Adapted Excerpts from
Crowdsourcing 101: Putting the WSDM of Crowds to Work for You

Omar Alonso
Microsoft
9 February 2011

Matthew Lease
University of Texas at Austin
Crowdsourcing

• Take a job traditionally performed by a designated agent (usually an employee)
  – Outsource it to an undefined, generally large group of people via an open call

• New application of many open source principles
Crowdsourcing

• Outsource micro-tasks
• Success stories
  – Wikipedia
  – Apache
• Power law
• Attention
• Incentives
• Diversity
AMT

- Amazon Mechanical Turk (AMT, www.mturk.com)
- Crowdsourcing platform
- On-demand workforce
- “Artificial artificial intelligence”: get humans to do hard part
- Named after “The Turk”, a fake chess playing machine
- Constructed by Wolfgang von Kempelen in 18th C.
Amazon Mechanical Turk
From Wikipedia, the free encyclopedia

- The Amazon Mechanical Turk (MTurk) is a crowdsourcing Internet marketplace that enables computer programmers (known as Requesters) to co-ordinate the use of human intelligence to perform tasks which computers are unable to do.
- The Requesters are able to pose tasks known as HITs (Human Intelligence Tasks), such as choosing the best among several photographs of a store-front, writing product descriptions, or identifying performers on music CDs.
- Workers (called Providers in Mechanical Turk's Terms of Service) can then browse among existing tasks and complete them for a monetary payment set by the Requester.
- To place HITs, the requesting programs use an open Application Programming Interface, or the more limited Mturk Requester site [1].
- Requesters can ask that Workers fulfill Qualifications before engaging a task, and they can set up a test in order to verify the Qualification.
- They can also accept or reject the result sent by the Worker, which reflects on the Worker's reputation.
- Payments for completing tasks can be redeemed on Amazon.com via gift certificate or be later transferred to a Worker's U.S. bank account. Requesters, which are typically corporations, pay 10 percent over the price of successfully completed HITs to Amazon.
Wisdom of Crowds

Requires
• Diversity
• Independence
• Decentralization
• Aggregation

Input: large, diverse sample
(to increase likelihood of overall pool quality)

Output: consensus or selection (aggregation)
vs. Ensemble Learning

• Use multiple models to obtain better performance than from any constituent model
• Often combines many weak learners to produce a strong learner
• Compensate for poor individual learning by performing a lot of extra computation
• Tend to work better when significant diversity
• Using less diverse strong learners has worked better than “dumbing-down” models to increase diversity (Gashler et al.’08)
Human Computation

• Use humans as processors in a distributed system
  – Perform tasks computers aren’t good at
  – Automated system can make “external calls” to the “HPU”
• Reverse: identify tasks computers can’t do (Captcha)
• Examples
  – Games with a purpose (e.g. ESP game)
  – ReCaptcha

Human Computation

- Not a new idea
- Computers before computers
A New World of Application Design

New man+machine hybrid applications blend automation with crowd interaction to achieve new capabilities exceeding components

• CrowdSearch (T. Yan et al., MobiSys 2010)


• Translation by Iteractive Collaboration between Monolingual Users, B. Bederson et al. GI 2010
Pay ($$$)

P. Ipeirotis March 2010

Crowdsourcing 101: Putting the WSDM of Crowds to Work for You.
Examples

• A closer look at previous work with crowdsourcing
• Includes experiments using AMT
• Subset of current research
  – Check the bibliography section for more references
• Wide range of topics
  – NLP, IR, Machine Translation, etc.
NLP

- AMT to collect annotations
- Five tasks: affect recognition, word similarity, textual entailment, event temporal ordering
- High agreement between workers and gold standard
- Bias correction for non-experts

Machine Translation

• Manual evaluation on translation quality is slow and expensive
• High agreement between non-experts and experts
• $0.10 to translate a sentence


B. Bederson et al. Translation by Iteractive Collaboration between Monolingual Users, GI 2010
Data quality

- Data quality via repeated labeling
- Repeated labeling can improve label quality and model quality
- When labels are noisy, repeated labeling can preferable to a single labeling
- Cost issues with labeling

Quality control on relevance assessments

- INEX 2008 Book track
- Home grown system (no AMT)
- Propose a game for collecting assessments
- CRA Method

Page Hunt

• Learning a mapping from web pages to queries
• Human computation game to elicit data
• Home grown system (no AMT)
• More info: pagehunt.msrlivelabs.com

Snippets

- Study on summary lengths
- Determine preferred result length
- Asked workers to categorize web queries
- Asked workers to evaluate the quality of snippets
- Payment between $0.01 and $0.05 per HIT

Timeline annotation

- Workers annotate timeline on politics, sports, culture
- Given a timex (1970s, 1982, etc.) suggest something
- Given an event (Vietnam, World cup, etc.) suggest a timex

Twitter

• Detecting uninteresting content text streams
  — Alonso et al. SIGIR 2010 CSE Workshop.
• Is this tweet interesting to the author and friends only?
• Workers classify tweets
• 5 tweets per HIT, 5 workers, $0.02
• 57% is categorically not interesting
AMT – How it works

• Requesters create “Human Intelligence Tasks” (HITs) via web services API or dashboard
• Workers (sometimes called “Turkers”) log in, choose HITs, perform them
• Requesters assess results, pay per HIT satisfactorily completed
• Currently >200,000 workers from 100 countries; millions of HITs completed
The Worker

• Sign up with your Amazon account
• Tabs
  – Account: work approved/rejected
  – HIT: browse and search for work
  – Qualifications: browse and search for qualifications
Example – Relevance and ads

How relevant are these 25 advertisements to a search term?

Instructions

In this task, you will be given a search term and a small advertisement. Please rate how relevant the advertisement is to the search terms. The scale is from 1 to 4, where 1 is not relevant at all and 4 is completely relevant. Below is a description of each rating.

4 - Completely Relevant Ad
These are often the exact item

3 - Closely Related Ads
An ad for iPod cases would be

2 - Somewhat Related Ads
For instance, an ad for speakers

1 - Irrelevant Ads
Ads that have nothing to do with

Tips

A search query of “**underwear**”
### Example – Spelling correction

#### Evaluate a Spelling Correction for a Product Search Query

**Instructions**

Imagine that a user is searching for products at an online shopping website. When the user searches for a term, the site suggests a spelling correction, such as "Did you mean: XYZ?" Evaluate whether this spelling correction is **GOOD** or **BAD**. If you aren't sure if the suggestion gives the proper spelling or are not familiar with the search terms, select **I DON'T KNOW**.

When evaluating corrections, ignore capitalization. All search terms and corrections are shown in lower case. A correction can be good even if a space is used instead of a hyphen. For example, "blu-ray" and "blu-ray" are both good spelling corrections for "blue ray", even though the trademarked term is "blue-ray".

Sample search results are provided for context. However, you should base your response on the accuracy of the spelling correction, not the relevance of the results.

**Note:** We pay bonuses for high-quality responses! You will earn a bonus if your answer is consistent with the majority of respondents. However, if you consistently disagree with the majority, you will be blocked from participating in our future experiments. (An answer is considered to be the majority response when it’s selected by two-thirds or more of the workers who complete the HIT.)

#### Task

Please evaluate the following spelling correction, using the provided results for context:

<table>
<thead>
<tr>
<th>User’s search query</th>
<th>Suggested correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>enemax</td>
<td>enema</td>
</tr>
</tbody>
</table>

#### Example

- **Photographic Print of Coloured X-ray of the colon from Science Photo Library (kitchen)**
  - productType: HOME_FURNITURE_AND_DECOR
  - superSaver: false
  - averageRating: 5.0

- **Reports on Publications Issued and Registered in the Several Provinces of British India (Paperback)**
  - productType: ANGL_SBOOK
  - superSaver: true
  - averageRating: 3.6

- **Enema Kit: (Personal Care)**
  - productType: HEALTH_PERSONAL_CARE
  - superSaver: false
  - averageRating: 4.0

*Is the correction of enemax to enema GOOD or BAD?*

- **GOOD.** Yes, the suggested spelling correction corrects a misspelling.
- **BAD.** No, the suggested spelling correction is incorrect or unnecessary.
- **I DON'T KNOW.** Not sure if the suggested spelling correction gives the proper spelling, or not familiar with the search terms.
Help Classify Arabic into Dialects!

This task is for Arabic speakers who understand the different local Arabic dialects (اللغات المحلية أو التأريخ)، and can distinguish them from Fusha Arabic (المساحي).

Below, you will see several Arabic sentences. For each sentence:

1. Tell us how much dialect (لغة) is in the sentence, and then
2. Tell us which Arabic dialect the writer intends.

This following map explains the dialects:

PLEASE READ the following. You MUST understand the classifications, otherwise your work might be rejected!!

- **Levantine** (شامي) does NOT mean 'Syrian' only. It includes: Syrian, but ALSO: Jordanian is Levantine, Palestinian is Levantine, and Lebanese is Levantine. That's why all these countries are green in the map.

- **Maghrebi** (مغربى) does NOT mean 'Moroccan' only. It includes: Moroccan, but ALSO: Algerian is Maghrebi, Tunisian is Maghrebi, and Libyan is Maghrebi. That's why all these countries are purple in the map.

- The word 'dialect' does NOT mean 'spelling mistake' (نحو إبلاطه). If the writer was trying to write in 100% standard Arabic, classify it as No dialect, even if it has some spelling mistakes.
Who are the workers?

A. Baio, November 2008

P. Ipeitorotis. March 2010
Who are the workers?

P. Ipeitorotis. March 2010

- 47% US, 34% India, 19% other
The Requester

- Sign up with your Amazon account
- Amazon payments
- Purchase prepaid HITs
- There is no minimum or up-front fee
- AMT collects a 10% commission
- The minimum commission charge is $0.005 per HIT
Dashboard - II

Crowdsourcing 101: Putting the WSDM of Crowds to Work for You.
Quality control - II

• Approval rate
• Qualification test
  – Problems: slows down the experiment, difficult to “test” relevance
  – Solution: create questions on topics so user gets familiar before starting the assessment
• Still not a guarantee of good outcome
• Interject gold answers in the experiment
• Identify workers that always disagree with the majority

Crowdsourcing 101: Putting the WSDM of Crowds to Work for You.
Filtering bad workers

• Approval rate
• Qualification test
  – Problems: slows down the experiment, difficult to “test” relevance
  – Solution: create questions on topics so user gets familiar before starting the assessment
• Still not a guarantee of good outcome
• Interject gold answers in the experiment
• Identify workers that always disagree with the majority
More on quality

• Lots of ways to control quality:
  – Better qualification test
  – More redundant judgments
  – More than 5 workers seems not necessary

• Various methods to aggregate judgments
  – Voting
  – Consensus
  – Averaging
Recent Workshops

• Human Computation: [HCOMP 2009](#) & [HCOMP 2010](#) at KDD

• IR: [Crowdsourcing for Search Evaluation](#) at SIGIR 2010

• NLP
  – [Creating Speech and Language Data With Amazon's Mechanical Turk](#). NAACL 2010
  – [Maryland Workshop on Crowdsourcing and Translation](#). June, 2010

• ML: [Computational Social Science and the Wisdom of Crowds](#). NIPS 2010

• [Advancing Computer Vision with Humans in the Loop](#) at CVPR 2010

• Conference: [CrowdConf 2010](#) (organized by CrowdFlower)
News and Upcoming Events


Special issue of Information Retrieval journal on Crowdsourcing (papers due 4/30)

Upcoming Conferences & Workshops
• HCOMP workshop at AAAI (papers due 4/22)
• SIGIR workshop? (in review)
• CrowdConf 2011 (TBA)

Events & Resources: http://ir.ischool.utexas.edu/crowd
Marketplaces

- Mturk (www.mturk.com)
- Crowdflower (www.crowdflower.com)
- Cloudcrowd, domystuff...
- Other resources:
  - http://blog.turkalert.com/
  - http://www.turkalert.com/
  - http://turkers.proboards.com
Tools and Packages

Common infrastructure layers atop or in place of MTurk or other platforms

• Turklt
• Get Another Label (& qmturk)
• Turk Surveyor
• Ushandi
Thank You!

For questions about tutorial or crowdsourcing, email:

omar.alonso@microsoft.com
ml@ischool.utexas.edu

Cartoons by Mateo Burtch  (buta@mindspring.com)


Bibliography - III

- S. Mizzaro. Measuring the agreement among relevance judges, MIRA 1999