

# Perspectives on Crowdsourcing Annotations for Natural Language Processing

Aobo Wang · Cong Duy Vu Hoang · Min-Yen Kan

Received: date / Accepted: date

**Abstract** Crowdsourcing has emerged as a new method for obtaining annotations for training models for machine learning. While many variants of this process exist, they largely differ in their method of motivating subjects to contribute and the scale of their applications. To date, however, there has yet to be a study that helps the practitioner to decide what form an annotation application should take to best reach its objectives within the constraints of a project. We first provide a faceted analysis of existing crowdsourcing annotation applications. We then use our analysis to discuss our recommendations on how practitioners can take advantage of crowdsourcing and discuss our view on potential opportunities in this area.

**Keywords** Human Computation · Crowdsourcing · NLP · Wikipedia · Mechanical Turk · Games with a Purpose · Annotation

## 1 Introduction

It is an accepted tradition in natural language processing (NLP) to use annotated corpora to obtain machine learned models for performing many tasks: machine translation, parsing, and summarization. Given that machine learners can only perform tasks as good as their

---

This research is done for CSIDM Project No. CSIDM-200805 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

Aobo Wang  
AS6 04-13 Computing 1, 13 Computing Drive National University of Singapore Singapore 117417  
Tel.: +65-9644-4455  
E-mail: wangaobo@comp.nus.edu.sg

Cong Duy Vu Hoang  
AS6 04-13 Computing 1, 13 Computing Drive National University of Singapore Singapore 117417  
E-mail: hcdvu@comp.nus.edu.sg

Min-Yen Kan  
AS6 05-12 Computing 1, 13 Computing Drive National University of Singapore Singapore 117417  
Tel.: +65-6516-1885  
Fax: +65-6779-4580  
E-mail: kanmy@comp.nus.edu.sg

input annotation, much work in annotation centered on defining high quality standards that were reliable and reproducible, and finding appropriately trained personnel to carry out such tasks. The Penn Treebank and SemCor are probably the most visible examples in this community. Even now, this high quality route continues to be used in other high-profile annotation projects, such as the Penn Discourse TreeBank (Prasad et al, 2008), FrameNet (Baker et al, 1998), PropBank (Kingsbury and Palmer, 2002) and OntoNotes (Pradhan et al, 2007), among others.

An alternative to high quality annotation is to make use of quantity and the rule that redundancy in large data would act to filter out noise. The emergence of the Web made this a real possibility, where raw monolingual and parallel corpora, term counts and user generated content enabled the mining of large amounts of statistical data to train NLP models, both in supervised and unsupervised machine learning modes. In Web 2.0, it is also clear that the Web has made *people* available as resources to take advantage of. This trend reaches one logical conclusion when the web serves to network human service providers with those seeking their services. Although this process is described by many different terms, we use the term *crowdsourcing* throughout this paper. Crowdsourcing is a strategy that combines the effort of the public to solve one problem or produce one particular thing. “Crowdsourcing” has been used in the popular press to emphasize that the workers need not be experts but laymen or amateurs. While human subjects can be used to provide data or services in many forms, we limit our attention in this work on annotations for data useful to NLP tasks, and do not focus on the distributed nature of crowdsourcing.

Crowdsourcing takes many forms that require different forms of motivation to achieve the end goal of annotation. In *Games with a Purpose* (hereafter, GWAP), the main motivator is fun (von Ahn and Dabbish, 2008a,b). Annotation tasks are designed to provide entertainment to the human subject over the course of short sessions. In *Amazon Mechanical Turk* (MTurk), the main motivator is profit. Providers create and list batches of small jobs termed *Human Intelligence Tasks* (HITs) on Amazon’s Mechanical Turk website, which may be done by the general public. Workers who fulfill these tasks get credited in micropayments. While certainly not the only paid labor sourcing environment, Mechanical Turk’s current ubiquity make “MTurk” a useful label to refer to this and other forms of computer mediated labor. *Wisdom of the Crowds* (WotC) is another form of crowdsourcing. WotC deployments allow members of the general public to collaborate to build a public resource, or to predict event outcomes or to estimate difficult to guess quantities. Wikipedia, the most well-known WotC application, has different motivators that have changed over time. Initially, altruism and indirect benefit were factors: people contributed articles to Wikipedia to help others but also to build a resource that would ultimately help themselves. As Wikipedia matured, prestige of being a regular contributor or editor also slowed the ranks of contributors from the crowd to a more stable formalized group (Suh et al, 2009).

It is important to recognize that these different motivators crucially shape each form of crowdsourcing, changing key characteristics. Equally important is to note that the space of possible motivations and dimensions of crowdsourcing have not been fully explored. Given raw linguistic data, what vehicle for annotation would be most fruitful to pursue? Thus far, there has been no systematic scheme for potential projects to follow. We attempt to address these issues in depth. In particular, we:

- deconstruct crowdsourced annotation applications into pertinent dimensions and then give our subjective analysis of published crowdsourced NLP applications using the above dimensions (Section 3);

- 
- analyze the characteristic of the three genres of crowdsourcing approaches and make recommendations for the most effective means of obtaining annotations for sample NLP tasks (Section 4);
  - analyze these crowdsourcing instances and propose better crowdsourcing platforms (Section 5);
  - discuss the future of crowdsourcing annotations in the conclusion (Section 6).

## 2 Related Survey Work on Crowdsourcing

Although crowdsourcing is a fairly recent development, it is recognized as a growing and burgeoning research area, as evidenced by this journal’s special issue as well as several works that have produced an overview of these methods from different perspectives.

Most related in scope is Quinn and Bederson (2009) who describe a taxonomy of crowdsourcing (termed “Distributed Human Computation” in their paper). They divide crowdsourcing into seven genres: GWAP, Mechanized Labor (which we term “MTurk”), Wisdom of Crowds, Crowdsourcing, Dual-Purpose Work, Grand Search, Human-based Genetic Algorithms and Knowledge Collection from Volunteer Contributors. They deconstruct these genres along six dimensions and also discuss future directions towards utilizing crowdsourcing to solve other computational problems. While certainly a useful starting point, the genres are generally defined and do not specifically address the task of annotation. As such, some of their genres are irrelevant and we feel are better combined from an annotation perspective.

In parallel, Yuen et al (2009) also surveyed crowdsourcing applications, categorizing them into six classes: initiatory human computation, distributed human computation, social game-based human computation with volunteers, paid engineers and online players. Their survey analyzes crowdsourcing from a social gaming perspective, differentiating the classes based on game structure and mechanisms. They also touch on the performance aspects in such systems, presenting references to primary studies that describe methods for best measuring the reliability of results coming from crowdsourcing frameworks. Yuen *et al.*’s work, however, does not go beyond its survey to suggest any critiques or analyses of the existing crowdsourcing literature.

Aside from these two surveys that examined crowdsourcing in its wider definition, studies have also analyzed specific theoretical aspects. von Ahn and Dabbish (2008a) present general design principles for GWAPs. They articulate three game classes (or “templates”). Each template defines the game’s basic rules and winning conditions such that it is in the players’ interest to perform the intended computation. They also describe a set of design principles to complement the templates and propose metrics to define GWAP computation success in terms of maximizing the utility obtained per player hour spent. In a similar vein, Jain and Parkes (2009) surveyed existing game-theoretic models for their use in crowdsourcing models, outlining challenges towards advancing theory that can enable better design.

Focusing on the user’s perspective, Roinestad et al (2009) suggests that participation can be improved through either useful tools or entertaining distractions that result in the production of desired data. They developed tools to assist users in managing their web bookmarks in a social setting, providing a tagging game for users to connect resources from each other. This environment helps the system acquire additional knowledge about tags and resources implicitly.

Kazai et al (2009) proposes a method for the collective gathering of relevance assessments using a social game model to instigate users’ engagement. They discuss the approach in detail, and present the results of a pilot study conducted on a book corpus. The analysis

revealed the relationships between the affordances of the system, the incentives of the social game, and the behavior of the assessors.

Although there has been clear interests in harnessing crowdsourcing, the summative work thus far has concentrated on its mechanisms and design. Surveys have described past work in each crowdsourcing mechanism separately, but have yet to compare application instances using a uniform set of criteria. Also, there has been little work to unify these frameworks with respect to annotating natural language data, although there clearly is much explicit interest in this area, as evidenced by recent workshops on the MTurk framework (Callison-Burch and Dredze, 2010) and on collaboratively constructed semantic resources (Gurevych and Zesch, 2009, 2010), aside from this issue.

To our knowledge there has been little that describes which crowdsourcing model is appropriate in which situation or task. It is also not clear when the traditional form of manual annotation by local experts is appropriate – are the days of highly-paid expert annotation over? A NLP practitioner who conducts research and development in need of annotations to train models cannot rely on the existing literature to recommend which form of annotation her specific task should take.

### 3 Comparing Crowdsourcing Applications

Given these limitations of previous work, we first revisit crowdsourcing applications to distill a set of dimensions suitable for characterizing them. This work differs from Quinn and Bederson (2009), as we construct our five dimensions from the practitioner’s perspective. However, there is certainly overlap, as some of their dimensions are preserved in our analysis, which are marked with asterisks in the subsequent headers. There is also certainly correlation between dimensions. Many of our five dimensions – motivation, annotation quality, setup effort, human participation and data character – have internal facets, which we briefly discuss in turn.

A key part of our work is to also assign a range of possible values for a facet. This allows us to “grade” a crowdsourcing example with our subjective opinion, and compare them with traditional, in-house annotation efforts. While opinions will certainly differ on the exact value that should be assigned, the introduction of facet values allows us to compare across applications to uncover trends for success and identify areas that could be better exploited; low scores are not meant to demean an application.

To create the ratings below, each of the authors ( $n=3$ ) independently skimmed and assessed each of the 53 publications that described instantiations of crowdsourcing. Using the description of each of the dimensions that follow, we graded each instance on a scale from 0 to 1, where 0 is given to dimensions that are “not applicable” to the instance, and increasingly positive scores indicate increasing favor to the practitioner. We calculated inter-annotator agreement using Kappa, achieving an overall average agreement for all dimensions of 0.40, which indicates fair to moderate agreement between annotators.

Table 1 shows our classification and scores a few representative crowdsourcing applications based on our analysis below. Due to space limitations, we have omitted the full ratings table in this paper, which we have made available online<sup>1</sup>. Rather than discuss all of the instances in the table, we limit our discussion to the several well-known instances: Phrase Detectives (Chamberlain et al, 2009), a GWAP created to annotate relationships between words and phrases; the set of 5 NLP tasks performed with MTurk (Snow et al, 2008)

<sup>1</sup> <http://wing.comp.nus.edu.sg/crowdsourcing-lrej/>

**Table 1** Sample subjective average scores as assigned by the authors for different instances of the crowdsourcing frameworks. Scores are normalized to unity. Higher scores indicate better value from the practitioner’s perspective. Values are indicative only; numbers have been truncated to one decimal place and are not meant to be definitive. Inter-annotator agreement ( $\kappa$ ) shown in first data column.

Dimensions		$\kappa$	Sample Applications								
			GWAP			MTurk			WotC		
			Phrase Detectives	ESP Game	Average of 24 GWAPs	5 NLP Tasks (Snow et al.)	TREC Blog Assessment	Average of 21 MTurks	Open Mind Initiative	Wikipedia	Average of 8 WotCs
<b>Motivation*</b>	Fun	0.30	0.8	1.0	0.7	0.6	0.2	0.3	0.4	0.3	0.4
	Profit	0.51	0.0	0.0	0.0	0.9	0.5	0.6	0.0	0.0	0.1
	Altruism	0.94	0.2	0.1	0.1	0.0	0.0	0.0	1.0	1.0	0.9
<b>Annotation Quality*</b>		0.62	0.6	0.9	0.7	1.0	0.7	0.7	0.7	0.9	0.7
<b>Setup Effort</b>	Usability	0.80	0.7	1.0	0.8	0.8	0.8	0.7	0.7	1.0	0.7
	Practitioner’s Implementation Cost	0.34	0.3	0.8	0.4	0.7	1.0	0.9	0.6	0.6	0.5
<b>Human Participation</b>	Recognition	0.49	0.3	0.8	0.4	0.9	0.9	0.9	0.7	1.0	0.7
	Worker Base	0.49	0.3	1.0	0.5	0.7	0.4	0.5	0.9	0.9	0.8
<b>Task Character</b>	Data Character	0.40	0.7	1.0	0.6	0.6	0.6	0.5	1.0	1.0	0.9
	Specialization	0.48	0.4	0.3	0.4	0.4	0.5	0.6	0.4	0.9	0.6

(specifically, affective text analysis, word similarity, textual entailment, event annotation and word sense disambiguation); and the WotC Open Mind Initiative, a project to collect commonsense facts for automated, real-world reasoning.

### 3.1 Motivation\*

The dimension that distinguishes crowdsourcing variants is also one of the primary issues a practitioner will need to consider: motivation. Different “personas” (in the marketing sense) of the public are driven to participate by different drivers. We characterize the approaches systems use to engage participants based on how internally motivated they are.

- **Fun** is a significant motivator, and is heavily aligned with the GWAP form of crowdsourcing. Fun applications encourage re-playability and when coupled with a social aspect (users play with/against each other) can drive additional participation. Fun can also hide an ulterior annotation task that could be tedious or complicated (von Ahn and Dabbish, 2008b; Vickrey et al, 2008).

- **Profit** is also another driver, exemplified best by the MTurk framework. In Amazon’s Mechanical Turk, payment is determined by the MTurk practitioner, enabling a tradeoff between participation and cost. Since many MTurk workers come from third-world countries, the amount of payment per annotation is often very low in comparison to in-house or hired annotations (Kittur et al, 2008; Callison-Burch and Dredze, 2010; Snow et al, 2008; Ipeirotis, 2010).

- **Altruism** drives users to annotate for the sake of the system itself. People may annotate or provide information to gain indirect benefit later, as in the case of WotC applications, such as review sites and Wikipedia, where the community benefits as a whole as more users contribute. In some sense, interest and importance can be seen as drivers for the examples in this group. Productively passing time in between television commercial breaks, as noted as a motivator in MTurk (Ipeirotis, 2010), can also be seen as altruism.

Scoring *motivation* in crowdsourcing is dependent on each facet. GWAPs thus tend to score higher on fun, such as the ESPGame (von Ahn and Dabbish, 2004), which we feel is

social, is visually appealing, has levels and time pressure, all which add to the excitement. MTurk tasks vary in how fun they are by task, (*e.g.*, visual tasks for computer vision (Sorokin and Forsyth, 2008) get higher ratings than textual tasks for NLP (Snow et al, 2008; Kaisser and Lowe, 2008)) based on our subjective opinion but the interface to the tasks are largely common and utilitarian, making them more uniform. Profit is only an element of the MTurk form of crowdsourcing, but it is the primary motivator in traditional annotation efforts. Altruistic motivation, if manifested, usually serves as a secondary incentive, and can replace payment in certain cases (*e.g.*, a small charitable donation will be made by the practitioner if the worker performs the annotation).

### 3.2 Annotation Quality\*

A practitioner also needs to choose a framework that matches her minimum level of acceptable quality standards. This aspect has been a central concern in crowdsourcing frameworks, and as crowdsourcing frameworks evolve to become more centrally mediated by computer algorithms, data quality has also become a dimension that can be traded off for other factors. For instance, tradeoff between quality and quantity is commonly involved by many crowdsourcing frameworks. High quality of data collection sometimes means strict or complex annotation rules to follow, which may lower the annotation speed or even prevent annotators from attempting.

Annotation quality is important as the purpose of the annotation is to obtain a high-quality quality annotated resource, to train a machine learned model. One strategy is to have multiple annotators independently agree on the annotation as measured using standard agreement metrics, in the task itself or in a pilot task<sup>2</sup>, or by asking the crowd to validate the acquired annotations in a separate task (a two-stage annotation process), or adjusting the system’s notion of trust of particular workers online (Sheng et al, 2008; Feng et al, 2009). Different thresholds can be set to determine correctness of the output with an arbitrarily high probability (von Ahn and Dabbish, 2004; Vickrey et al, 2008; Snow et al, 2008). Another method is to impose constraints on who may do a task.

A computer mediator (such as those used in GWAP) can be imbued with abilities to track how well an annotator performs, allowing effective pairing between annotators and tasks. In medium- or small-scale tasks, such complexity may not be justified, unless the cost of adoption is minimal; in such cases, a simple thresholding or qualification task (*a la* MTurk) may suffice. Small tasks may opt to perform post-processing agreement calculations on a pilot evaluation, to better tune the expected final annotation quality, as is often done in traditional annotation.

### 3.3 Setup Effort

This dimension measures the effort in creating the annotation interface. Design should keep in mind the end objective of creating a dataset or learned model. Facets for this dimension involve both the worker and the practitioner.

- **Usability** is viewed from the worker’s perspective. A crowdsourcing application must exhibit a minimum level of usability to be able to collect data transparently, without hassle to the worker. These requirements are highest for the GWAP genre, as most games must be

---

<sup>2</sup> In MTurk, the notion of a “qualification test” can be viewed this way.

eminently usable (von Ahn, 2006; von Ahn and Dabbish, 2008b); games that are not usable (or appealing) are simply not played. GWAP applications are also more sensitive to usability in general; ones that are more user-friendly do attract more users. In contrast, the MTurk framework is less sensitive to setup effort, by virtue of its centralization on Amazon. MTurk provides a standardized user interface to potential workers in finding HITs at virtually no cost to the practitioner. While individual HITs can be designed in inspired ways to attract workers, usability enhancements are not immediately visible to workers before starting a HIT, lessening the impact of this factor. Usability impacts WotC and traditional annotation on a case by case basis. We do note that WotC applications often require workers to register and log in, which can be a barrier to providing annotations. Allowing anonymous participation or delayed registration (*i.e.*, after a few annotations are given) can mitigate this issue. A substantial annotation environment also aids usability; toolkits and raw HTML provide good support for text annotation, but provide less well-accepted methods for supporting area or freeform annotation.

In all cases, simplifying usability often decreases task completion times and improves overall satisfaction, yielding a synergistic relationship with participation and annotation efficiency. Difficult annotation tasks – in terms of lengthy instructions or complex tasks – cause dropout and discourage repetition, adversely affecting participation or success rate, as observed by Madnani et al (2010); Koller et al (2010) and others. Simple tasks do best for both motivators of entertainment (GWAP) and micropayments (MTurk). Complex tasks that can be decomposed into a series of simple tasks, when properly checkpointed between stages, also work well in these frameworks (Le et al, 2010; Siorpaes and Hepp, 2008).

- **Practitioner’s Implementation Cost** measures the overhead in creating the annotation platform. This includes overhead for technical issues, including server and database maintenance as well as the software driving the annotation user interface. Most negligible in cost are traditional paper-based surveys, common in small-scale tasks. MTurk, with its centralized system, also brings a low cost to creating generic tasks (Sorokin and Forsyth, 2008; Kittur et al, 2008). These can be fulfilled using HTML (*e.g.*, radio buttons and text form fields (Note that MTurk is not limited to HTML, as Koblin (2006) illustrates, but it is the implementation of least cost). MTurk offers tools and guidelines to help practitioners provide qualification tests and structure HITs appropriately. The MTurk API further allows the practitioner to monitor and manipulate task statistics in their own application. Standalone web-applications and GWAPs, in contrast, have a high cost – the practitioner must define annotation tasks and tools as well as maintain the framework.

In many ways, GWAPs have the greatest cost in both facets: no publicly-available toolkits currently support GWAP creation in both usability/playability (needed to attract participants (von Ahn, 2005)), nor in implementing its underlying support. WotC and web-based local manual annotations also have high costs, but online tools such as web survey hosts and editor plugins mitigate this somewhat. MTurk benefits from its scale, making it a competitive environment versus traditional one-off web-based annotation tasks.

### 3.4 Human Participation\*

A larger user base drives more participation and labeling, and through repetitive annotation, data quality. This dimension assesses the worker scale.

- **Recognition** measures how visible the annotation task is to the general public. While large-impact applications like Wikipedia exist, they are the exception. Using a popular portal

as a bridge is quite helpful for most applications. MTurk wins this dimension as it aggregates thousands of workers and tasks in a single site. Other frameworks are recognized less; GWAP instances by von Ahn, a prime contributor, are grouped in a single website, but does not yet integrate GWAPs from other scholars. Because a worker base is needed for success, this factor measures the public relations necessary to convert the public to annotators. Instances of smaller applications, in GWAPs or web-based or manual annotations depend critically on the recruiting techniques of the practitioner.

- **Worker Base** measures the size and qualifications of the annotator base. The overriding factor here is the worker base size. As an upper bound, Wikipedia has over 150,000 active users, users that have edited a page within the last 30 days<sup>3</sup>. As example GWAPs, the ESP Game has approximately 23,000 players; Phrase Detectives reports over 500 users (von Ahn and Dabbish, 2008a). For MTurk, there are more than 100,000 workers in over one hundred countries<sup>4,5</sup>. A large-scale worker base makes it possible to connect specialized workers to specific tasks. In the NLP domain, only large-scale crowdsourcing solutions like WotC and MTurk, allow practitioners the possibility of obtaining data from under-resourced languages (Zesch et al, 2007; Loup and Ponterio, 2006; Irvine and Klementiev, 2010).

Two demographic studies by Ipeirotis (2008, 2010) of MTurk workers correlated well with the general demographics of Internet users. General demographic studies on Wikipedia users also exist (Voss, 2005), and show that due to its public nature, users self-select to edit articles, which may be an indication of the quality of its contributors.

### 3.5 Task Character

Complementary to the qualities of the workers is the character of the of the annotation task. Here, the scale and subject matter are important facets for consideration.

- **Data Character** measures the size of the data to be annotated, both in terms of items to be annotated as well as the size of individual items, as well as its characteristics relevant to annotation. Depending on the task's purpose, data may be used for a small, individual study, re-used for multiple projects and communities, or large Web-scale dataset such as Google Images, to be visible and useful to more than just the annotation task. Large scale data requires special considerations to be annotated, requiring much active participation and the public's general awareness. Small datasets usually are better served by using some existing infrastructure to alleviate setup costs.

When individual item sizes are large or in a particular medium, they may cause difficulties for the worker. In NLP, such has been the case of using audio data, either in terms of serving audio data to be annotated or asking individuals to provide audio sample recordings (Kunath and Weinberger, 2010; Novotney and Callison-Burch, 2010). Such specialized data requires careful implementation; otherwise worker participation may be adversely affected.

Visual tasks are usually quick, while many monolingual NLP tasks that require reading and assimilating meaning take a longer period of time. Translation tasks and full fledged writing and editing are most exhaustive, requiring a high level of concentration and qualification. A formal training phase, similar to that proposed by Chamberlain et al (2008) can help mitigate this by allowing hands-on experience guide annotators, rather than require the reading of laborious training materials.

<sup>3</sup> [http://en.wikipedia.org/wiki/Main\\_Page](http://en.wikipedia.org/wiki/Main_Page)

<sup>4</sup> <https://www.mturk.com/mturk/welcome>

<sup>5</sup> These statistics for worker base size were current as of July 2010

• **Specialization** measures the specificity of knowledge that an annotation task requires. General human recognition or knowledge can be used for visual tasks or for reverse Turing tests – tests that aim to distinguish humans from automated robots (von Ahn et al, 2008). Many NLP annotations (part of speech tagging) may be acceptable for native speakers of the target language and sometimes not for non-native speakers. The quality and type of annotation may need to consider the target audience in these specialized cases; for example, assessing chunk level dependencies may be plausible for native speakers and to create a parser model returning the same type of output, but the general public is likely not skilled enough to deliver quality Penn Treebank tagging or formal definitions for a dictionary. Specialized tasks such as human translation which prefers a bilingual translator may need to find component tasks that are suitable for a wider audience, in order to lower the barrier of entry for annotation.

In manual annotation tasks, control over these factors is at the practitioner’s sole discretion, whereas in the GWAP and WotC forms, these controls are largely non-existent for the practitioner. However, WotC environments, being community-driven and shaped by human social activities, large initiatives often self-organize and self-employ active, trusted workers as editors that enforce may enforce specialization to a degree.

Being profit driven, MTurk allows some flexibility for the practitioner to introduce filters for the qualifications for workers. This also allows a worker to be more certain that their efforts will be paid (a practitioner may veto work, if she finds it unsatisfactory). Qualifications are often in terms of explicit skills or knowledge, but could be intrinsic qualities of the potential worker. Of particular interest to NLP tasks are a worker’s native and second languages as well as the geographic location, which can figure greatly in multilingual machine translation, speech processing tasks (e.g., Bloodgood and Callison-Burch (2010); Kunath and Weinberger (2010)).

While most MTurk framework profess to have general workers without catering to specific skills, we note that some crowdsourcing sites have become hosts for specific communities of workers, especially programmers<sup>6</sup> and copyediting<sup>7</sup>.

## 4 Recommendations

These dimensions help reason about the efficacy of certain approaches for future crowdsourcing practitioners. Table 1 shows that each of the three crowdsourcing frameworks and manual annotations have some distinguishing values and that each framework’s instances form clusters. For each framework, we assign specific dimensions a “Pro” or “Con” value using our subjective opinion, and then make recommendations on suitable NLP tasks. Our values are from the perspective of the practitioner, where a “Pro” value indicates that the methodology is well-suited on this dimension versus competing frameworks.

### 4.1 Recommendation on GWAP

- **Pros:** Fun, Altruism, Usability
- **Cons:** Practitioner’s Implementation Cost, Recognition, Worker Base, Specialization

<sup>6</sup> e.g., [elance.com](http://elance.com) and [rentacoder.com](http://rentacoder.com).

<sup>7</sup> [editz.com](http://editz.com), formerly [goosegrade.com](http://goosegrade.com).

Although several particular GWAP games have numerous players, most can certainly benefit from more participation. As GWAPs are implemented by many different research labs and decentralized, recruiting workers-as-players for GWAPs is largely a one-off task, and (currently) difficult. Submitting GWAP to free game sites may help jump-start participation; practitioners could also find seed players through social networks or by recruiting players through MTurk (as was done in (Law et al, 2007)’s TagATune and (Ho et al, 2009)’s KissKissBan). While they compete with normal on-line games for players, the added satisfaction of knowing that your game playing goes to help worthy cause may impact participation and the demographics, to some extent.

For GWAP to pay off, a large-scale annotation task is needed to offset the *setup costs*, as the overhead in GUI creation and the software infrastructure needed to maintain artifacts for encouraging fun are expensive. These include high score listings, invitations to encourage a worker’s contacts to participate, and ranks for experienced players. A key missing ingredient to make GWAPs more competitive is the current absence of a free uniform development platform that would cater to many of these mundane tasks. However, UI design and game-play still need to be done individually to give a game its own individuality and its players a sense of ownership.

NLP tasks need to appeal to a general audience to be amenable for GWAP, as the barrier to starting the task must be low and require only a short attention span to complete. We have seen examples of *Coreference Annotation*(Chamberlain et al, 2008, 2009), *Paraphrase Corpora Collection*(Chklovski, 2005) and *Semantic Relations Collection*(Vickrey et al, 2008) that have been moderately successful at this. As GWAPs are primarily motivated by fun, the task should occasionally contain surprises. We believe that problems whose answers could change with different contextual information may be good candidates (*e.g.*, *Named Entity Resolution* and *Discourse Parsing*).

To encourage a larger degree of participation, GWAPs in other fields have relied in part on the viral nature of getting others to participate. Social networking platforms could play a role in creating useful applications for NLP annotations. Short quizzes such as “*How good is your English grammar/vocabulary?*” could serve to compile statistics on common syntactic errors (for *Machine Translation*) and vocabulary familiarity (for *Readability Studies*).

GWAPs have been characterized as taking on three different game structures: output-agreement, inversion-problem, and input-agreement (von Ahn and Dabbish, 2008a). Output-agreement force a player and their partner to agree on their produced outputs to score; input-agreement asks whether the two players received the same stimuli from the system (*e.g.*, in Tag-a-Tune, whether the two player received the same song). The inversion-problem scenario ask one player (the *Guesser*) to reproduce the input of the second player (the *Describer*) using the second player’s output.

In demonstrated GWAPs, we observe that inputs may be complex multimedia (such as pictures or songs) that are infeasible for a human to generate as an answer to a task. Outputs, on the other hand, can be quite simple, as they are to be generated by a player – *e.g.*, a descriptive phrase. This is the case of how a majority of NLP tasks might be cast as well. In such cases, the inversion-problem game is infeasible. An adapted inversion-problem task that asks the guesser to choose among a confusion set may work instead. However, the input- and output-agreement problems are easily catered to such tasks. In many cases, both of these scenarios can be applied to a game and we recommend that both be used within a single game session to encourage more styles of playing, which may lead to more fun.

## 4.2 Recommendation on MTurk

- **Pros:** Recognition, Practitioner’s Implementation Cost, Profit, Specialization
- **Cons:** Fun, Data Character
- **Caveat:** Annotation Quality

While Amazon’s Mechanical Turk is not the only example of the MTurk crowdsourcing framework, its prevalence in research and studies have made it a de facto standard for mechanized labor. MTurk’s hallmarks (low *setup costs*, large *worker base*, and controllable *specialization* of a task’s workforce) allows it to compete very well with traditional and web-based annotations for many tasks. Other MTurk frameworks that serve specific populations, also exhibit these qualities to a lesser degree.

However, tasks requiring true experts or long training periods may not be well-served by such a workforce, and may force a practitioner to go up the continuum of pay to hire contractors at a higher pay rate. There are a growing number of companies – InforSense<sup>8</sup>, 2PiRad<sup>9</sup> and ifiCLAIMS<sup>10</sup> – that now serve this general outsourcing model. Whether the intended study’s funds allows an external party to be paid to broker this task is also a possible issue for public-funded research. Small one-off studies also may still be better served by paper- or web-based surveys.

This leads us to discuss a caveat concerning Annotation Quality. Being uniquely motivated by profit and a viable substantial source of income for certain demographics of workers, there is the financial incentive to cheat on tasks. Almost every annotation design in the MTurk framework needs controls to ensure annotation quality. This can be achieved in ways already discussed: screening workers using acceptance ratings thresholds, using multiple annotators with agreement threshold based on differing criteria, inserting known gold-standard questions to detect spam workers, and using other workers to rate the quality of initial worker annotation.

Cheating is an especially important factor in certain NLP tasks where freely available programs or services can simulate worker competencies. For example, workers asked to provide sentence translations may simulate competence by using services like Google Translate or translation software, defeating the purpose of the annotation task (Bloodgood and Callison-Burch, 2010; Ambati and Vogel, 2010). Paying higher rates for annotation, while adding motivation for genuine workers to complete the annotation task quickly, also incentivizes cheating (Le et al, 2010; Mason and Watts, 2009). For cheating to have little incentive, doing the task properly and cheating must take close to the same level of effort.

A few companies – Crowdfunder<sup>11</sup>, Samasource<sup>12</sup> and CloudCrowd<sup>13</sup> – have created a trust layer over MTurk, by incorporating such safeguards into their system so practitioners can concentrate on their task specification. These companies’ interfaces make it easier to assign gold standard answers to tasks and to view and monitor analytics on a task. These services essentially add points in the MTurk service curve, trading monetary cost for implementation cost relief and annotation quality assurance.

With respect to NLP applications, *POS Tagging* and *Parsing* problems are short but tedious and difficult tasks perhaps requiring little specialization that could benefit from

<sup>8</sup> [www.inforsense.com](http://www.inforsense.com)

<sup>9</sup> <http://www.2pirad.com>

<sup>10</sup> <http://www.ificlaims.com>

<sup>11</sup> [www.crowdfunder.com](http://www.crowdfunder.com)

<sup>12</sup> [www.samasource.com](http://www.samasource.com)

<sup>13</sup> [www.cloudcrowd.com](http://www.cloudcrowd.com)

an MTurk instance. With MTurk’s qualification and use of repetition to achieve a base annotation threshold, difficult tasks requiring specialization or specific linguistic expertise are possible. Also possible are user evaluations of NLP system outputs designed for end users, of which we have recently seen in *Summarization*(Filatova, 2009) and *Machine Translation*(Callison-Burch, 2009).

### 4.3 Recommendation on WotC

- **Pros:** Annotation Quality, Recognition, Altruism, Specialization
- **Cons:** Usability, Fun, Data Character

While Wikipedia and Wiktionary have figured prominently in NLP research, the majority of these studies have studied how to utilize existing WotC resources rather than creating the annotations themselves. Examined from this perspective, existing WotC instances are similar to other existing datasets and corpora – they need to be filtered and adapted for use in a target task.

When we focus on WotC creation, we find a strong bias to compiling resources as opposed to annotations. This is because outputs of WotC applications should have direct relevance to its workers. We find WotC scores highly for *Recognition* and *Specialization*. This implies the annotation tasks in this category may be able to solve tasks that require intensive effort or expertise. To entice annotators to a task, they must feel the indirect benefit of contributing to community knowledge and learning. Many users of a WotC application learn of its existence while looking for information rather than hoping to contribute, which leads to the decreased enthusiasm for contribution. According to Huberman et al (2009), the recognition and visibility of a project is paramount to a WotC application’s survival; ones that fail to sustain an active and growing user base can die off quickly, as workers can quickly sense their efforts are not rewarded.

As a result, a key factor in the success of a WotC instance is whether it can maintain and increase its workers’ enthusiasm, which determines the scale of annotation. To maximize the chance of success, WotC instances should make the most of attention and competition among its worker base. In WotC, productivity exhibits a strong positive dependence on the attention the public have paid to the contribution or contributors. Moreover, contributors compare themselves to others when having low productivity and to themselves when exceeding a personal milestones.

These findings suggest that WotC practitioners should make good use of the two factors of public attention and competitive attitude, to stimulate the contributors’ enthusiasm and increase the productivity in a large scale. Examples of such additional motivation include showing the contributor how many users have benefited from the annotation, displaying a worker’s contribution history and ranking among peers. Searches for annotations within the WotC resource that do not yet exist can prompt a user to provide their take, and search logs that end up with such non-existent resources can be filled in by pushing the tasks to active users. Greater enthusiasm for active contributors could be enabled by ranking the list of the providers by their (different types of) contributions, appealing to their prestige.

Due to the fact that many users of a WotC application learn of its existence while looking for information rather than hoping to contribute, another way to increase the productivity is to convert as many passive users to active contributors as possible. Although the passive users should not be given limited access to the WotC application, measures can be taken to entice them to be a contributor. A promising way is to build an social group among the users

and contributors especially who are involved in the same individual task. When users get used to the contribution system, get noticed by others, receive feedback on their contribution and create relationships with others, they will increase their participation.

From a certain perspective, WotC NLP applications have existed for a while. Informal multilingual and domain-specific dictionaries and thesaurii have been compiled by avid amateurs for many years, accepting emailed contributions. With Web 2.0, technology has eliminated the incorporation time for new contributions<sup>14</sup> Where there may be public and community interest in a resource, a WotC application may be useful. Such annotation tasks should not be time-sensitive, but long-lived, as contributions to WotC often are over a long term, in proportion to the level of *specialization* and current size of the application's existing annotation. Thus we feel *Ontological and Lexical Resource Construction* and *Domain Specific Machine Translation* may be suitable NLP tasks for WotC.

## 5 Discussion

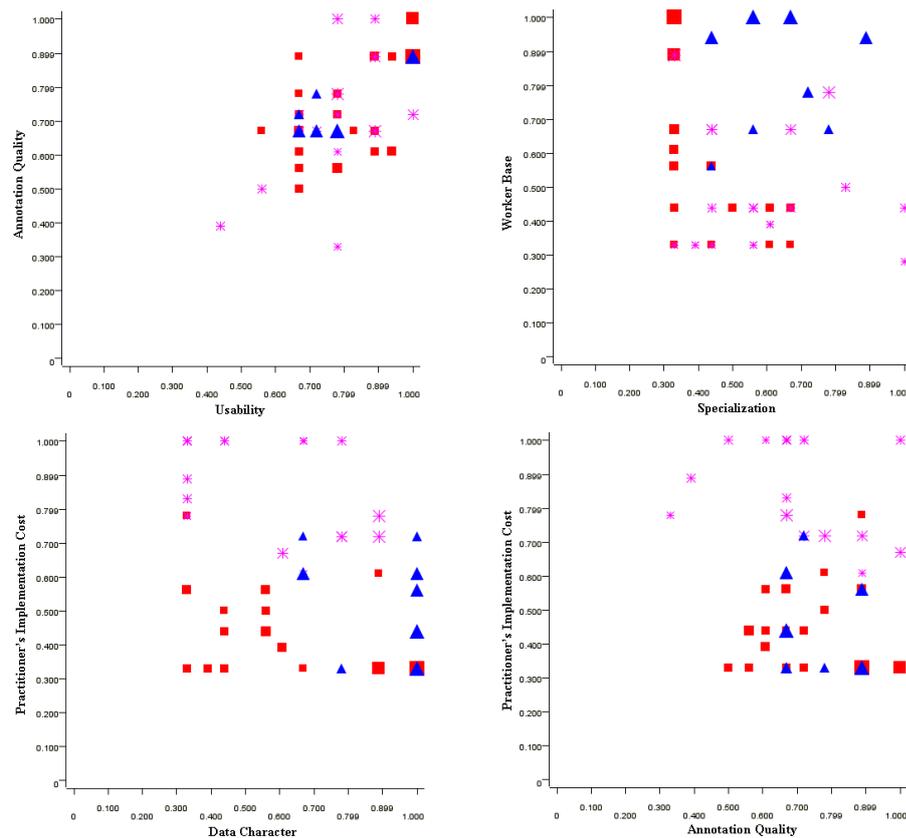
In developing our dimensions, we manually removed dimensions that overlap, favoring orthogonal axes. For example, annotation efficiency of a framework can be determined from an understanding of the participants and the data. Even when we removed these overlaps, there is still considerable correlation, giving us an opportunity to analyze these crowdsourcing instances to see whether better crowdsourcing platforms can be provided. While it is hard to visualize instances in multiple dimensions, plots of our judgments of these crowdsourcing instances in two dimensions are still instructive. Figure 1 plots several of the dimensions against each other for the 53 surveyed crowdsourcing instances. In all of our dimensions, higher scores are interpreted to be better for the practitioner; such that the upper right corner (1.0, 1.0) of the plots represent ideal conditions.

In the upper left plot, we see that annotation quality and usability are highly correlated. Practitioners need to ensure that the tasks are simple and easy, while ensuring that the annotation framework is easy-to-use and can avert common errors through validation. The other three plots highlight the particular strengths of the different crowdsourcing methods. In the upper right plot, GWAP and WotC tasks have stayed true to the layman-as-crowd property, not requiring sophisticated abilities for the most part. MTurk is the exception in which qualification tests allow the recruitment of a "specialist crowd", which can perform difficult tasks. Difficult tasks do correlate with a smaller worker base in general, but in the case of well-known WotC instances, the prestige of contribution may still attract a large worker base to contribute their expert opinion. In both of the bottom plots, we can see the distinct advantage that MTurk has in offering a standardized, centralized framework in decreasing the cost for the practitioner. In the bottom left, we also see that (successful) WotC applications can annotate or build large resources as compared to GWAPs. In the final plot in the bottom right, our study yields a note of concern: we unfortunately do not see a positive correlation between Implementation Cost and Annotation Quality. Practitioners need to be aware that costly setup does not have a correlation with quality annotation; on the contrary, costly implementation may be an artifact of difficult tasks or annotation media, which may result in poor annotation.

Aside from the above plots, we highlight a few other observations from our study.

---

<sup>14</sup> cf Wordnik <http://www.wordnik.com/> and Quora <http://www.quora.com/>.



**Fig. 1** Selected crowdsourced instances plotted along two dimensions. (Clockwise, from upper left) Annotation Quality versus Usability, Worker Base versus Specialization, Practitioner's Implementation Cost versus Annotation Quality, and Practitioner's Implementation Cost versus Worker Base. Red squares, magenta stars, and blue triangles denote GWAP, MTurk and WotC, respectively. Symbols that are larger denote multiple overlapping instances. Scores are author-assigned and subjective and normalized; for all scores, higher is better from the view of the practitioner (e.g., a high Practitioner Implementation Cost means that it is easier for the practitioner to set up the annotation task).

From our judgments, we see that MTurk currently beats out both GWAP and WotC forms in terms of setup costs. Practitioners looking for lightweight, small, one-off annotation tasks, should consider MTurk as a viable alternative to traditional annotation recruitment tasks. Setup costs for the latter two frameworks can be certainly enhanced with the introduction of better toolkits.

Standardization in MTurk allows for low setup costs, but makes the range of tasks a bit more limiting. Where MTurk really does win for annotators is in its Worker Base, a factor that lets annotation tasks complete quickly. This factor alone makes it plausible to design a series of pilot annotation tasks before running the final, well-designed and calibrated task at a large scale.

Worker availability and sheer numbers in MTurk allow practitioners to get a large number of items annotated but by non-experts. Can many non-experts reach the quality level of experts? The answer to this in the MTurk literature is mixed, but most studies have con-

cluded “yes”. Especially when efforts are made to filter out cheaters and obviously noisy annotations, the use of smart, multi-annotator merging strategy such as ROVER (Fiscus, 1997) can deliver performance close to or exceeding intra-expert agreement levels (Lawson et al, 2010; Mellebeek et al, 2010).

GWAPs require a large initial effort to create, especially in GUI and game strategy design. While interesting, this form of crowdsourcing still requires more work for practitioners currently. However, the number of players may not be adequate for some tasks, which may be due to their lack of visibility, but also because the games are not as entertaining as compared to their professionally-designed kin. Promotion of GWAPs through social networks Kuo et al (2009), or via mobile platforms may be viable solutions.

Finally, the primary constraints for a practitioner are often time and money. Strong time constraints make both pure GWAP and WotC forms impractical, as there is little direct influence a practitioner can leverage to increase participation. Monetary budget can be used to incentivize these types of annotation tasks, either directly or indirectly by charity or lottery.

## 6 Conclusions and Outlook

We have examined crowdsourcing in its wider meaning, as a vehicle for obtaining annotations from the general public. We have paid particular attention towards understanding these tasks from the perspective of the practitioner who needs to get data annotated, with special attention on natural language processing (NLP) tasks. In particular, we have assessed many crowdsourcing instances in this paper and assigned subjective scores along dimensions of important to the practitioner. While the scores are only indicative<sup>15</sup>, it has uncovered particular strengths and weaknesses of the different crowdsourcing methodologies.

In the current state of affairs, Amazon’s Mechanical Turk and others of its ilk are strong contenders in performing one-off annotation tasks as well as complex tasks that can be easily decomposed, thanks to their unified presentation, tools and large, centralized worker base. Games with a Purpose (GWAP) and Wisdom of the Crowds (WotC) applications also work for niche applications where it is possible for the annotation task to be entertaining or useful to the community as a whole.

The space of crowdsourcing is no doubt evolving, and our study points out that each framework is distinct in character. A clear trend in the development of crowdsourcing is that the space of possible annotation platforms is expanding to include many more points that allow practitioners to trade off costs in one dimension for another. Such hybrid methods may address weaknesses of individual frameworks while synergistically retaining the advantages. The literature already documents instances where the dimensions of annotation quality, quantity and cost can be traded off. As the community adapts work from other areas where human judgment has played a more central role – e.g., trust models in collaborative filtering (O’Donovan and Smyth, 2005; Massa and Avesani, 2007) – we expect formal models of user annotation to supplant the heuristic methods currently being reported.

Some forms of crowdsourcing have weaknesses that we feel could be addressed in the near future. For example, we believe that a GWAP toolkit may alleviate the current prohibitive cost of entry to the genre. New mediums of interchange and annotation have already started that do not use the static web as the vehicle: using the mobile phone (Eagle, 2009) and the web browser itself – in the form of a browser extension (Griesi et al, 2007) – are platforms to be utilized.

<sup>15</sup> It would have been an interesting exercise to crowdsource the ratings task itself and achieve statistically significant sample size to give more definitive results, but our time and budget did not allow this.

The space of possible configurations also applies to motivation. Current crowdsourcing frameworks, as we have defined it, largely differentiate by a single motivation factor, but that does not prevent future applications from fusing some of these motivators together. A few instances of crowdsourcing have incentivized users by combining two of the three motivational dimensions of fun, profit and altruism. We note that NLP has a unique attribute that lends for the motivation factor of *self-enrichment*: language learning. Although it is hard for language learner to correctly annotate text in a language they are not native speaker of, novel methods may assist such learners in providing useful annotation or translation. For example, instead of demanding full-text translations or annotations, candidate answers provided by other users or machine translation tools can be improved by a worker who may be a language learner.

The above observation leads us to an interesting conclusion that machine systems and humans can work synergistically in certain problem areas where systems have general competencies in coverage but where performance is lacking in specific sub-portions of the task. Similar active learning, where data may be abundant but labels are scarce or expensive to obtain (Settles, 2009), tighter integration between learning and annotation will lead to models where annotation data is specifically chosen to address weaknesses in the learned model. Crowdsourced work explicitly aimed at helping develop machine agents in the guise of the Semantic Web are also beginning to take shape Siorpaes and Hepp (2008). In the translation scenario above, identifying problematic areas for translation systems could be done by crowdsourced inspection of translation output. Reinforcing examples in these areas can then be solicited from workers to fix such errors. Research in this area is in its nascent stage but both toolkits for specific application areas, and integration of crowdsourcing directly into statistical learning framework seem promising (Chang, 2010; Quinn et al, 2010).

We note in closing that Web 2.0 made the web social, connecting people with people. Current crowdsourcing frameworks play along this line, connecting workers to practitioners. Akkaya et al (2010) shows that MTurk workers are anonymous – coming and going, generally not learning nor contribution beyond their atomic interaction with the tasks. Future crowdsourcing is likely to also connect workers to workers and practitioners to practitioners, incorporating more robust reputation models. We feel this emphasis on the social aspects will make social networking platforms and API key features in the next crowdsourcing framework generation.

**Acknowledgements** We would like to thank many of our colleagues who have taken time off their tight schedules to help review and improve to this paper, including Yee Fan Tan, Jesse Prabawa Gozali, Jun-Ping Ng, Jin Zhao and Ziheng Lin.

## References

- von Ahn L (2005) Human computation. PhD thesis, CMU, USA, URL <http://reports-archive.adm.cs.cmu.edu/anon/2005/CMU-CS-05-193.pdf>
- von Ahn L (2006) Invisible computing - games with a purpose. IEEE Computer Magazine, 2006; 39 (6) pp 92–94, URL <http://www.cs.cmu.edu/~biglou/ieee-gwap.pdf>
- von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA, pp 319–326, DOI <http://doi.acm.org/10.1145/985692.985733>

- 
- von Ahn L, Dabbish L (2008a) Designing games with a purpose. *Commun ACM* 51(8):58–67, DOI 10.1145/1378704.1378719, URL <http://dx.doi.org/10.1145/1378704.1378719>
- von Ahn L, Dabbish L (2008b) General techniques for designing games with a purpose. *Communications of the ACM* 51(8):58–67, DOI <http://doi.acm.org/10.1145/1378704.1378719>, URL [http://www.cs.cmu.edu/~biglou/GWAP\\_CACM.pdf](http://www.cs.cmu.edu/~biglou/GWAP_CACM.pdf)
- von Ahn L, Maurer B, McMillen C, Abraham D, Blum M (2008) reCAPTCHA: Human-based character recognition via web security measures. *Science* p 1160379, URL [http://www.cs.cmu.edu/~biglou/reCAPTCHA\\_Science.pdf](http://www.cs.cmu.edu/~biglou/reCAPTCHA_Science.pdf)
- Akkaya C, Conrad A, Wiebe J, Mihalcea R (2010) Amazon mechanical turk for subjectivity word sense disambiguation. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Association for Computational Linguistics, Los Angeles, pp 195–203, URL <http://www.aclweb.org/anthology/W10-0731>
- Ambati V, Vogel S (2010) Can crowds build parallel corpora for machine translation systems? In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Association for Computational Linguistics, Los Angeles, pp 62–65, URL <http://www.aclweb.org/anthology/W10-0710>
- Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet Project. In: *Proceedings of COLING-ACL*, Montreal, Canada, pp 86–90
- Bloodgood M, Callison-Burch C (2010) Using mechanical turk to build machine translation evaluation sets. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Association for Computational Linguistics, Los Angeles, pp 208–211, URL <http://www.aclweb.org/anthology/W10-0733>
- Callison-Burch C (2009) Fast, cheap, and creative: Evaluating translation quality using amazon’s mechanical turk. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, Singapore, pp 286–295
- Callison-Burch C, Dredze M (2010) Creating speech and language data with amazon’s mechanical turk. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Association for Computational Linguistics, Los Angeles, pp 1–12, URL <http://www.aclweb.org/anthology/W10-0701>
- Chamberlain J, Poesio M, Kruschwitz U (2008) Phrase detectives: A web-based collaborative annotation game. In: *In proceeding of the International Conference on Semantic Systems, iSemantics 2008*, Austria, URL [http://www.anawiki.org/phrasedetectives\\_isem08.pdf](http://www.anawiki.org/phrasedetectives_isem08.pdf)
- Chamberlain J, Kruschwitz U, Poesio M (2009) Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In: *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*, Association for Computational Linguistics, Suntec, Singapore, pp 57–62, URL [http://www.aclweb.org/anthology/W/W09/W09-3309\\_n](http://www.aclweb.org/anthology/W/W09/W09-3309_n)
- Chang J (2010) Not-so-latent dirichlet allocation: Collapsed gibbs sampling using human judgments. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Association for Computational Linguistics, Los Angeles, pp 131–138, URL <http://www.aclweb.org/anthology/W10-0720>

- Chklovski T (2005) Collecting paraphrase corpora from volunteer contributors. In: K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture, ACM, New York, NY, USA, pp 115–120, DOI <http://doi.acm.org/10.1145/1088622.1088644>
- Eagle N (2009) txteagle: Mobile crowdsourcing. In: Internationalization, Design and Global Development, Lecture Notes in Computer Science, vol 5623, Springer
- Feng D, Besana S, Zajac R (2009) Acquiring high quality non-expert knowledge from on-demand workforce. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Association for Computational Linguistics, Suntec, Singapore, pp 51–56, URL <http://www.aclweb.org/anthology/W/W09/W09-3308>
- Filatova E (2009) Multilingual wikipedia, summarization, and information trustworthiness. In: SIGIR Workshop on Information Access in a Multilingual World., Boston, Massachusetts, URL <http://storm.cis.fordham.edu/~filatova/PDFfiles/FilatovaCLIR2009.pdf>
- Fiscus JG (1997) A post-processing system to yield word error rates: Recognizer output voting error reduction (rover). In: ASRU
- Griesi D, Paziienza MT, Stellato A (2007) Semantic turkey – a semantic bookmarking tool. In: The Semantic Web: Research and Applications, 4th European Semantic Web Conference (ESWC 2007), Lecture Notes in Computer Science, vol 4519, Springer, pp 779–788, system Description
- Gurevych I, Zesch T (eds) (2009) Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. Association for Computational Linguistics, Suntec, Singapore, URL <http://www.aclweb.org/anthology/W/W09/W09-33>
- Gurevych I, Zesch T (eds) (2010) Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. COLING, Beijing, China
- Ho CJ, Chang TH, Lee JC, Hsu JYj, Chen KT (2009) Kisskissban: a competitive human computation game for image annotation. In: Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09), ACM, New York, NY, USA, pp 11–14, URL <http://doi.acm.org/10.1145/1600150.1600153>
- Huberman B, Romero D, Wu F (2009) Crowdsourcing, attention and productivity. *Journal of Information Science* 35(6):758–765
- Ipeirotis P (2008) Mechanical Turk: The Demographics. URL <http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html>
- Ipeirotis P (2010) New demographics of Mechanical Turk. URL <http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html>
- Irvine A, Klementiev A (2010) Using mechanical turk to annotate lexicons for less commonly used languages. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 108–113, URL <http://www.aclweb.org/anthology/W10-0717>
- Jain S, Parkes DC (2009) The role of game theory in human computation systems. In: Bennett P, Chandrasekar R, Chickering M, Ipeirotis PG, Law E, Mityagin A, Provost FJ, von Ahn L (eds) KDD Workshop on Human Computation, ACM, pp 58–61, URL <http://dblp.uni-trier.de/db/conf/kdd/hcomp2009.html#JainP09>
- Kaisser M, Lowe J (2008) Creating a research collection of question answer sentence pairs with Amazon's Mechanical Turk. In: European Language Resources Association

- 
- (ed) Proceedings of the Sixth International Language Resources and Evaluation (LREC '08), Marrakech, Morocco, URL [http://www.lrec-conf.org/proceedings/lrec2008/pdf/565\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/565_paper.pdf)
- Kazai G, Milic-Frayling N, Costello J (2009) Towards methods for the collective gathering and quality control of relevance assessments. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009), ACM, pp 452–459, URL <http://dx.doi.org/10.1145/1571941.1572019>
- Kingsbury P, Palmer M (2002) From treebank to propbank. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC '02), Las Palmas, Spain
- Kittur A, Chi EH, Suh B (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, NY, USA, pp 453–456, URL <http://doi.acm.org/10.1145/1357054.1357127>
- Koblin A (2006) The sheep market: Two cents worth. Master's thesis, Univ. of California, Los Angeles (UCLA)
- Koller A, Striegnitz K, Gargett A, Byron D, Cassell J, Dale R, Moore J, Oberlander J (2010) Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In: Proceedings of the 6th International Natural Language Generation Conference (INLG), Dublin, Ireland
- Kunath S, Weinberger S (2010) The wisdom of the Crowds Ear: Speech accent rating and annotation with Amazon Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 168–171, URL <http://www.aclweb.org/anthology/W10-0726>
- Kuo YL, Lee JC, Chiang KY, Wang R, Shen E, Chan CW, Hsu JYJ (2009) Community-based game design: experiments on social games for commonsense data collection. In: Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09), ACM, New York, NY, USA, pp 15–22, URL <http://doi.acm.org/10.1145/1600150.1600154>
- Law ELM, von Ahn L, Dannenberg RB, Crawford M (2007) TagATune: A game for music and sound annotation. In: In Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR, URL <http://www.cs.cmu.edu/~elaw/papers/ISMIR2007.pdf>
- Lawson N, Eustice K, Perkowitz M, Yetisgen-Yildiz M (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, Association for Computational Linguistics, Los Angeles, pp 71–79, URL <http://www.aclweb.org/anthology/W10-0712>
- Le A, Ajot J, Przybocki M, Strassel S (2010) Document image collection using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 45–52, URL <http://www.aclweb.org/anthology/W10-0707>
- Loup J, Ponterio R (2006) On the net - Wikipedia: A multilingual treasure trove. *Language Learning and Technology* 10:4–7
- Madnani N, Boyd-Graber J, Resnik P (2010) Measuring transitivity using untrained annotators. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and

- Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 188–194, URL <http://www.aclweb.org/anthology/W10-0730>
- Mason W, Watts DJ (2009) Financial incentives and the “performance of crowds”. In: Bennett P, Chandrasekar R, Chickering M, Ipeirotis PG, Law E, Mityagin A, Provost FJ, von Ahn L (eds) KDD Workshop on Human Computation, ACM, pp 77–85, URL <http://dblp.uni-trier.de/db/conf/kdd/hcomp2009.html#MasonW09>
- Massa P, Avesani P (2007) Trust-aware recommender systems. In: Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07), ACM, New York, NY, USA, pp 17–24, URL <http://doi.acm.org/10.1145/1297231.1297235>
- Mellebeek B, Benavent F, Grivolla J, Codina J, R Costa-Jussà M, Banchs R (2010) Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 114–121, URL <http://www.aclweb.org/anthology/W10-0718>
- Novotney S, Callison-Burch C (2010) Crowdsourced accessibility: Elicitation of wikipedia articles. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, pp 41–44, URL <http://www.aclweb.org/anthology/W10-0706>
- O'Donovan J, Smyth B (2005) Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05), ACM, New York, NY, USA, pp 167–174, URL <http://doi.acm.org/10.1145/1040830.1040870>
- Pradhan SS, Hovy E, Marcus M, Palmer M, Ramshaw L, Weischedel R (2007) OntoNotes: A unified relational semantic representation. In: Proceedings of the First IEEE International Conference on Semantic Computing, Irvine, CA, USA
- Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B (2008) The Penn Discourse Treebank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)
- Quinn AJ, Bederson BB (2009) A taxonomy of distributed human computation. Tech. rep., Univ. of Maryland, College Park
- Quinn AJ, Bederson BB, Yeh T, Lin J (2010) CrowdFlow: Integrating machine learning with mechanical turk for speed-cost-quality flexibility. Tech. Rep. HCIL-2010-09, University of Maryland, College Park,
- Roinestad H, Burgoon J, Markines B, Menczer F (2009) Incentives for social annotation. In: Proceedings of the Twentieth ACM Conference on Hypertext and Hypermedia (HT '09), ACM, New York, NY, USA, URL <http://informatics.indiana.edu/files/Papers/ht09-gal-demo.pdf>
- Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison
- Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. In: KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp 614–622, URL <http://doi.acm.org/10.1145/1401890.1401965>
- Siorpaes K, Hepp M (2008) OntoGame: Weaving the semantic web by online games. In: The Semantic Web: Research and Applications, Springer Berlin / Heidelberg, Lecture Notes in Computer Science, vol 5021/2008, pp 751–766, URL <http://www.springerlink.com/content/k0q415u721011510/>

- 
- Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, pp 254–263, URL <http://www.aclweb.org/anthology-new/D/D08/D08-1027.pdf>
- Sorokin A, Forsyth D (2008) Utility data annotation with amazon mechanical turk. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08, Anchorage, AK., pp 1–8, URL [http://vision.cs.uiuc.edu/~sorokin2/papers/cvpr08\\_annotation.pdf](http://vision.cs.uiuc.edu/~sorokin2/papers/cvpr08_annotation.pdf)
- Suh B, Convertino G, Chi EH, Pirolli P (2009) The singularity is not near: slowing growth of wikipedia. In: WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, ACM, New York, NY, USA, pp 1–10, DOI <http://doi.acm.org/10.1145/1641309.1641322>
- Vickrey D, Bronzan A, Choi W, Kumar A, Turner-Maier J, Wang A, Koller D (2008) Online word games for semantic data collection. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, pp 533–542, DOI <http://www.aclweb.org/anthology/D08-1056>, URL <http://www.stanford.edu/~dvickrey/game.pdf>
- Voss J (2005) Measuring Wikipedia. International Conference of the International Society for Scientometrics and Informetrics : 10th, Stockholm (Sweden), 24-28 July 2005
- Yuen MC, Chen LJ, King I (2009) A survey of human computation systems. Computational Science and Engineering, IEEE International Conference on 4:723–728, DOI <http://doi.ieeecomputersociety.org/10.1109/CSE.2009.395>
- Zesch T, Gurevych I, Mühlhäuser M (2007) Analyzing and accessing wikipedia as a lexical semantic resource. In: Biannual Conference of the Society for Computational Linguistics and Language Technology, pp 213–221