Crowdsourcing

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

26 Sep 2019

Announcements, logistics

- Reminder: Early exam next Thursday
  - Anything up to an including today *may* be on the exam
  - Majority of focus on things covered through HW1 and HW2
  - Any non-communicating calculation device allowed
  - 1 two-sided sheet of notes that you create yourself allowed (to be handed in)
- HW3 will be posted by next class period
- JupyterHub = ☹
  - HW2 server was upgraded this morning (new URL)

Last time

- Measurement arises in many aspects of NLP applications:
  - How you define your features
  - How you define your outcome
  - Your system *as* a measurement device

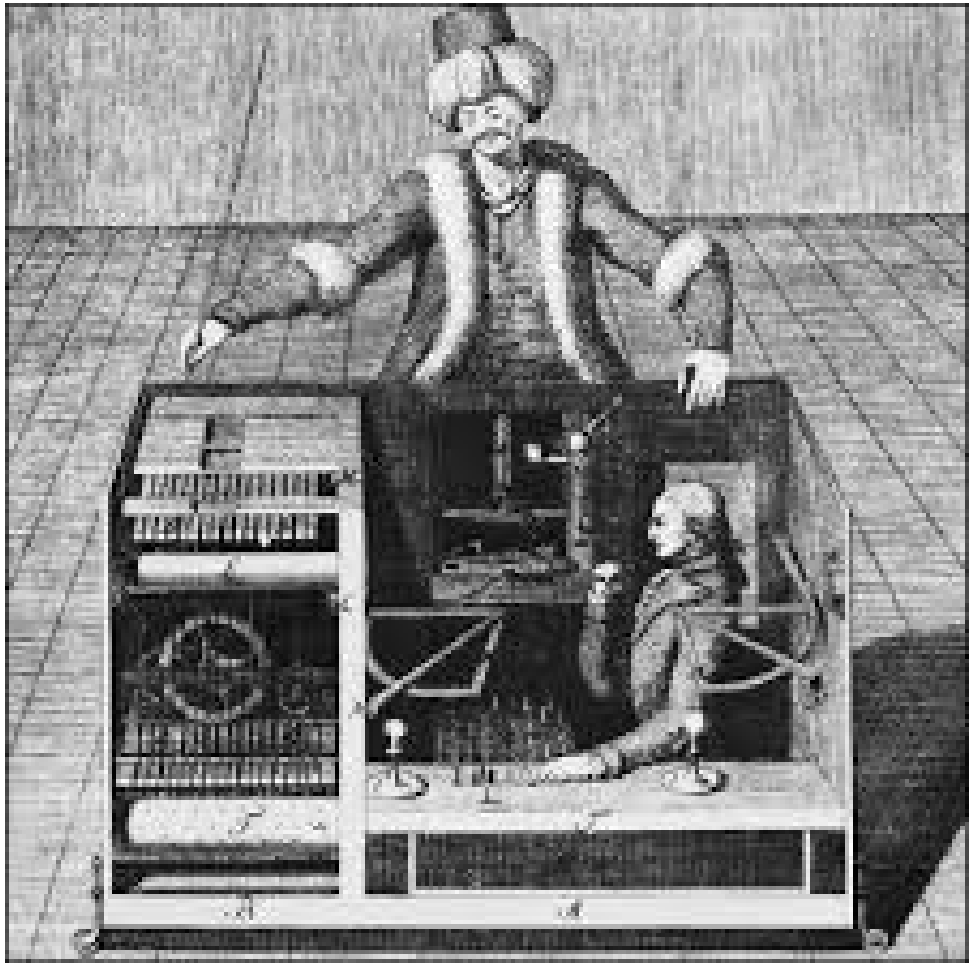
- Thinking specifically about measurement matters especially *when the construct in question is essentially contested*

- Different validity measures capture different things; top three categories:
  - Construct validity
  - Content validity
  - Convergent/Criterion validity

Today

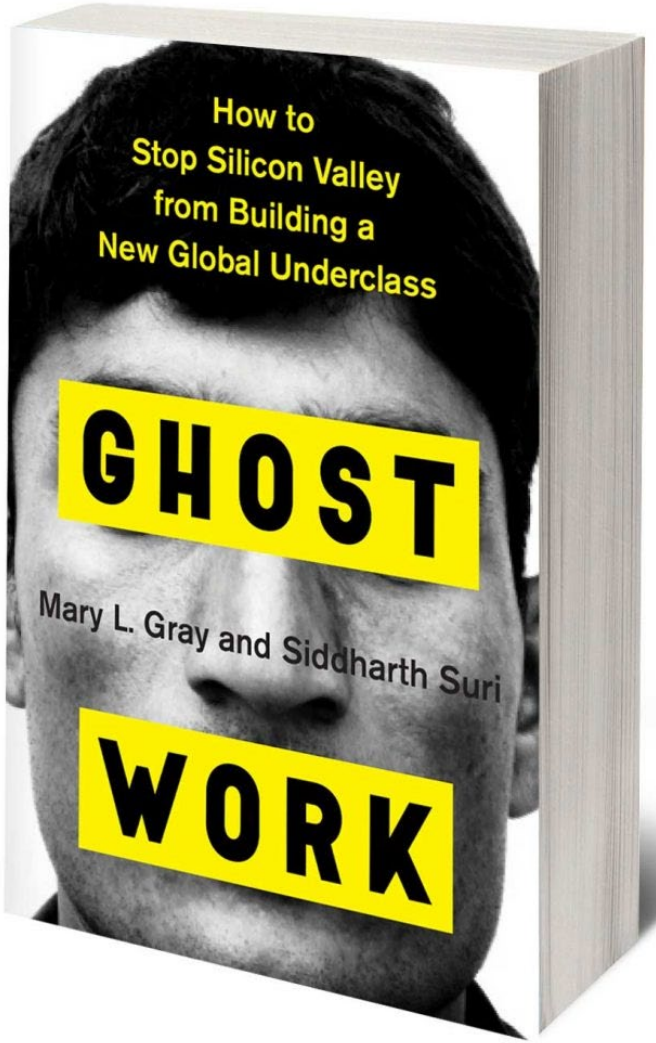
- Crowdsourcing as a method of collecting…
  - Labels (the usual)
  - Text (gaining momentum)
  - Features (less common)
  - Evaluation (pretty common)

Requisite Mechanical Turk image
(now you don’t need to use it in your talks)



Really important reading
for those who do lots of crowdsourcing

GhostWork
by Mary Gray and Sid Suri



"AI's dirty secret is just how much human labor lurks inside apparently automated systems. [This labor market] is firmly rooted in a century-long history of struggles for labor protections. Their eye-opening analysis…offers not just a detailed diagnosis of current conditions but also a path towards a better future."

—**Paul Dourish**, Chancellor's Professor, UCI

What is crowdsourcing?

A technology for connecting people who need labor done (you)

with those who are willing to perform that labor (the crowd)

for reasonable compensation

General promise of crowdsourcing

- Can hire people to perform tasks quickly and easily
- Contrary to some belief, crowdsourcing is not usually *cheaper*
  - (unless you massively underpay your workers)
- But it *is* a lot faster, and allows anyone (with $) to conduct experiments
- ….and you can get access to speakers of many languages/varieties

Many varieties of crowdsourcing exist

- Mechanical Turk, Crowdflower, etc.
  - Generally small tasks or surveys, complete API
  - Generally without specialized expertise
  - Many “anonymous” workers for a single task
- Upwork, etc.
  - Larger tasks, specialized knowledge
  - More akin to actually hiring a small number of contractors
- Others….

Corpus annotation through Crowdsourcing: Toward Best Practice Guidelines

[Marta Sabou, Kalina Bontcheva, Leon Derczynski, Arno Scharl]




I. Project Definition

1a. Select NLP Problem and crowdsourcing genre
1b. Decompose NLP problem into tasks
1c. Design crowdsourcing task

II. Data Preparation

2a. Collect and pre-process corpus
2b. Build or reuse annotator and management interfaces
2c. Run pilot studies

III. Project Execution

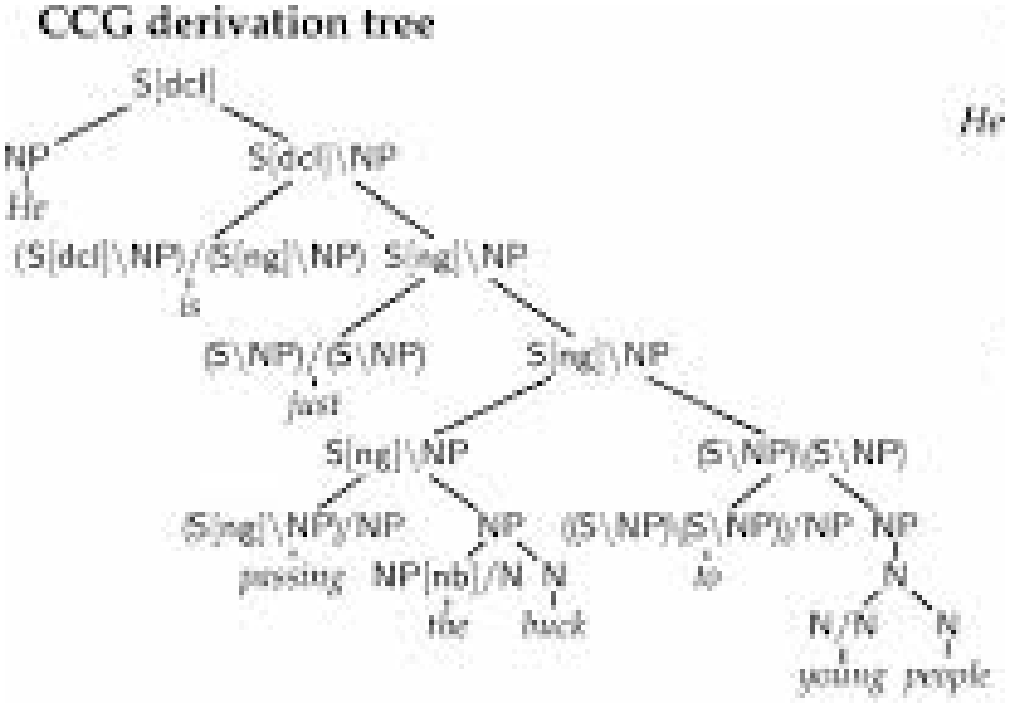
3a. Recruit and screen contributors
3b. Train, profile and retain contributors
3c. Manage and monitor crowdsourcing tasks

IV. Data Evaluation and Aggregation

4a. Evaluate and aggregate annotations
4b. Evaluate overall corpus characteristics

Project definition

- Need to break down task into small, easy, general bits [He et al., EMNLP'16]



Temple also said Sea Containers' plan raises numerous legal, regulatory, financial and fairness issues, but didn't ***elaborate***.	
Q:	What didn't ***elaborate***?
[1] ****	Temple
[2] *	Sea Containers' plan
[3]	None of the above.

Project definition

- Need to break down task into small, easy, general bits [He et al., EMNLP'16]

- Need to decide how long workers have to annotate
  - (Longer is better.)

Data preparation

- Generally a good idea to do a small, inexpensive pilot before full task

- Generally good to have “test cases” where you know the right answer

Project execution

- If you’re doing a task where language ability is very important, include a test for that

Legal & Ethical Issues

- Keep things private: delete annotators ids ASAP
- Need to decide how much to pay
  - Suggestion: Time yourself at the task, pay at least minimum wage ($10-$15/hr)
- Ensure consent
  - There's disagreement around how consent work; at least think about it
- Think about crowdworker wellbeing
  - E.g., when annotating data that, in reading it, may itself cause harm

Adjudication

- You have five crowdworkers annotate an example, and need a ground truth
- Some things you can do:
  - Majority (or plurality) vote
  - Keep collecting more annotations
  - Maintain the uncertainty
  - Remove annotations from “unreliable” annotators
  - Discard examples with too much disagreement (bad idea)
- After you’ve done these things,
  inter-annotator agreement is a much trickier notion

Crowdsourcing for linguistic data

- There's increasing use of crowdworkers to provide "natural" language responses
  - have dialogs with systems
  - translate text
  - write reading comprehension questions
  - write "inferences"
- This can often lead to *annotation artifacts*
  (not unique to crowdsourced language, but worse)
  e.g., Gugurangan, Swayamdipta et al., NAACL'18

Premise	Two dogs are running through a field.
Entailment	There are **animals outdoors**.
Neutral	Some puppies are running **to catch a stick**.
Contradiction	The pets are **sitting on a couch**.

Model	**SNLI**
majority class	34.3
`fastText`	**67.0**

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the* maybe correct *category because it's impossible for the dogs to be both running and sitting.*

Figure 1: The instructions used on Mechanical Turk for data collection.

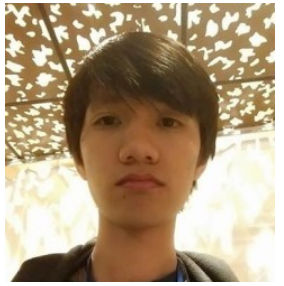
Elicit and verify

- When eliciting language from crowdworkers
- Common to have separate crowdworkers verify that language
- Example from Nguyễn Xuân Khánh for collecting cross-lingual summaries

Language	ISO 639-1	gv-snippet	gv-crowd
Number of articles			
English	en	4,573	529
Spanish	es	3,921	487
Malagasy	mg	2,680	374
Bengali	bn	2,253	352
French	fr	2,130	352
Portuguese	pt	798	162
Russian	ru	795	139
Arabic	ar	745	191
Italian	it	718	135
Macedonian	mk	701	138
Greek	el	694	128
German	de	647	204
Japanese	ja	424	75
Swahili	sw	418	84
Dutch	nl	348	87
Other statistics			
Summarized by		GV authors/translators	MTurkers
Summary languages		All versions	English
Summary lengths (words)		(*)	40-50
Article lengths (words)		150-500	150-350

- First, collect summaries
- Then have other workers evaluate summaries
  - Accuracy
  - Coverage
  - Understandability
- Evaluation criteria given to first set of workers

(New Frontiers in Summarization 2019)



Crowdsourcing features

(e.g., Zou, Chaudhuri & Kalai, 2015)

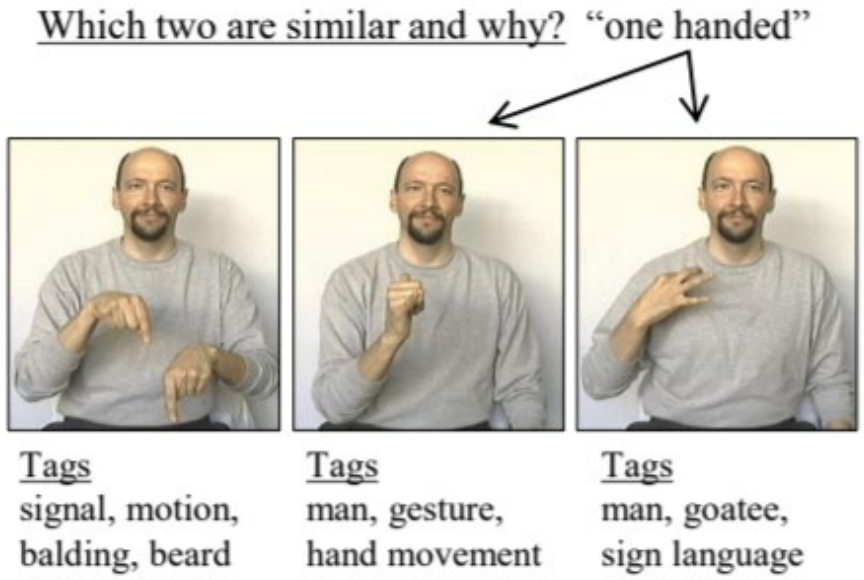


Figure 3: The first five features obtained from a representative run of the Adaptive Triple algorithm on the signs (left), faces (middle) and products (right) datasets. Each triple of images is shown in a row beside the proposed feature, and the two examples declared to have that feature are shown on the left, while the remaining example is shown on the right.

Crowd evaluation

(Nenkova, Passonneau, McKeown, ACM'07;
Shapira et al., NAACL'19)

Summary content units

A1. The industrial espionage case involving GM and VW began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.

B3. However, he left GM for VW under circumstances, which along with ensuing events, were described by a German judge as "potentially the biggest-ever case of industrial espionage".

C6. He left GM for VW *in March 1993*.

D6. The issue stems from the alleged recruitment of GM's eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.

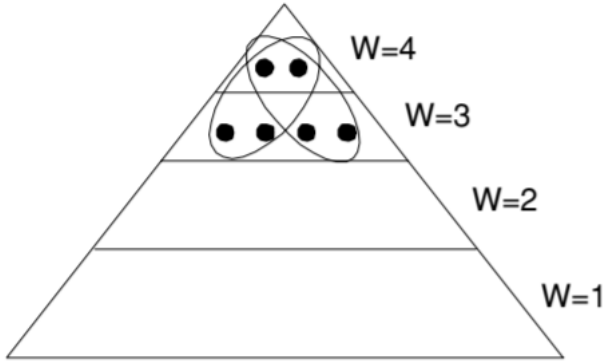
E1. On March 16, 1993, with Japanese car import quotas to Europe expiring in two years, renowned cost-cutter, Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing and Production director.

F3. In March 1993, Lopez and seven other GM executives moved to VW overnight.

SCU1 (w=6): *Lopez left GM for VW*
A1. the hiring of Jose Ignacio Lopez, an employee of GM . . . by VW
B3. he left GM for VW
C6. He left GM for VW
D6. recruitment of GM's . . . Jose Ignacio Lopez
E1. Agnacio Lopez De Arriortua, left his job . . . at General Motor's Opel . . . to become Volkswagen's . . . director
F3. Lopez . . . GM . . . moved to VW

SCU2 (w=3) *Lopez changes employers in March 1993*
C6 in March, 1993
E1. On March 16, 1993
F3. In March 1993

Pyramid Evaluation



Crowd Version

1. Extract Units
   - Avoid merging SCUs
   - Filter "noisy" workers
2. Evaluate
   - Present summary
   - Present SCUs
   - Count matches

Interactive crowd experiments (cf nodegame.org, Stefano Balietti)

What Is an Experiment?

- An experiment is a ***methodological procedure*** carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis.
- A ***test*** under ***controlled conditions*** that is made to demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried.
- An experiment is an empirical method that ***arbitrates between competing hypotheses***.

Interactive crowd experiments (cf nodegame.org, Stefano Balietti)

What Is a Synchronous Experiment?

- It is an experiment where participants
  - (i) interact in a **common environment**,
  - (ii) at the **same time**,
  - (iii) and their actions have an ***immediate* effect** on the decision process, and experimental outcome (e.g., monetary payoff)

- Synchronous experiments can be **turn-based** or **real-time**

Today

- Crowdsourcing as a method of collecting…
  - Labels (the usual)
  - Text (gaining momentum)
  - Features (less common)
  - Evaluation (pretty common)