University of BRISTOL

Crowdsourced Data Management:Industry and Academic Perspectives, Marcus and Parameswaran, 2015.

10.1 Annotation and Crowdsourcing

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

Dataset Construction

- Annotations: the correct, gold standard labels.

[('The movie was enjoyable', 1), ('I couldn't finish watching', -1), …]

[('The', DT), ('movie', NN), ('was', VBD)…]

- What do we need annotations for?
  - Training machine learning models.
  - Computing evaluation metrics (also for rule-based systems) before deploying.
  - Performing error analysis: what kinds of errors does it make?

Expert Annotators

- Traditional approach to annotation:
  - Hire a linguist for annotations like parts of speech.
  - Hire a domain expert for specialised annotations like labelling ‘process’ entities in scientific papers.
- Is an expert always right?

Expert Annotators

- Traditional approach to annotation: hire ~~a linguist~~ some linguists or domain experts.
- If one makes a mistake, the annotators disagree.
- Solutions:
  - Re-annotate examples with disagreements.
  - Hire three or more experts and take the majority vote.

Expert Annotators

- Traditional approach to annotation: hire ~~a linguist~~ some linguists.
- But experts are expensive, their time is limited!
- Deep learning is data-hungry, so is thorough evaluation.
- How can we obtain large datasets at reasonable speed and cost?

Crowdsourcing

Demographics and Dynamics of Mechanical Turk Workers, Difallah et al., 2018.
One Million for Zooniverse – and One for Galaxy Zoo! Simmons, 2014.

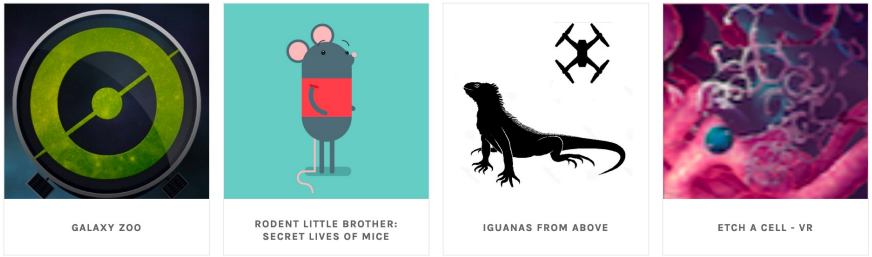
- Recruit a large number of non-expert annotators to provide the annotations!

amazon mechanical turk

AMT: Pay a few cents per task to each worker.

bristol.ac.uk

Zooniverse: Volunteer citizen scientists analyse interesting data.



Crowdsourcing

Demographics and Dynamics of Mechanical Turk Workers, Difallah et al., 2018.
One Million for Zooniverse – and One for Galaxy Zoo! Simmons, 2014.

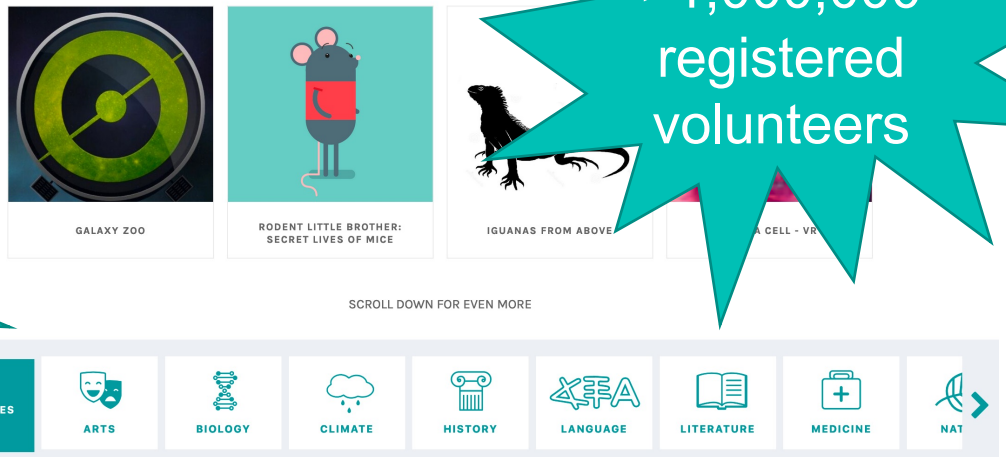
- Recruit a large number of non-expert annotators to provide the annotations!

amazon mechanical turk

100,000 workers available at any time

Zooniverse:

>1,000,000 registered volunteers



bristol.ac.uk

Crowdsourcing

[Demographics and Dynamics of Mechanical Turk Workers](), Difallah et al., 2018.
[One Million for Zooniverse – and One for Galaxy Zoo!]() Simmons, 2014.

- Recruit a large number of non-expert annotators to provide the annotations!

amazon mechanical turk

Zooniverse:

100,000 workers available at any time

Crowdsourced annotation can be fast!

>1,000,000 registered volunteers

bristol.ac.uk

An AMT Task

Annotate the emotions associated with each turn in a conversation.

Figure from EmotionLines: An Emotion Corpus of Multi-Party Conversations, Hsu et al., (2018).

bristol.ac.uk

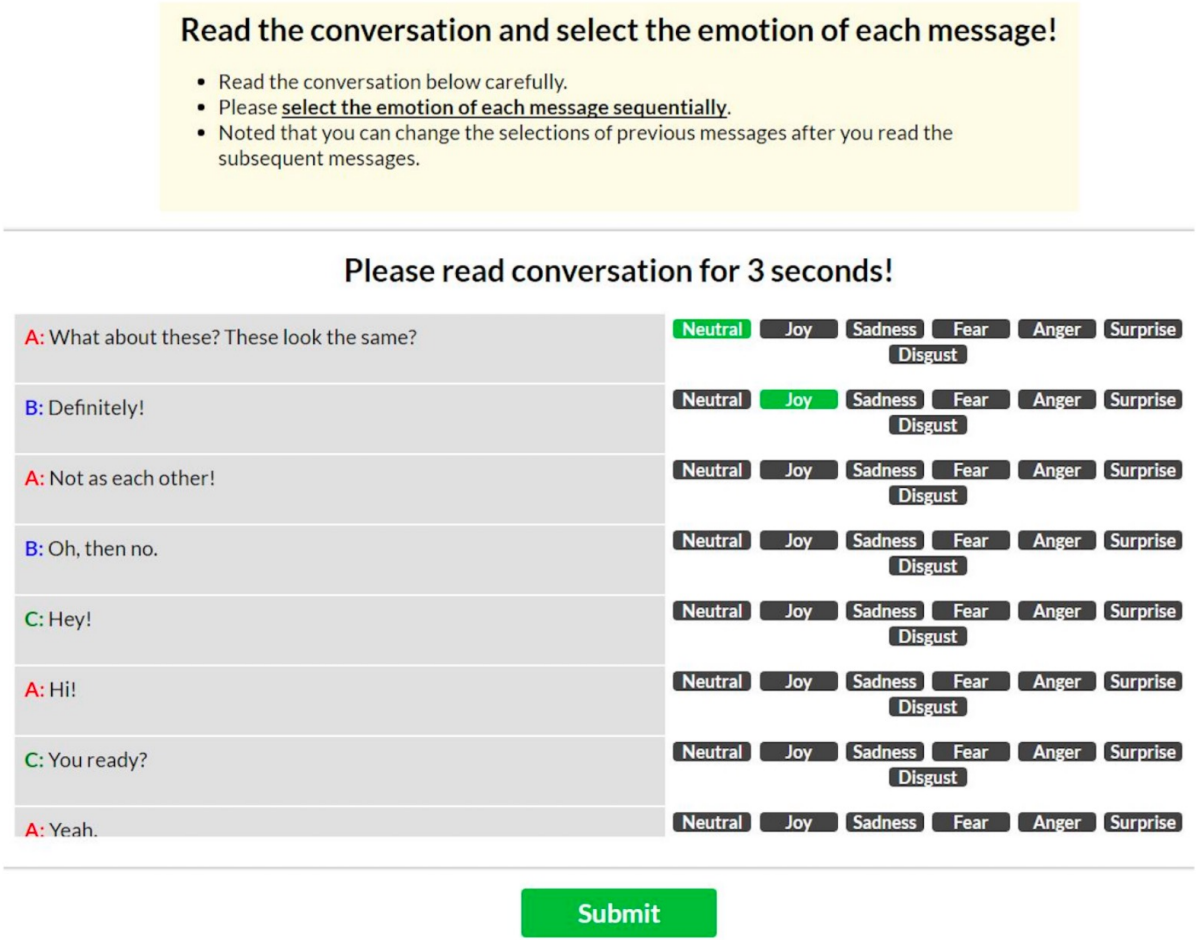


Figure 1: Worker interface on Amazon Mechanical Turk

Summary

- Annotating data is a fundamental part of text analytics, required for training and evaluation.
- Expert annotations are reliable but expensive so hard to scale up to large datasets.
- Crowdsourcing offers a faster, cheaper annotation process by connecting with untrained annotators.
- Tasks need to be described using clear instructions and no jargon.

University of BRISTOL

Crowdsourced Data Management:Industry and Academic Perspectives, Marcus and Parameswaran, 2015.

10.2 Crowdsourcing Challenges

Edwin Simpson

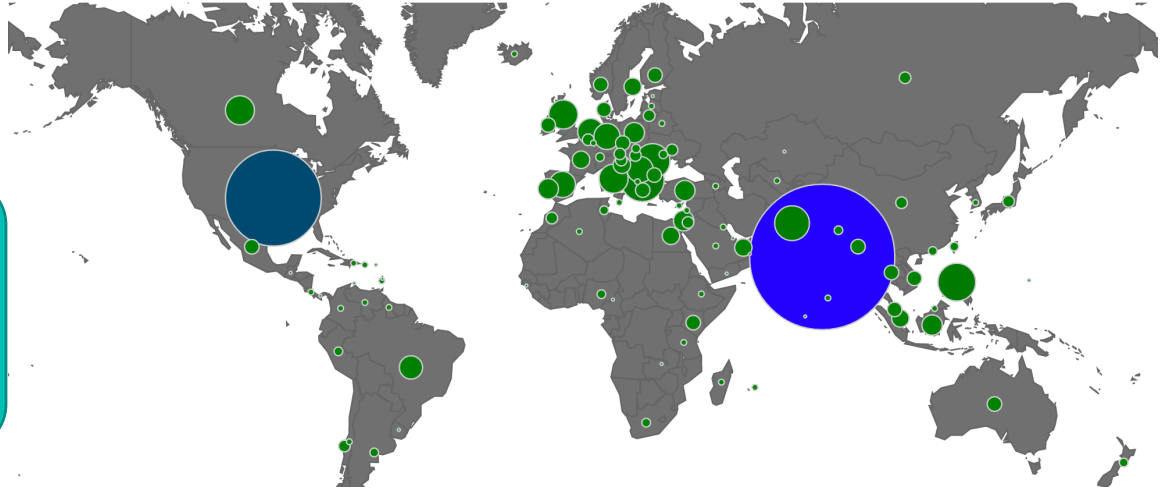
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

Challenges of Crowdsourcing

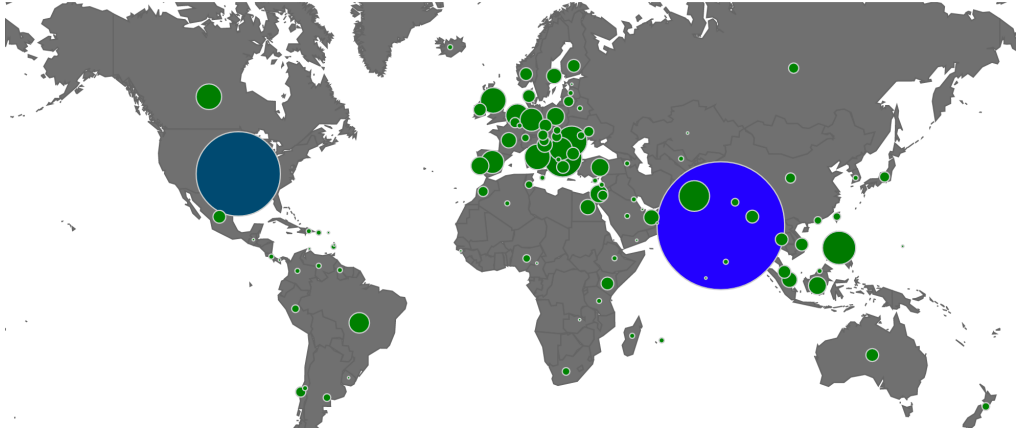
- Language diversity:
  - Can only legally hire workers on AMT in USA and India;
  - Other platforms specialise in other countries, like Clickworker in Germany;
  - But very difficult to obtain data in less widely-spoken languages.



The number of workers per country on AMT, from [The Language Demographics of Amazon Mechanical Turk](), Pavlick et al., 2014

Challenges of Crowdsourcing

- Biases:
  - Many text annotation tasks like sentiment analysis are not purely objective;
  - Can pick up the biases of the population;
- Ensure that we pay an ethical minimum wage according to how long each annotation task takes and worker's location.



The number of workers per country on AMT, from [The Language Demographics of Amazon Mechanical Turk](), Pavlick et al., 2014

Challenges of Crowdsourcing

- Quality! Errors are caused by…
  - Spammers: people who want to get paid without doing the task properly;
  - Misunderstandings: difficult tasks and ambiguous instructions;
  - Skills and expertise that are below expert level.
- How can we produce high quality datasets if crowdworkers make a lot of mistakes?

Wisdom of the Crowd

- Guess the weight:

- In 1906, Francis Galton observed that the median guess was accurate to within 1%.

Wisdom of the Crowd

- The errors of different annotators cancel out!

$$E_{crowd} = \frac{1}{M} E_{averageIndividual}$$

  - Where *E* is expected error and *M* is the number of annotators.
- This holds under strong assumptions:
  - Errors are not correlated – each person makes different mistakes.
  - Confusion between class labels is random
- In practice, many annotators make the same mistakes, but this insight still helps us to increase quality.

Redundant Labelling

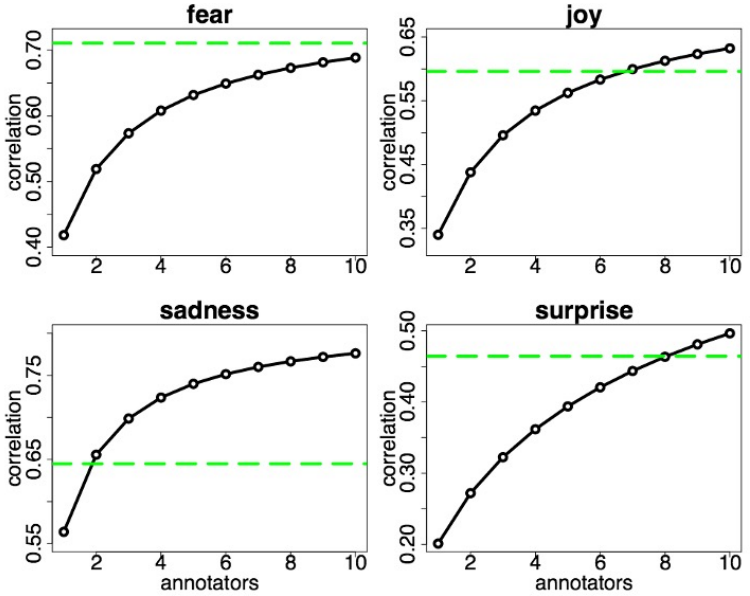
- Multiple annotators label each document.
- We then aggregate the labels:
  - Mean of values
  - Majority vote
  - Machine learning methods that weight annotators by their accuracy and bias.
- The high-quality, aggregated labels can then be provided as a gold-standard for training and evaluation.

Cheap and Fast – But Is It Good?

- Systematic comparison of workers to experts on various NLP tasks
- E.g. rate headlines to reflect emotions
- Y-axis: correlation with mean of experts
- Green dashed lines: 1 expert
- Black solid lines: increasing number of workers per task
- On average, 4 workers ≈ 1 expert

Outcry at N Korea 'nuclear test'

$(Anger, 30), (Disgust,30), (Fear,30), (Joy,0),$
$(Sadness,20), (Surprise,40), (Valence,-50).$



Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks., Snow et al., 2008.

Crowdsoucing Design

- Good task design also improves label accuracy:
  - Break complex tasks into small steps;
  - Provide simple instructions and examples;
  - Bonuses for good work.
- Develop crowdsourcing tasks using an iterative process:
  - Test a proposed crowdsourcing task on a small amount of data;
  - Compare the aggregated crowd labels to expert labels for the small sample;
  - Modify the task design to reduce misunderstandings and errors;
- Use small expert-labelled dataset to find out how many annotators we need to reach our desired accuracy.

Task Design Example

Goal: extract information about who the CEOs of companies are.

Bad:

"Annotate the named entities that refer to CEOs or companies, then type in the pairs of entities that are related."

Better:

- "Step 1: Highlight all references to CEOs in this text."
- "Step 2: Highlight all references to companies by their names."
- "Step 3: The name of a CEO and a company is highlighted in this text. Is this person the CEO of this company?

Summary

- Crowdsourcing has issues with quality, language diversity, biases and ethical treatment of workers.
- The most common ways to improve quality are:
  - Breaking annotation tasks into small steps;
  - Redundant labelling, which reduces some kinds of error.