Beyond Mechanical Turk: Using Techniques from Meta Learning to Compare Crowdsourcing Platforms Across Languages

Sarah K. K. Luger
Navera, Inc
101 Montgomery Street, Ste. 350, SF, CA 94104
sarah.luger@navera.com

Abstract
Successful natural language processing (NLP) annotation tasks utilize crowdsourced worker pools that are optimized for the quality, cost, and speed of task completion. Recent work has shown that it is difficult to compare worker pools across platforms because there are insufficient numbers of workers in the geographical locations speaking the languages that are required. This problem is only getting worse with limits on worker location on the dominant crowdsourced worker platform, Amazon’s Mechanical Turk. Preliminary work suggests a general method for measuring the cost in time, money, and worker quality across platforms. This paper suggests a strategy for comparing a series of annotation tasks that seek a more definite, empirical approach. This work in progress proposes running a series of parallel experiments across popular crowdsourcing platforms and using strategies from image processing tasks to compare task difficulty.

Background
Crowdsourcing projects have moved through three phases. The first phase consisted of allocating tasks to colleagues who in the case of academic projects usually included university students. In industrial environments, this meant recording user behavior in both trial and actual settings. With an increase in the utility of annotations in developing machine learning-based algorithms, gathering large-scale annotations from diverse crowds was an invaluable research resource.

The advent of Amazon’s Mechanical Turk, (AMT), (Amazon 2013) ushered in the second phase of crowdsourcing, extending the possible physical location of the crowd workers completing tasks outside of the academic lab or industrial product users. This allowed crowdsourcing tasks to take on much more ambitious problems efficiently, cost-effectively, and at-scale. With the dominance of Mechanical Turk came competitors that provided specialized crowds including those that allowed task managers to optimize for cost, speed, and quality. Crowd management strategies for annotators reaching agreement on judgments and correctly gauging task difficulty, among other issues in human-computer interaction continued to be challenges for researchers.

The third phase began when Mechanical Turk reduced its new crowd member sign-ups to people who are based in the United States or may be paid legally in the United States (new worker registration requires a Social Security Number). The Mechanical Turk platform initially provided a unique avenue to gather human judgments on myriad tasks and its large online presence was vital to NLP annotation projects (Callison-Burch 2009). Thus, for tasks that require international, cultural, or linguistic expertise these limitations have forced project managers to explore the Mechanical Turk alternatives more intently.

The third phase of crowdsourced work strategies must still contend with questions regarding task efficiency especially in regard to maintaining cognitive loads that keep annotators honestly participating and mentally engaged until the end of tasks. Building tasks that are engaging to the human annotators and produce high-quality judgments are difficult. This paper does not attempt to solve this problem, but emphasizes that well-constructed and interesting tasks are ideal.

Identifying Mechanical Turk competitors that best replace Amazon’s offering is challenging. This requires analyzing which specific task components or skills each competitor optimizes. The task components discussed in this paper are those most applicable to NLP-focused annotation
Preliminary Research

A series of annotation tasks covering noun phrase extraction, named entity recognition (NER), part-of-speech (POS) tagging, sentiment analysis, opinion mining, resource allocation, language naturalness, and disaster response in Korean, Spanish, Brazilian Portuguese, and English presented data on which crowd annotation platform provided the best results in each of a series of categories for each task. Not all of the annotation projects were completed in each language, although the breadth of tasks and languages provided a good baseline for testing related efforts. The three crowd platforms used in these tasks were Upwork (Upwork 2015), Amazon Mechanical Turk, and Crowdflower (Biewald and Van Pelt 2013). In addition, the three platforms did not support crowds for all of the languages this research covers. The meta-information recorded for these tasks were speed of completion, cost of completion, availability of a viable annotator pool, and quality of annotators.

The preliminary results showed that across Korean, Spanish, and Brazilian Portuguese, UpWork (formerly Elance-oDesk) provided the highest quality, most cost-effective, and quickest task completion. Mechanical Turk was the best value platform for English, with the cost and quality of the annotators’ work both being very low. (NB: When the tasks began, Crowdflower did not have a large enough Korea-based, Korean language crowd for the tasks. That has now changed.) The low cost of Mechanical Turk annotators allowed the tasks to be cost-effectively annotated by more workers, and the large size of the crowd provided a timely completion of the tasks. Mechanical Turk was the best option of the three platforms for all English language tasks.

Crowdflower provided a value proposition between Mechanical Turk and UpWork. The higher cost of Crowdflower’s workers corresponded with timely task completion but completion time for English was not as fast as with Mechanical Turk. Crowdflower also provided useful, if expensive, metrics on the task completion status and some insight on worker skill. Setting up tasks on Crowdflower takes moderate initial time investment from the task manager’s perspective.

In machine learning, the techniques of “meta learning” (Lemke et al. 2013) are often used to weight metadata in discovering the best solution to a problem where the goal is to optimize the payout and minimize the costs. The hypothesis that each of the three platforms is best in discrete human-in-the-loop annotation use-cases is not incorrect, but using techniques from meta learning to minimize costs and maximize results for web-based annotation has value even outside of NLP applications. The metadata associated with the preliminary annotation tasks provide an interesting constraint satisfaction problem. Decomposing the multiple initial annotation projects into a single task in two languages across three platforms may be first step in exploring how best to optimize annotation tasks. Using meta learning techniques to produce actionable insights promises to provide direction when deciding which crowd to use for which tasks as well as for maintaining quality-consistent results for similar tasks across languages.

Current Work in Progress

In the preliminary experiments, there were no direct comparisons for each task in each language because of the lack of appropriate crowds on the three crowd platforms. That has now changed with Crowdflower gaining a Korean language team. With the loss of Mechanical Turk as a viable option for non-English language workers, a competitive replacement may be the Poland-based 10Clouds, a worker recruitment platform similar to Upwork (10Clouds 2009). In fact, the current lack of one perfect crowdsourcing platform is an opportunity for new data solutions providers and crowd work platforms. These options also include LeadGenius (Kulkarni et al. 2011), Information Evolution (Manning and Ghosh 2007), Spare5 (Spare5 2016), and DefinedCrowd (Braga 2016).

A series of experiments tracking the cost, speed, and annotators’ work quality across two named entity recognition tasks: one English and one Spanish language has been undertaken using crowds on all three platforms. The target data for annotation is 3,000 English and 1,000 Spanish question-answer pairs. The questions come from freely available human resources frequently asked question (FAQ) sections of websites. The questions are 1 sentence in length and the answers are at most 5 sentences long, as shown in Examples 1 and 2.

What is an FSA?

An FSA, or a flexible spending account (also known as a flexible spending arrangement) is a special account you put money into that you use to pay for certain out-of-pocket health care costs. You don’t pay taxes on this money. This means you’ll save an amount equal to the taxes you would have paid on the money you set aside. Employers may make contributions to your FSA, but aren’t required to.

Example 1: A human resources FAQ from an online resource.

For instance, the most interesting components of the Example 1 question-answer pair include the domain-specific concept as an abbreviation (“FSA”), as its common spelling (“flexible spending account”), and its para-
phrase (“flexible spending arrangement”). In Example 2, the definition provided in the answer is specific to how payroll is dispersed and presents a paraphrase (“26 paychecks a year”) and a contrasting option with its paraphrase (“semi-monthly” or “24 times a year”).

**What does biweekly mean?**

In payroll terms, biweekly means you are paid every two weeks, or a total of 26 paychecks a year. Don’t confuse this with semi-monthly, which means you receive paycheck twice each month, or 24 times a year.

**Example 2: A human resources FAQ from an online resource.**

Annotation costs and speed are being compared directly, but the annotator work quality is being examined using Item Response Theory (IRT) (Luger and Bowles 2013) and supervised learning methods such as maximum likelihood estimation (Whitehall et al. 2009) (Raykar et al. 2010). IRT is a performance distribution measure from educational testing theory that has been used in recent crowd research to classify individual worker performance within annotator pools (Christoforaki and Ipeirotis 2014).

A second set of experiments attempts to build a taxonomy from a set of parent nodes and a list of key terms. The data covers terminology and functional areas within corporate human resources. There are 8 parent nodes and 571 key terms to be organized in a tree structure. The data is in English, but the taxonomies will also need to be translated to Spanish and Chinese. There is a gold standard taxonomy that has been authored by two domain experts covering this data set.

The annotators will use the Cascade approach to build the taxonomies (Chilton et al. 2013) from question-answer pairs and then compare them to the gold standard created by the experts. The three steps in taxonomy building with Cascade are annotators generating multiple categories for each question-answer pair, annotators choosing the best category for each pair from the generated list, and then automatting building the taxonomy structure using what Chilton et al. (2013) call global structure inference. Examples 1 and 2 present the type of question-answer pairs used to build the taxonomy.

### Data and Comparisons

The experiments present multiple task types that inherently introduce varying task difficulty. One way to measure task difficulty relies on comparing the different tasks to each other on a spectrum. This “triplet” method uses a three-way comparison approach. Triplets are three items, one from each task and ask “Based on these items, is task a more similar in difficulty to task b than to task c?”

These comparisons are asking humans to provide constraints of the form $d(a, b) < d(a, c)$, where $d(x, y)$ represents some perceptual distance between $x$ and $y$. We refer to these constraints as triplets. Each triplet provides a small unit of information about one point’s location with respect to the other points in human perceptual space. (Wilber et al. 2014)

This method was effectively used for comparing images, but using the same underlying logic to compare NLP and knowledge management tasks would allow a sense of relative difficulty that may be a first step in generating similarly difficult task classes. The comparison will be ranked out of the number of tasks being compared. Ideally, this approach will be formalized into describing classes of crowdsourced tasks.

Examples of the 571 key human resources terms may be found in Table 1.

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Exit Interview</th>
<th>Incentive Pay</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSA</td>
<td>Coaching</td>
<td>Salary</td>
</tr>
<tr>
<td>Bereavement Leave</td>
<td>Accident Compensation</td>
<td>Total Remuneration</td>
</tr>
</tbody>
</table>

**Table 1: Table of examples of unstructured corporate human resources terminology.**

The dimensions of comparison for the crowdsourcing platforms may be seen in Table 2. The platforms are listed on the left and the column directly to the right is annotator work quality. An expert is considered a “1”, thus the work quality metric is based on how many errors an average member of this crowd makes as compared to the gold standard. Next, speed is a measure of the average number of annotations completed in an hour. The cost is per annotation. Then, task difficulty is based on how many tasks have been compared via the triplet method. In this example, there are two other tasks that are more difficult.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Work Quality</th>
<th>Speed</th>
<th>Cost</th>
<th>Task Diff.</th>
<th>GWC Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMT</td>
<td>.4</td>
<td>35/hour</td>
<td>.10</td>
<td>1/3</td>
<td>N</td>
</tr>
<tr>
<td>Crowdflower</td>
<td>.8</td>
<td>40/hour</td>
<td>.12</td>
<td>1/3</td>
<td>N</td>
</tr>
<tr>
<td>Upwork</td>
<td>.7</td>
<td>32/hour</td>
<td>.11</td>
<td>1/3</td>
<td>N</td>
</tr>
</tbody>
</table>

**Table 2: Dimensions of crowdsourced task and platform comparison for the example Spanish NER experiment.**

Finally, the last dimension attempts to highlight whether or not the platforms have a code of ethics with which they manage their workers. The Good Work Code has been adopted by several crowdsourcing platforms as a way to ensure to their customers that the workers they are employing indirectly are being treated fairly (National Domestic
Workers Alliance 2016). This code advocates for workers’ rights to a safe work environment and fair wages. While cost and speed of labor should be measured, this code assures that they are not being optimized at the expense of human dignity. Speed and cost are numbers that are produced by human labor.

Continuing Research

There are numerous dimensions of human computation task measurement that cannot be addressed within this initial method. The goal of this work is to present a methodology that can be replicated in future experiments to measure additional dimensions of crowd work and additional types of tasks. The types of work that members of the crowd perform are much more diverse than NLP-oriented ones. Thus, while initial research is focused on differentiating the comparative difficulty of a series of named entity recognition and taxonomy building tasks, the subsequent experiments will include not only a wider breadth of NLP but also broader classification tasks.

Again, there are many features of human annotation task and worker pool management that this work does not attempt to answer. Research areas including task queuing, user interface optimization, worker behavior monitoring, payment incentivization, and workflow management are usually examined on one platform and the research community would benefit from more extensible, cross-platform analyses. Additionally, since many tools such as AutoMan (Berger et al. 2012) and TurKit (Little et al. 2010), have been built to interface with Amazon’s Mechanical Turk, future work should include comparisons utilizing these optimization tools.

Choosing a crowdsourced labor platform should be a matter of reflecting on the relative goals of the task and knowing which platforms are best suited to fulfilling these goals. While not attempting to consolidate all of the features of the available crowd work platforms or to more broadly produce a general human computation interface, there are distinct parallels to the goal of a Crowdsourcing Compiler (Kearns 2011). The human computation research community has made incremental inroads in identifying strategies to optimize crowdsourced tasks but has yet to formalize comparative platform utility and task difficulty. This work describes a methodology, presents data for current experiments, and suggests metrics for comparing platforms in an effort to formalize multiple dimensions of task measurement.

References


