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

Rion Snow
Brendan O’Connor
Daniel Jurafsky
Andrew Ng
2008
Mechanical Turk

- Amazon

- Workers and Requesters

- Human Intelligence Task (HIT)
Experiments

- Mechanical Turk Annotation Tasks
  - Replications of five NLP tasks using MT annotators

- Bias correction annotation tasks

- Supervised machine learning on non-expert annotations for affect recognition task
Interannotator Agreement (ITA)

- ITA = Pearson’s correlation coefficient between two measures

- E vs. E
  - Compute n ITAs between one expert and the average of the remaining and take average to get average expert ITA

\[
\begin{align*}
\text{ITA}_{e,1} &= \text{pearsons}(e_1, \text{avg}(e_2, e_3, ..., e_n)) \\
\text{ITA}_{e,2} &= \text{pearsons}(e_2, \text{avg}(e_1, e_3, ..., e_n)) \\
\vdots \\
\text{ITA}_{e,n} &= \text{pearsons}(e_n, \text{avg}(e_1, e_2, ..., e_{n-1})) \\
\text{ITA}_e &= \text{ITA}_{e,1} + \text{ITA}_{e,2} + ... + \text{ITA}_{e,n} / n
\end{align*}
\]
Interannotator Agreement

- NE vs. E
  \[ \text{ITA}_{ne} = \text{ITA}_{ne,1} + \text{ITA}_{ne,2} + \ldots + \text{ITA}_{ne,n} / n \]

- E vs. All
  \[ \text{ITA}_{e,all} = \text{avg}( \text{ITA}(e_1, \text{ITA}_e), \text{ITA}(e_1, \text{ITA}_{ne}), \text{ITA}(e_2, \text{ITA}_e), \text{ITA}(e_2, \text{ITA}_{ne}), \ldots \text{ITA}(e_n, \text{ITA}_e), \text{ITA}(e_n, \text{ITA}_{ne})) \]

- NE vs. All
  \[ \text{ITA}_{ne,all} = \text{avg}( \text{ITA}(ne_1, \text{ITA}_e), \text{ITA}(ne_1, \text{ITA}_{ne}), \text{ITA}(ne_2, \text{ITA}_e), \text{ITA}(ne_2, \text{ITA}_{ne}), \ldots \text{ITA}(ne_n, \text{ITA}_e), \text{ITA}(ne_n, \text{ITA}_{ne})) \]
Annotation Task 1—Affective Text Analysis

- Strapparava and Mihalcea (2007)
- Headline: Outcry at N Korea ‘nuclear test’

- [1,100]
  - Anger, disgust, fear, joy, sadness, surprise

- [-100,100]
  - Overall positive or negative valence
Annotation Task 1—Affective Text Analysis

- Average expert and non expert ITA

<table>
<thead>
<tr>
<th>Emotion</th>
<th>E vs. E</th>
<th>E vs. All</th>
<th>NE vs. E</th>
<th>NE vs. All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.459</td>
<td>0.503</td>
<td>0.444</td>
<td>0.573</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.583</td>
<td>0.594</td>
<td>0.537</td>
<td>0.647</td>
</tr>
<tr>
<td>Fear</td>
<td>0.711</td>
<td>0.683</td>
<td>0.418</td>
<td>0.498</td>
</tr>
<tr>
<td>Joy</td>
<td>0.596</td>
<td>0.585</td>
<td>0.340</td>
<td>0.421</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.645</td>
<td>0.650</td>
<td>0.563</td>
<td>0.651</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.464</td>
<td>0.463</td>
<td>0.201</td>
<td>0.225</td>
</tr>
<tr>
<td>Valence</td>
<td>0.759</td>
<td>0.767</td>
<td>0.530</td>
<td>0.554</td>
</tr>
<tr>
<td>Avg. Emo</td>
<td>0.576</td>
<td>0.603</td>
<td>0.417</td>
<td>0.503</td>
</tr>
<tr>
<td>Avg. All</td>
<td>0.580</td>
<td>0.607</td>
<td>0.433</td>
<td>0.510</td>
</tr>
</tbody>
</table>
Annotation Task 1—Affective Text Analysis

- What is minimum number of non-expert annotators to beat expert?

<table>
<thead>
<tr>
<th>Emotion</th>
<th>1-Expert</th>
<th>10-NE</th>
<th>( k )</th>
<th>( k )-NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.459</td>
<td>0.675</td>
<td>2</td>
<td>0.536</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.583</td>
<td>0.746</td>
<td>2</td>
<td>0.627</td>
</tr>
<tr>
<td>Fear</td>
<td>0.711</td>
<td>0.689</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Joy</td>
<td>0.596</td>
<td>0.632</td>
<td>7</td>
<td>0.600</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.645</td>
<td>0.776</td>
<td>2</td>
<td>0.656</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.464</td>
<td>0.496</td>
<td>9</td>
<td>0.481</td>
</tr>
<tr>
<td>Valence</td>
<td>0.759</td>
<td>0.844</td>
<td>5</td>
<td>0.803</td>
</tr>
<tr>
<td>Avg. Emo.</td>
<td>0.576</td>
<td>0.669</td>
<td>4</td>
<td>0.589</td>
</tr>
<tr>
<td>Avg. All</td>
<td>0.603</td>
<td>0.694</td>
<td>4</td>
<td>0.613</td>
</tr>
</tbody>
</table>
Annotation Task 2—Word Similarity

- Miller and Charles (1991)
- [0, 10]
  - 30 pairs of words
  - {boy, lad}, {noon, string}
Annotation Task 2—Word Similarity
Annotation Task 3—Recognizing Textual Entailment

- Dagan et al. (2006)
- Shown two sentences
- Asked if the second hypothesis can be inferred from the first [T/F]

1. “Crude Oil Prices Slump”
2. “Oil prices drop” => True

1. “The government announced last week that it plans to raise oil prices”
2. “Oil prices drop” => False
Annotation Task 3—Recognizing Textual Entailment

![Graph showing RTÉ-ITA accuracy with increasing annotations]
Annotation Task 4—Event Annotation

- Pustejovsky et al. (2003) TimeBank corpus
- Strictly before or strictly after
- Verbs only

“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”
Annotation Task 4—Event Annotation

Temp. Ordering ITA

accuracy

annotators
Annotation Task 5—Word Sense Disambiguation

- Pradhan et al. (2007)
- SemEval Word Sense Disambiguation Lexical Sample task

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

1. executive officer of a firm, corporation, university
2. head of a country (other than the U.S.)
3. head of the U.S., President of the United States
Annotation Task 5—
Word Sense Disambiguation

![Graph showing accuracy over annotators](image)
## Cost

<table>
<thead>
<tr>
<th>Task</th>
<th>Labels</th>
<th>Cost (USD)</th>
<th>Time (hrs)</th>
<th>Labels per USD</th>
<th>Labels per hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect</td>
<td>7000</td>
<td>$2.00</td>
<td>5.93</td>
<td>3500</td>
<td>1180.4</td>
</tr>
<tr>
<td>WSim</td>
<td>300</td>
<td>$0.20</td>
<td>0.174</td>
<td>1500</td>
<td>1724.1</td>
</tr>
<tr>
<td>RTE</td>
<td>8000</td>
<td>$8.00</td>
<td>89.3</td>
<td>1000</td>
<td>89.59</td>
</tr>
<tr>
<td>Event</td>
<td>4620</td>
<td>$13.86</td>
<td>39.9</td>
<td>333.3</td>
<td>115.85</td>
</tr>
<tr>
<td>WSD</td>
<td>1770</td>
<td>$1.76</td>
<td>8.59</td>
<td>1005.7</td>
<td>206.1</td>
</tr>
<tr>
<td>Total</td>
<td>21690</td>
<td>25.82</td>
<td>143.9</td>
<td>840.0</td>
<td>150.7</td>
</tr>
</tbody>
</table>
Bias Correction for Non-Expert Annotators

- Problem: varying reliability for Turkers

- Solution: create a model to enhance quality
  - Dawid and Skene (1979)
Bias Correction for Non-Expert Annotators

- All workers’ probability of response for example $i$ - multinomial

$$P(y_{i1}, \ldots, y_{iW}, x_i) = \left( \prod_w P(y_{iw} | x_i) \right) p(x_i)$$
Bias Correction for Non-Expert Annotators

- Probability of true label for new example
  - Posterior log-odds

\[
\log \frac{P(x_i = Y | y_{i1} \ldots y_{iW})}{P(x_i = N | y_{i1} \ldots y_{iW})} = \sum_w \log \frac{P(y_{iw} | x_i = Y)}{P(y_{iw} | x_i = N)} + \log \frac{P(x_i = Y)}{P(x_i = N)}
\]
Bias Correction Example
Supervised Affect Recognition System

- Train a system on expert and another on non-expert annotations

- For each token \( t \) in any of the headlines

\[
Score(e, t) = \frac{\sum_{H \in H_t} Score(e, H)}{|H_t|}
\]

- For a new headline \( H \)

\[
Score(e, H) = \sum_{t \in H} \frac{Score(e, t)}{|H|}
\]
## Supervised Affect Recognition System

<table>
<thead>
<tr>
<th>Emotion</th>
<th>1-Expert</th>
<th>10-NE</th>
<th>$k$</th>
<th>$k$-NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.084</td>
<td>0.233</td>
<td>1</td>
<td>0.172</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.130</td>
<td>0.231</td>
<td>1</td>
<td>0.185</td>
</tr>
<tr>
<td>Fear</td>
<td>0.159</td>
<td>0.247</td>
<td>1</td>
<td>0.176</td>
</tr>
<tr>
<td>Joy</td>
<td>0.130</td>
<td>0.125</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.127</td>
<td>0.174</td>
<td>1</td>
<td>0.141</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.060</td>
<td>0.101</td>
<td>1</td>
<td>0.061</td>
</tr>
<tr>
<td>Valence</td>
<td>0.159</td>
<td>0.229</td>
<td>2</td>
<td>0.146</td>
</tr>
<tr>
<td>Avg. Emo</td>
<td>0.116</td>
<td>0.185</td>
<td>1</td>
<td>0.135</td>
</tr>
<tr>
<td>Avg. All</td>
<td>0.122</td>
<td>0.191</td>
<td>1</td>
<td>0.137</td>
</tr>
</tbody>
</table>
Thank you

• Questions?
Bias Correction Example

<table>
<thead>
<tr>
<th>i</th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

\[
\log \frac{P(x_3 = Y | Y, N, N)}{P(x_3 = N | Y, N, N)} = \log \frac{P(Y | x_3 = Y)}{P(Y | x_3 = N)} + \log \frac{P(N | x_3 = Y)}{P(N | x_3 = N)} + \log \frac{P(N | x_3 = Y)}{P(N | x_3 = N)} + \log \frac{P(x_3 = Y)}{P(x_3 = N)}
\]