Improving the quality of crowdsourced annotations

Gianluca Demartini
University of Sheffield
gianlucaodemartini.net/crowdsourcing
Gianluca Demartini

- BSc, MSc at U. of Udine, Italy
- PhD at U. of Hannover, Germany
  - Entity Retrieval
- Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- **Lecturer in Data Science** at the iSchool, U. of Sheffield
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 13, ISWC 13, SearchSolutions 2015

www.gianlucaademartini.net
Research Interests

• **Entity-centric Information Access** (2005-now)
  – Structured/Unstructured data (SIGIR 12), TRank (ISWC 13)
  – NER in Scientific Literature (WWW 14) Prepositions (CIKM 14)

• **Hybrid Human-Machine Systems** (2012-now)
  – ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  – Memory-based Information Systems (WWW 14, PVLDB)

• **Better Crowdsourcing Platforms** (2013-now)
  – Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  – Scale-up Crowdsourcing (HCOMP 14), Dynamics (WWW 15)
Aims and objectives

• Quality assurance techniques in crowdsourcing platforms

• Crowd answer aggregation

• The human dimension of crowdsourcing
  – Worker behaviors
  – Modeling skills and knowledge
Schedule

- Design HITs with quality in mind
- Crowd answer aggregation techniques
- Worker profiling and selection
- Malicious worker behaviors
- Worker types
Design HITs with quality in mind
A Task on MTurk

Choose the best category for this image

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.
High-level Issues in Crowdsourcing

• Process
  – Experimental design, annotation guidelines, iteration

• Choose crowdsourcing platform (or roll your own!)

• Human factors
  – Payment / incentives, interface and interaction design, communication, reputation, recruitment, retention

• Quality Control / Data Quality
  – Trust, reliability, spam detection, consensus labeling
Task Design

• Ask the right questions
• Workers may not be experts so don’t assume the same understanding in terms of terminology
• Instructions matter!
• Show examples
• Hire a technical writer
  – Engineer writes the specification
  – Writer communicates
Task Design - UI

• Generic tips
  – Experiment should be self-contained.
  – Keep it short and simple. Brief and concise.
  – Be very clear with the task.
  – Engage with the worker. Avoid boring stuff.
  – Always ask for feedback (open-ended question) in an input box.
Bad Example

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff

Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the Film & Theater how-to video => 332492 Get a 35mm film look with a depth of field adapter
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete 25 video descriptions in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is not necessary to repeat the headline in your entry. It will NOT count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is IRRELEVANT.
10. Do NOT describe the video in the following manner. “She turns around to face the camera. Then she faces left.” Follow the examples below.

Current Word Count: 0 Current Character Count: 0 / 2000

Criteria for REJECTION:

1. Entries with obvious and multiple spelling or grammatical errors will be rejected.
2. Entries with fewer than 100 words will be automatically rejected.
3. Entries copied from the web or other places will be rejected. Multiple plagiarized answers will lead to being BLOCKED. You may use a quotation, but the majority of your content must be ORIGINAL.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being blocked.
5. Tasks submitted without descriptions will be rejected.
6. Tasks submitted with inaccurate descriptions will be rejected as well.
7. Do NOT add any personal opinions. Entries with personal opinions or reviews will be automatically REJECTED.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in rejection.
9. Entries that transcribe the video will be REJECTED.
Good Example

• All information is available
  – What to do
  – Search result
  – Question to answer
Form and Metadata

• Form with a close question (binary relevance) and open-ended question (user feedback)
• Clear title, useful keywords
• Workers need to find your task
How Much to Pay?

• Price commensurate with task effort
  – Ex: $0.02 for yes/no answer + $0.02 bonus for optional feedback

• Ethics & market-factors
  – e.g. non-profit SamaSource contracts workers refugee camps

• Uptake & time-to-completion vs. Cost & Quality
  – Too little $$, no interest or slow
  – too much $$, attract spammers

• Accuracy & quantity
  – More pay = more work, not better (W. Mason and D. Watts, 2009)
Quality Control

• Extremely important part of the experiment
• Approach as “overall” quality; not just for workers
• Bi-directional channel
  – You may think the worker is doing a bad job.
  – The same worker may think you are a lousy requester.
Quality Control

• Approval rate: easy to use, & just as easily defeated
• Mechanical Turk Masters
  – Recent addition, only for specific tasks
• Qualification test
  – Pre-screen workers’ ability to do the task (accurately)
• Assess worker quality as you go
  – Trap questions with known answers (“honey pots”)
  – Measure inner-annotator agreement between workers
Qualification tests: pros and cons

• Advantages
  – Great tool for controlling quality
  – Adjust passing grade

• Disadvantages
  – Extra cost to design and implement the test
  – May turn off workers, hurt completion time
  – Refresh the test on a regular basis
  – Hard to verify subjective tasks like judging relevance

• Try creating task-related questions to get worker familiar with task *before* starting task in earnest
Other quality heuristics

• Justification/feedback as quasi-captcha
  – Should be optional
  – Automatically verifying feedback was written by a person may be difficult (classic spam detection task)

• Broken URL/incorrect object
  – Leave an outlier in the data set
  – Workers will tell you
  – If somebody answers “excellent” for a broken URL => probably spammer
Dealing with bad workers

• Pay for “bad” work instead of rejecting it?
  – Pro: preserve reputation, admit if poor design at fault
  – Con: promote fraud, undermine approval rating system

• Use bonus as incentive
  – Pay the minimum $0.01 and $0.01 for bonus
  – Better than rejecting a $0.02 task

• If spammer “caught”, block from future tasks
  – May be easier to always pay, then block as needed
Build Your Reputation as a Requestor

- **Word of mouth effect**
  - Workers trust the requester (pay on time, clear explanation if there is a rejection)
  - Experiments tend to go faster
  - Announce forthcoming tasks (e.g. tweet)
Crowd Worker Communities

Turkopticon.com
Mturkforum.com
Turkernation.com
What can go wrong?

• Low-quality results can be due to:
  – Bad instructions
  – Pay not high enough or too high
  – Not enough assignments: ask multiple times
Schedule

• Design HITs with quality in mind
• **Crowd answer aggregation techniques**

• Worker profiling and selection
• Malicious worker behaviors
• Worker types
Crowd Answer Aggregation Techniques
Redundancy

• Assign the same HIT to multiple workers (e.g., 3 or 5)

• Answer aggregation
  – Majority vote
  – Weighted average of answers
  – ZenCrowd (learn weights for workers)
  – Aggregate based on worker similarity
Majority Vote

• Ask N workers and pick the most popular answer
• Works for multiple-choice questions
  – Relevance judgments
  – Sentiment analysis / supervised machine learning
• For other task use iterations
  – Audio transcription
  – Ask one worker to transcribe, the next to correct, etc.
• Learning weights for workers
Entity Factor Graphs

- **Graph components**
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints

- **Probabilistic Inference**
  - Select all links with posterior prob >\(\tau\)

2 workers, 6 clicks, 3 candidate links
Aggregation based on worker similarity

- “Community-Based Bayesian Aggregation Models for Crowdsourcing”, Venanzi et al., WWW2014.
- Community-based Bayesian aggregation model
- Group workers by the type of errors they do
SQUARE

• A benchmark for crowd answer aggregation
  – Binary choices (e.g., sentiment)
  – Multiple-choices (e.g., relevance, word-sense disambiguation)

• Compares a number of aggregation techniques over a number of tasks

http://ir.ischool.utexas.edu/square/
Other benchmarks

• Simulations
  – BATC - A Benchmark for Aggregation Techniques in Crowdsourcing
  – Understand effect on efficiency and effectiveness
  – Set aggregation parameters
Schedule

• Design HITs with quality in mind
• Crowd answer aggregation techniques

• Worker profiling and selection
• Malicious worker behaviors
• Worker types
Worker profiling and selection
Pull (Traditional) Crowdsourcing

• In MTurk HITS are published on the market
• The first worker willing to do it can take it
• Pro: Fast
• Con: Not necessarily optimal / not the best worker for the task
Push Crowdsourcing

• Pick-A-Crowd: A system architecture that uses Task-to-Worker matching:
  – The worker’s social profile
  – The task context

• Workers can provide higher quality answers on tasks they relate to
In: 22nd International Conference on World Wide Web (WWW 2013)
Matching Models (1/3)– Category Based

• The requester provides a list of categories related to the batch
• We create a subset of pages whose category is in the category list of the batch
• Rank the workers by the number of liked pages in the subset
Matching Models—Expert Finding

- Build an inverted index on the pages’ titles and description
- Use the title/description of the tasks as a key word query on the inverted index and get a subset of pages
- Rank the workers by the number of liked pages in the subset
Matching Models (3/3) – Semantic Based

- Link the context to an external knowledge base (e.g., DBPedia)
- Exploit the underlying graph structure to determine the Hits and Pages similarity
  - Assumption that a worker who likes a page is able to answer questions about related entities
  - Worker who likes a page is able to answer questions about entities of the same type
- Rank the workers by the number of liked pages in the subset

```
SELECT ?x WHERE { <uri(a_i)> ?x <uri(p_i)> }.
```

```
SELECT ?x WHERE { <uri(a_i)> <rdf:type> ?x . <uri(p_i)> <rdf:type> ?x }.
```
Experimental Evaluation

• The Facebook app **OpenTurk** implements part of the Pick-A-Crowd architecture:
  – More than **170 registered workers** participated
  – Over **12k pages** crawled

• Covered both multiple answer questions as well as open-ended questions
  – 50 images with multiple choice question and 5 candidate answers (Soccer, Actors, Music, Authors, Movies, Animes)
  – Answer 20 open-ended questions related to the topic (Cricket)
### My customized list of batches:

<table>
<thead>
<tr>
<th>Batch description</th>
<th>Challenge</th>
<th>Number of tasks</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football players identifications</td>
<td>Facebook Recommend</td>
<td>5</td>
<td>Completed</td>
</tr>
<tr>
<td>What movie is this scene from?</td>
<td>Facebook Recommend</td>
<td>9</td>
<td>31 available</td>
</tr>
<tr>
<td>Comics, mangas and characters</td>
<td>Facebook Recommend</td>
<td>5</td>
<td>41 available</td>
</tr>
</tbody>
</table>

### List of all batches:

<table>
<thead>
<tr>
<th>Batch description</th>
<th>Challenge</th>
<th>Number of tasks</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors identification</td>
<td>Facebook Recommend</td>
<td>8</td>
<td>40 available</td>
</tr>
<tr>
<td>Music bands identification</td>
<td>Facebook Recommend</td>
<td>4</td>
<td>31 available</td>
</tr>
<tr>
<td>Book authors identification</td>
<td>Facebook Recommend</td>
<td>5</td>
<td>48 available</td>
</tr>
<tr>
<td>Cricket questions.</td>
<td>Facebook Recommend</td>
<td>8</td>
<td>11 available</td>
</tr>
</tbody>
</table>
Like vs Accuracy

Worker Accuracy vs Relevant Likes

- Soccer
- Actors
- Music
- Book Authors
- Movies
- Anime
- Cricket
## Evaluation - Comparison With Mechanical Turk

<table>
<thead>
<tr>
<th>Assignment Method</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMT 3</td>
<td>0.66</td>
</tr>
<tr>
<td>AMT 5</td>
<td>0.62</td>
</tr>
<tr>
<td>AMT Masters 3</td>
<td>0.54</td>
</tr>
<tr>
<td>Category-based 3</td>
<td>0.79</td>
</tr>
<tr>
<td>Category-based 5</td>
<td>0.83</td>
</tr>
<tr>
<td>Voting Model $t_i$ 3</td>
<td>0.80</td>
</tr>
<tr>
<td>Voting Model $t_i$ 5</td>
<td><strong>0.85</strong></td>
</tr>
<tr>
<td>Voting Model $A_i$ 3</td>
<td>0.69</td>
</tr>
<tr>
<td>Voting Model $A_i$ 5</td>
<td>0.72</td>
</tr>
<tr>
<td>En. type 3</td>
<td>0.66</td>
</tr>
<tr>
<td>En. type 5</td>
<td>0.79</td>
</tr>
<tr>
<td>1-step 3</td>
<td>0.66</td>
</tr>
<tr>
<td>1-step 5</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Discussion

- Pull vs. Push methodologies in Crowdsourcing
- Pick-A-Crowd system architecture with Task-to-Worker recommendation
- Experimental comparison with AMT shows a consistent quality improvement

“Workers Know what they Like”
OpenTurk.com

- Yet another a platform? Build on top of Mturk!
- Chrome Extension for push / notification
- 400+ users
- Open source: https://github.com/openturk/extension
Schedule

• Design HITs with quality in mind
• Crowd answer aggregation techniques
• Worker profiling and selection
• **Malicious worker behaviors**
• Worker types
Malicious worker behaviours

Challenges

➢ Quality Control Mechanisms
  ○ Diverse pool of crowd workers
  ○ Wide range of behavior
  ○ Various motivations
Malicious Workers

“workers with ulterior motives, who either simply sabotage a task, or provide poor responses in an attempt to quickly attain task completion for monetary gains”

➢ Typically adopted solution to prevent/flag malicious activity: Gold-Standard Questions
➢ Flourishing Crowdsourcing markets, advances in malicious activity

Need to understand workers behavior and types of malicious activity.
We focus on analyzing the malicious behavior of workers in **SURVEYS**
- Subjective nature
- Open-ended questions
- Gold-standards are not easily applicable

---

A Taxonomy of Microtasks on the Web.
**Research Questions**

**RQ#1**
Do untrustworthy workers adopt different methods to complete tasks, and exhibit different kinds of behavior?

**RQ#2**
Can behavioral patterns of malicious workers in the crowd be identified and quantified?

**RQ#3**
How can task administrators benefit from the prior knowledge of plausible worker behavior?
Survey Design

- CrowdFlower Platform to deploy survey
- Survey questions
  - Demographics
  - Educational & general background
- 34 Questions in total
  - Open-ended
  - Multiple Choice
  - Likert-type
- Responses from 1000 crowd workers
  - Monetary Compensation per worker: 0.2 USD
➢ Questions regarding previous tasks that were successfully completed

1. What is the title of a previous task/job you completed on any micro-task platform?

1 (a). What was the description of this task?

1 (b). Please identify at least 5 keywords or tags that represent this task?

➢ 2 Attention-check questions
  ○ Engage workers
  ○ Gold-standard to separate Trustworthy/Untrustworthy workers (we found 568 trustworthy, 432 untrustworthy)

How many times did you slip and fall during your last visit to planet Mars?

0 5 10 15 20
Analyzing Malicious behavior in the Crowd

Based on the following aspects, we investigate the behavioral patterns of crowd workers.

I. eligibility of a worker to participate in a task
II. conformation to the pre-set rules
III. satisfying expected requirements fully
### Behavioral Patterns

| Ineligible Workers (IW) | **Instruction:** Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.  
| **Response:** ‘this is my first task’ |
| Fast Deceivers (FD) | eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.  
| **Response:** ‘What’s your task?’ , ‘adasd’, ‘fgfgf gsd ljlkj’ |
| Rule Breakers (RB) | **Instruction:** Identify 5 keywords that represent this task (separated by commas).  
| **Response:** ‘survey, tasks, history’, ‘previous task yellow’ |
| Smart Deceivers (SD) | **Instruction:** Identify 5 keywords that represent this task (separated by commas).  
| **Response:** ‘one, two, three, four, five’ |
| Gold Standard Preys (GSP) | These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions! |
Observations

We manually annotated each response from the 1000 workers.

- 568 workers passed the gold-standard: Trustworthy workers (TW)
- 432 workers failed to pass the gold-standard: Untrustworthy workers (UW)
- 335 trustworthy workers gave perfect responses: Elite workers
- 665 non-elite workers (233 TW, 432 UT) were manually classified into the different classes according to their behavioral patterns.
Workers Classification

- 73 untrustworthy workers and 93 trustworthy workers were classified into 2 different classes, while the rest were uniquely classified.
- Inter-rater agreement between the experts (according to Krippendorf’s Alpha) : 0.94

Acceptability of Responses

- Inter-rater agreement between the experts (according to Krippendorf’s Alpha) : 0.89
Distribution of Workers

- Untrustworthy Workers
- Trustworthy Workers
- All Workers

Types of Behavior:
- Ineligible Workers
- Fast Deceivers
- Rule Breakers
- Smart Deceivers
- Gold-Standard Preys

Percentage of Workers
Acceptability: “The acceptability of a response can be assessed based on the extent to which a response meets the priorly stated expectations.”

E.g.

Instruction: Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.
Response: ‘survey, tasks, history’ ⇒ ‘0’
Response: ‘previous, job, finding, authors, books’ ⇒ ‘1’

We consider open-ended questions.

$$M_{worker} = 1 - \left(1/n \sum_{i=1}^{n} A_{ri}\right)$$

where, $n$ is the total number of responses from a worker and $A_{ri}$ represents the acceptability of response ‘i’
Degree of maliciousness of trustworthy (TW) and untrustworthy workers (UW) and their average task completion time.
Task Completion Time vs Worker Maliciousness

- Ineligible Workers
- Fast Deceivers
- Rule Breakers
- Smart Deceivers
- Gold-Standard Preys

Types of Malicious Behavior:
- UW Average Time
- TW Average Time
- UW Maliciousness
- TW Maliciousness
Where are the workers from?

![Chart showing the distribution of workers per country]
Tipping Point

“the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response”

Table 1. Relationship between the Maliciousness and Tipping Point of untrustworthy and trustworthy workers (percentage of workers having tipping point @ R).

<table>
<thead>
<tr>
<th>Maliciousness</th>
<th>UW</th>
<th>TW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; M \leq 0.2$</td>
<td>40.9% @ R-7</td>
<td>28.5% @ R-7</td>
</tr>
<tr>
<td></td>
<td>31.8% @ R-6</td>
<td>28.5% @ R-5</td>
</tr>
<tr>
<td>$0.2 &lt; M \leq 0.4$</td>
<td>43.47% @ R-3</td>
<td>30% @ R-5</td>
</tr>
<tr>
<td></td>
<td>21.73% @ R-6</td>
<td>30% @ R-3</td>
</tr>
<tr>
<td>$0.4 &lt; M \leq 0.6$</td>
<td>66.19% @ R-3</td>
<td>88% @ R-4</td>
</tr>
<tr>
<td></td>
<td>25.35% @ R-2</td>
<td>5.1% @ R-3</td>
</tr>
<tr>
<td>$0.6 &lt; M \leq 0.8$</td>
<td>71.05% @ R-2</td>
<td>60% @ R-3</td>
</tr>
<tr>
<td></td>
<td>28.95% @ R-3</td>
<td>40% @ R-2</td>
</tr>
<tr>
<td>$0.8 &lt; M \leq 1$</td>
<td>100% @ R-2</td>
<td>100% @ R-2</td>
</tr>
</tbody>
</table>
Task Design Guidelines

➢ Using the ‘Tipping Point’ for early detection of malicious activity.
➢ Using ‘Malicious Intent’ as a measure to discard unreliable responses from workers and improve the quality of results.

Ineligible Workers
Pre-screening to tackle Ineligible Workers (IW).

Fast Deceivers
Stringent and persistent validators and monitoring worker progress to tackle Fast Deceivers (FD) and Rule Breakers (RB).

Rule Breakers

Smart Deceivers
Psychometric approaches to tackle Smart Deceivers (SD).

Gold Standard Preys
Post-processing to accommodate fair responses from Gold-standard Preys (GSP).
Contributions

✓ Identified different types of malicious behavior exhibited by crowd workers.

✓ Measuring ‘maliciousness’ of workers to quantify their behavioral traits, and ‘tipping point’ to further understand worker behavior.

✓ This understanding helps requesters in effective task design, ensures adequate utilization of the crowdsourcing platform(s).

✓ Guidelines for effective design of Surveys by limiting malicious activity.
Schedule

• Design HITs with quality in mind
• Crowd answer aggregation techniques
• Worker profiling and selection
• Malicious worker behaviors
• Worker types
Worker types
Overview

• Tracking worker behavior to automatically classify worker types
• Use predicted worker type for worker pre-selection
• Train workers who under-perform
• Optimize efficiency and effectiveness of crowd work
Worker behavior (on-going work)

• Behavioral Traces
  – Task completion time
  – Time before first click and keypress
  – Browser Tab switching/opening
  – Window focus
  – Mouse/touchpad scrolling
  – Mouse moves
Worker types

• High quality
  – Diligent Workers (DW)
  – Competent Workers (CW)

• Low quality
  – Ineligible Workers (IW)
  – Fast Deceivers (FD)
  – Smart Deceivers (SD)
  – Rule Breakers (RB)
  – Incompetent Workers (IncW)
  – Sloppy Workers (SW)
Worker performances

![Bar chart showing worker performances](chart.png)
Worker dashboards

(a) A competent worker (CW) in CC task 40x3.
(b) A fast deceiver (FD) in CC task 40x3.
Conclusions

• Crowdsourcing is naturally influenced by human behaviors
• Quality is a 2-sides process that requires effort both from requesters as well as from workers
• The first time you try the results will look bad
• Get feedback and iterate!

gianlucaudemartini.net/crowdsourcing