Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington
Week 8: 5/14-5/16

Overview

- Agreement, reliability, validity
- Reliability in sociophonetics
- Reading questions
- Chance-corrected agreement
- >2 coders
- Weighted metrics
- Krippendorff's methodological recommendations
- Artstein & Poesio's recommendations
- Lab preview

Validity and reliability of annotations

- Reliability: Is the annotation methodology reproducible?
- Validity: Do the annotations capture some aspect of reality ("ground truth")?
- How does inter-annotator agreement relate to each of these?
- How do considerations of reliability and validity come up in our research?

Reliability in sociophonetics

- Annotations can be auditory impressions (categorical) or acoustic measurements (continuous)
- Measure can be between different coders or annotations at different times by the same coder
- Outliers: Remove or correct before measuring reliability

Reading questions

- Consistency of auditory coding between raters of different dialectal backgrounds: How much should we worry about this? Can we target this particular issue in training, and if so, how?

Reading questions

- Segmentation: In addition to needing to be consistent, it seems like the choice of segmentation can have profound effects on annotation quality. Should corpora have standard segmentations so that different annotation projects all use the same ones? How do you choose the right segmentation?

Reading questions

- What if the set of labels is small, but the markables vary over a big range? (E.g., Riebold's creaky voice study.) How can we use these measures of reliability?

Reading questions

- Does intra-rater reliability rely on the presence of another rater? In other words, does intra-rater reliability mean anything if there is no inter-rater reliability to compare it to?

- How do we ensure that we are achieving validity and not just reliability?

Reading questions

- On page 557 (A&P) there's some set notation for the sets of items, categories, and coders. Could we maybe unpack that?

- The set of **items** is $\{ i \mid i \in I \}$ and is of cardinality **i**.
- The set of **categories** is $\{ k \mid k \in K \}$ and is of cardinality **k**.
- The set of **coders** is $\{ c \mid c \in C \}$ and is of cardinality **c**.

Terminology

- Items (i): things to code (also: "markables")
- Categories (k): labels in the annotation scheme
- Coder (c): annotators
- Ao: observed agreement
- Ae: expected agreement

Agreement

- First pass: What percentage of the labels are the same?
- Problems with this:
  - Favors coding schemes with fewer categories (why?)
  - Does not correct for the distribution of items among categories (why?)
- When would we want to know it anyway?

Chance-corrected coefficients

- How much agreement over and above chance is found?

$$S, \pi, \kappa = \frac{\mathrm{A}_o - \mathrm{A}_e}{1 - \mathrm{A}_e}$$

- Ao is the same for all of them (percentage agreement). Difference is in Ae.

Estimating expected agreement (Ae)

- S: Chance = uniform distribution of items among categories
- π: Chance = not necessarily uniform distribution, but same for all coders
- κ: Chance = separate distributions for each coder

Estimating expected agreement

- S: Depends only on the number of categories
- π: Estimate from annotated data
- κ: Estimate from annotated data

Relationship among S, π, κ

- π <= S
- π <= κ
- κ, S: no fixed relationship

S is problematic

- Can be artificially increased by adding spurious categories that won't appear in the data
- Uniform distribution is an okay guess if we have no info, but we do have (post-hoc) info

π vs. K

- π: chance of agreement among arbitrary coders
- κ: chance of agreement among this set of coders
- Choice between the two should depend on the desired interpretation of chance agreement
- Arstein & Poesio recommend π (or α) for most CL tasks, but note that π and κ are often actually very close

>2 coders

- Simplistic version: Calculate pairwise agreement and average
- Better: multi-π and multi-κ
- Measure both Ao and Ae in terms of pairwise agreement

Weighted Agreement Coefficients

- In π and κ, all disagreements are treated equally
- But in many coding systems, some labels are more similar than others
  - Examples?

Krippendorff's α

- Similar to π in that Ae is based on one distribution for all coders
- Applies to multiple coders
- Allows for different magnitudes of disagreement

Krippendorff's α

- Different ways to measure dissimilarity lead to different values for the coefficient
- Makes it even harder to interpret those values

Krippendorff's methodological recommendations

- To use observed agreement as a measure of reproducibility, the study must:
  - Employ an exhaustively formulated, clear, and usable coding scheme
  - Use clearly specified criteria concerning the choice of coders
  - Ensure that coders generate the labels independently

Does work in CL do this? Should it? What about sociolx?

- To use observed agreement as a measure of reproducibility, the study must:
  - Employ an exhaustively formulated, clear, and usable coding scheme
  - Use clearly specified criteria concerning the choice of coders
  - Ensure that coders generate the labels independently

Establishing significance

- Can be done with the z statistic
- But that only shows whether the agreement is different from chance
- We really want to know how much it differs from perfect agreement
- More relevant measure is confidence interval

A&P conclusions: methodology

- #1: Measuring IAA is important
- #2: Use a chance-corrected measure
- #3: Using expert coders is ok
- #4: Also perform some kind of other independent evaluation of the corpus (based on the task it is meant for, for example)

A&P conclusions: choosing coefficients

- π if the labels are equally distinct from each other
- Attenuate annotator bias by using more annotators
- α if the labels aren't equally distinct

A&P conclusions: Interpreting values

- One single cut-off point for all types of tasks doesn't seem reasonable
- The field needs to put more thought into what these values mean
  - (esp. with weighted metrics)
- More important than the actual value is the discussion of methodology

Lab 7 preview