the set of semantic forms of interest are known to the system developer. For other tasks, such as developing a spoken dialog system, the distribution of instantiated semantic forms, which is required as an input for our methods, is an important aspect of the corpus that must be collected. This distribution over semantic forms may still be authored, or may be collected by transcribing and annotating interactions of a dialog system with real users. We also foresee the benefits of applying our methods to other

areas such as semantic machine translation, semantic image annotation, and semantic parsing.

SLTC: *What do you feel are the logical next steps for this line of work?*

William: A key direction for fielding more robust automated interactive systems is endowing them with the ability to detect when they do not know or understand some aspects of the world--versus misidentifying things unknown as known and then performing poorly. So following this line of work, we are interested in using the crowdsourced data to build open-world dialogue systems that can detect unknown classes of input, especially in the dynamic, complex, and unconstrained environments. We also plan to study the life-long simulations of the open-world dialogue systems.

SLTC: *Finally, I know that you did this work while on an internship at Microsoft Research. How would you describe that experience?*

William: There is no doubt that MSR is one of the best places for internships. Mentors and collaborators here at MSR are very easy to talk to, and they are very helpful on any questions that you might have. They respect you as a co-worker, listen to your ideas, and actively discuss research topics with you. Also, I had so much fun with other interns in outdoor activities during the Summer -- you will just love this experience.

We look forward to hearing more about crowdsourcing data collection in future work!

REFERENCES

- [1] W. Wang, D. Bohus, E. Kamar, E. Horvitz (2012). "Crowdsourcing the Acquisition of Natural Language Corpora: Methods and Observations," in Proceedings of SLT, Miami, Florida.

If you have comments, corrections, or additions to this article, please contact the author: Matthew Marge, mrma...@cs.cmu.edu.

Matthew Marge is a doctoral student in the Language Technologies Institute at Carnegie Mellon University. His interests are spoken dialogue systems, human-robot interaction, and crowdsourcing for natural language research.