Crowdsourcing for NLP

Vered Shwartz

April 8, 2018
Motivation

YOU GET DATA, YOU GET DATA

EVERYBODY GETS DATA!
Outline

What is Crowdsourcing?
  What can be Crowdsourced?
  Workflow

Crowdsourcing Tips and Tricks
  Aggregating Worker Answers
  Quality Control
  More Tips

Ethical Issues

Amazon Mechanical Turk Demo
What is Crowdsourcing?

- Blend word: crowd (intelligence) + (out-)sourcing
What is Crowdsourcing?

- Blend word: crowd (intelligence) + (out-)sourcing
- Take a task normally performed by experts
What is Crowdsourcing?

Blend word: **crowd (intelligence) + (out-)sourcing**

Take a task normally performed by experts

- e.g. translation, annotating semantic relations between words/sentences, asking and answering questions, etc.
What is Crowdsourcing?

Blend word: crowd (intelligence) + (out-)sourcing

Take a task normally performed by experts

- e.g. translation, annotating semantic relations between words/sentences, asking and answering questions, etc.

Outsource it to a large crowd of non-experts (workers)
Experts vs. Crowd-sourcing Workers

Benefits of using non-experts:

- Reduced cost
  - e.g. translation in 12% the cost [Zaidan and Callison-Burch, 2011]
Experts vs. Crowd-sourcing Workers

Benefits of using non-experts:

- Reduced cost
  - e.g. translation in 12% the cost [Zaidan and Callison-Burch, 2011]
- Workers are easily available via crowdsourcing platforms
Experts vs. Crowd-sourcing Workers
Benefits of using non-experts:

- Reduced cost
  - e.g. translation in 12% the cost [Zaidan and Callison-Burch, 2011]
- Workers are easily available via crowdsourcing platforms
- Shorter time to complete task (many people working in parallel)
Experts vs. Crowd-sourcing Workers
How can a non-expert achieve the same quality as the expert?

- It’s not a single worker that replaces the expert, it’s the crowd.
Experts vs. Crowd-sourcing Workers
How can a non-expert achieve the same quality as the expert?

- It’s not a single worker that replaces the expert, it’s the crowd.
- Assign each task to $k$ workers (typically 3-5).
Experts vs. Crowd-sourcing Workers
How can a non-expert achieve the same quality as the expert?

- It’s not a single worker that replaces the expert, it’s the crowd.
- Assign each task to $k$ workers (typically 3-5).
- Aggregate their answers.
Experts vs. Crowd-sourcing Workers
How can a non-expert achieve the same quality as the expert?

- It's not a single worker that replaces the expert, it's the crowd.
- Assign each task to $k$ workers (typically 3-5).
- Aggregate their answers.
- Assumption: on average, the crowd thinks like the expert.
Some tasks require experts and can’t be crowdsourced
  - e.g. syntactic parsing annotations
What can be Crowdsourced?

- Some tasks require experts and can’t be crowdsourced
  - e.g. syntactic parsing annotations

- **Crowdsourcability rules of thumb:**
  1. Keep it simple
  2. One possible answer
  3. Short tasks
  4. Bonus: fun
Keep It Simple

- The task is easy to explain
Keep It Simple

- The task is easy to explain
- Instructions don’t rely on expert knowledge
Keep It Simple

- The task is easy to explain
- Instructions don’t rely on expert knowledge
- Instructions are short
  - Don’t expect workers to read a 50 page manual
Keep It Simple

- The task is easy to explain
- Instructions don’t rely on expert knowledge
- Instructions are short
  - Don’t expect workers to read a 50 page manual
- Instructions include examples
One possible answer

- People can easily agree on the “correct” answer

Good:
- is there a cat in this image?

Bad:
- what is the meaning of life?

When multiple answers are possible, do a second-pass

Task #1: describe the following image
Task #2: does this text describe the image?

Other workers validate the image descriptions
One possible answer

- People can easily agree on the “correct” answer

- Good: *is there a cat in this image?*
One possible answer

- People can easily agree on the “correct” answer

- Good: *is there a cat in this image?*
- Bad: *what is the meaning of life?*
One possible answer

- People can easily agree on the “correct” answer

  - Good: is there a cat in this image?
  - Bad: what is the meaning of life?

- When multiple answers are possible, do a second-pass
  - Task #1: describe the following image
One possible answer

- People can easily agree on the “correct” answer

  - Good: *is there a cat in this image?*
  - Bad: *what is the meaning of life?*

- When multiple answers are possible, do a second-pass
  - Task #1: *describe the following image*
  - Task #2: *does this text describe the image?*
One possible answer

- People can easily agree on the “correct” answer
  - Good: *is there a cat in this image?*
  - Bad: *what is the meaning of life?*

- When multiple answers are possible, do a second-pass
  - Task #1: *describe the following image*
  - Task #2: *does this text describe the image?*
  - Other workers validate the image descriptions
Short tasks

- Assume workers have a short attention span
**Short tasks**

- Assume workers have a short attention span.
- Ask one question per task (HIT).
Short tasks

- Assume workers have a short attention span
- Ask one question per task (HIT)
- If the task seems too complicated for crowdsourcing, can we break it into crowdsourcable subtasks?
## Short tasks

- Group few questions when they otherwise require re-doing a task, e.g. reading the same sentence

```markdown
<table>
<thead>
<tr>
<th>Original sentence</th>
<th>New sentence</th>
<th>Describes same event</th>
<th>Adds Information</th>
<th>Not correct / grammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man and a woman sitting near small boats.</td>
<td>A man and a woman sitting <strong>far from</strong> small boats.</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>A man and a woman sitting <strong>far away from</strong> small boats.</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A man and a woman sitting near <strong>little</strong> boats.</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A man and a woman <strong>standing</strong> near small boats.</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A man and a woman sitting near <strong>old</strong> boats.</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Would you like to add a comment?
Making Annotations Fun

- Get workers to be more engaged and they’ll perform better
Making Annotations Fun

- Get workers to be more engaged and they’ll perform better

- The extreme case: Gamification
  - Instead of paying for annotations, develop a game
What is Crowdsourcing?
What can be Crowdsourced?

Making Annotations Fun

- Get workers to be more engaged and they’ll perform better

- The extreme case: Gamification
  - Instead of paying for annotations, develop a game
  - “Workers” play for fun, you get annotations
Crowdsourcing Platforms

- Amazon Mechanical Turk
- CrowdFlower
- Many others...
- We will focus on Amazon Mechanical Turk (AMT, MTurk)
**Terminology**

- **Requester**: the person paying to collect data
- **Worker**: the person getting paid to provide data
- **HIT**: single task (question) in Amazon Mechanical Turk
Before Getting Started...

Before spending money, try it on your friends: https://requestersandbox.mturk.com
Before Getting Started...

- Before spending money, try it on your friends: https://requestersandbox.mturk.com
- Get feedback and improve task & instructions
Before Getting Started...

- Before spending money, try it on your friends: https://requestersandbox.mturk.com
- Get feedback and improve task & instructions
- Time the HITs to decide how much to pay
What is Crowdsourcing?

Workflow

- Plan task
- Prepare data
- Decide on worker qualification
- Launch small batch
- Evaluate results & worker feedback
- Are results satisfactory?
  - Yes
    - Publish another batch
    - Done?
      - Yes: Stop
      - No: Workflow continues
  - No: Workflow continues
Tips and Tricks
Aggregating Worker Answers

- Most common: majority voting
Aggregating Worker Answers

- Most common: majority voting
- Things to consider:
  - How to break ties?
Aggregating Worker Answers

- Most common: majority voting
- Things to consider:
  - How to break ties?

- More advanced: weighting answers by worker quality
  - MACE (Multi-Annotator Competence Estimation)
    [Hovy et al., 2013]
Quality Control

Questions to be asked:

- How do we know if the annotations are good?
Quality Control

Questions to be asked:
- How do we know if the annotations are good?
- Can we trust workers?
Quality Control

Questions to be asked:

- How do we know if the annotations are good?
- Can we trust workers?
- What if they just cheat to get easy money?
Quality Control

Questions to be asked:
- How do we know if the annotations are good?
- Can we trust workers?
- What if they just cheat to get easy money?

Quality control solutions:
1. Qualification requirements
2. Control questions
3. Qualification test
4. Second-pass reviewing
5. Carrot and stick
Qualification Requirements
Who can participate in your task?

- Masters qualification (not recommended)
Qualification Requirements
Who can participate in your task?

- Masters qualification (not recommended)
- Statistics about previous work:
  - Approval rate
  - Number of completed HITs
Qualification Requirements
Who can participate in your task?

Masters qualification (not recommended)
Statistics about previous work:
  - Approval rate
  - Number of completed HITs
- Location, other properties
Qualification Requirements
Who can participate in your task?

- Masters qualification (not recommended)
- Statistics about previous work:
  - Approval rate
  - Number of completed HITs
- Location, other properties
- Custom qualification
Qualification Test

- Worker must pass a qualification test to work on the task
Qualification Test

- Worker must pass a qualification test to work on the task
- Consists of a few questions
  - Similar (not necessarily identical) to the task
  - Predicts worker’s ability to perform the task well
Qualification Test

- Worker must pass a qualification test to work on the task
- Consists of a few questions
  - Similar (not necessarily identical) to the task
  - Predicts worker’s ability to perform the task well
- Requester defines gold answers and minimum score requirement
Qualification Test

- Worker must pass a qualification test to work on the task
- Consists of a few questions
  - Similar (not necessarily identical) to the task
  - Predicts worker’s ability to perform the task well
- Requester defines gold answers and minimum score requirement
- Workers are not paid to take the test
Qualification Test

- Worker must pass a qualification test to work on the task
- Consists of a few questions
  - Similar (not necessarily identical) to the task
  - Predicts worker’s ability to perform the task well
- Requester defines gold answers and minimum score requirement
- Workers are not paid to take the test
- Built-in mechanism in AMT (using the API)
Qualification Test

Please read the instructions below and answer the question based on the instructions.

Instructions:

A political leader can make "moves," or decisions, to pursue an action. Write a sentence using these criteria to describe the move below:

The sentence should mimic how the move is discussed in the news.
The sentence needs to include a subject, verb, action, and predicate object that show how the move is expressed. Time, location, and clauses are necessary in order to model actual news.
Sentences should involve countries outside the United States.
Sentences must differ from the examples provided.
Sentences in languages other than English (especially Middle Eastern languages such as Arabic, Farsi, and Turkish) may receive bonuses.

Move: Appeal for military cooperation

Description of Move: Make an appeal for, request, or suggest initiating or expanding military ties.

Sample sentences: South Korea has requested to lease a Russian military training ground, military officers in Seoul said on Thursday. | NATO suggested expanding its military ties with the Euro Zone Friday afternoon.

Specific Clarification (if applicable): Use this code for requests to develop or expand military relations by engaging in acts such as joint military maneuvers or exercises. Appeals for provision of military aid—not mutual exchange—are coded as 0232 instead. Actual events of military cooperation are coded as 062.

Example of a Bad Sentence: "Trump spoke yesterday." This sentence is not acceptable because it is non-specific to any move and refers to the United States.

1. Which of these sentences best fits the above criteria?

- The president wants troops.
- President Assad requested deepening military ties between Russia and Syria Friday.
- Russia sent troops to Syria last Wednesday in response to threats of terrorism.
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer

...and the answers are unambiguous

If a worker fails to answer too many control questions, don't use their answers. They are either not good or trying to cheat. Reject their work/block them.
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer
  - ...and the answers are unambiguous
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer
  - ...and the answers are unambiguous

- If a worker fails to answer too many control questions, don’t use their answers
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer
  - ...and the answers are unambiguous

- If a worker fails to answer too many control questions, don’t use their answers
  - They are either not good or trying to cheat
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer
  - ...and the answers are unambiguous

- If a worker fails to answer too many control questions, don’t use their answers
  - They are either not good or trying to cheat
  - Reject their work/block them
Control Questions

- Incorporate special questions in the task:
  - for which you know the correct answer
  - ...and the answers are unambiguous

- If a worker fails to answer too many control questions, don’t use their answers
  - They are either not good or trying to cheat
  - Reject their work/block them

- Built-in mechanism in CrowdFlower
Second-pass Reviewing

- Second task in which workers validate previous workers’ answers
Carrot and Stick
Good Workers (that's the easy one)

- Bonus them
Carrot and Stick
Good Workers (that’s the easy one)

- Bonus them
- They will keep working for you
Carrot and Stick
Good Workers (that’s the easy one)

- Bonus them
- They will keep working for you
- They will write good things about you in Turker forums
Carrot and Stick
Bad Workers

- Don’t reject work
  - They are not paid for the work they did
Carrot and Stick
Bad Workers

- Don't reject work
  - They are not paid for the work they did
  - Unfair and unethical
Carrot and Stick
Bad Workers

- Don’t reject work
  - They are not paid for the work they did
  - Unfair and unethical
  - Misunderstanding ≠ cheating
  - Maybe you didn’t explain the task clearly?
Carrot and Stick
Bad Workers

- Don’t reject work
  - They are not paid for the work they did
  - Unfair and unethical
  - Misunderstanding ≠ cheating
  - Maybe you didn’t explain the task clearly?
  - You’ll get bad reviews in Turker forums
Carrot and Stick
Bad Workers

- Don’t reject work
  - They are not paid for the work they did
  - Unfair and unethical
  - Misunderstanding $\neq$ cheating
  - Maybe you didn’t explain the task clearly?
  - You’ll get bad reviews in Turker forums

- Alternative #1: Blocking a worker
  - Worker can’t work on your tasks anymore
  - Doesn’t see your task, doesn’t know they are blocked
Carrot and Stick

Bad Workers

- Don’t reject work
  - They are not paid for the work they did
  - Unfair and unethical
  - Misunderstanding ≠ cheating
  - Maybe you didn’t explain the task clearly?
  - You’ll get bad reviews in Turker forums

- Alternative #1: Blocking a worker
  - Worker can’t work on your tasks anymore
  - Doesn’t see your task, doesn’t know they are blocked

- Alternative #2: Revoking qualification
  - Knows they are unqualified
  - Can re-take qualification test
Carrot and Stick

Recognize bad/good workers

- Look at the most productive workers
**Carrot and Stick**

**Recognize bad/good workers**

- Look at the most productive workers
- They are potentially good engaged workers
Carrot and Stick
Recognize bad/good workers

- Look at the most productive workers
- They are potentially good engaged workers
- ...or spammers.
General Tips
Make Workers’ Lives Easier

- Make sure they don’t have to scroll (yes, seriously!)
General Tips
Make Workers’ Lives Easier

- Make sure they don’t have to scroll (yes, seriously!)
- Repeat important guidelines and don’t expect workers to remember them
General Tips
Make Your Task Cheater-Proof

- Control for the balance of expected answers
  - e.g. in validation tasks, if you expect most answers to be “yes”, add some negative samples
General Tips
Make Your Task Cheater-Proof

- Control for the balance of expected answers
  - e.g. in validation tasks, if you expect most answers to be “yes”, add some negative samples
- Otherwise workers exploit this bias
General Tips

You are bound to make mistakes

- Work in small batches
General Tips
You are bound to make mistakes

- Work in small batches
- Look at worker answers and comments between batches
General Tips
You are bound to make mistakes

- Work in small batches
- Look at worker answers and comments between batches
- Use it to improve the task
General Tips
Do you want your task to finish quickly?

- For language tasks in English, publish batches in morning time in the US
General Tips
Do you want your task to finish quickly?

- For language tasks in English, publish batches in morning time in the US
- Advertise your task in Turker Nation
Ethical Issues
Don’t Underpay!
(Can’t stress that enough!)

- Workers are first of all people
Don’t Underpay!
(Can’t stress that enough!)

- Workers are first of all people
- Many rely on crowdsourcing as their main income
Ethical Issues

Don’t Underpay!
(Can’t stress that enough!)

- Workers are first of all people
- Many rely on crowdsourcing as their main income
- They have no job security
Don’t Underpay!
(Can’t stress that enough!)

- Workers are first of all people
- Many rely on crowdsourcing as their main income
- They have no job security
- Don’t take advantage of new workers
  - May do lower-paid work to build up their reputation
Give Workers The Benefit of the Doubt

- Don’t reject work, use “auto-approve”
Give Workers The Benefit of the Doubt

- Don’t reject work, use “auto-approve”
- Even bad workers spent time working and should be paid
Give Workers The Benefit of the Doubt

- Don’t reject work, use “auto-approve”
- Even bad workers spent time working and should be paid
- Are you sure you explained the task well?
  - Maybe it’s your fault if workers performed badly?
The Good News: Paying well pays off

- When you pay enough, you get the good workers
- Experienced workers don’t work for little money
The Good News: Paying well pays off

- When you pay enough, you get the good workers
- Experienced workers don’t work for little money

- Underpaying means:
  - Task will take longer to complete
The Good News: Paying well pays off

- When you pay enough, you get the good workers
- Experienced workers don’t work for little money

- Underpaying means:
  - Task will take longer to complete
  - Annotations may be of low quality
How much should you pay?

- Rule of thumb: 6-8 USD per hour
How much should you pay?

- Rule of thumb: 6-8 USD per hour
- Time your task using the sandbox
  - Don’t trust AMT statistics; workers open multiple HITs
How much should you pay?

- Rule of thumb: 6-8 USD per hour

- Time your task using the sandbox
  - Don’t trust AMT statistics; workers open multiple HITs

- Not less than 2 cents per (very small) HIT
How much should you pay?

- Rule of thumb: 6-8 USD per hour
- Time your task using the sandbox
  - Don’t trust AMT statistics; workers open multiple HITs
- Not less than 2 cents per (very small) HIT
- Take into account ~20% fees to AMT
Amazon Mechanical Turk Demo
Links

- Chris Biemann’s crowdsourcing course at ESSLLI 2016: http://esslli2016.unibz.it/?page_id=346

References
