Inter annotator agreement

CS140

February 28, 2017

+ Annotations are like squirrels…

Annotation Diff helps with "spot the difference"

Thanks to University of Sheffield

Which things are most similar?

Thanks to University of Sheffield

+ Validity vs. Reliability

Artstein and Poesio, 2008

- We are interested in the validity of the manual annotation
  - i.e. whether the annotated categories are correct
- But there is no “ground truth”
  - Linguistic categories are determined by human judgment
  - Consequence: we cannot measure correctness directly
- Instead measure reliability of annotation
  - i.e. whether human annotators consistently make same decisions
    → they have internalized the scheme
  - Assumption: high reliability implies validity
- How can reliability be determined?

Slides from Karen Fort, inist, 2011

+ Achieving Reliability (consistency)

- Approaches:
  - each item is annotated by a single annotator, with random checks (≈ second annotation)
  - some of the items are annotated by two or more annotators
  - each item is annotated by two or more annotators - followed by reconciliation
  - each item is annotated by two or more annotators - followed by final decision by superannotator (expert)
- In all cases, measure of reliability: ***coefficients of agreement***

Slides from Karen Fort, inist, 2011

+ Precision/Recall: back to basics

- **Recall**: measures the quantity of found annotations

$$\text{Recall} = \frac{\text{Nb of } \textbf{correct found} \text{ annotations}}{\text{Nb of } \textbf{correct expected} \text{ annotations}}$$

  - **Silence**: *complement* of recall (correct annotations not found)

- **Precision**: measures the quality of found annotations

$$\text{Precision} = \frac{\text{Nb of } \textbf{correct found} \text{ annotations}}{\textbf{Total} \text{ Nb of } \textbf{found} \text{ annotations}}$$

  - **Noise**: *complement* of precision (incorrect annotations found)˜

Slides from Karen Fort, inist, 2011

+ F-measure: back to basics

Wikipedia Dec. 10, 2010

- Harmonic mean of precision and recall
  - or balanced F-score

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

  - ... aka the F1 measure: recall and precision are evenly weighted.

- It is a special case of the general F measure:
  - The value of allows to favor:
    - recall ($\beta = 2$)
    - precision ($\beta = 0.5$)

$$F\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

Slides from Karen Fort, inist, 2011

+ Easy and Hard Tasks

Objective tasks

- Decision rules, linguistic tests
- Annotation guidelines with discussion of boundary cases
- POS tagging, syntactic annotation, segmentation, phonetic transcription, . . .
  - IAA POS tagging = 98.5%
  - IAA Syntax = 93%

Subjective tasks

- Based on speaker intuitions
- Short annotation instructions
- Lexical semantics (subjective interpretation!), discourse annotation & pragmatics, subjectivity analysis, . . .
  - IAA word senses = 70%

Slides from Karen Fort, inist, 2011
[Brants, 2000] for POS and Syntax, [Veronis, 2001] for WSD.

Existing Reliability Measures

- Cohen's Kappa (Cohen, 1960)
- Fleiss's Kappa
- Scott's π (Scott, 1955)
- Krippendorff's α (Krippendorff, 1980)
- Rosenberg and Binkowski, 2004
  - Annotation limited to two categories

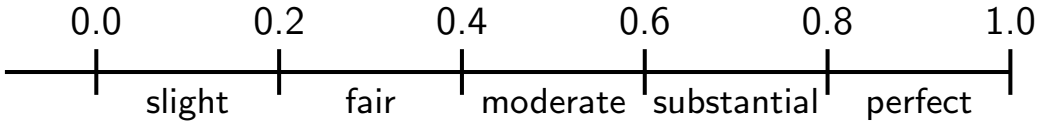
+ Cohen's Kappa (κ)

- Measures the agreement between two annotators, while taking into account the possibility of chance agreement. The equation is:
- Pr(a) actual agreement
- Pr(e) expected agreement

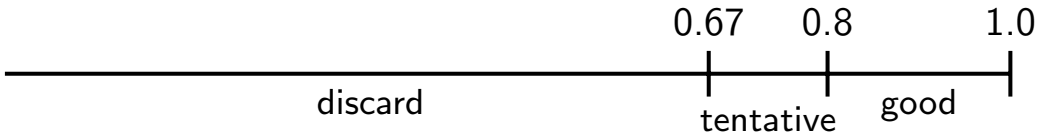
$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Scales for the interpretation of Kappa

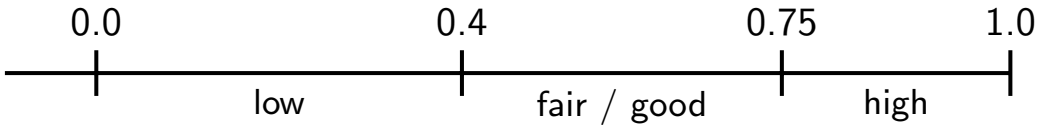
- "It depends"
  - Landis and Koch, 1977



  - Krippendorff, 1980



  - Green, 1997



- "If a threshold needs to be set, 0.8 us a good value [Arstein & Poesio, 2008

Slides from Karen Fort, inist, 2011

		B	B	B	
		pos	neut	neg	TOT
A	pos	54	28	3	85
A	neut	31	18	23	72
A	neg	0	21	72	93
	TOT	85	67	98	250

+ Example: Pr(a)

		B	B	B	TOT
		pos	neut	neg	
A	pos	54	28	3	85
A	neut	31	18	23	72
A	neg	0	21	72	93
	TOT	85	67	98	250

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Pr(a)= (54 + 18 + 72) / 250 = **.576** (57.6%)

+ Example: Pr(e)

		B	B	B	TOT
		pos	neut	neg	
A	pos	54	28	3	85
A	neut	31	18	23	72
A	neg	0	21	72	93
	TOT	85	67	98	250

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- A used the label "positive" 85 times (54 + 28 + 3), or .425% of the time.
- B also used the "positive" label 85 times (54 + 31), which is also .425.
- .425 x .425 = .180

$$Pr(e) = .180 + .077 + .146 = \mathbf{.403}$$

+ Example, computing K

- Putting *Pr(a)* and *Pr(e)* into the equation gives us:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$$K = \frac{.576 - .403}{1 - .403}$$

$$K = \frac{.173}{.597} = .29$$

+ Same agreement, different results

		B	B	TOT
		pos	neg	60
A	pos	45	15	40
A	neg	25	15	
	TOT	70	30	100

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$$K = \frac{.60 - .54}{1 - .54} = 0.1304$$

		B	B	TOT
		pos	neg	60
A	pos	25	35	40
A	neg	5	35	
	TOT	30	70	100

$$K = \frac{.60 - .46}{1 - .46} = 0.2593$$

Fleiss's Kappa (K)

- Cohen's Kappa only addresses 3 annotators
- Fleiss's Kappa allows for more
  - P is actual agreement
  - P_e is expected agreement

$$K = \frac{P - P_e}{1 - P_e}$$

~	pos	neut	neg
A	85	72	93
B	85	67	98
C	68	99	83
D	88	88	74
E	58	120	72
Tot	384	446	420

First step

- Calculate how many assignments went (proportionally) to each category (P_c)

$$Pc = \frac{1}{Aa} A \sum_{i=1}^{A} a_{ic}, 1 = \frac{1}{a} \sum_{c=1}^{k} a_{ic}$$

If you add up all the annotations that an annotator made and divide by the number of annotator each annotator made, you'll get 1

Number of annotators

Number of annotations each annotator created

Sum of the values in the columns

Computing Pc

- So, if we apply the *Pc* equation to the first annotation category, we get the following equation:
- *P(positive)*

  = (85 + 85 + 68 + 88 + 58) / (5 × 250)

  = 384 / 1250

  = **.3072**

~	pos	neut	neg
A	85	72	93
B	85	67	98
C	68	99	83
D	88	88	74
E	58	120	72
Tot	384	446	420
Pc	.307	.357	.336

Computing Pi

- Pi represents each annotators agreement with other annotators, compared to all possible agreement values
  - a: number of annotations per annotator
  - k: number of categories
  - c: current category
  - i: current annotator

$$P_i = \frac{\left(\sum_{c=1}^{k} a_{ic}^2\right) - (a)}{a(a-1)}$$

Application to the example

- *P(Annotator A)* = ($(85^2 + 72^2 + 93^2) - 250$) / $250\,(250-1)$
  - = $21058 - 250$ / 62250
  - = $20808/62250$
  - = $.3343$

$$P_i = \frac{\left(\sum_{c=1}^{k} a_{ic}^2\right) - (a)}{a(a-1)}$$

~	pos	neut	neg	Pi
A	85	72	93	.334
B	85	67	98	.338
C	68	99	83	.338
D	88	88	74	.333
E	58	120	72	.345
Tot	384	446	420	

Final steps

- *Take the average:*
  - $P(e) = (.3343 + .3384 + .3384 + .3328 + .3646) / 5$
    $= 1.7085 / 5$
    $= .3417$
- calculate *P(e)* by summing the squares of the *Pc* values,
- $P(e) = .30722 + .35682 + .3362$
  $= .335$
- plug these values into Fleiss's Kappa equation and calculate our IAA score:
  - $K = (.3417 - .335) / (1 - .335)$
    $= .0067 / .665$
    $= .004$

What's it mean?

- 2 tasks, 2 different values of : .29 and .004.
- Landis and Koch 1977 provide these guidelines for interpreting κ and other agreement metrics:

K	Agreement level
<0	poor
0.01 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.81 – 0.80	substantial
0.81 – 1.00	perfect

+ Interpreting the numbers

- Example 1: Kappa .29 (fair)
  - Only 3 categories
  - BUT: problem is with the "neutral" tag and one annotator had more problem than others
  - SO: beware of averages
- Example 4: Fleis Kappa: .004
  - Annotators A, B, and D all have nearly the same number of positive reviews
  - Annotator E is so far off from everyone else

		B	B	B
		pos	neut	neg
A	pos	54	28	3
A	neut	31	18	23
A	neg	0	21	72

~	pos	neut	neg	Pi
A	85	72	93	.334
B	85	67	98	.338
C	68	99	83	.338
D	88	88	74	.333
E	58	120	72	.345
Tot	384	446	420	