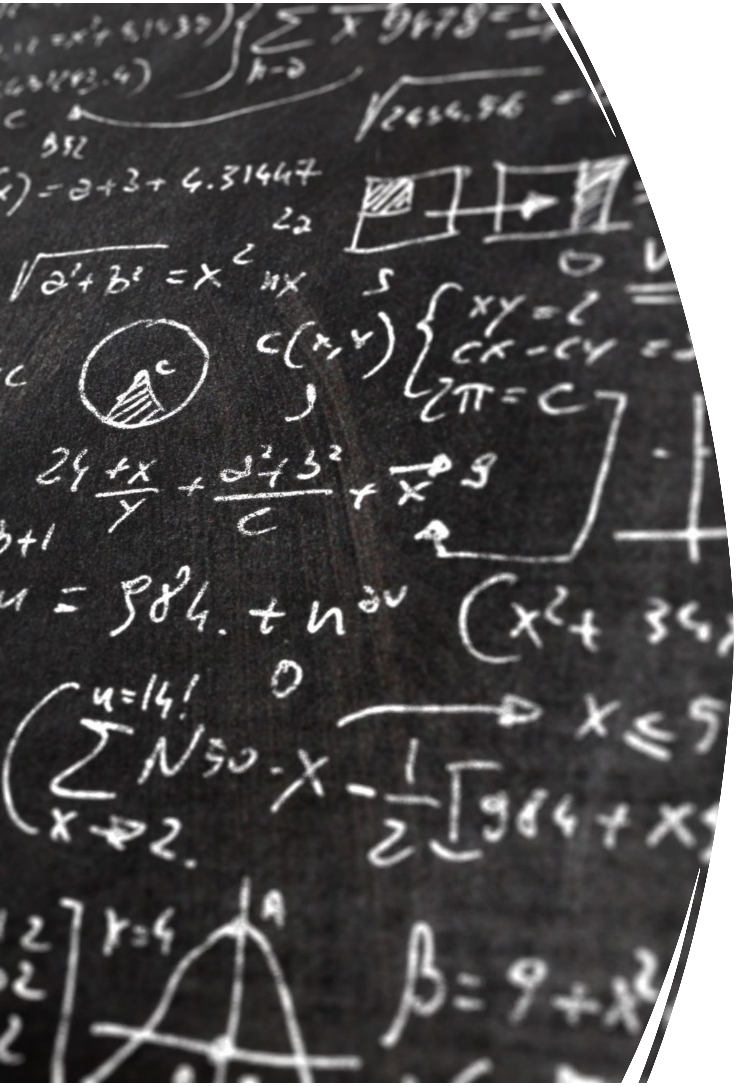


CS 383 – Computational Text Analysis

Lecture 18
Evaluation metrics
Crowdsourcing

Adam Poliak

03/27/2023

Slides adapted from Dan Jurafsky

Announcements

- HW06:
  - Should be ready tonight
  - Hopefully you’ve been collecting the tweets

Midterm - Format

Multiple Choice

Short Answer

Problems to work out by hand

Machine Learning in a nutshell

In a ML model, what are we training?

- **Parameters!**

How do we train parameters in supervised learning?

train parameters == figure out values for the parameters

- Update weights by using them to make predictions and seeing **how far off our predictions** are
  - **Loss function!**

Algorithm to learn weights?

- **SGD**
- Others exist but not covering them

Outline

Evaluation metrics (classification)

Where do labels come from?

Classify a tweet as viral or not

Accuracy

- Model A performs 60% accuracy, would you say this is good, decent, or awful?
- Model A performs 80% accuracy, would you say this is good, decent, or awful
- Model A performs 98% accuracy, would you say this is good decent or awful?

Evaluation: Accuracy

- Imagine we saw 1 million tweets
  - 100 of them were viral
  - 999,900 were not
- We could build a dumb classifier that just labels every tweet "not viral"
  - It would get 99.99% accuracy!!! Wow!!!!
  - But useless! Cant find the viral tweets!
- When should we not we use **accuracy** as our metric?
  - When data isn’t balanced across labels/classes

The 2-by-2 confusion matrix

true positive	**false positive**
false negative	**true negative**

The 2-by-2 confusion matrix

		gold standard labels	
		gold positive	gold negative
system output labels	system positive	**true positive**	**false positive**
	system negative	**false negative**	**true negative**

The 2-by-2 confusion matrix

gold standard labels

		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{\text{tp}}{\text{tp+fn}}$		**accuracy** $= \frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Evaluation: Precision

- % of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Evaluation: Recall

- % of items actually present in the input that were correctly identified by the system.

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}$$

Why Precision and recall

- Our dumb viral-classifier
  - label no tweets as ”viral"

Accuracy=99.99%

but

Recall = 0

- (it doesn't get any of the 100 viral tweets)

Precision and recall, unlike accuracy, emphasize true positives:

- finding the things that we are supposed to be looking for.

A combined measure: F

- F measure: a single number that combines P and R:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- We almost always use balanced F_1 (i.e., $\beta = 1$)

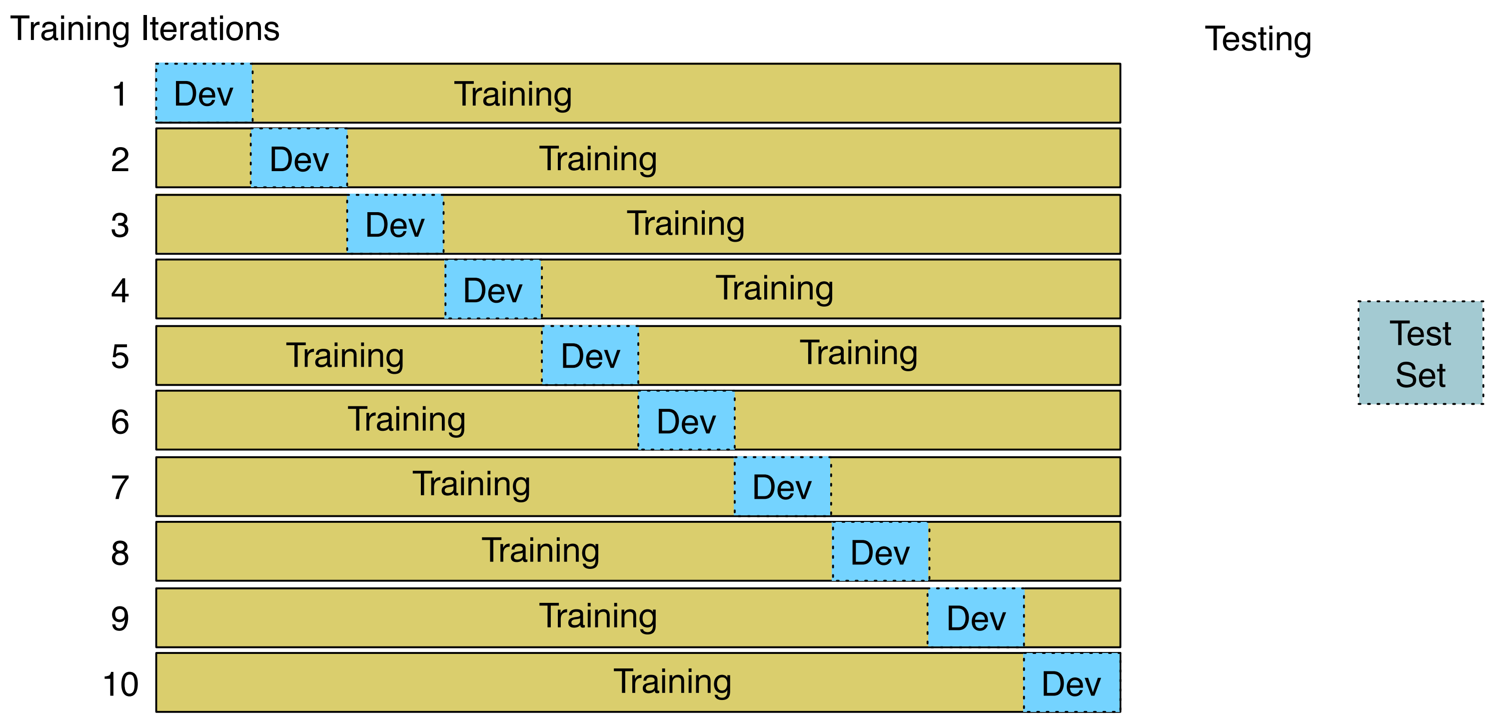
$$\mathrm{F}_1 = \frac{2PR}{P + R}$$

Development Test Sets ("Devsets") and Cross-validation

- Train on training set, tune on devset, report on testset
  - This avoids overfitting (‘tuning to the test set’)
  - More conservative estimate of performance
  - But paradox: want as much data as possible for training, and as much for dev; how to split?

Cross-validation: multiple splits

- Pool results over splits, Compute pooled dev performance



Confusion Matrix for 3-class classification

		gold labels		
		urgent	normal	spam
system output	urgent	8	10	1
	normal	5	60	50
	spam	3	30	200

precision$_u$ $= \dfrac{8}{8+10+1}$

precision$_n$ $= \dfrac{60}{5+60+50}$

precision$_s$ $= \dfrac{200}{3+30+200}$

recall$_u$ $= \dfrac{8}{8+5+3}$

recall$_n$ $= \dfrac{60}{10+60+30}$

recall$_s$ $= \dfrac{200}{1+50+200}$

How to combine Precision/Recall from 3 classes to get one metric

- **Macro**averaging:
  - compute the performance for each class, and then average over classes
- **Micro**averaging:
  - collect decisions for all classes into one confusion matrix
  - compute precision and recall from that table.

Macroaveraging and Microaveraging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

In classification: where do the labels come from?

Crowdsourcing to the rescue

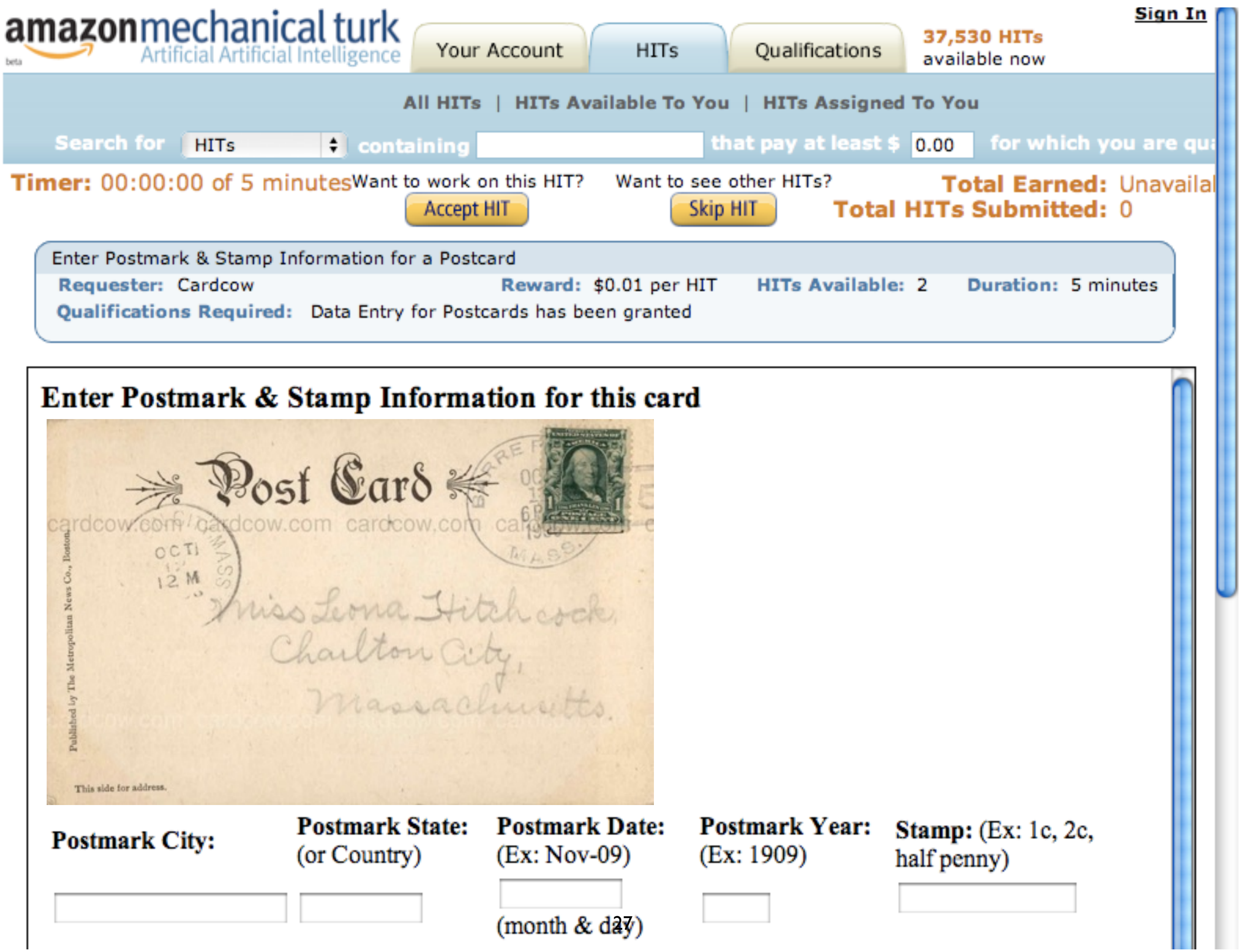
Outline

Evaluation metrics

Crowdsourcing

Example: Optical Character Recognition

- Destination City
- Destination State
- Destination Zip
- Post Mark
- Stamp



Example from Ellie Pavlick

Amazon Mechanical Turk

- https://worker.mturk.com/

Crowdsourcing Companies

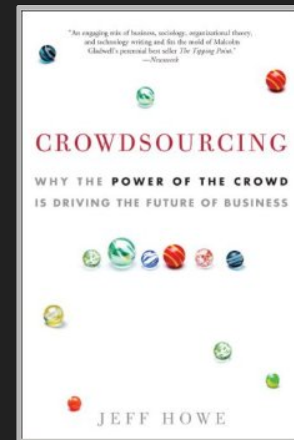
What is crowdsourcing?

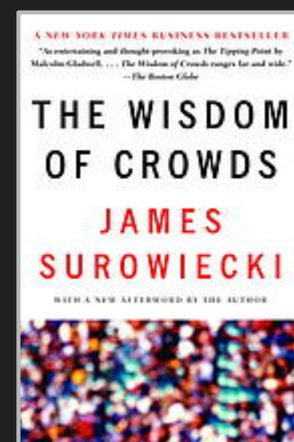
What is “Crowdsourcing”?

- An open call to a group of people

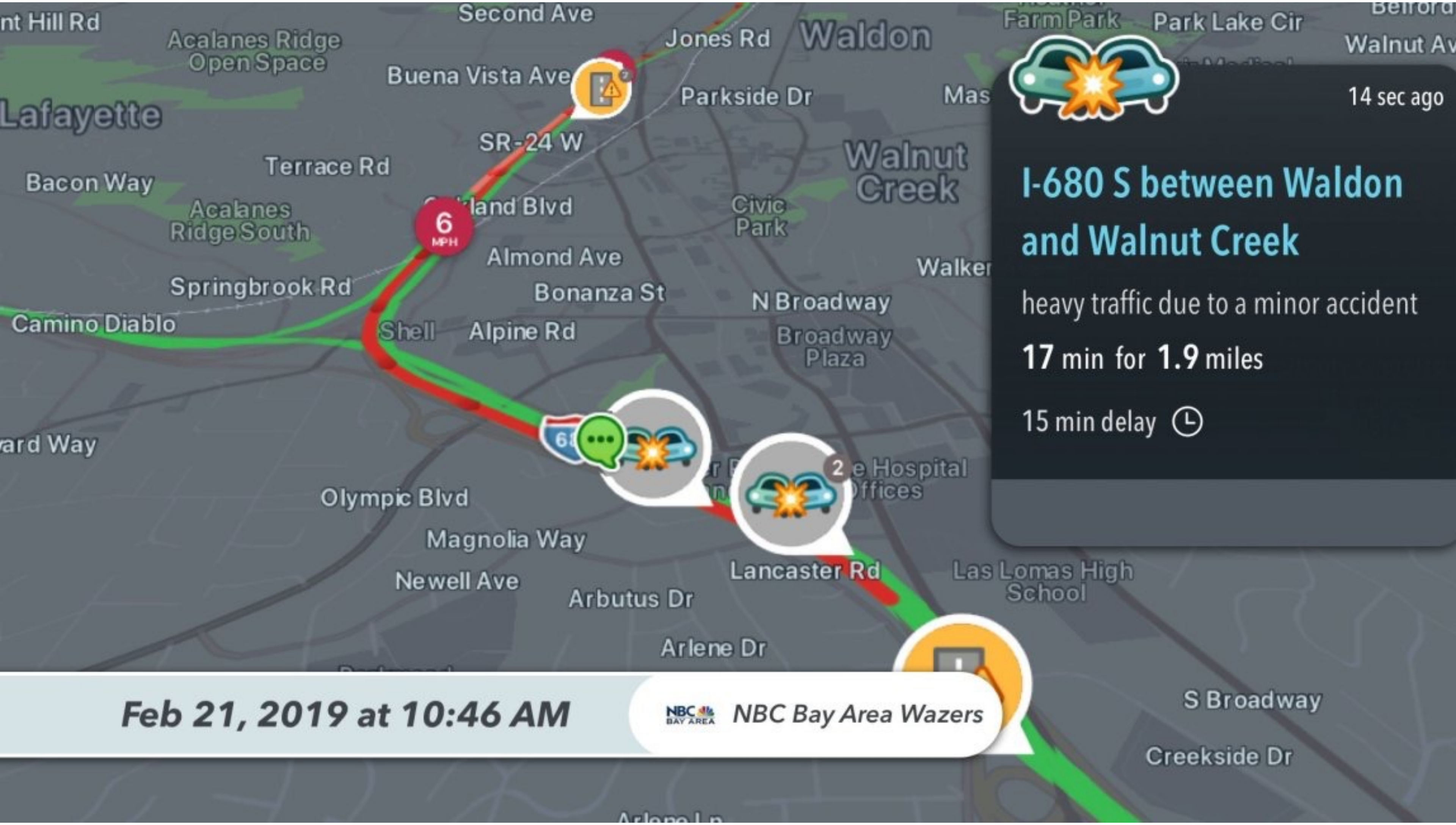
- “Crowdsourcing”
  - *“Crowdsourcing is the act of taking a job traditionally performed by a designated agent ... and outsourcing it to…a large group of people in the form of an open call.”*
  - [Jeff Howe, Wired]

- Books
  - Jeff Howe: *Crowdsourcing*
  - James Surowiecki: *The Wisdom of Crowds*





Is this Crowdsourcing????

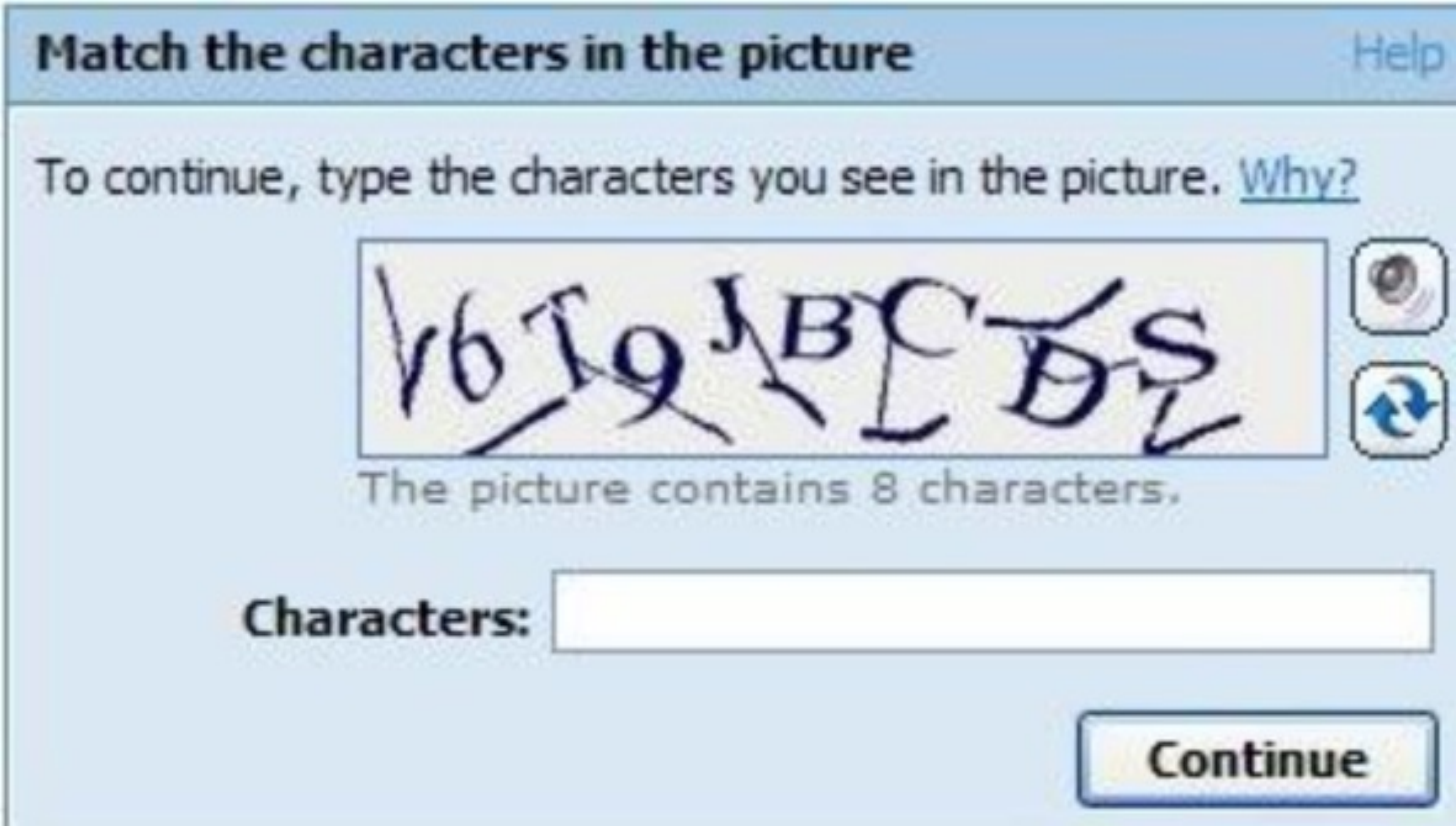

I-680 S between Waldon and Walnut Creek
14 sec ago
heavy traffic due to a minor accident
17 min for 1.9 miles
15 min delay
Feb 21, 2019 at 10:46 AM
NBC Bay Area Wazers
Lafayette
Waldon
Walnut Creek
SR-24 W
Second Ave
Jones Rd
Buena Vista Ave
Parkside Dr
Acalanes Ridge Open Space
Terrace Rd
Bacon Way
Acalanes Ridge South
Almond Ave
Civic Park
Springbrook Rd
Bonanza St
N Broadway
Camino Diablo
Alpine Rd
Broadway Plaza
Olympic Blvd
Magnolia Way
Newell Ave
Arbutus Dr
Lancaster Rd
Las Lomas High School
Arlene Dr
S Broadway
Creekside Dr
Park Lake Cir
6 MPH

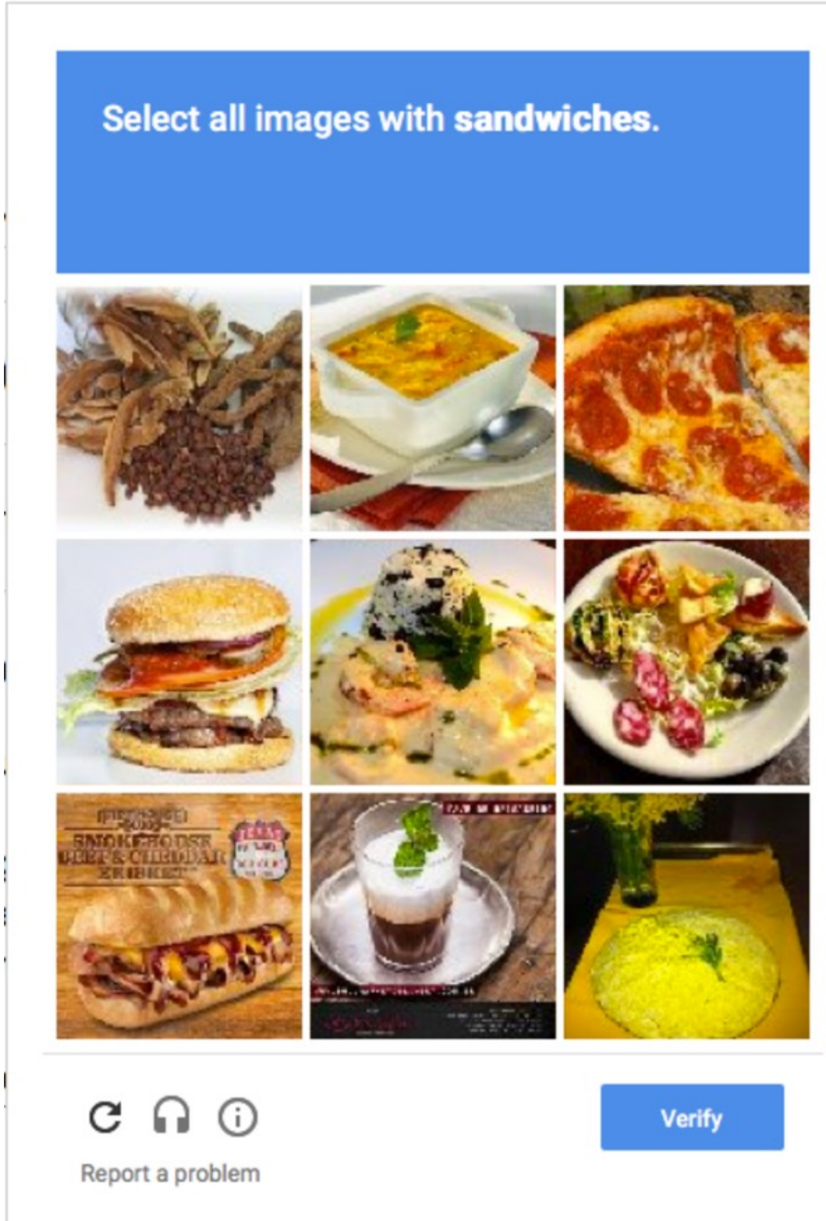

airbnb


lyft
DO NOT
OPEN DOOR
IT'S AUTOMATIC

Why Crowdsourcing?

- No one worker will *always* be available
- Open call allows for more available human intelligence
  - Allow for the creation of on-demand systems
  - Even real-time becomes possible — 1s responses or less with multiplexing
- Any individual has a chance of error
  - With groups of workers, we might be able to reduce this error rate
  - Especially for ephemeral workers
- Collectively, we can get pieces that work together in parallel


Match the characters in the picture
Help
To continue, type the characters you see in the picture. Why?
The picture contains 8 characters.
Characters:
Continue


Select all images with sandwiches.
Verify
Report a problem

CAPTCHA

CAPTCHA

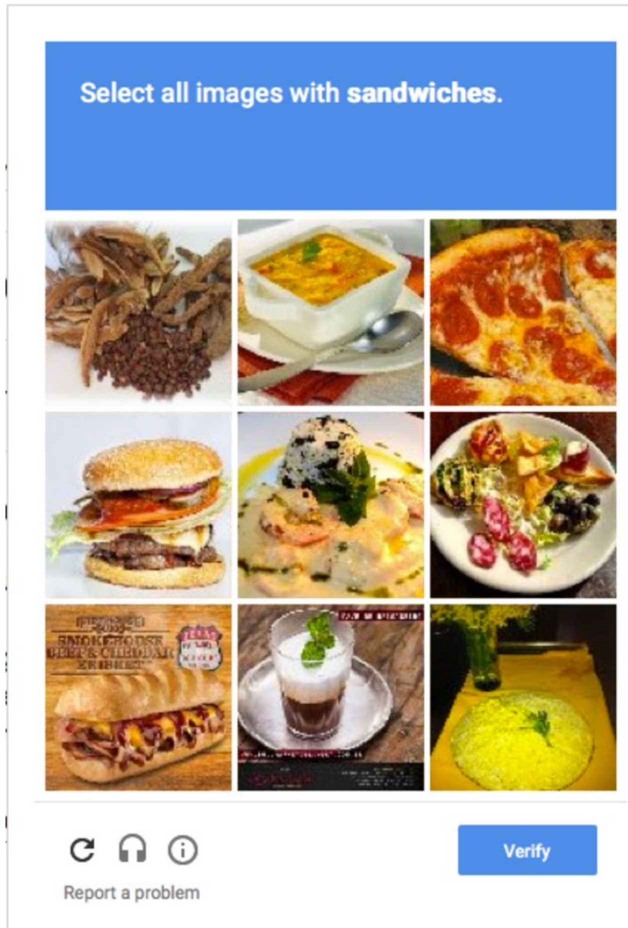
- **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part

CAPTCHA

- **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part
- Verify users are humans, not bots

reCAPTCHA

reCAPTCHA



reCAPTCHA

reCAPTCHA: The Genius Who's Tricking the World Into Doing His Work

Turns out reCAPTCHAs are doing more than 'proving you're a human' and being annoying. Introducing human-based computation.

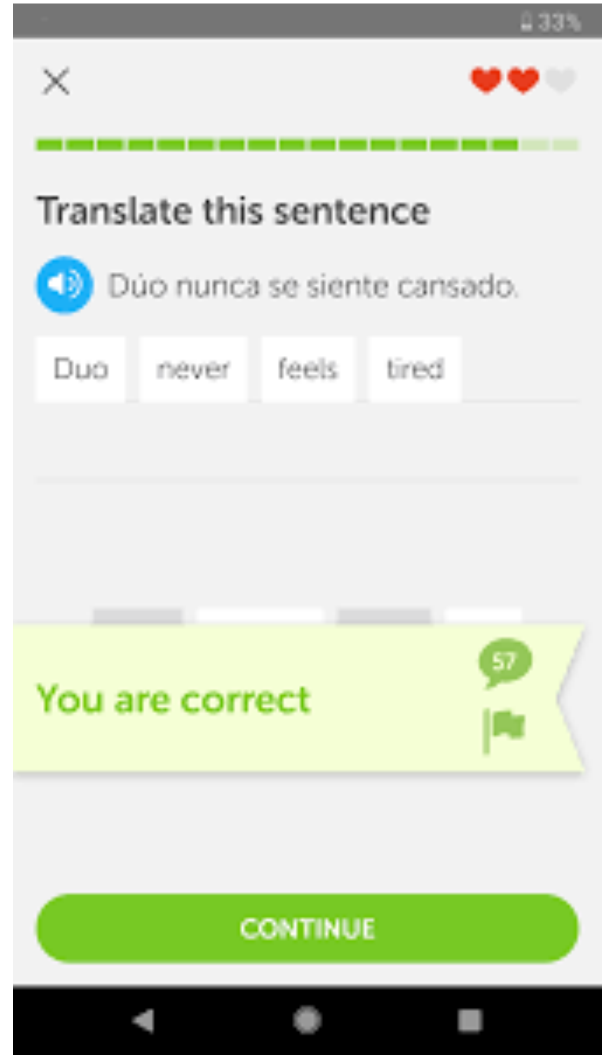
BY JOHN HAVEL
DECEMBER 3, 2015

- Planet money podcast with Luis Von Ahn:

- ~4:30 until 7:00 - https://www.npr.org/templates/transcript/transcript.php?storyId=716827880

- *KING: The New York Times ended up being reCAPTCHA's first client. Now when you solve a CAPTCHA, next to a few random letters and numbers, there was also a picture of a word from an old issue of the Times that computers couldn't read. When you typed in that word, you weren't just protecting the Internet from spam. You were also helping to turn a hundred years of old newspapers into a searchable digital archive.*

duolingo



33%
Translate this sentence
Dúo nunca se siente cansado.
Duo
never
feels
tired
You are correct
57
CONTINUE

Duolingo now translating BuzzFeed and CNN

Luis PLUS 25 18 11 9 5 1401

This week we're excited to announce partnerships with BuzzFeed and CNN to have their content translated by our community of language learners.

duolingo
+
BuzzFeed
+
CNN

Goals

- Understanding Crowdsourcing for AI

- Examples of Crowdsourcing

- Issues of Crowdsourcing

Crowdsourcing for AI


A.I. Is Learning From Humans. Many Humans.
Artificial intelligence is being taught by thousands of office workers around the world. It is not exactly futuristic work.

iMerit employees must learn unusual skills for their labeling, like spotting a problematic polyp on a human intestine. Rebecca Conway for The New York Times

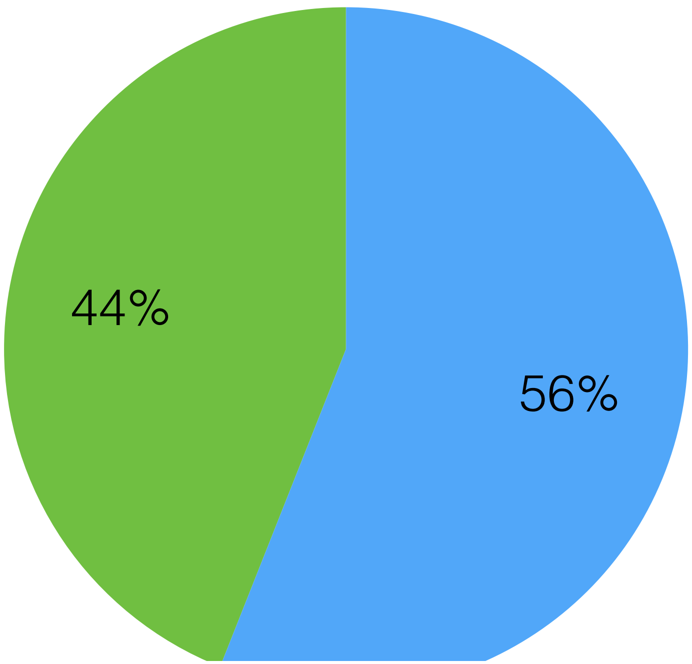
Crowdsource Workers

Who is doing all this work for us?

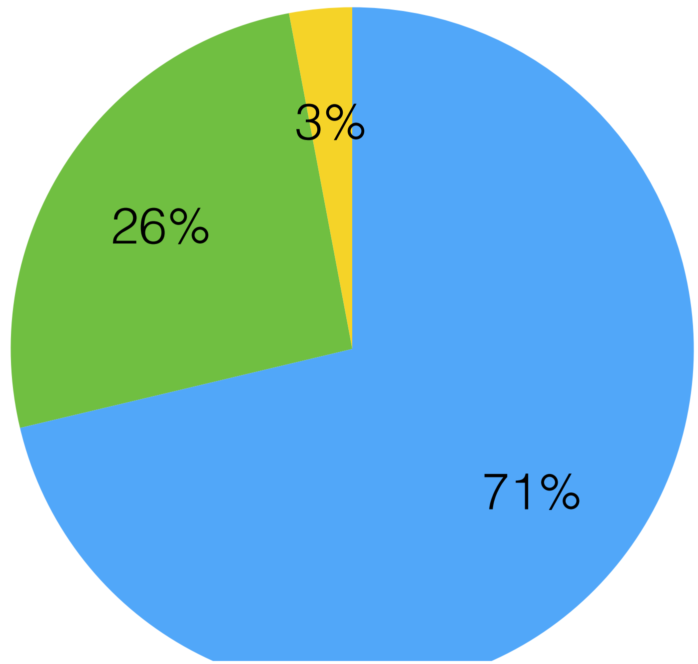
76

Who is doing all this work for us?

Mechanical Turk



CrowdFlower



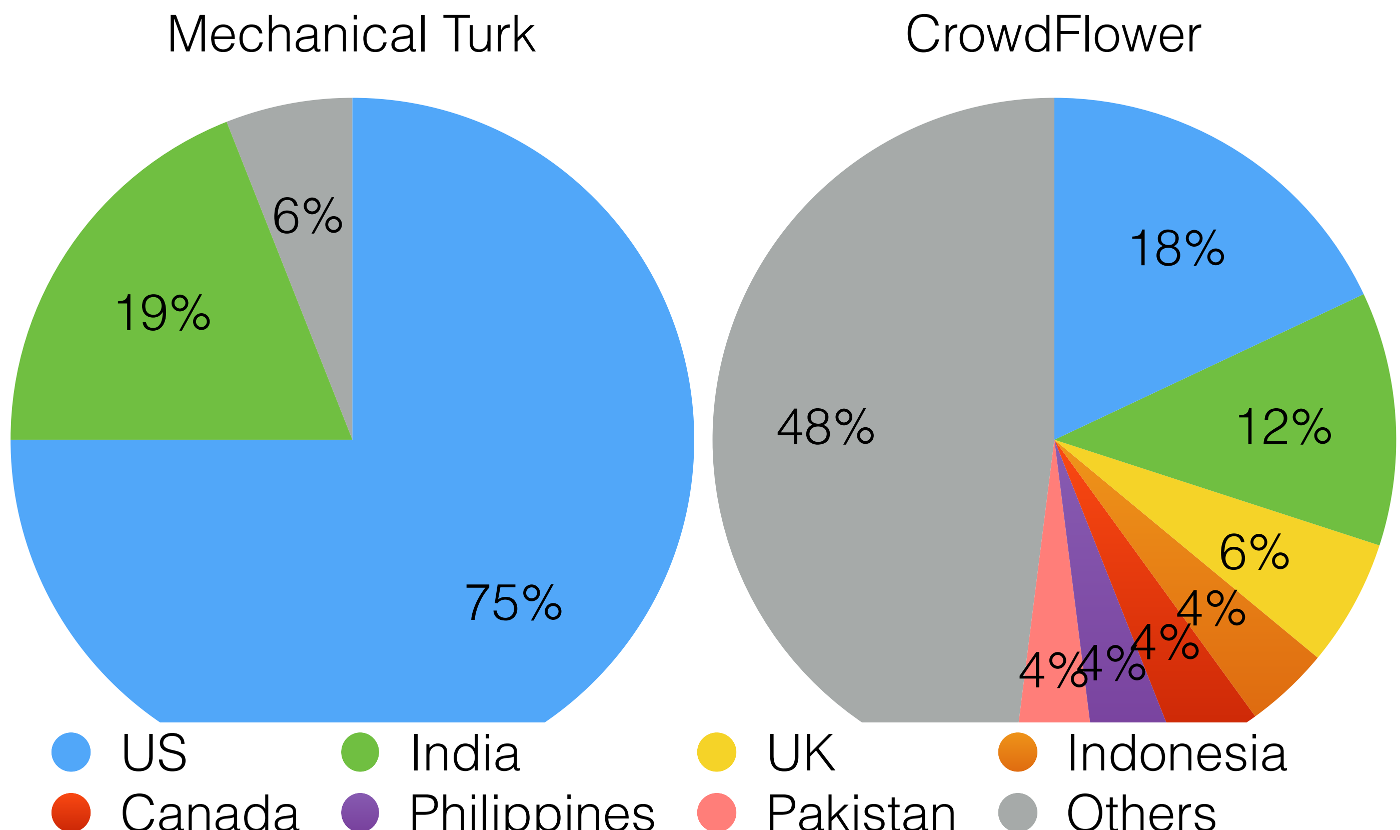
Male Female Unspecified

Slide from Ellie Pavlick

- Panos Ipeirotis's site: http://demographics.mturk-tracker.com
- CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

77

Who is doing all this work for us?

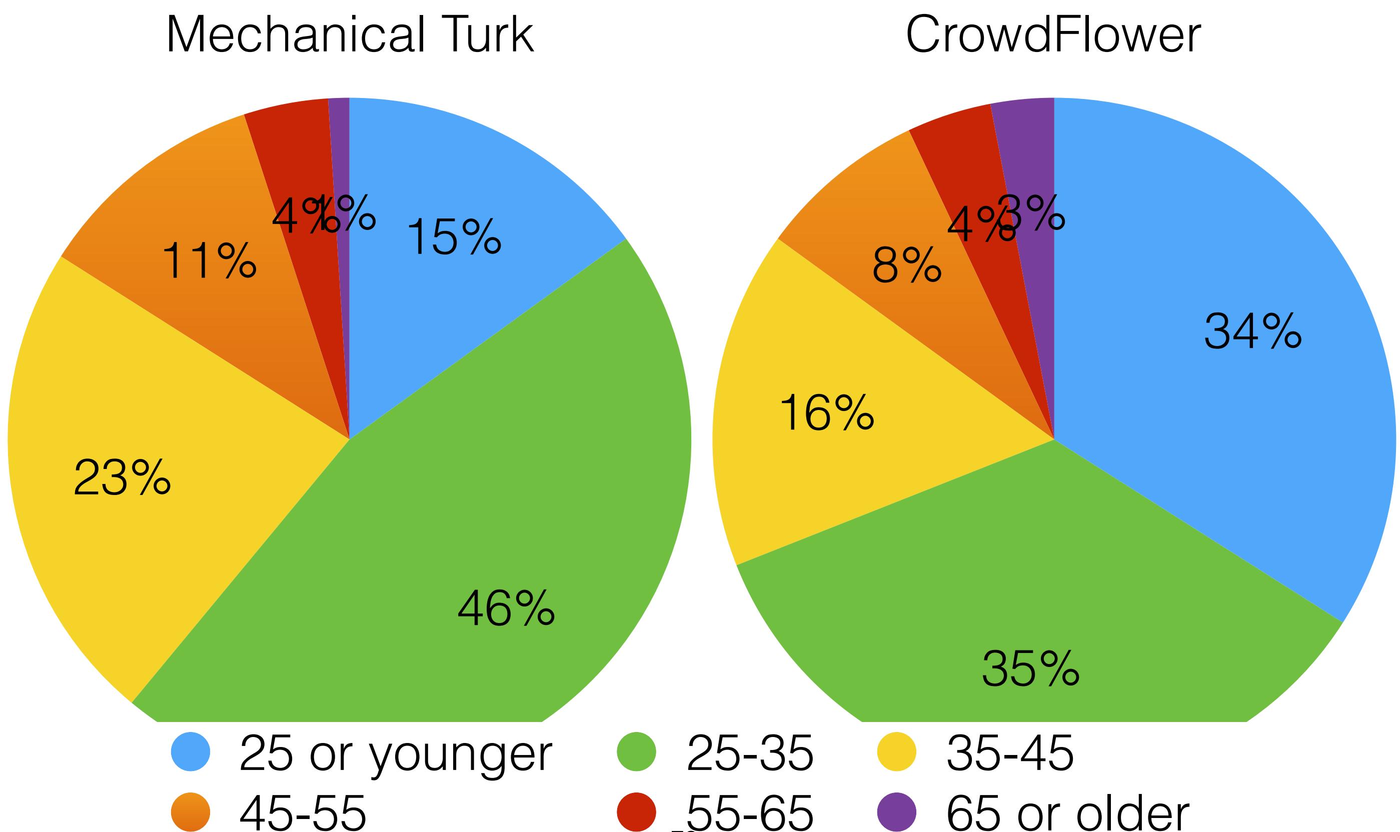


Slide from
Ellie Pavlick

78

• CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

Who is doing all this work for us?

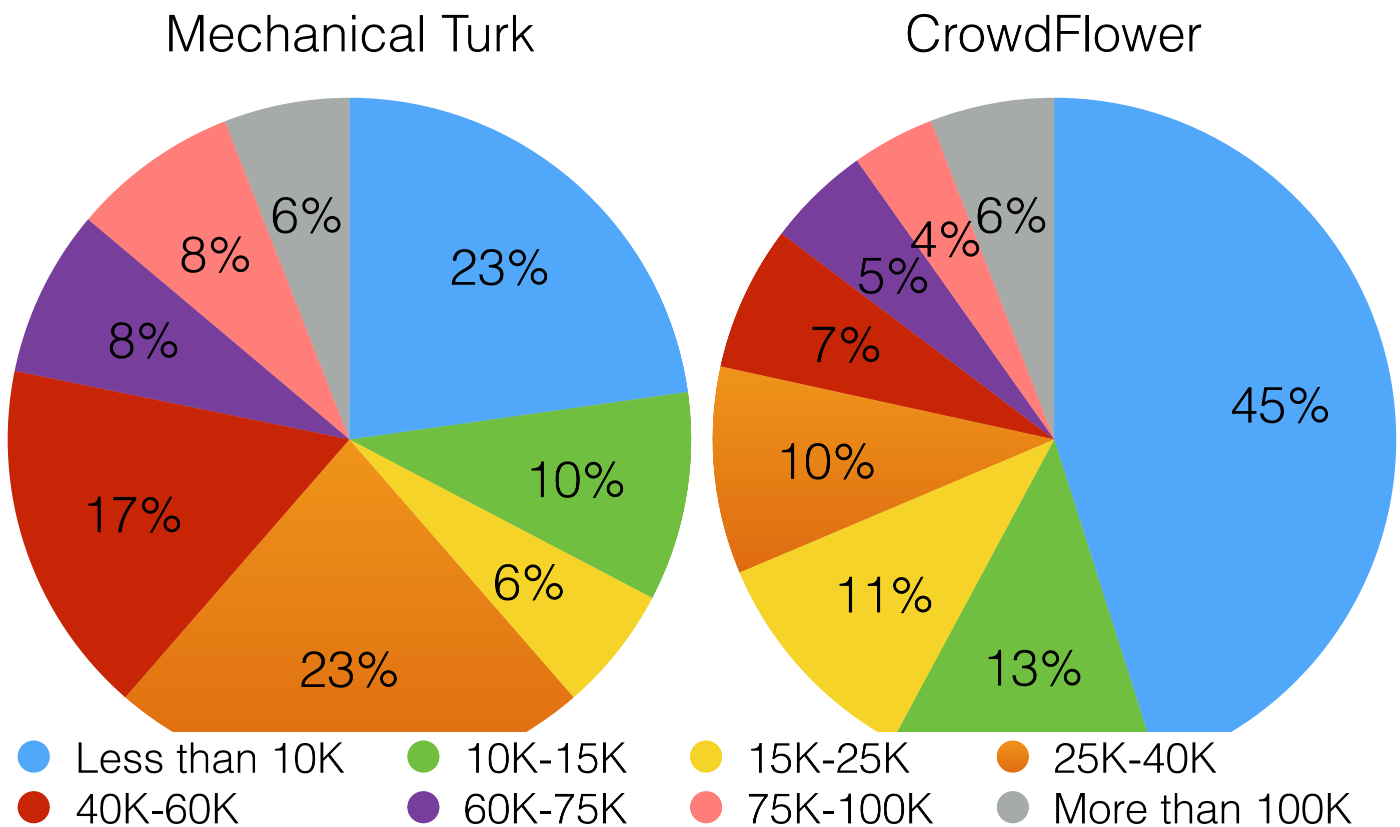


Slide from
Ellie Pavlick

79

- CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

Who is doing all this work for us?



Slide from
Ellie Pavlick

- 80 http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html
- http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

Why are they doing all this work for us?

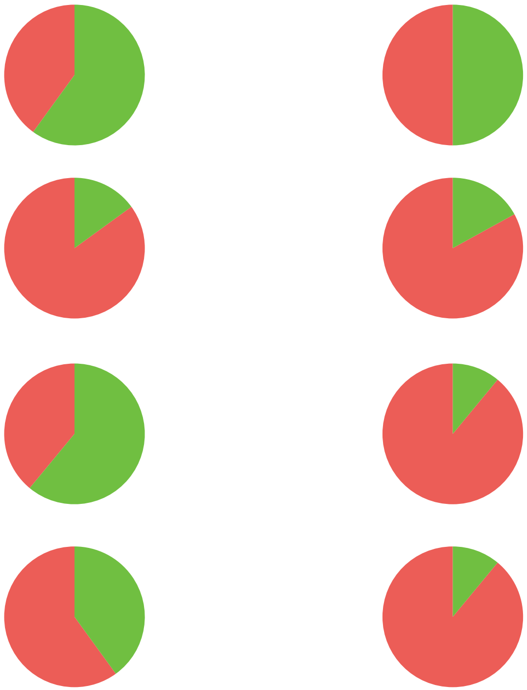
Mechanical Turk* CrowdFlower

Good way to spend free time and earn money (e.g. instead of TV)

As a primary source of income

As a secondary source of income/ pocket change

It is just so much fun!!



Slide from Ellie Pavlick

*Data is from 2010 and reflects only US workers

- 82 http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html
- http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

Common Misconceptions:

Slide from
Ellie Pavlick

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled

Slide from
Ellie Pavlick

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed

Slide from
Ellie Pavlick

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed
- Isolated, anti-social

Slide from
Ellie Pavlick

Common Misconceptions:

- Only from developing countries, non-native English speakers, uneducated, unskilled
- Work for $1/hour, doing it for fun in our PJs, unemployed
- Isolated, anti-social
- Cheaters, lazy, satisficers, inattentive

Slide from
Ellie Pavlick

How Turkers Work

- 10-20% of workers do 80% of the work
- Want large batches with high throughput
- Often dislike one-off HITs, e.g. surveys

Slide from
Ellie Pavlick

- Musthag, M., & Ganesan, D. (2013). Labor dynamics in a mobile micro-task market. Proceedings of the SIGCHI Conference on …, 641. http://doi.org/10.1145/2470654.2470745
- Chandler, J., Mueller, P. A., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behavior Research Methods, 46, 112–130. http://doi.org/10.3758/s13428-013-0365-7

89

How Turkers Work

- Online communities: Turkopticon, TurkerNation, Reddit, Facebook
- Scripts: IndiaTurkers, GreasyFork, HitDB, TurkMaster, HIT Scraper
- Websites and plugins: Turk Alert, mTurk List, CrowdWorkers

Slide from
Ellie Pavlick

More info about Turkers

A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk

Kotaro Hara[1,2], Abigail Adams[3], Kristy Milland[4,5], Saiph Savage[6], Chris Callison-Burch[7], Jeffrey P. Bigham[1]
[1]Carnegie Mellon University, [2]Singapore Management University, [3]University of Oxford
[4]McMaster University, [5]TurkerNation.com, [6]West Virginia University, [7]University of Pennsylvania
kotarohara@smu.edu.sg

CHI 2018

ABSTRACT
A growing number of people are working as part of on-line crowd work. Crowd work is often thought to be low wage work. However, we know little about the wage distribution in practice and what causes low/high earnings in this setting. We recorded 2,676 workers performing 3.8 million tasks on Amazon Mechanical Turk. Our task-level analysis revealed that workers earned a median hourly wage of only ~$2/h, and only 4% earned more than $7.25/h. While the average requester pays more than $11/h, lower-paying requesters post much more work. Our wage calculations are influenced by how unpaid work is accounted for, *e.g.*, time spent searching for tasks, working on tasks that are rejected, and working on tasks that are ultimately not submitted. We further explore the characteristics of tasks and working patterns that yield higher hourly wages. Our analysis informs platform design and worker tools to create a more positive future for crowd work.

temporarily out-of-work engineers to work [1,4,39,46,65].

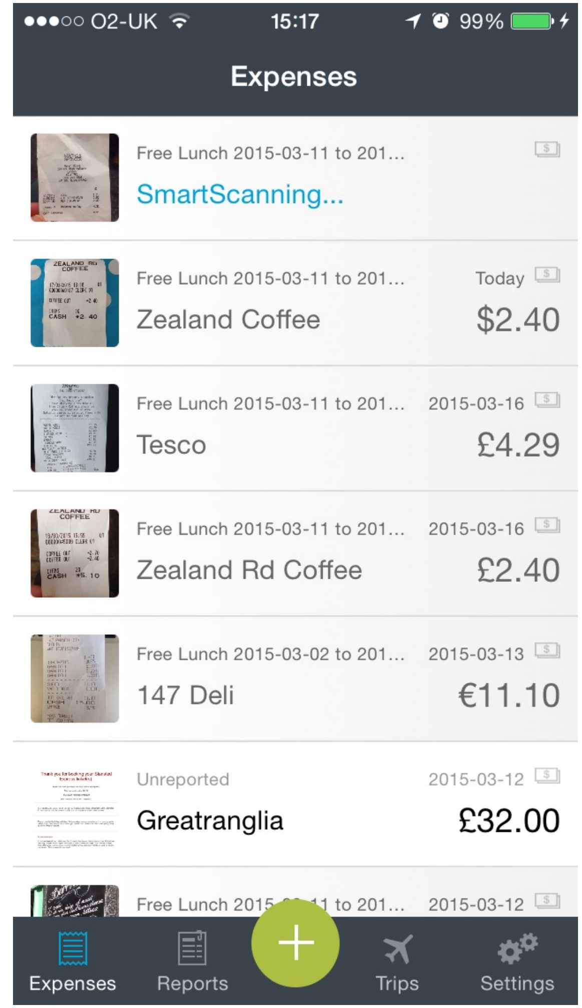
Yet, despite the potential for crowdsourcing platforms to extend the scope of the labor market, many are concerned that workers on crowdsourcing markets are treated unfairly [19,38,39,42,47,59]. Concerns about low earnings on crowd work platforms have been voiced repeatedly. Past research has found evidence that workers typically earn a fraction of the U.S. minimum wage [34,35,37–39,49] and many workers report not being paid for adequately completed tasks [38,51]. This is problematic as income generation is the primary motivation of workers [4,13,46,49].

Detailed research into crowd work earnings has been limited by an absence of adequate quantitative data. Prior research based on self-reported income data (*e.g.*, [4,34,49]) might be subject to systemic biases [22] and is often not sufficiently granular to facilitate a detailed investigation of earnings dispersion. Existing data-driven quantitative work in crowdsourcing research has taken the employers'

Link to paper

Issues with Crowdsourcing

Expensify


●●●○○ O2-UK
15:17
99%
Expenses
Free Lunch 2015-03-11 to 201...
SmartScanning...
Free Lunch 2015-03-11 to 201...
Today
Zealand Coffee
$2.40
Free Lunch 2015-03-11 to 201...
2015-03-16
Tesco
£4.29
Free Lunch 2015-03-11 to 201...
2015-03-16
Zealand Rd Coffee
£2.40
Free Lunch 2015-03-02 to 201...
2015-03-13
147 Deli
€11.10
Unreported
2015-03-12
Greatranglia
£32.00
2015-03-12
Expenses
Reports
Trips
Settings

Leaking data

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE

BIZ & IT —

Expensify sent images with personal data to Mechanical Turkers, calls it a feature

Expensify announces "private" transcription on Mechanical Turk as "Turkers" report seeing sensitive data.

SEAN GALLAGHER - 11/27/2017, 12:54 PM

Rochelle

@Rochelle

I wonder if Expensify SmartScan users know MTurk workers enter their receipts. I'm looking at someone's Uber receipt with their full name, pick up, and drop off addresses.

1,035 10:00 PM - Nov 22, 2017

729 people are talking about this

Fake AI

The rise of 'pseudo-AI': how tech firms quietly use humans to do bots' work

Using what one expert calls a 'Wizard of Oz technique', some companies keep their reliance on humans a secret from investors

Gregory Koberger
@gkoberger

How to start an AI startup

1. Hire a bunch of minimum wage humans to pretend to be AI pretending to be human

2. Wait for AI to be invented

585 3:08 PM - Mar 1, 2016 · California, USA

289 people are talking about this

Beyond Labeling: Text Generation

Crowdsourcing for 2 NLP tasks

- Story Cloze

- Natural Language Inference

Story Cloze Test

Goal: Design an evaluation schema for story understanding and narrative structure learning

Proposed Task: Given a context of four sentences, predict the endings of the story

Slide from Nasrin Mostafazadeh

An Examples Story Cloze Test

- **Context**: Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down one knee.

What do you think happens next?

What do you think likely doesn't happen next?

Slide from Nasrin Mostafazadeh

An Examples Story Cloze Test

- **Context**: Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down one knee.

- **Right Endings by Two Turkers**:
  - He proposed to Sheryl and she said Yes!
  - Tom asked Sheryl to marry him.
- **Wrong Endings by Two Turkers**:
  - He wiped mud off of his boot.
  - Tom tied his shoe and left Sheryl.

* We have collected 3,744 **doubly human-verified** Story Cloze Test instances.

16

Slide from Nasrin Mostafazadeh

Creating Story Cloze Dataset

- Ask Turkers to write 5 sentence story

- Ask Turkers to write incorrect ending

Story Cloze Test

Given a story (context of four sentences) and 4 possible endings, choose the most likely ending of the story?

How would you model this problem?

Approach 1: Language Modeling

$$e^* = \operatorname*{argmax}_{e \in \{e_1, e_2\}} p_{lm}(e|\text{prefix})$$

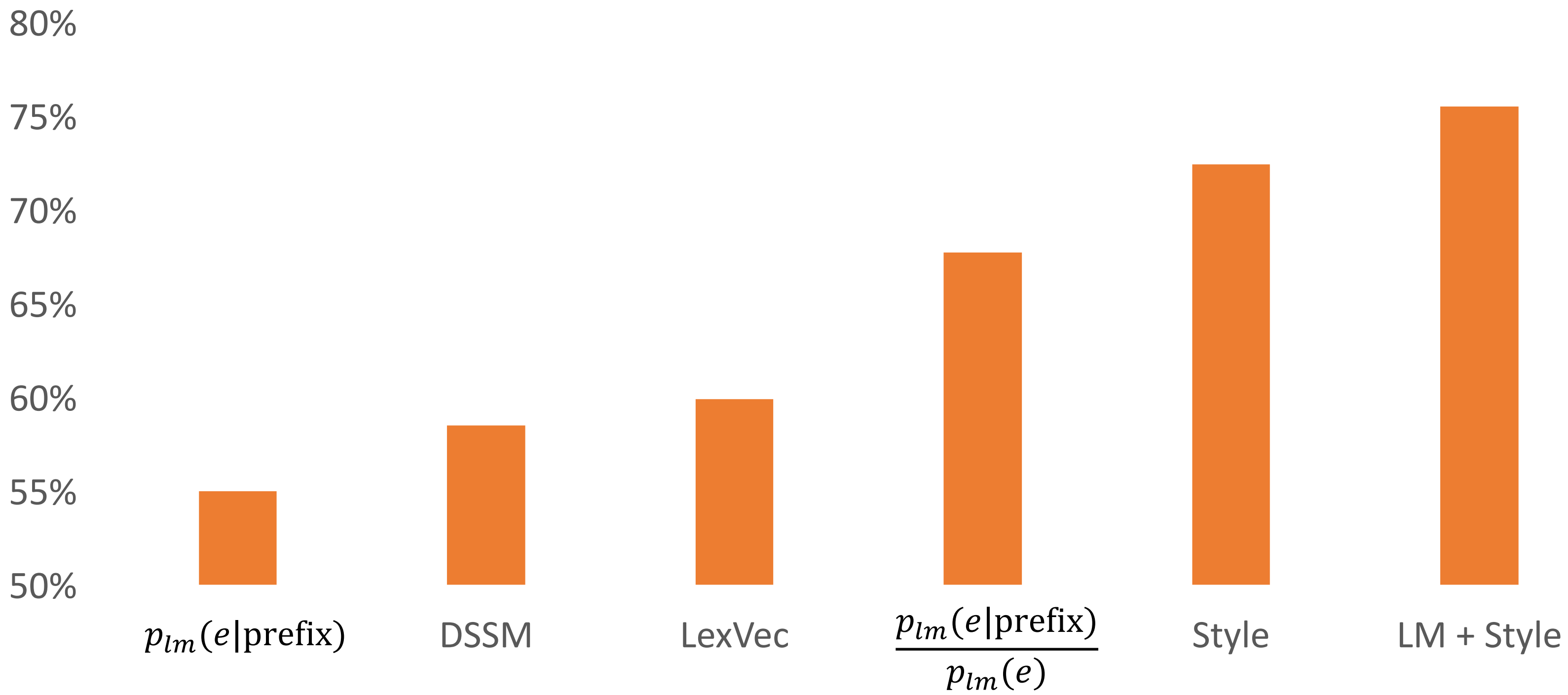
Approach 1.1: Language Modeling+

$$e^* = \underset{e \in \{e_1, e_2\}}{\operatorname{argmax}} \frac{p_{lm}(e|\text{prefix})}{\boldsymbol{p_{lm}(e)}}$$

Approach 2.0: Style

- Intuition: authors use different **style** when asked to write *right* vs. *wrong* story ending
- We train a style-based classifier to make this distinction
- Features are computed using **story endings only**
  - Without considering the story prefix

Results


80%
75%
70%
65%
60%
55%
50%
$p_{lm}(e|\text{prefix})$
DSSM
LexVec
$\frac{p_{lm}(e|\text{prefix})}{p_{lm}(e)}$
Style
LM + Style

Story Cloze Test

Goal: Design an evaluation schema for story understanding and narrative structure learning

Proposed Task: Given a context of four sentences, predict the endings of the story

Does this dataset test this goal?

Why yes? Why not?

Slide from Nasrin Mostafazadeh

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Multiple labeling schemas

entailment neutral contradiction

entailed not-entailed

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Stanford Natural Language Inference (SNLI)

Stanford Natural Language Inference (SNLI)

- Turker is:

1. shown context (premise)

Stanford Natural Language Inference (SNLI)

- Turker is:

1. shown context (premise)
2. generates hypothesis for each label:

- *entailed, neutral, contradiction*

SNLI - Example

Premise: ***A woman is reading with a child***

entailment neutral contradiction

SNLI - Example

Premise: ***A woman is reading with a child***

~~entailment~~ ~~neutral~~ contradiction

SNLI - Example

Premise: ***A woman is reading with a child***

Hypothesis: ***A woman is sleeping***

~~entailm~~ent ~~neutral~~ contradiction

SNLI - Example

Premise: ***A woman is reading with a child***

entailment ~~neutral~~ con~~tradiction~~

SNLI - Example

Premise: ***A woman is reading with a child***

Hypothesis: ***A woman has a book***

entailment ~~neutral~~ con~~tradiction~~

Hypothesis Only NLI

Hypothesis Only NLI

Hypothesis: ***A woman is sleeping***

Hypothesis Only NLI

Premise: $\varnothing$

Hypothesis: ***A woman is sleeping***

Hypothesis Only NLI

Premise: $\emptyset$

Hypothesis: ***A woman is sleeping***

entailment neutral contradiction

Hypothesis Only NLI

Premise: ∅

Hypothesis: ***A woman is sleeping***

entailment neutral contradiction

Add UW paper too

Hypothesis Only Baselines in Natural Language Inference

Best Paper Award

Adam Poliak[1] **Jason Naradowsky**[1] **Aparajita Haldar**[1,2]
Rachel Rudinger[1] **Benjamin Van Durme**[1]
[1]Johns Hopkins University [2]BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

*SEM 2018

Abstract

We propose a *hypothesis only* baseline for diagnosing Natural Language Inference (NLI). Especially when an NLI dataset assumes inference is occurring based purely on the relationship between a context and a hypothesis, it follows that assessing entailment relations while ignoring the provided context is a degenerate solution. Yet, through experiments on ten distinct NLI datasets, we find that this approach, which we refer to as a hypothesis-only model, is able to significantly outperform a majority-class baseline across a number of NLI datasets. Our analysis suggests that statistical irregularities may allow a model to perform NLI in some datasets beyond what should be achievable without access to the context.

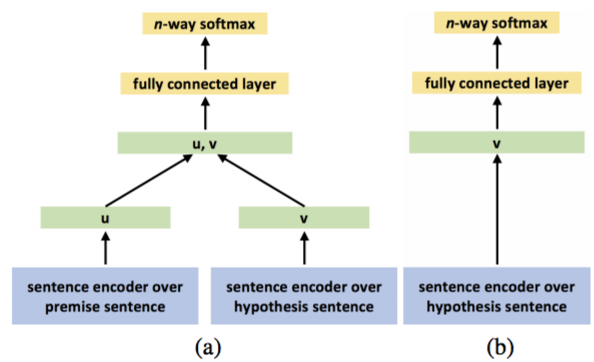
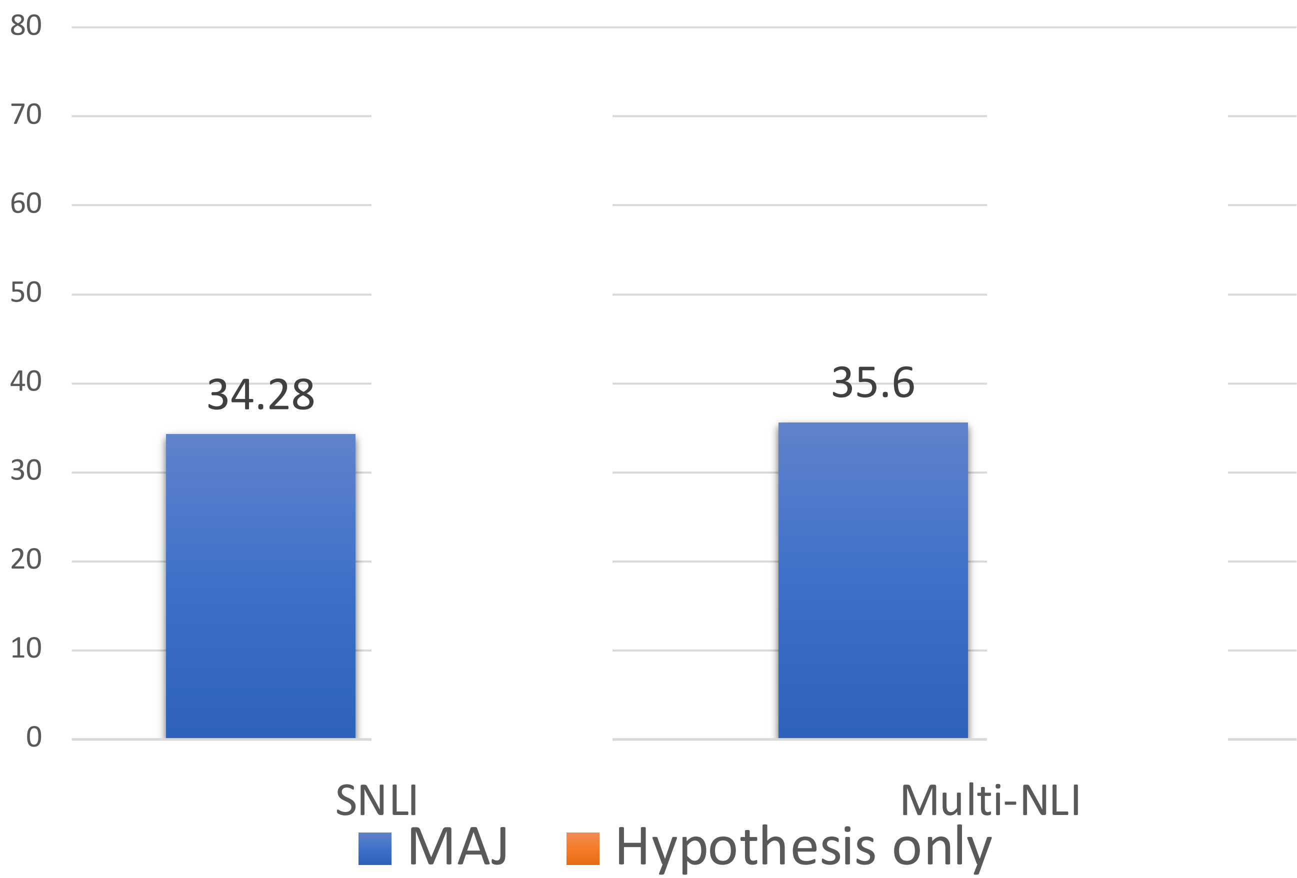
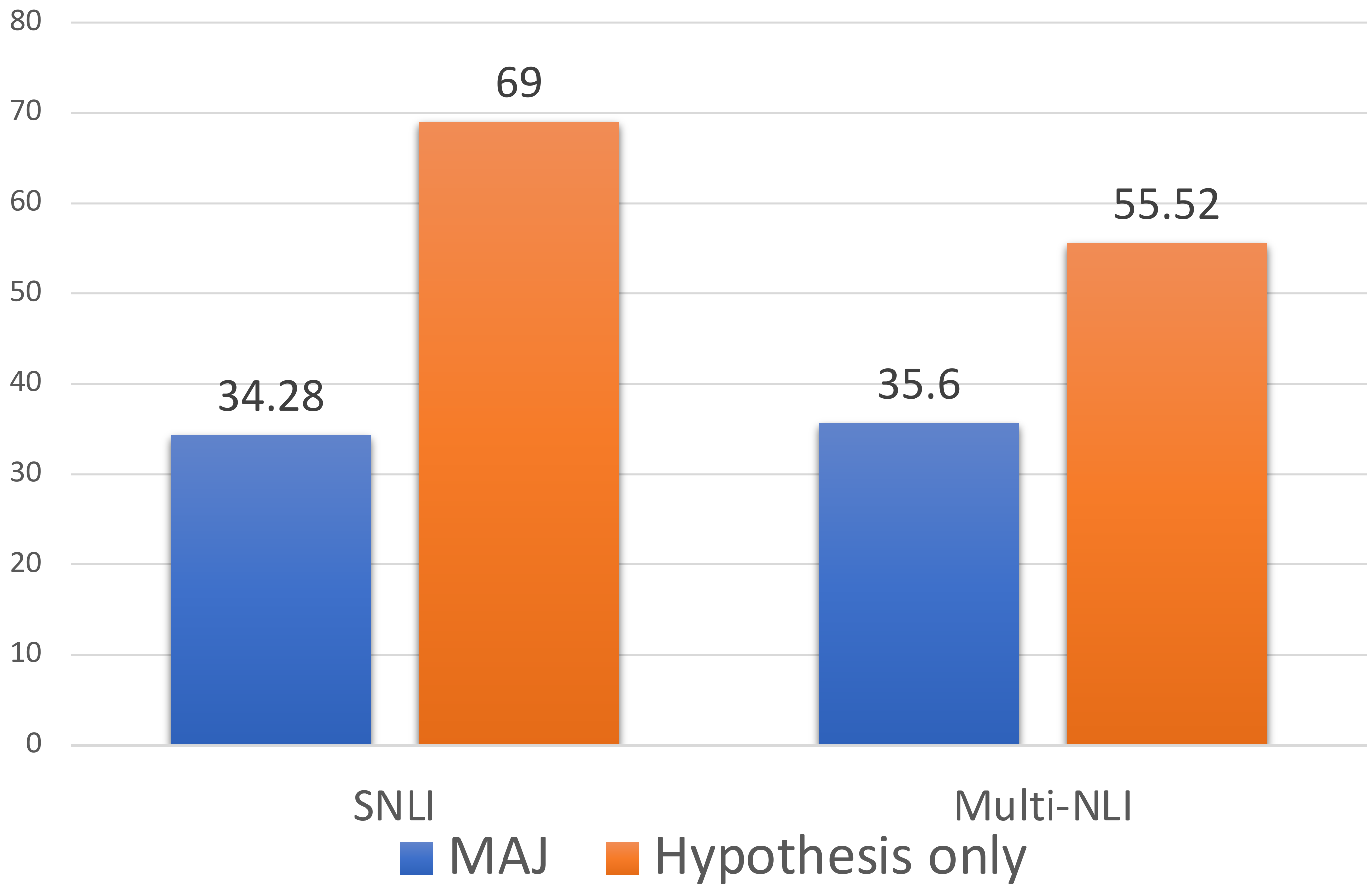


Figure 1: (1a) shows a typical NLI model that encodes the premise and hypothesis sentences into a vector space to classify the sentence pair. (1b) shows our hypothesis-only baseline method that ignores the premise and only encodes the hypothesis sentence.

Human Elicited Results



Human Elicited Results



Origin of SNLI

Origin of SNLI

flickr

- (Young et. al. 2014)

Origin of SNLI

- (Young et. al. 2014)

Origin of SNLI

- (Young et. al. 2014)

Origin of SNLI

flickr

- (Young et. al. 2014)

A woman is sleeping

Premises:

Hypothesis: A woman is sleeping

Premises:

A woman sings a song while playing piano

Hypothesis: A woman is sleeping

Premises:

This woman is laughing at her baby shower

Hypothesis: A woman is sleeping

Premises:

A woman with glasses is playing jenga

Hypothesis: A woman is sleeping

Why is she sleeping?

Elicitation Bias

- Descriptions of "dog":

-- McRae et al. (2005)

Elicitation Bias

- Descriptions of "dog":
- - barks

-- McRae et al. (2005)

Elicitation Bias

- Descriptions of “dog”:
- - barks
- - has a tail

-- McRae et al. (2005)

Elicitation Bias

- Descriptions of “dog”:
- - barks
- - has a tail
- - larger than a tulip
-

-- McRae et al. (2005)

Elicitation Bias

- Descriptions of "dog":
- - barks
- - has a tail
- - larger than a tulip
- - moves faster than an infant

-- McRae et al. (2005)

Elicitation Bias

- *"Features such as* ***is larger than a tulip*** *or* ***moves faster than an infant,*** *although logically possible, do not occur in human responses … people are capable of* ***verifying*** *that a* ***dog is larger than a pencil****."*

-- McRae et al. (2005)

Studies in eliciting norming data are prone to **repeated responses across subjects**

Other issues with SNLI generation

Bias and Stereotyping in Elicited Natural Language Inferences

Highly Associative Terms in SNLI

Boys: *fighting, muddy, guns, misbehaving*

Girls: *ballerinas, slumber, jumprope, laughing, bikinis, selfies*

Women: *gossiping, dresses, affection, chat, smile, hairdresser, receptionist, grieving*

Men: *cowboys, dock, dudes, workers, computers, boxers, lumberjacks, supervisors, thieves, roofers, engineers, surfboards, supervisors*

Explicitly Prejudicial Instances

Premise: An African American man looking at some butchered meat that is hanging from a rack outside a building.
Hypothesis (Contradiction): **A black man is in jail.**

Premise: Adult with red boots and purse walking down the street next to a brink wall.
Hypothesis (Neutral): **A whore looking for clients.**

Premise: Several Muslim worshipers march towards Mecca.
Hypothesis (Neutral): **The Muslims are terrorists.**

Rudinger, May, and Van Durme, 2017

Summary

- Crowdsourcing helps get labeled data for AI/ML
- Examples of Crowdsourcing, Companies, and Workers
- Issues of Crowdsourcing
  - Ethical, data integrity
  - Textual elicitation

Courses on Crowdsourcing

- Cornell: *Crowdsourcing and Human Computation*
- Upenn: *Crowdsourcing and Human Computation*
- Virginia Tech: *Crowdsourcing and Human Computation*
- CMU: *Crowd Programming*
- Umichigan: *Human Computation and Crowdsourcing Systems*