Lecture 11b: Annotation

Nathan Schneider
(with material from Henry Thompson, Alex Lascarides)

ENLP | 3 March 2022

Annotation

Why “gold” $\neq$ perfect

Quality Control

Factors in Annotation

Suppose you are tasked with building an annotated corpus. (E.g., with part-of-speech tags.) In order to estimate **cost** in time and money, you need to decide on:

- Source data (genre? size? licensing?)
- Annotation scheme (complexity? guidelines?)
- Annotators (expertise? training?)
- Annotation software (graphical interface?)
- Quality control procedures (multiple annotation, adjudication?)

Annotation Scheme

- Assuming a competent annotator, some kinds of annotation are straightforward for most inputs.
- Others are not.
  - Text may be ambiguous
  - There may be gray area between categories in the annotation scheme

You play annotator

Noun or adverb?

- **Yesterday** was my birthday .
- **Yesterday** I ate a cake .
- He was fired **yesterday** for leaking the information .
- I read it in **yesterday** 's news .
- I had not heard of it until **yesterday** .

You play annotator

Verb, noun, or adjective?

- We had been **walking** quite briskly
- **Walking** was the remedy, they decided
- In due time Sandburg was a **walking** thesaurus of American folk music.
- we all lived within **walking** distance of the studio
- a woman came along carrying a folded umbrella as a **walking** stick
- The **Walking** Dead premiered in the U.S. on October 31, 2010, on the cable television channel AMC

Annotation: Not as easy as you might think

Pretty much any annotation scheme for language will have some difficult cases where there is gray area, and multiple decisions are plausible.

- Because human language needs to be **flexible**, it cuts corners and is reshaped over time.
- Not just syntax: wait till we get to semantics!

Annotation Guidelines

However, we want a dataset's annotations to be as clean as possible so we can use them reliably in systems.

Documenting conventions in an annotation manual/standard/guidelines document is important to help annotators produce **consistent** data, and to help end users interpret the annotations correctly.

Annotation Guidelines

- Penn Treebank: 36 POS tags (excluding punctuation).
- Tagging guidelines (3rd Revision): 34 pages
  - "The temporal expressions *yesterday*, *today* and *tomorrow* should be tagged as nouns (NN) rather than as adverbs (RB). Note that you can (marginally) pluralize them and that they allow a possessive form, both of which true adverbs do not." (p. 19)
  - An entire page on nouns vs. verbs.
  - 3 pages on adjectives vs. verbs.
- Penn Treebank bracketing (tree) guidelines: >300 pages!

Annotation Quality

But even with extensive guidelines, human annotations won't be perfect:

- Simple error (hitting the wrong button)
- Not reading the full context
- Not noticing an erroneous pre-annotation
- Forgetting a detail from the guidelines
- Cases not anticipated by or not fully specified in guidelines (room for interpretation)

"Gold" data will have some tarnish. How can we measure its quality?

Inter-annotator agreement (IAA)

- An important way to estimate the reliability of annotations is to have multiple people independently annotate a common sample, and measure **inter-annotator**/coder/rater **agreement**.
- **Raw agreement rate**: proportion of labels in agreement
- If the annotation task is perfectly well-defined and the annotators are well-trained and do not make mistakes, then (in theory) they would agree 100%.
- If agreement is well below what is desired (will differ depending on the kind of annotation), examine the sources of disagreement and consider additional training or refining guidelines.
- The agreement rate can be thought of as an upper bound (**human ceiling**) on accuracy of a system evaluated on that dataset.

IAA: Beyond raw agreement rate

- Raw agreement rate counts all annotation decisions equally.
- Some measures take knowledge about the annotation scheme into account (e.g., counting singular vs. plural noun as a minor disagreement compared to noun vs. preposition).
- What if some decisions (e.g., POS tags) are far more frequent than others?
    - If 2 annotators both tagged *hell* as a noun, what is the chance that they agreed **by accident**? What if they agree that it is an interjection (rare tag)—is that equally likely to be an accident?
    - **Chance-corrected** measures such as Cohen's kappa (κ) adjust the agreement score based on label probabilities. (Cohen's assumes 2 raters, categorical labels)
    - ... but they make modeling assumptions about how "accidental" agreement would arise; important that these match the reality of the annotation process!

Cohen's κ

- 2 raters (annotators A and B), categorical labels $(y_1, y_2, \dots)$
- From interannotator confusion matrix, compute:
  - Observed probability of agreement (i.e., raw agreement rate): $p_o = \hat{P}(A = B = y_1) + \hat{P}(A = B = y_2) + \cdots$
  - Expected agreement **by chance** if annotators' decisions were independent: $p_e = \hat{P}(A = y_1)\hat{P}(B = y_1) + \hat{P}(A = y_2)\hat{P}(B = y_2) + \cdots$
- Agreement above chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- Interpretation of κ is subjective.
  - Landis and Koch (1977): 0–0.20 is "slight" agreement, 0.21–0.40 is "fair", 0.41–0.60 is "moderate", 0.61–0.80 is "substantial", and 0.81–1 is "almost perfect"
- Assumes that chance is random guessing according to one's overall preferences—not always realistic!
- Tends to underestimate agreement for rare labels.

Crowdsourcing

- Quality control is even more important when eliciting annotations from “the crowd”.
- E.g., **Amazon Mechanical Turk** facilitates paying anonymous web users small amounts of money for small amounts of work (“Human Intelligence Tasks”).
- Need to take measures to ensure annotators are qualified and taking the task seriously.
  - Redundancy to combat noise: Elicit 5+ annotations per data point.
  - Embed data points with known answers, reject annotators who get them wrong.