

Crowdsourcing NLP data

CS 685, Fall 2020

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

many slides from Chris Callison-Burch

stuff from last time...

- Topics you want to see covered?

Crowdsourcing

- Useful when you have a short, simple task that you want to scale up
 - Sentiment analysis: SST-2 (label a sentence as pos/neg)
 - Question answering: SQuAD, etc (write a question about a paragraph)
 - Textual entailment: SNLI, MNLI (write a sentence that entails or contradicts a given sentence)
 - Image captioning: MSCOCO (write a sentence describing a given image)
 - etc.

Why are we learning about this?

- We've learned about all of the state-of-the-art models at this point
- How do we test the limits of these models?
 - We design newer more challenging tasks... these tasks require new datasets
- Data collection is perhaps even more important than modeling these days
 - and it's often not done properly, which negatively impacts models trained on them

Amazon Mechanical Turk

- www.mturk.com
- Pay workers to do your tasks (called “human intelligence tasks” or HITs)!
- Most common crowdsourcing platform for collecting NLP datasets (and also in general)

Building your own HIT

(for easy tasks)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with $\{\text{variables}\}$
- Upload a CSV file to populate the variables
- Pre-pay Amazon for the work
- Approve/reject work from Turkers
- Analyze results

[Home](#)[Create](#)[Manage](#)[Developer](#)[Help](#)[New Project](#)[New Batch with an Existing Project](#)[Create HITs individually](#)

Start a New Project

Categorization

[Data Collection](#)[Moderation of an Image](#)[Sentiment](#)[Survey](#)[Survey Link](#)[Tagging of an Image](#)[Transcription from A/V](#)[Transcription from an Image](#)[Writing](#)[Other](#)

Example of Categorization

Choose the best category for this image



- kitchen
- living
- bath
- bed
- outside

[View Instructions](#) ↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

You must ACCEPT the HIT before you can submit the results.

[Create Project »](#)

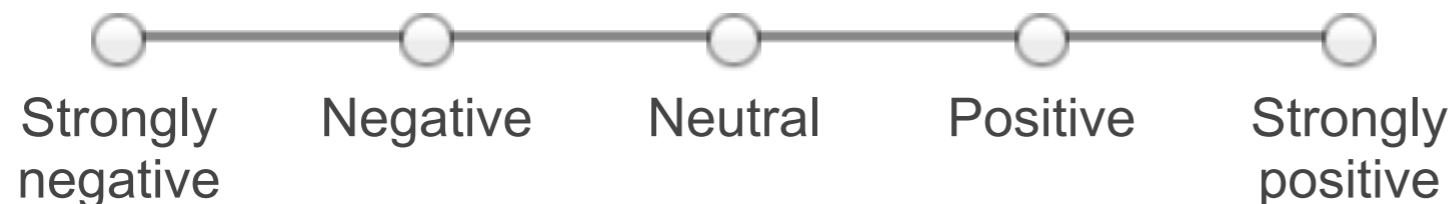
Sentiment

Pick the best sentiment based on the following criterion.

Strongly positive	Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"
Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
Strongly negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are terrific... NOTTTT!!!!!" or "I will NEVER shop there again!!!"

Judge the sentiment expressed by the following item toward: Amazon

If you loved Firefly TV show, amazing Amazon price for entire series: about \$27 BlueRay & \$17 DVD.



Sentiment Project

Create Project

Provide Instructions

Upload Data File

Preview

Checkout

Instructions for Workers

For each point on the scale, define the criteria Workers should follow when selecting it. We have provided sample instructions that you can customize below.

Strongly positive:

Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"

Positive:

Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."

Neutral:

Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"

Negative:

Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."

Strongly negative:

Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are terrific... NOTTTT!!!!!!" or "I will NEVER shop there again!!!!"

Number of Responses

How many Workers do you want to rate sentiment for each item? (Help me choose)

1
2
3
4
✓ 5
6
7
8
9

Home

Create

Manage

Developer

Help

New Project

New Batch with an Existing Project

Create HITs individually

Obama sentiment 9/13 — Judge the sentiment expressed by the following item toward: President Obama

How does it work?

Sentiment Project

Create Project

Provide Instructions

Upload Data File

Preview

Checkout

Upload Data File

Please provide a .csv file of your data so that we can create your work items. [\(learn more\)](#)

Don't have one? You can [view a sample](#).

Choose File

tweets.csv

Upload

After your file is uploaded, you can specify which columns to show Workers.

Next »

Leave feedback for this page.

Obama sentiment 9/13 2 — Judge the sentiment expressed by the following item toward: President Obama

[How does it work?](#)

Sentiment Project

[Create Project](#)[Provide Instructions](#)[Upload Data File](#)[Preview](#)[Checkout](#)

Upload Data File - Choose Fields

Each column from your .csv file is shown as a field below. Tell us which fields you want to show Workers and whether the fields contain text, a link to an image, or a link to a website.

Optionally, you can add a label to appear next to your data. For instance, if the field is a "tweet," you might want to add a label of "Tweet."

Field	Show Workers?	Type of Data	Label
Tweet	<input checked="" type="checkbox"/>	Text	Tweet
user	<input type="checkbox"/>	Text	username

If this doesn't look right, you can [upload a different file](#).

[Next »](#)[Leave feedback for this page.](#)

Project Name: This name is not displayed to Workers.

Describe your HIT to Workers

Title

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description

Give more detail about this task. This gives Workers a bit more information before they decide to view your HIT.

Keywords

Provide keywords that will help Workers search for your HITs.

This project may contain potentially explicit or offensive content, for example, nudity. ([See details](#))

Setting up your HIT

Reward per assignment

Tip: Consider how long it will take a Worker to complete each task. A 30 second task that pays \$0.05 is a \$6.00 hourly wage.

Number of assignments per HIT

How many unique Workers do you want to work on each HIT?

Time allotted per assignment

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

HIT expires in

Maximum time your HIT will be available to Workers on Mechanical Turk.

Results are automatically approved in

After this time, all unreviewed work is approved and Workers are paid.

Obama sentiment 9/13 — Judge the sentiment expressed by the following item toward: President Obama

[How does it work?](#)

Sentiment Project

[Create Project](#)[Provide Instructions](#)[Upload Data File](#)[Preview](#)[Checkout](#)

Checkout

Number of Items		1,000
Number of Workers per Item	x	5
Number of Worker Submissions	=	5,000
Reward per Submission (details)	x	\$0.020 
Total Worker Rewards	=	\$100.000
Total Mechanical Turk fees (details)	+	\$45.000
Total cost	=	\$145.000

Items Completed 46 of 1000

■ Negative 2%
 ■ Neutral 2%
 ■ Positive 0%
 ■ Incomplete 95%
[Download Results](#)[Add Time](#)[Cancel](#)**Sentiment Project**[Answer Summary](#)[Results](#)[Cost](#)**Details**

Project Obama sentiment 9/13

Status in Progress

Question

Judge the sentiment expressed by the following item toward: President Obama

Created 09/08/13 13:10

Time elapsed 28 minutes

Est. completion 09/08/13 14:35

Expiration time 09/12/13 13:12

Worker time limit 1 hour

Effective hourly rate \$4.000

Number of Workers 5

Input file tweets.csv

Results [download results](#)

Filter: Positive

[Instructions](#) →**Judge the sentiment expressed by the following item toward: President Obama****Tweet:** @MeatSauce1 Obama and the NFL has saved Detroit!! #ford #HardcorePawAverage Sentiment rating of **0.6** based on **5 responses**

Strongly Positive (+2)	<div style="width: 0%;"></div>	(0)
Positive (+1)	<div style="width: 60%;"></div>	(3)
Neutral (0)	<div style="width: 40%;"></div>	(2)
Negative (-1)	<div style="width: 0%;"></div>	(0)
Strongly Negative (-2)	<div style="width: 0%;"></div>	(0)

Purpose of redundancy

- MTurk lets you set the number of assignments per HIT
- That gives you different (redundant) answers from different Turkers
- This lets you conduct surveys (num assignments = num respondents)
- Also, lets you take votes and do tie-breaking, or do quality control
- Redundancy $\geq 10x$ incurs higher fees on MTurk

Worker Requirements

Advanced

[Worker requirements](#) «

Worker requirements:

Specify ALL the qualifications Workers must meet to work on your HITs:

is [remove](#)

greater than or equal to [remove](#)

greater than or equal to [remove](#)

[\(+\)](#) Add another criterion (up to 5)

Only Workers who qualify to do my HITs can preview my HITs.

Yes No

Also critical for model evaluation!

Instructions Shortcuts What is the paraphrase relationship between the rewritten sentence and the original? ⚙️

original sentence: Damn ... I think it 's impossible to choose just one .

rewritten sentence: Damn . . . I don 't have one I can choose just by one .

Select an option

- no paraphrase relationship 1
- approximately the same meaning, but the rewritten sentence is ungrammatical 2
- approximately the same meaning and the rewritten sentence is grammatical 3

Why might we prefer human evaluation over automatic evaluation (e.g., BLEU score)?

Collecting data from MTurk can have unintended consequences for models if you're not careful!

strategies used by crowd workers

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family .
Contradiction	A woman is not taking money for any of her sticks.

The result: models can predict the label without seeing the premise sentence!

Model	SNLI	MultiNLI	
		Matched	Mismatched
majority class	34.3	35.4	35.2
fastText	67.0	53.9	52.3

Table 2: Performance of a premise-oblivious text classifier on NLI. The MultiNLI benchmark contains two test sets: matched (in-domain examples) and mismatched (out-of-domain examples). A majority baseline is presented for reference.

Were workers misled by the annotation task examples?

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 .

Were workers misled by the annotation task examples?

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

Were workers misled by the annotation task examples?

Premise	Two dogs are running through a field.	
Entailment	There are animals outdoors.	generic words
Neutral	Some puppies are running to catch a stick.	Add cause /
Contradiction	The pets are sitting on a couch.	purpose clause

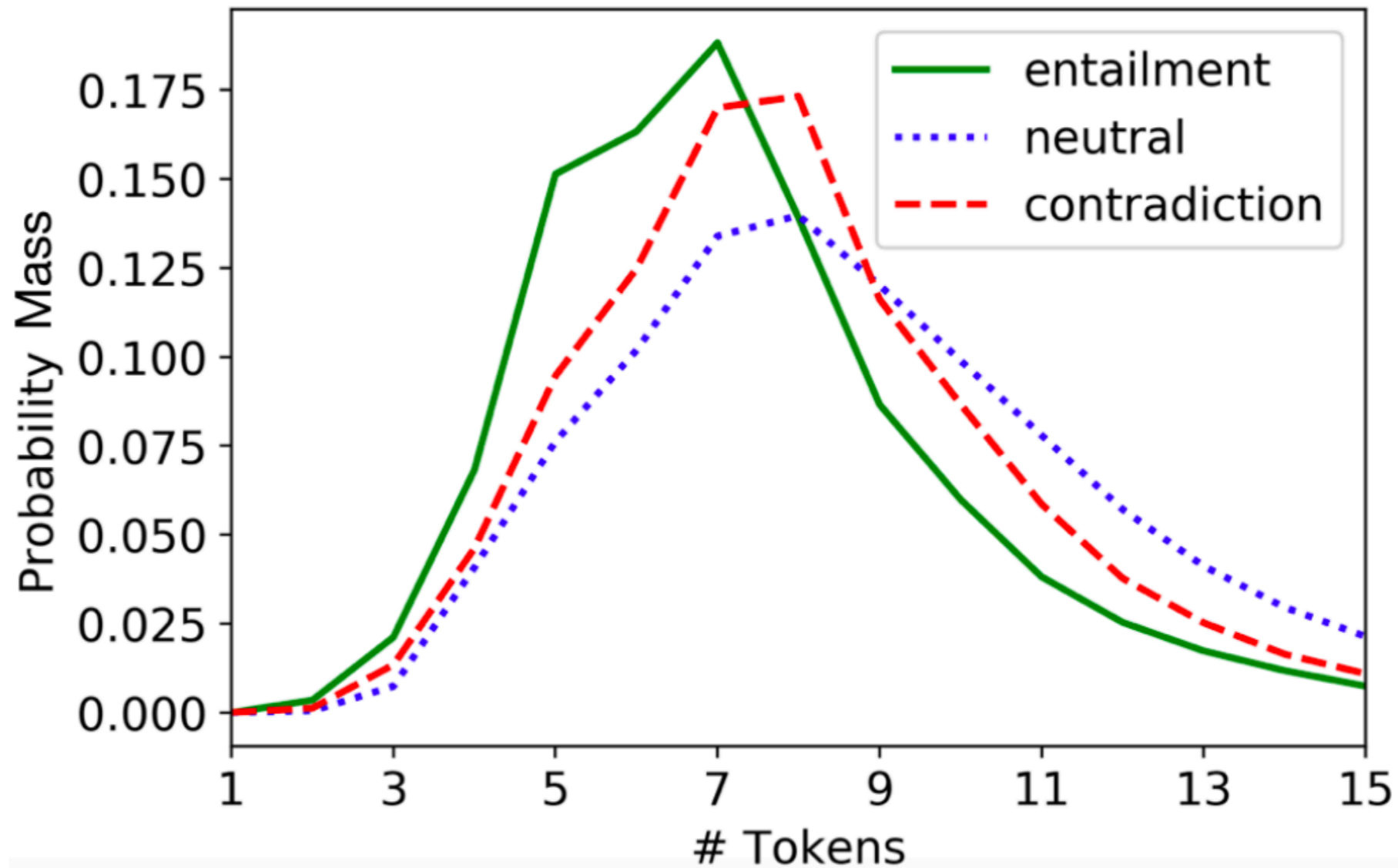
Were workers misled by the annotation task examples?

Premise	Two dogs are running through a field.	
Entailment	There are animals outdoors.	generic words
Neutral	Some puppies are running to catch a stick.	Add cause / purpose clause
Contradiction	The pets are sitting on a couch.	Add words that contradict any activity

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%



Table 4: Top 5 words by $\text{PMI}(\text{word}, \text{class})$, along with the proportion of *class* training samples containing *word*. MultiNLI is abbreviated to MNLI.

Sentence length is correlated to the label



Entailments are shorter than neutral sentences!

Issues with SQuAD

	Image Classification	Reading Comprehension
Possible Input		Tesla moved to the city of Chicago in 1880.
Similar Input		Tadakatsu moved to the city of Chicago in 1881.
Semantics	Same	Different
Model's Mistake	Considers the two to be different	Considers the two to be the same
Model Weakness	Overly sensitive	Overly stable

Issues with SQuAD

Article: Super Bowl 50

Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Crowdsourcing works for tasks that are

- Natural and easy to explain to non-experts
- Decomposable into simpler tasks that can be joined together
- Parallelizable into small, quickly completed chunks
- Well-suited to quality control (some data has correct gold standard annotations)

Crowdsourcing works for tasks that are

- Robust to some amount of noise/errors (the downstream task is training a statistical model)
- Balanced and each task contains the same amount of work
 - Don't have tons of work in one assignment but not another
 - Don't ask Turkers to annotate something occurs in the data $\ll 10\%$ of the time

Guidelines for your own tasks

- Simple instructions are required
- If your task can't be expressed in one paragraph + bullets, then it may need to be broken into simpler sub-tasks

Guidelines for your own tasks

- Quality control is paramount
 - Measuring redundancy doesn't work if people answer incorrectly in systematic ways
 - Embed gold standard data as controls
- Qualification tests v. no qualification test
 - Reduce participation, but usually ensures higher quality

More complex tasks?

- You can host your own task on a separate server, which Turkers can then join
- They complete tasks, and then receive a code which they can paste into the Amazon MT site to get paid

QuAC dialog QA example

turker 1

student

turker 2

teacher

QuAC dialog QA example

turker 1

student

- provided with a topic to ask questions about (e.g., *Daffy Duck - origin & history*)
- asks questions to learn as much as they can about this topic

**Q: what is the origin of
Daffy Duck?**

turker 2

teacher

QuAC dialog QA example

turker 1

student

- provided with a topic to ask questions about (e.g., *Daffy Duck - origin & history*)
- asks questions to learn as much as they can about this topic

Q: what is the origin of Daffy Duck?

turker 2

teacher

- provided full text of Wikipedia section on Daffy Duck's origin

Origin and history [edit]

Daffy first appeared in *Porky's Duck Hunt*, released on April 17, 1937. The cartoon was directed by **Tex Avery** and animated by **Bob Clampett**. *Porky's Duck Hunt* is a standard hunter/prey pairing for which **Leon Schlesinger's** studio was famous, but Daffy (barely more than an unnamed bit player in this short) was something new to moviegoers: an assertive, completely unrestrained, combative protagonist. Clampett later recalled:

"At that time, audiences weren't accustomed to seeing a cartoon character do these things. And so, when it hit the theaters it was an explosion. People would leave the theaters talking about this daffy duck."^[3]

This early Daffy is less **anthropomorphic** and resembles a "normal" **black duck**. In fact, the only aspects of the character that have remained consistent through the years are his voice characterization by **Mel Blanc**; and his black feathers with a white neck ring. Blanc's characterization of Daffy once held the world record for the longest characterization of one animated character by his or her original actor: 52 years.

The origin of Daffy's voice, with its notable **lateral lisp**, is a matter of some debate. One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp. However, in Mel Blanc's autobiography, *That's Not All Folks!*, he contradicts that conventional belief, writing, "It seemed to me that such an extended mandible would hinder his speech, particularly on words containing an s sound. Thus 'despicable' became 'destpicable.'"

Daffy's slobbery, exaggerated lisp was developed over time, and it is barely noticeable in the early cartoons. In *Daffy Duck & Egghead*, Daffy does not lisp at all except in the separately drawn set-piece of Daffy singing "*The Merry-Go-Round Broke Down*" in which just a slight lisp can be heard.

In *The Scarlet Pumpernickel* (1950), Daffy has a middle name, Dumas as the writer of a swashbuckling script, a nod to **Alexandre Dumas**. Also, in the *Baby Looney Tunes* episode "The Tattletale", Granny addresses Daffy as "Daffy Horatio Tiberius Duck". In *The Looney Tunes Show* (2011), the joke middle names "Armando" and "Sheldon" are used.

Aliases	Duck Dodgers
Species	American black duck
Gender	Male
Significant other(s)	Melissa Duck Tina Russo (<i>The Looney Tunes Show</i>) Mrs. Daffy Duck
Nationality	American

A: first appeared in Porky's Duck Hunt

QuAC dialog QA example

- External server handles worker matching, student / teacher assignment, and facilitates the dialogue
- We used Stanford's *cocoa* library to set up this data collection
 - <https://github.com/stanfordnlp/cocoa>
- Roughly \$65k spent on MTurk to collect QuAC

Problems Encountered

- so many!
- **lag time:** most important issue when two workers are interacting w/ each other
- **quality control:** unresponsive, low-quality questions, cheating > report feature
- **pay:** devised a pay scale to encourage longer dialogs
- **instructions:** workers don't read them!
we joined turker forums to pilot our task
- **validation:** expensive but necessary