Anveshan
A Framework for Multiple Annotators’ Labeling Behavior

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi
Columbia University in the City of New York

Nancy Ide
Vassar College – Poughkeepsie, NY
Outline

• Motivation
• What does Anveshan do?
• Trained Annotators and Mechanical Turkers
• Anveshan Methodology
• Experiments and Results
• Conclusion and Future Work
Motivation

• Annotation – essential for supervised machine learning problems on linguistic data
• Usually evaluated by a single score (2-3 annotators)
  – Percent agreement
  – IA : Kappa (Cohen), Alpha (Krippendorff)
• Subjectivity in annotation
  – Differences in interpretation inherent in ordinary language
  – Observable when we have multiple annotators, categorical labels.
  – Single score is too atomic
Sources of Variation

• Same set of annotators have largely varying agreement over different tasks
• Several factors contribute to the variation
• Variation can occur across
  – Annotators (In Expertise)
  – Instances (In Complexity)
  – Label Sets (Size and Similarity)
  – Tasks (Across Individual Types)
Anveshan – Looking Deeper

• Subjective annotation task: Word-Sense Disambiguation

• Agreement score varies from word to word
  – Due to annotator expertise
  – Number of word senses
  – Word/Instance complexity

• Issues
  – Presence of outliers
  – Confusible labels
  – Systematic differences in annotator behavior
Word-Sense Annotation Data

• The Manually Annotated Sub-Corpus (MASC) project (Ide et al., 2010)
  – American English texts drawn from OANC
  – As of May 2010 release date, consists of 82K words
    http://www.americannationalcorpus.org/MASC/Home.html
  – 10 fairly frequent, polysemous words
  – 100 occurrences of each word annotated by 5/6 trained annotators (TA)

• Amazon’s Mechanical Turk (MT)
  – 3 polysemous words (adjectives)
  – 150 occurrences of each word annotated by 14 untrained annotators
Alpha Score for Agreement

- Reliability coefficient to measure quality of annotation
- Alpha given by: \( \alpha = 1 - \frac{D_o}{D_e} \)
  - \( D_o \) is the observed disagreement
  - \( D_e \) is the disagreement expected by chance
  - \(-1 \leq \alpha \leq 1\)
- Advantages:
  - Has the same interpretation for different datasets
  - Independent of
    - Data skew,
    - Number of annotators
    - Number of categories
## MASC – IA (Alpha) Scores

<table>
<thead>
<tr>
<th>Word-pos</th>
<th>Senses</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avail.</td>
<td>Used</td>
<td>Ann</td>
<td>Alpha</td>
<td></td>
</tr>
<tr>
<td>long-j</td>
<td>9</td>
<td>4</td>
<td>6</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>fair-j</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>quiet-j</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>time-n</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>work-n</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>land-n</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>show-v</td>
<td>12</td>
<td>10</td>
<td>5</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>tell-v</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>know-v</td>
<td>11</td>
<td>10</td>
<td>5</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>say-v</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>0.37</td>
<td></td>
</tr>
</tbody>
</table>
Mech. Turk – IA (Alpha) Scores

<table>
<thead>
<tr>
<th>Word-pos</th>
<th>Senses</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avail.</td>
<td>Used</td>
<td>Ann</td>
<td>Alpha</td>
</tr>
<tr>
<td>long-j</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>0.15</td>
</tr>
<tr>
<td>fair-j</td>
<td>10</td>
<td>10</td>
<td>14</td>
<td>0.25</td>
</tr>
<tr>
<td>quiet-j</td>
<td>6</td>
<td>6</td>
<td>15</td>
<td>0.08</td>
</tr>
</tbody>
</table>

- Same words have much lower agreement among mechanical turkers as compared to trained annotators
- Can this be used?
Anveshan - Methods

- An approach to more subtle analysis of annotation variation
- Statistical measures
  - Probabilities of Sense usage
    \[ P_a(S = s_i) = \frac{\text{count}(s_i, a)}{\sum_{j=1}^{m} \text{count}(s_j, a)} \]
  - Kullbach-Leibler Divergence
    \[ KLD(P, Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \]
  - Jensen-Shannon Divergence
    \[ JSD(P, Q) = \frac{1}{2} KLD(P, M) + \frac{1}{2} KLD(Q, M) \quad M = (P + Q)/2 \]
Anveshan - Methods

• Methodology
  – For each annotator $a_i$, compute $P(a_i)$
  – Compute $P_{avg}$, which is average of all $P(a_i)$
  – Compute $JSD(P(a_i), P(a_j))$ for all pairs $(a_i, a_j)$
  – Lastly, a distance measure for each annotator, the KLD between each annotator and the average of the remaining annotators.
Results – Outlier Detection

- KLD scores and Sense usage stats for “Long”. Annotator 108 is the outlier, using sense 999 much more than other annotators.
Dropping Outliers – IA Scores

<table>
<thead>
<tr>
<th>Word</th>
<th>Old Alpha</th>
<th>Ann Dropped</th>
<th>New Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>0.67</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>land</td>
<td>0.49</td>
<td>1</td>
<td>0.54</td>
</tr>
<tr>
<td>know</td>
<td>0.37</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>tell</td>
<td>0.45</td>
<td>2</td>
<td>0.52</td>
</tr>
<tr>
<td>say</td>
<td>0.37</td>
<td>2</td>
<td>0.44</td>
</tr>
<tr>
<td>fair</td>
<td>0.54</td>
<td>2</td>
<td>0.63</td>
</tr>
</tbody>
</table>
• **Systematic Disagreement for the word “Show”**
  – (A105,A102);(A107,A108) are similar in behavior.
  – A101 varies most from the mean.
JSD vs. IA Score

- Comparison between JSD and Alpha scores for pairs of annotators ("Show")
- Low JSD : Similar sense usage distribution
- High Alpha : Higher than chance agreement
Confusable Senses

- Confusability between Sense 101 and Sense 102.
- Sense 1 is given as “expressing words”; Sense 2 as “report or maintain”.

![Confusability Graph](image-url)
MTs vs. TAs – Noise

- Mech. Turkers sense usage distribution for “Long”
- Mech. Turkers used much more senses than trained annotators.
- Lot of noise is sense usage distribution
MTs and TAs together

- Set of 5 MTs gave high agreement for “Fair” – IA:0.6
- Put together with 5 TAs for “Fair” – IA:0.54
- Overall agreement falls to 0.43
Conclusion and Future Work

- Deeper analysis of annotation is necessary for subjective annotation tasks
- Anveshan’s methodology identifies sources of disagreement
- Dropping outliers can give a much better IA score
- Untrained annotators increase noise
- Future Work
  - Evaluation and learning issues when there is no known true label
  - Incorporating hidden knowledge (e.g. annotator expertise)
  - Feature engineering to predict variation in annotation