Cheap and Fast – But is it Good? Evaluating Non-Expert Annotation for Natural Language Tasks

Rion Snow Brendan O'Connor Daniel Jurafsky Andrew Y.Ng

Yifang Fu
2016/10/05

Introduction

- Human linguistic annotation is crucial for many natural language processing tasks but can be expensive and time-consuming.
- Explore the use of Amason's Mechanical Turk system to determine whether non-expert labelers can provide reliable natural language annotations.
  - Affect Recognition
  - Word Similarity
  - Recognizing Textual Entailment
  - Event Temporal Ordering
  - Word Sense Disambiguation

Contributions

- Show high agreement between Machanical Turk non-expert annotations and existing gold standard labels
- For task of affect recognition, show that using non-expert labels for training machine learning algorithms can be effective as using gold standard annotations from experts
- Propose a technique for bias correction that significantly improves annotation quality on two tasks.

Task Design

- Platform: AMT

- Some Tricks:
- Keep task descriptions as succinct as possible
  - —— Easy to understand the Task
- Task require only a multiple-choice response or numeric input within a fixed range
  - —— Easy to accomplish the Task
- For every task, collect ten independent annotation for each unique item
  - —— study how data quality improves with the number of independent annotations

Affective Text Analysis

Outcry at N Korea 'nuclear test'

[0,100]

Emotions: Anger, Disgust, Fear, Joy, Sadness, Surprise

[-100,100]

Overall positive or negative valence

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

Experiment Data:

- 100-headline sample
- Collect 10 affect annotations for each of the seven label types
- Total 7000 affect labels

Evaluation

- How well the non-experts agreed with the experts?
- Compare interannotator agreement(ITA)
- ITA is measured by calculating the Pearson Correlation of one labels with the average of other five labels

Emotion	E vs. E	E vs. All	NE vs. E	NE vs. All
Anger	0.459	0.503	0.444	0.573
Disgust	0.583	0.594	0.537	0.647
Fear	0.711	0.683	0.418	0.498
Joy	0.596	0.585	0.340	0.421
Sadness	0.645	0.650	0.563	0.651
Surprise	0.464	0.463	0.201	0.225
Valence	0.759	0.767	0.530	0.554
Avg. Emo	0.576	0.603	0.417	0.503
Avg. All	0.580	0.607	0.433	0.510

Table 1: Average expert and non-expert ITA on test-set

E vs.E
ITA(Expert, Expert)

E vs.All
ITA(Expert, Non-expert + Expert)

NE vs.E
ITA(Non-expert, Expert)

NE vs.All
ITA(Non-expert, Non-expert + Expert)

Results

- Experts are better labelers: experts agree with experts more than non-experts agree with experts.
- Adding non-experts to the gold standard (E vs.All) improves agreement,.

Emotion	E vs. E	E vs. All	NE vs. E	NE vs. All
Anger	0.459	0.503	0.444	0.573
Disgust	0.583	0.594	0.537	0.647
Fear	0.711	0.683	0.418	0.498
Joy	0.596	0.585	0.340	0.421
Sadness	0.645	0.650	0.563	0.651
Surprise	0.464	0.463	0.201	0.225
Valence	0.759	0.767	0.530	0.554
Avg. Emo	0.576	0.603	0.417	0.503
Avg. All	0.580	0.607	0.433	0.510

Table 1: Average expert and non-expert ITA on test-set

Evaluation

- How many averaged non-experts it would take to rival the performance of a single expert ?
- ‘ meta-labeler ’ : average the labels of each possible subset of n non-expert annotations, for value of n in {1, 2, …, 10}.
- Compute the ITA with ‘ meta-labeler ’ and expert annotators.

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.459	0.675	2	0.536
Disgust	0.583	0.746	2	0.627
Fear	0.711	0.689	–	–
Joy	0.596	0.632	7	0.600
Sadness	0.645	0.776	2	0.656
Surprise	0.464	0.496	9	0.481
Valence	0.759	0.844	5	0.803
Avg. Emo.	0.576	0.669	4	0.589
Avg. All	0.603	0.694	4	0.613

Table 2: Average expert and averaged correlation over 10 non-experts on test-set. k is the minimum number of non-experts needed to beat an average expert.

K is the minimum number of non-expert annotations from with we can create a meta-labeler that has equal or better ITA than an expert annotator

Results

- For all tasks except "Fear", we are able

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.459	0.675	2	0.536
Disgust	0.583	0.746	2	0.627

we paid US$2.00 in order to collect the 7000 non-expert annotations

3500 non-expert labels per USD

as at least 875 expert-equivalent labels per USD.

It is so cheap!!!

- expert annotations achieve the equivalent ITA across all 7 tasks.

Table 2: Average expert and averaged correlation over 10 non-experts on test-set. k is the minimum number of non-experts needed to beat an average expert.

Word Similarity

Replicate the word similarity task used in

(Miller and Charles 1991))

{ boy, lad }

[0,10], fraction

Highly similar { boy, lad } —— unrelated { noon, string }

Experiment Data:

- 30 word pairs
- Collect 10 annotations for each of the 30 word pairs
- Total 300 annotations

Evaluation

- Average the numeric responses from each possible subset of n annotators and computing the ITA correlation with respect to the gold scores reported in (Miller and Charles, 1991)
- The horizontal line is (Resnik, 1999)'s 0.958 correlation

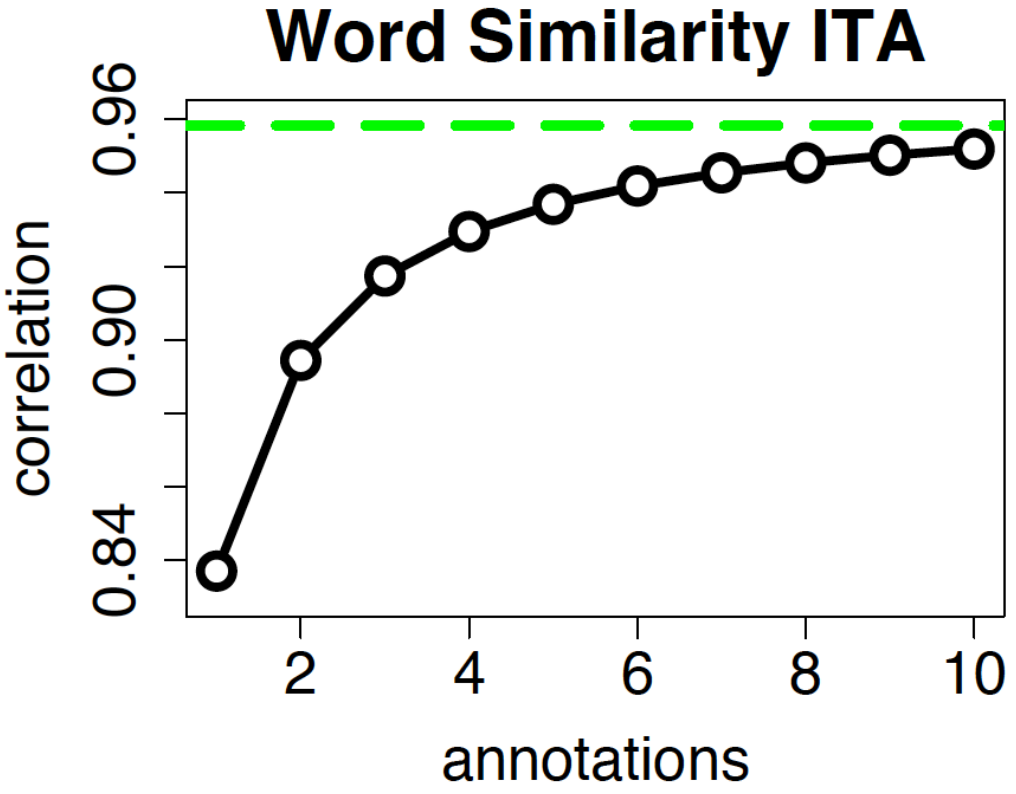


Figure 2: ITA for word similarity experiment

Results

- At 10 annotators, we achieve a correlation of 0.952, well within the range of other studies of expert and non-expert annotations.
- The Task of 300 annotations was completed by 10 annotators in less than 11minutes,at the rate of 1724 annotations / hour.

It is so fast!!!

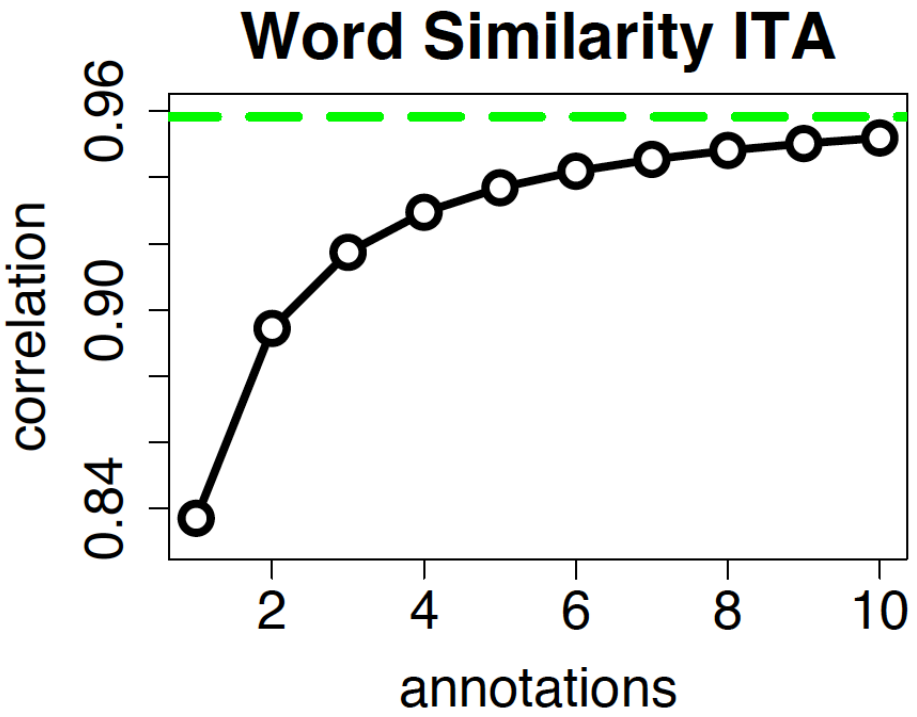
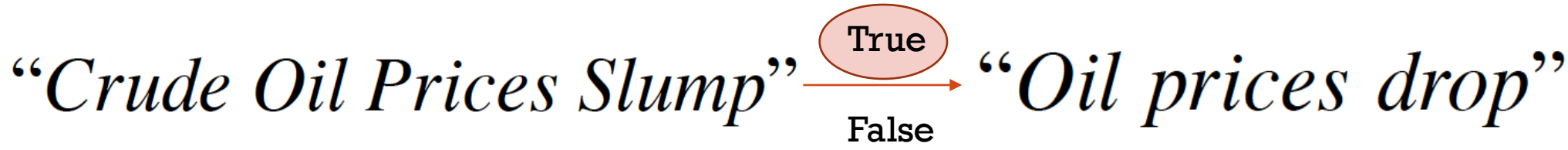


Figure 2: ITA for word similarity experiment

Recognizing Textual Entailment

- Replicates the Recognizing Textual Entailment task proposed in PASCAL Recognizing Textual Entailment task (Dagan et al. 2006)



Evaluation & Result

- Collect 10 annotations for each sentence pair.
- Use simple majority voting when considering multiple non-expert annotations.
- At 10 annotators, we achieve a maximum accuracy of 89.7%.

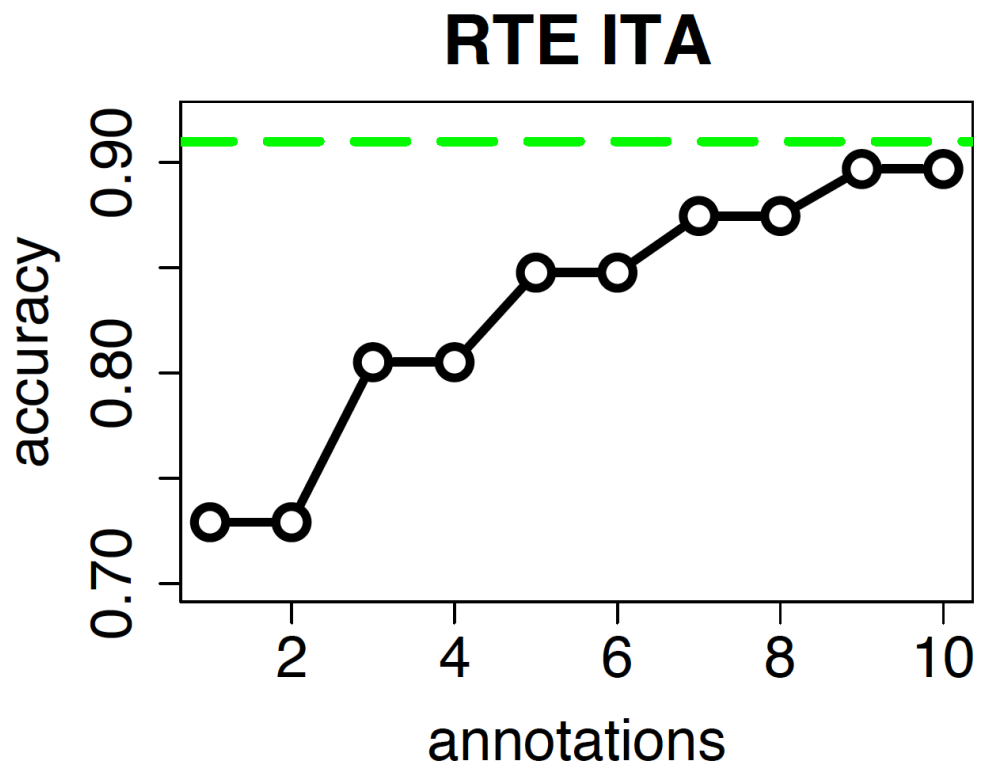


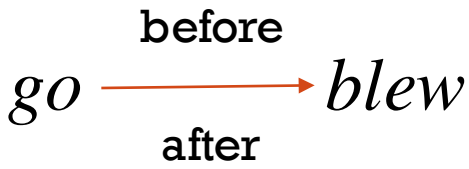
Figure 3: Inter-annotator agreement for RTE experiment

Event Annotation

Event-Pairs —— verb events only

{before, after} —— temporal relation

"It just blew up in the air, and then we saw two fireballs go down to the, to the water, and there was a big small, ah, smoke, from ah, coming up from that"



Result

- Achieve high agreement for this task, at a rate of 0.94 with simple voting over 10 annotators
- No expert ITA numbers have been reported for this simplified temporal ordering task.

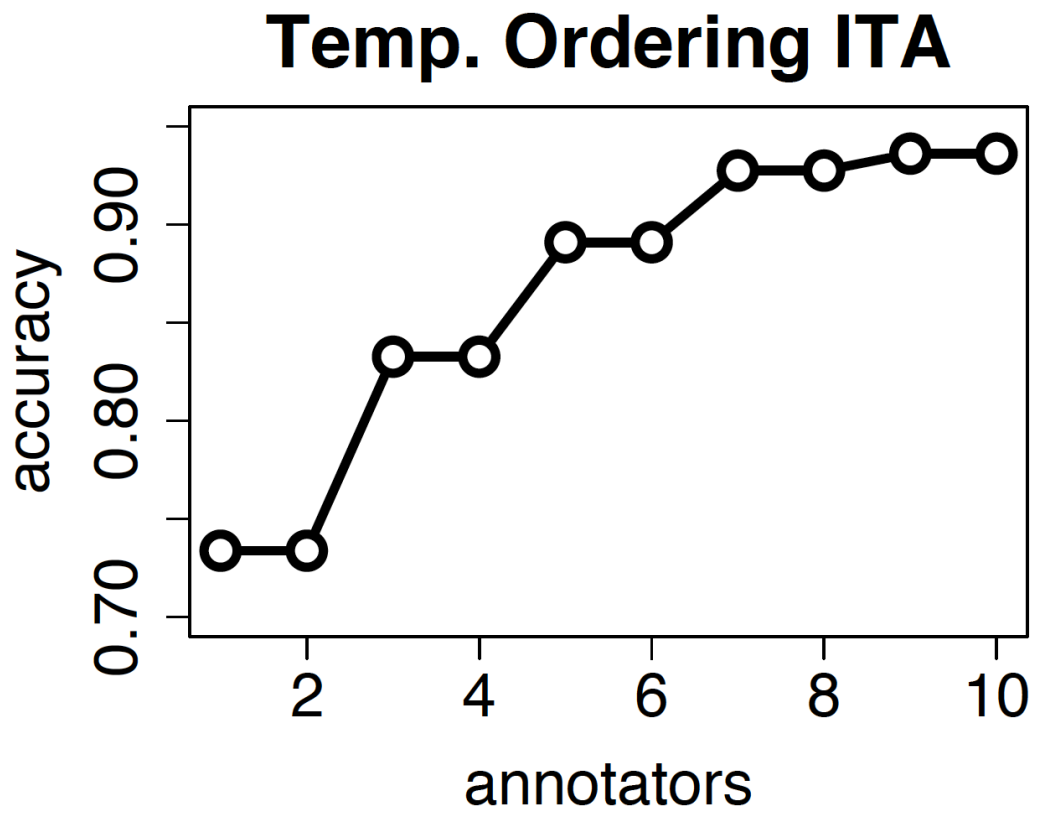


Figure 4: ITA for temporal ordering experiment

Word Sense Disambiguation

- A paragraph containing

"Robert E. Lyons III…was appointed president and chief operating officer…"

President:

1) executive officer of a firm, corporation, or university

2) head of a country (other than the U.S.)

3) head of the U.S., President of the United States

Experiment Data:

- ❖ 177 examples of the noun "president" for three senses
- ❖ Collect 10 annotations for each "president"

Result

- Achieve a very high rate of 0.994 accuracy.
- The best automatic system performance, with an accuracy of 0.98
- An error in the original gold standard
- After correcting this error, the non-expert accuracy rate is 100%.

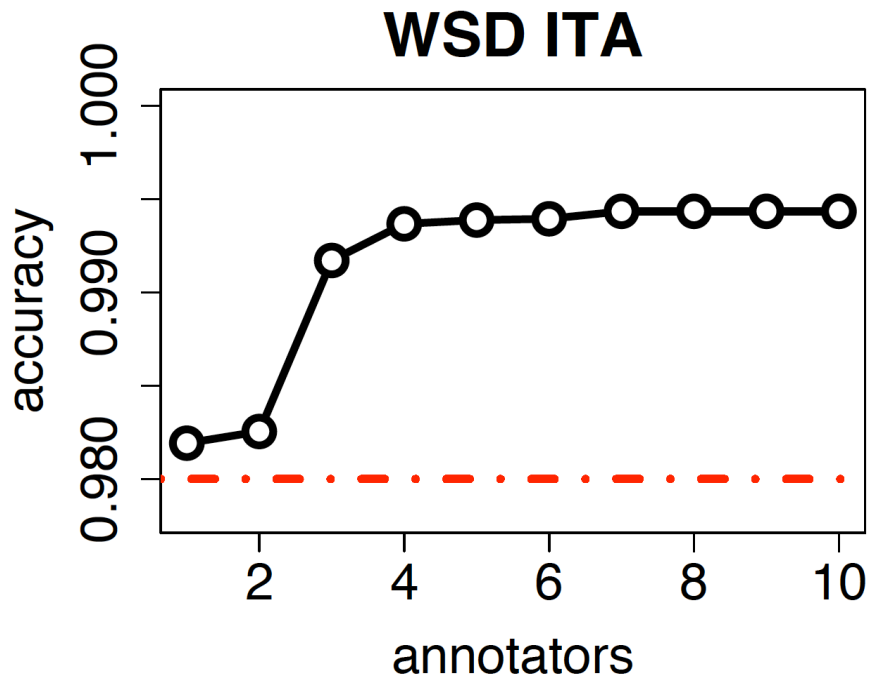


Figure 5: Inter-annotator agreement for WSD experiment

Summary

Task	Labels	Cost (USD)	Time (hrs)	Labels per USD	Labels per hr
Affect	7000	$2.00	5.93	3500	1180.4
WSim	300	$0.20	0.174	1500	1724.1
RTE	8000	$8.00	89.3	1000	89.59
Event	4620	$13.86	39.9	333.3	115.85
WSD	1770	$1.76	8.59	1005.7	206.1
Total	21690	25.82	143.9	840.0	150.7

Table 3: Summary of costs for non-expert labels

Bias Correction

- Problem: The reliability of individual workers varies. A small few give very noisy responses.

- Solution: Recalibrate worker's responses to more closely match expert behavior using a small amount of expert-labeled training data.

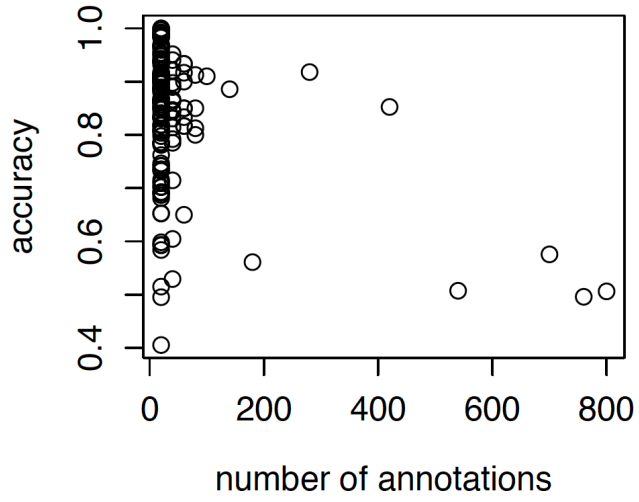


Figure 6: Worker accuracies on the RTE task. Each point is one worker. Vertical jitter has been added to points on the left to show the large number of workers who did the minimum amount of work (20 examples).

Bias Correction In Categorical Data

- Example i has true label x_i
- Different workers give labels $y_{i1}, y_{i2}, \ldots y_{iW}$
- To infer the posterior probability of the true label for a new example
- Bayes rules

$$\log \frac{P(x_i = Y | y_{i1} \ldots y_{iW})}{P(x_i = N | y_{i1} \ldots y_{iW})}$$

$$= \sum_w \log \frac{P(y_{iw} | x_i = Y)}{P(y_{iw} | x_i = N)} + \log \frac{P(x_i = Y)}{P(x_i = N)}$$

- Worker response likelihoods $P(y_w | x = Y)$ and $P(y_w | x = N)$ can be directly estimated from frequencies of worker performance on gold standard examples.
- Weighted Voting Rule: each worker's vote is weighted by their log likelihood ratio for their given response.

Example Task

- Recognizing Textual Entailment has an average +4.0% accuracy increase, averaged across 2 through 10 annotators.
- Event annotation gets +3.4%accuracy increase.

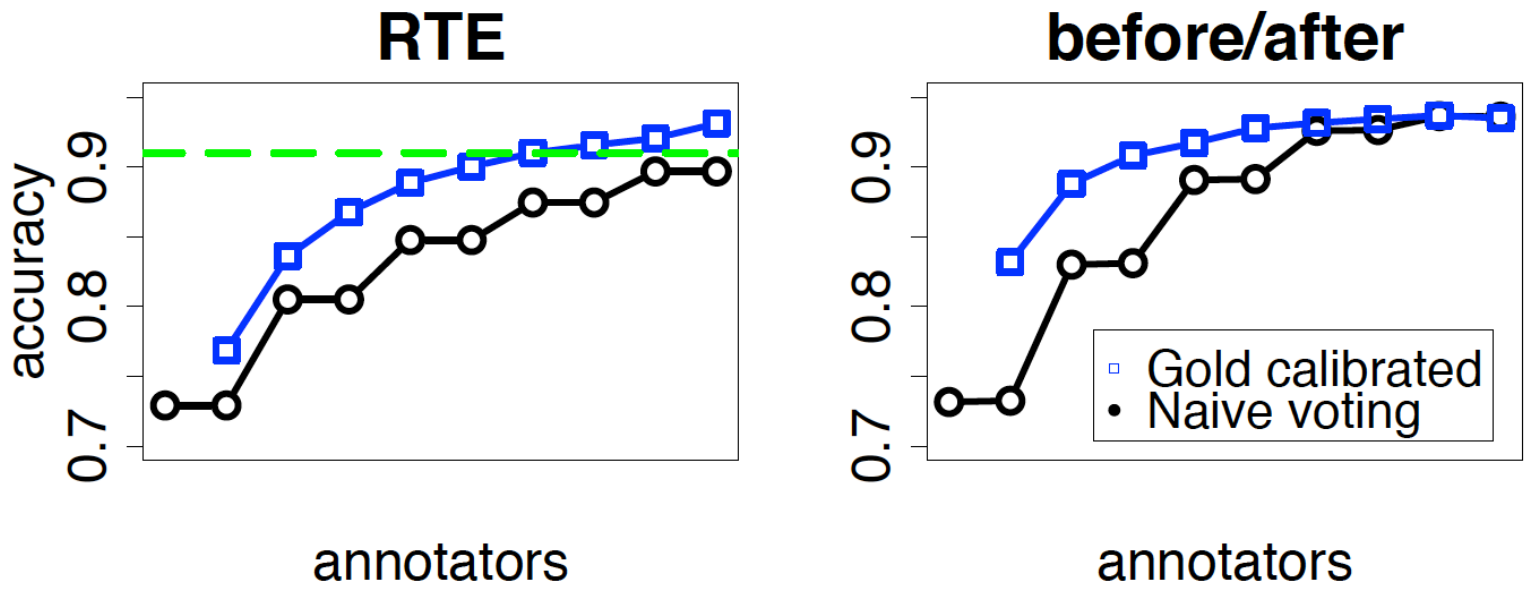


Figure 7: Gold-calibrated labels versus raw labels

Training System

- Affect Recognition
- 100 headline as a training set, 900 headlines as test set
- Each expert annotator we train a system and create a gold standard test set using the average the remaining five labelers
- For each possible subset of n non-expert labels annotators, for n = {1, 2, . . . , 10} we train a system, and evaluate by calculating Pearson correlation with the same set of gold standard datasets.

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.084	0.233	1	0.172
Disgust	0.130	0.231	1	0.185
Fear	0.159	0.247	1	0.176
Joy	0.130	0.125	–	–
Sadness	0.127	0.174	1	0.141
Surprise	0.060	0.101	1	0.061
Valence	0.159	0.229	2	0.146
Avg. Emo	0.116	0.185	1	0.135
Avg. All	0.122	0.191	1	0.137

Table 4: Performance of expert-trained and non-expert-trained classifiers on test-set. k is the minimum number of non-experts needed to beat an average expert.

Training System

- K is the minimum number of non-expert annotations required to achieve similar performance to the expert annotations.
- For five of the seven tasks, k value is one.
- With a single set of non-expert annotations outperforms the average system trained with the labels from a single expert.

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.084	0.233	1	0.172
Disgust	0.130	0.231	1	0.185
Fear	0.159	0.247	1	0.176
Joy	0.130	0.125	–	–
Sadness	0.127	0.174	1	0.141
Surprise	0.060	0.101	1	0.061
Valence	0.159	0.229	2	0.146
Avg. Emo	0.116	0.185	1	0.135
Avg. All	0.122	0.191	1	0.137

Table 4: Performance of expert-trained and non-expert-trained classifiers on test-set. k is the minimum number of non-experts needed to beat an average expert.

Conclusion

- Demonstrate the effectiveness of using Amazon Mechanical Turk for a variety of natural language annotation tasks.
- Evaluation of non-expert labeler data vs. expert annotations for five tasks found that for many tasks only a small number of non-expert annotations per item are necessary to equal the performance of an expert annotator.
- Demonstrate significant improvement by controlling for labeler bias

THANK YOU!

Q&A