Evaluating Annotation Reliability

Outline

- Why we care
- ITA
  - How it is calculated
  - Confusion matrices
  - What about chance agreement?
- Agreement coefficients
  - How they work in general
  - Types of chance agreement
  - Types of coefficients
  - Problems for semantic annotation

Concerns about manual annotation

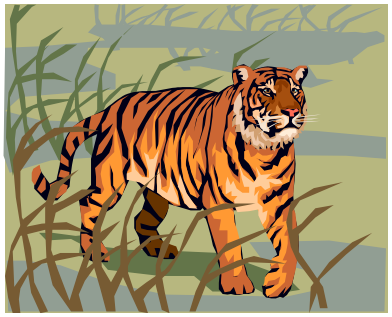
- Are the annotators doing a good job?
  - Do they understand guidelines?
  - Are they paying attention and/or capable of doing the job?
- Are the chosen categories good ones?
  - Are they missing a category?
  - Is it hard to tell the difference between the categories?
  - Are they totally inappropriate categories?

InterTagger Agreement (ITA)

- Simplest method
- Percentage of time annotators agree on the labels they have given the instances
- If you have 2 annotators and 10 tokens, and the annotations for 8 tokens match, ITA = .80, or 80%

Low ITA means

- The guidelines were unclear, or
- The annotators were watching TV while working
- There was a category missing, or
- Some of the categories are indistinguishable in the data, or
- The categories are entirely wrong for the data
- Something is wrong



High ITA means

- ~~Your data is great!~~

- Your data is consistent.
- Your data could be consistently wrong.
- Example: label all open class words, using labels “noun” and “verb”
- Both annotators decided to annotate adjectives with “noun”

Confusion matrices

- See whether there is any pattern to the disagreements
- Tell you where guidelines are obscure or categories are bad, at least bad for your data

RECOVER-V (greenlm, ahoward : 0.94)

	1	2	3	4	5
1	6	1	0	0	0
2	0	21	0	0	0
3	1	0	2	0	0
4	0	0	0	0	0
5	0	0	0	0	0

No pattern

```
COMMAND-V (greenlm, laesecke : 0.76)
      1   2   3   4   5
   ====================
1  | 11   2   2   2   0
2  |  1  11   1   0   0
3  |  1   0  10   1   0
4  |  0   0   0   0   0
5  |  0   0   0   0   0
```

pattern

```
FORGE-V (kelleya, crooksk : 0.36)
      1   2   3   4   5   6   7   8
   ================================
1  |  1   0   0   0   0   0   0   0
2  |  0   8   0   0   0   0   0   0
3  |  0   0   4   0   0   0   0   0
4  |  0   0   0   4   0   0   0   0
5  |  0   0   0   0   0   0   0   0
6  |  0   0  32   0   0   1   0   0
7  |  0   0   0   0   0   0   0   0
8  |  0   0   0   0   0   0   0   0
```

pattern

```
INDUCE-V (kelleya, greenlm : 0.86)
      1   2   3   4   5
   ===================
1  | 23   0   0   0   0
2  |  6  19   0   1   0
3  |  0   0   0   0   0
4  |  0   0   0   0   0
5  |  0   0   0   0   0
```

What do you see?

SEAT-V (greenlm, adamskm : 0.41)

	1	2	3	4	5
1	3	0	0	0	10
2	0	2	0	0	0
3	0	0	2	0	0
4	0	0	0	0	0
5	0	0	0	0	0

What do you see?

DEFY-V (laesecke, adamsjl : 0.86)

	1	2	3	4
1	37	2	1	0
2	1	3	0	0
3	2	1	3	0
4	0	0	0	0

EXPLOIT-V (crooksk, browneal : 0.59)

	1	2	3	4
1	27	14	6	0
2	0	2	0	0
3	0	0	0	0
4	0	0	0	0

If there is a pattern

- Say, category 1 and category 3 are confused with each other, then the problem lies with those categories, not the task as a whole
- If there is no pattern, look for a more general problem

What level of ITA is low?

- Must consider chance agreement
- If there are 2 annotators and 2 labels, we would expect them to agree at least 50%. Should be at least higher than that
- 2 annotators, 4 labels, chance agreement = 25% then 45% would be not bad.
- That assumes an even distribution of categories in the real world
  - Word sense (bank-financial vs. bank-river)
  - Semantic role labels (agent vs. instrument)

Coefficients

- Take chance agreement into account
- Differ in how they calculate the probability that a given coder will assign an item to a given category
- Basic types
  - S
  - Π
  - κ

What does the result mean?

-1 . . . . 0 1

Complete disagreement	Chance agreement	Perfect agreement

Always expect some agreement by chance. Coefficients will always be lower than the corresponding ITA, unless there is perfect agreement (1).

Basic formula

- C coder k category
- A_o observed agreement
- A_e expected agreement (agreement from chance)
- $1-A_e$ agreement above chance that is possible
- $A_o - A_e$ observed agreement above chance

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

- The difference lies in how they calculate A_e

S

- Assumes a uniform distribution across categories and coders
- All classes are equally likely
- 4 word senses, 25% chance of picking a particular sense

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = k \cdot \left(\frac{1}{k}\right)^2 = \frac{1}{k}$$

π

- Want to account for the "real" distribution of the categories in our data (some categories are much more likely)
- Uses the distribution of labels produced by the coders
- Same probability for each category across coders
- n_k number of items labeled with k by both coders
- i number of items

$$A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = \sum_{k \in K} \left(\frac{n_k}{2i} \right)^2 = \frac{1}{4i^2} \sum_{k \in K} n_k^2$$

- $\pi \leq S$; pi is almost always less than S

K

- Takes into account annotator bias
- Annotators may have different tendencies to use one category more than another
- Especially for semantic judgments

$$A_e^{\kappa} = \sum_{k \in K} \hat{P}(k|c_1) \cdot \hat{P}(k|c_2) = \sum_{k \in K} \frac{\mathbf{n}_{c_1 k}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 k}}{\mathbf{i}} = \frac{1}{\mathbf{i}^2} \sum_{k \in K} \mathbf{n}_{c_1 k} \mathbf{n}_{c_2 k}$$

- Most commonly used coefficient in NLP

Other options

- When there are more than 2 annotators
  - Fleiss's multi-π
  - Multi-κ
- When some disagreements are more important than others
  - Weighted agreement coefficients
  - Krippendorff's α

Just tell us which one to use

- No one uses S
- π and κ give very similar results
- A κ scores higher when there is a lot of variability in distribution between coders
- If testing with a small data set before single annotating the rest
  - Don't discount the variability
  - So use π
- If there are multiple annotators or weighted disagreements, see previous slide

Just tell us what a good score is

- $.67 \leq \kappa < .80$ for tentative reliability
- $\kappa \geq .80$ good reliability (Krippendorff, 1980)
- No, no, $\kappa \geq .80$ is a minimum (Krippendorff, 2004)
- Not testing to be sure annotators are better than chance, but to be sure they are not too far from perfect agreement
- Depends on task: Prevalence problem

Prevalence problem

- When 1 category is much more prevalent than another, almost impossible to get a high κ
- Rare categories then have great influence
- Average kappa across 40 verbs for word sense annotation .69
- *Boost* ITA .96
  κ -.18

Each annotator chose boost.02 once, but they disagreed on which one was boost.02

- Byrt, Bishop and Carlin (1993) : $2A_a - 1$
- Report ITA also

Resources

- If 8 or fewer categories: http://faculty.vassar.edu/lowry/kappa.html
- More categories and multiple annotators: http://cosmion.net/jeroen/software/kappao/
- Downloadable VBA program: http://agreestat.com/agreestat