CS224C: NLP for CSS

Data Annotation

Diyi Yang

Stanford CS

Lecture Overview

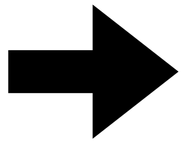
- Annotation
- Platform
- Quality
- Annotation agreement

Why Annotation Is Needed?

Modern machine learning or data science requires data, especially data with **labels** or **ground-truths**

Where Is Label or Ground-truth?

Modern machine learning or data science requires data, especially data with **labels** or **ground-truths**

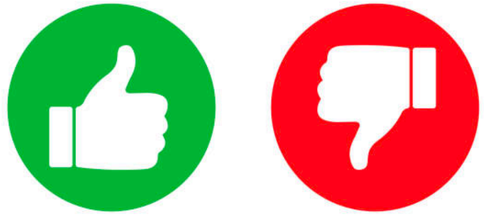


1. **Natural/existing annotation**
2. **New/human annotation**

Natural or Existing Data Annotation

Metadata about the text

User-contributed labels (e.g., like/dislike)



New or Human Annotation

Researchers or practitioners themselves

Recruit people for data annotation

* Local network
* Online crowdsourcing platforms

Platforms

Suhr, Alane, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel Bowman, and Yoav Artzi. "Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pp. 1-6. 2021.

Crowdsourcing Platforms

Amazon Mechanical Turk:

- Largest marketplace and flexible
- Oriented toward large scale annotation tasks
- Most workers in US or India, part-time, college-educated

Crowdsourcing Platforms

Upwork:

- Requesters hire workers individually and specifically
- Oriented around longer gigs or hiring specialists
- Higher typical pay, mostly larger than 25$ per hour

Crowdsourcing Platforms

Many others available!

Mechanical Turk Basics

1. Workers and requesters (i.e., researchers) join the platform. No training or experience required on either side.
2. A requester designs a simple UI (often an HTML form) to collect data.
3. The requester posts a batch of human intelligence tasks (HITs) using that UI, each representing individual small jobs that pay a fixed amount ($1?), and deposits money.
4. Over the following hours/days, workers choose HITs and complete them one-by-one.
5. Requesters quickly review submitted work and approve it (at their sole discretion), releasing payment.

Quality

Major Issues

Median hourly wage is only around **2$ per hour**

Only 4% earned more than $7.25/hr

Average requester pays > $11/hr

Lower-paying requesters post much more work

Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. "A data-driven analysis of workers' earnings on Amazon Mechanical Turk." In Proceedings of the 2018 CHI conference on human factors in computing systems, pp. 1-14. 2018.

Recruiting Trustworthy Workers

Amazon lets you filter by experience level: Common to limit HITs to experienced workers (>5,000 HITs completed) with low rejection rates (<2%).

- ➡ Be careful about needlessly high HIT counts: They push newer good workers into underpaid work.

- ➡ Amazon also lets you recruit its promoted ‘Master' workers. This is meaningless.

Kummerfeld, Jonathan K. "Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 343-349. 2021

Qualifications

You can assign manual qualifications to workers. Common setup:

- Post a public training/practice HIT that workers can only do once.
- Manually review work on that HIT, and use it to grant qualifications to work on the rest of the HITs.
- Periodically monitor work, and revoke qualifications if major problems arise.
- Don't reject work unless it's very clearly spam/fraud. This revokes payment for work that has already been done.

Quality Control

Use multiple HITs to ensure reasonable quality in test/validation data:

- When collecting test data for classification and annotation tasks, have several workers annotate each example.
- Fancy statistical methods can aggregate multiple annotations better than majority vote.
- When multiple parallel annotations can't be combined, consider building a second validation HIT to double check each data point.

Additional Quality Control

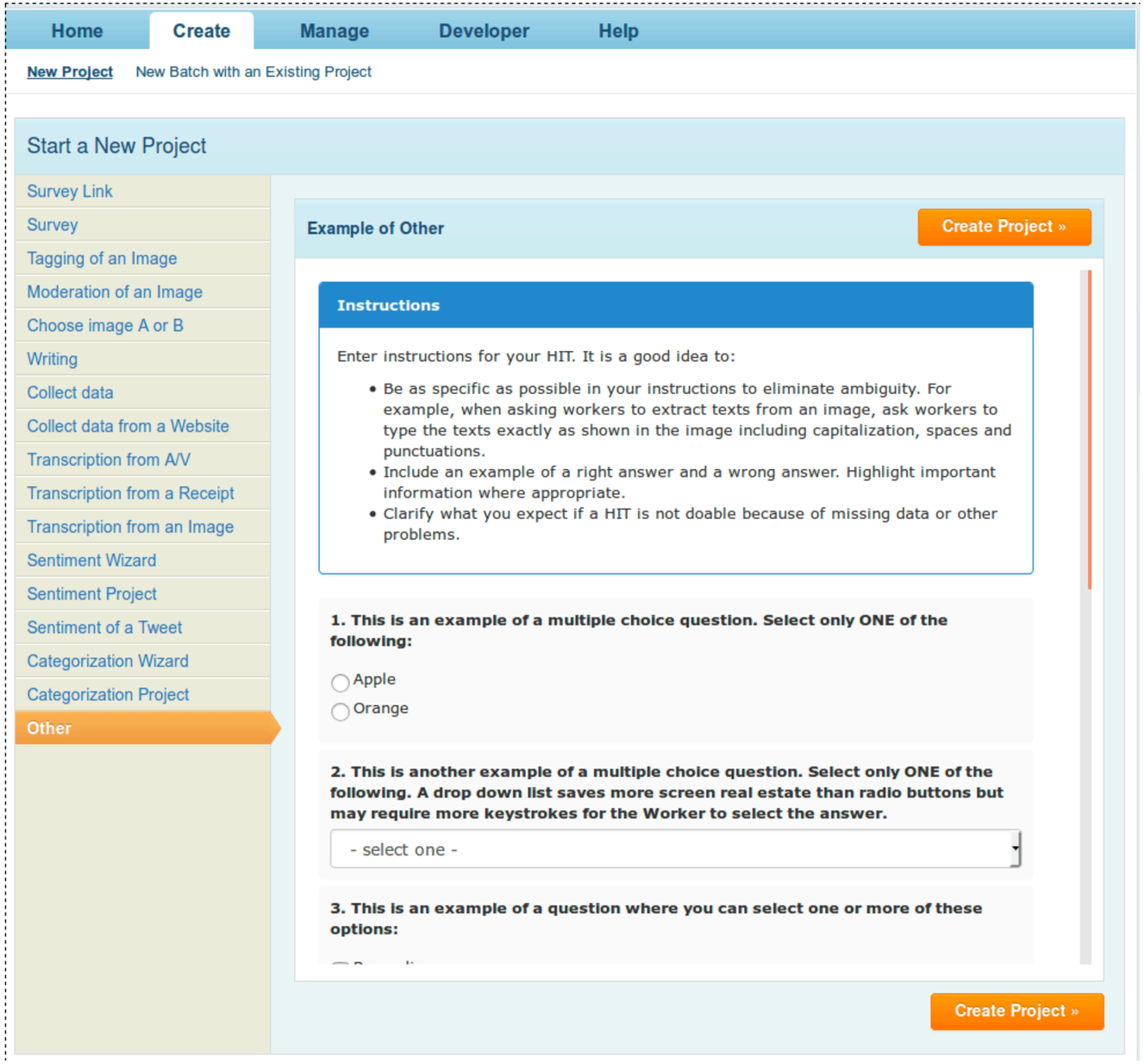
The initial annotation guideline may not work well.

- ✦Two researchers performs annotation on a subset of samples
- ✦Discuss their disagreements and agreements to refine the guideline
- ✦Repeat 2~3 times until a reasonable and sufficient agreement is reached

Building the UI

Amazon has simple HTML form templates you can edit, and it will let you upload/ download CSVs with data.

You can use javascript snippets to validate responses and add other interactive features.



Recommendations

1. Pay well.
2. Have clear, fair criteria for bonuses and rejection (typically in an FAQ doc).
3. Respond to worker questions quickly—daily at least.
4. Design your HITs to be usable and efficient.
5. Identify yourself clearly.
6. Give clear instructions, especially for how to handle weird/broken prompts.
7. Make sure HITs that pay the same take roughly the same amount of time.

Hourly Wage Estimation

Amazon's hourly wage estimate tool may not be trustworthy.

To start:

Do the work yourself for an hour and see how far you get.

Once your HIT is live:

Tools like Crowd-Workers and TurkOpticon let you see better estimates of actual time elapsed.

Annotation Agreement

Reliability and Validity

Validity is the **correctness** or **accuracy** of a measure

e.g., reflect what it's supposed to measure?

Reliability is the **consistency** of a measure

e.g., similar results under the same conditions multiple times?

Assumption: higher reliability might imply validity

How to Get Reliability or Consistency

The same item being annotated by two or more annotators!!!

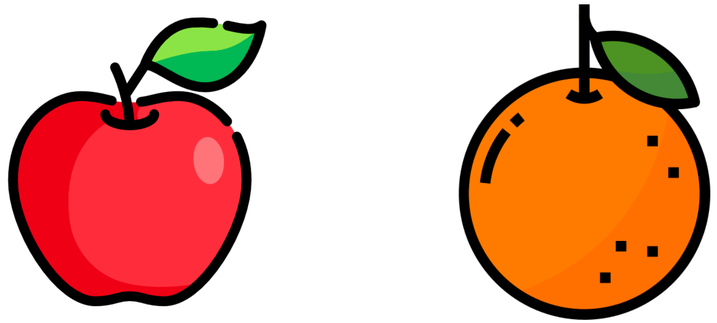
Key pieces:

- Same item
- Different annotators
- Independent annotation

Measuring Agreement

Observed Agreement: proportion of items on which the two annotators agree

	Apple	Orange	Total
Apple	30	15	45
Orange	10	45	55
Total	40	600	100



Artstein, Ron, and Massimo Poesio. "Inter-coder agreement for computational linguistics." *Computational linguistics* 34, no. 4 (2008): 555-596.

Measuring Agreement

Observed Agreement: proportion of items on which the two annotators agree

	Apple	Orange	Total
Apple	30	15	45
Orange	10	45	55
Total	40	60	100

Agreement:
(30 + 45) / 100 = 0.75

Artstein, Ron, and Massimo Poesio. "Inter-coder agreement for computational linguistics." *Computational linguistics* 34, no. 4 (2008): 555-596.

Chance Agreement?

Some agreement is expected by chance

> e.g., if two annotators are asked to pick Apple or Orange randomly, they might agree with each other half of the time

Chance-corrected agreement: agreement beyond chance

Chance-Corrected Agreement

Observed agreement A_o: proportion of actual agreement

Expected agreement A_e**:** expected value of A_o

Amount of agreement above chance: $A_o - A_e$

Maximum agreement above chance: $1 - A_e$

Proportion of the possible agreement beyond chance: $\frac{A_o - A_e}{1 - A_e}$

How to Obtain the Expected Agreement A_e

Expected agreement A_e is the probability of the two annotators c_1 and c_2 agreeing on any given category k

$$A_e = \sum_{k \in K} P(k \mid c_1) \cdot P(k \mid c_2)$$

Various coefficients available based on how these probabilities are calculated.

Cohen's Kappa

Cohen's κ assumes the random assignment of categories to items is governed by prior distributions that are unique to each other (annotator bias).

$$P(k \mid c_i) = \hat{P}(k \mid c_i) = \frac{\mathbf{n}_{c_i k}}{\mathbf{i}}$$

The actual number of assignment to k by c_i

The number of items

$$A_e^{\kappa} = \sum_{k \in K} \hat{P}(k \mid c_1) \cdot \hat{P}(k \mid c_2) = \sum_{k \in K} \frac{\mathbf{n}_{c_1 k}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 k}}{\mathbf{i}} = \frac{1}{\mathbf{i}^2} \sum_{k \in K} \mathbf{n}_{c_1 k} \mathbf{n}_{c_2 k}$$

Fleiss's Kappa: *more than two annotators*

Agreement is the proportion of agreeing pairs out of the total judgement pairs

Expected agreement:

the probability of agreement for an arbitrary pair of coders

Fleiss's Kappa: A_o

$$\frac{A_o - A_e}{1 - A_e}$$

Let $\mathbf{n}_{ik}$ stand for the number of items an item i is classified in category k

Each category k contributes $\binom{\mathbf{n}_{ik}}{2}$ pairs of agreeing judgements for item i

The amount of agreement agr_i for item i is

$$\text{agr}_i = \frac{1}{\binom{\mathbf{c}}{2}} \sum_{k \in K} \binom{\mathbf{n}_{ik}}{2} = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik} - 1)$$

The overall observed agreement is the mean of agr_i for all item i

$$A_o = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_{i \in K} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik} - 1)$$

Fleiss's Kappa: A_e

$$\frac{A_o - A_e}{1 - A_e}$$

Chance agreement: The agreement expected on the basis of a single distribution which reflects the combined judgements of all coders

$$\hat{P}(k) = \frac{1}{\mathbf{ic}}\mathbf{n}_k$$

Expected agreement: prob of each coder making their choice independently

$$A_e = \sum_{k \in K} \hat{P}(k)^2 = \frac{1}{(\mathbf{ic})^2} \sum_{k \in K} \mathbf{n}_k^2$$

Krippendorff's Alpha

The key limitation of prior measures is the disagreements are treated equally

For semantic and pragmatic features, ***disagreements are not all alike***.

$$\alpha = 1 - \frac{D_o}{D_e}$$

Intra-Class Correlation

Intra class correlation is usually defined as a ratio:

$$\frac{\text{variance of interest}}{\text{total variance}} = \frac{\text{variance of interest}}{\text{variance of interest + unwanted variance}}$$

ICC is used to assess the consistency, or conformity, of measurements made by multiple observers measuring the same quantity (exchangeable).

https://en.wikipedia.org/wiki/Intraclass_correlation

Choosing the Right Reliability Measure

Key factors:

- The number of annotators
- Annotator biases
- Do you have consistent raters?

Interpreting the Values

Rule of thumb for different agreement scores vary

- Less than 0.40—poor.
- Between 0.40 and 0.59—fair.
- Between 0.60 and 0.74—good.
- Between 0.75 and 1.00—excellent.