Computational Semantics and Pragmatics

Autumn 2014



Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam

Linguistic Annotation

Important for supervised learning methods and theory validation.

Can we rely on the judgements of one single individual?

> **From Carletta (1996):**
>
> "At one time, it was considered sufficient when working with such judgments to show examples based on the authors' interpretation. Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand, and reliably make, the judgments underlying the research. This is a reasonable requirement, because if researchers cannot even show that people can agree about the judgments on which their research is based, then there is no chance of replicating the research results."

Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254.

- an annotation is considered reliable if several annotators agree sufficiently – they consistently make the same decisions.

Inter-annotator Agreement

- Some terminology and notation:
  - set of *items* $\{i \mid i \in I\}$, with cardinality $\mathbf{i}$.
  - set of *categories* $\{k \mid k \in K\}$, with cardinality $\mathbf{k}$.
  - set of *coders* $\{c \mid c \in C\}$, with cardinality $\mathbf{c}$.

Observed Agreement

The simplest measure of agreement is observed agreement A_o:

- the percentage of judgements on which the coders agree, that is the number of items on which coders agree divided by total number of items.

Binary classification task (true / false): rhetorical question

items	coder A	coder B	agr
I mean, why not.	true	true	✓
How do you think we're going to pay for it?	false	true	×
Isn't that sad?	true	false	×
Did you use to live around here?	false	false	✓
Where's that?	false	false	✓
You ever go by Lucky Computer there?	false	false	✓

- $A_o = 4/6 = 66.6\%$

Contingency table:

	coder B		
coder A	true	false	
true	*1*	1	2
false	1	*3*	4
	2	4	6

Contingency table with proportions:
(each cell divided by total # of items **i**)

	coder B		
coder A	true	false	
true	*.166*	.166	.333
false	.166	*.5*	.666
	.333	.666	1

- $A_o = .166 + .5 = .666 = 66.6\%$

Observed vs. Chance Agreement

Problem: using observed agreement to measure reliability does not take into account agreement that is due to chance.

- In the above example, if annotators make random choices the expected agreement due to chance is 50%:
  - both coders randomly choose `true` $(.5 \times .5 = .25)$
  - both coders randomly choose `false` $(.5 \times .5 = .25)$
  - expected agreement by chance: $.25 + .25 = 50\%$
- An observed agreement of 66.6% is only mildly better than 50%

Factors that can lead to higher chance agreement

- Number of categories: fewer categories will result in higher agreement by chance.
  $\mathbf{k} = 2 \rightarrow 50\%$ $\mathbf{k} = 3 \rightarrow 33\%$ $\mathbf{k} = 4 \rightarrow 25\%$ ...

- Distribution of items among categories: if some categories are very frequent, observed agreement will be higher by chance.
  - both coders randomly choose `true` $(.95 \times .95 = 90.25\%)$
  - both coders randomly choose `false` $(.05 \times .05 = 0.25\%)$
  - expected agreement by chance $90.25 + 0.25 = 90.50\%$

  $\Rightarrow$ Observed agreement of 90% may be less than chance agreement.

In sum, observed agreement does not take chance agreement into account and hence is not a good measure of reliability.

Measuring Reliability

$\Rightarrow$ Reliability measures must be corrected for chance agreement.

- Let A_o be observed agreement, and A_e expected agreement by chance.
- $1 - A_e$: how much agreement beyond chance is attainable.
- $A_o - A_e$: how much agreement beyond chance was found.
- General form of chance-corrected agreement measure of reliability:

$$R = \frac{A_o - A_e}{1 - A_e}$$

The ratio between $A_o - A_e$ and $1 - A_e$ tells us which proportion of the possible agreement beyond chance was actually achieved.

- Some general properties of R:

perfect agreement	**chance agreement**	**perfect disagreement**
$R = 1 = \frac{A_o - A_e}{1 - A_e}$	$R = 0 = \frac{0}{1 - A_e}$	$R = \frac{0 - A_e}{1 - A_e}$

Measuring Reliability: *kappa*

Several agreement measures have been proposed in the literature

Arstein & Poesio (2008) Survey Article: Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics*, 34(4):555–596.

- The general form of R is the same for several measures $R = \frac{A_o - A_e}{1 - A_e}$
- They all compute A_o in the same way:
  - proportion of agreements over total number of items
- They differ on the precise definition of A_e.

We'll focus on the kappa (κ) coefficient (Cohen 1960; see also Carletta 1996)

- κ calculates A_e considering *individual* category distributions:
  - they can be read off from the marginals of contingency tables:

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

category distribution for coder A: $P(\texttt{true}|c_A) = .333$; $P(\texttt{false}|c_a) = .666$
category distribution for coder B: $P(\texttt{true}|c_B) = .333$; $P(\texttt{false}|c_B) = .666$

Chance Agreement for *kappa*

A_e: how often are annotators expected to agree if they make random choices according to their individual category distributions?

- we assume that the decisions of the coders are independent: need to multiply the marginals
- Chance of c_A and c_B agreeing on category k: $P(k|c_A) \cdot P(k|c_B)$
- A_e is then the chance of the coders agreeing on any k:

$$A_e = \sum_{k \in K} P(k|c_A) \cdot P(k|c_B|)$$

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

- $A_e = (.333 \cdot .333) + (.666 \cdot .666) = .111 + .444 = 55.5\%$

An Example

items	coder A	coder B	agr
I mean, why not.	true	true	✓
How do you think we're going to pay for it?	false	true	×
Isn't that sad?	true	false	×
Did you use to live around here?	false	false	✓
Where's that?	false	false	✓
You ever go by Lucky Computer there?	false	false	✓

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

- $A_o = .166 + .5 = .666 = 66.6\%$
- $A_e = (.333 \cdot .333) + (.666 \cdot .666) = .111 + .444 = 55.5\%$

$$\kappa = \frac{66.6 - 55.5}{100 - 55.5} = \frac{11.1}{44.5} = \mathbf{24.9\%}$$

kappa for more than two coders

kappa can be generalised to multiple coders.

- We need to compute pairwise observed agreement:
  - the amount of agreement for each item i is the proportion of agreeing judgements per pairs of coders out of the total number of judgment pair for i.
  - A_o is the mean of the amount of agreement for all items $i \in I$
- We need to compute pairwise expected agreement:
  - recall that k uses the individual category distributions $P(k|c) = \mathbf{n}_{kc}/\mathbf{i}$
  - the chance of two coders agreeing on k is $P(k|c_A) \cdot P(k|c_B)$
  - the chance of two arbitrary coders c_n and c_m agreeing on category k is the mean of $P(k|c_n) \cdot P(k|c_m)$ over all pairs of coders.
  - A_e is the sum of this join probability over all $k \in K$.
  - (this is equivalent to the mean of A_e for all pairs of coders)

Scales for the Interpretation of Kappa

- Landis and Koch (1977)

  0.0 – 0.2 : slight
  0.2 – 0.4 : fair
  0.4 – 0.6 : moderate
  0.6 – 0.8: substantial
  0.8 – 1.0 : perfect

- Krippendorff (1980)

  0.0 – 0.67 : discard
  0.67 – 0.8 : tentative
  0.8 – 1.0: good

- Green (1997)

  0.0 – 0.4 : low
  0.4 – 0.75 : fair / good
  0.75 – 1.0: high

- There are many other suggestions as well. . .

Weighted Disagreements

- The classic version of κ considers all types of disagreements equally.
- However, we may want to treat some disagreements as more important than others – some categories may be more similar than others.
- We can use weighted coefficients: Krippendorff's α and *weighted kappa* κ_w.
  - The formula for κ_w derives agreement from disagreement:

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

  - We'll see how to derive D_o and D_e from the confusion matrices; for details of the formulas see Arstein & Poesio (2008).

Weighted Disagreements – An Example

Consider this confusion matrix from Arstein & Poesio (2008):

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

We can calculate unweighted κ as described before:

- A_o : the sum of the cells in the diagonal
  $A_o = .46 + .32 + .10 = .88$
- A_e : the sum of the marginals for each category (multiplied)
  $A_e = .46 \times .52 + .44 \times .32 + .10 \times .16 = .396$
- $\kappa = (A_o - A_e)/(1 - A_e)$
  $\kappa = (.88 - .396)/(1 - .396) = .8013$

Weighted Disagreements – An Example

Suppose we weight the distances between the categories as shown in the RHS table: identical categories have 0 disagreement, while 1 denotes maximal disagreement.

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

	coder B		
coder A	Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

To calculate κ_w, we can derive D_o and D_e as follows:

- D_o : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).
- D_e : the sum of $D_e^{k_i k_j}$ for each category pair k_i, k_j, where
  - $D_e^{k_i k_j}$: the product of the marginals for k_i and k_j divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

Weighted Disagreements – An Example

coder A	coder B: Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

coder A	coder B: Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

- D_o : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).

$$\mathrm{D_o} = \frac{46\times 0 + 6\times 1 + 32\times 0 + 6\times 0.5 + 10\times 0}{100} = \frac{6+3}{100} = 0.09$$

- D_e : the sum of $D_e^{k_i k_j}$ for each category pair k_i, k_j, where
  - $D_e^{k_i k_j}$: the product of the marginals for k_i and k_j divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

$$\begin{aligned} &\tfrac{46\times 52}{100\times 100}\times 0 + \tfrac{44\times 52}{100\times 100}\times 1 + \tfrac{10\times 52}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 32}{100\times 100}\times 1 + \tfrac{44\times 32}{100\times 100}\times 0 + \tfrac{10\times 32}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{44\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{10\times 16}{100\times 100}\times 0 \end{aligned} \qquad 0.49$$

Weighted Disagreements – An Example

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

	coder B		
coder A	Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

$\kappa_w = 1 - (.09/.49) = .8163$

$\kappa = (.88 - .396)/(1 - .396) = .8013$

Different types of non-reliability

- Misinterpretation of annotation guidelines: may not result in disagreement $\rightarrow$ may not be detected
- Random slips: lead to chance agreement between annotators
- Different intuitions: lead to systematic disagreements

Gold-standard annotations

An annotated linguistic corpus typically is released with 1 annotation, considered the gold standard.

- often only a small part of the annotation is tested for reliability
- experts make tricky decisions
- annotators discuss and reach a consensus

The construction of annotated linguistic resources is rapidly and radically changing with the use of crowdsourcing platforms, like Amazon's Mechanical Turk.

- motivation: cheap, quick, more data
- challenges: how do we make sure the annotation is reliable? how to derive a gold standard?

Our website on Collective Annotation `http://www.illc.uva.nl/Resources/CollectiveAnnotation/`

Inter-annotator agreement with NLTK

Upon completing the annotation exercise (by Tue 30 latest), you will receive the annotation of a fellow student. You can use the following NLTK modules to compute inter-annotator agreement:

```
nltk.metrics.agreement
http://www.nltk.org/_modules/nltk/metrics/agreement.html
nltk.metrics.confusionmatrix
http://www.nltk.org/_modules/nltk/metrics/confusionmatrix.html
```

- Transform the annotation into the right format (a list of triples)
- Calculate kappa (and possibly other agreement measures)
- Extract the two lists of categories and print the confusion matrix
- What kinds of patterns of disagreement do you find? What refinements to the annotation guidelines would you propose to reduce the amount of disagreement?

We will discuss this in class on Thursday 2 Oct.

Example – does not show how to get the data into the right format.

