Creating annotated corpora

Giovanni Colavizza

Text Mining
Amsterdam University College

April 6, 2021

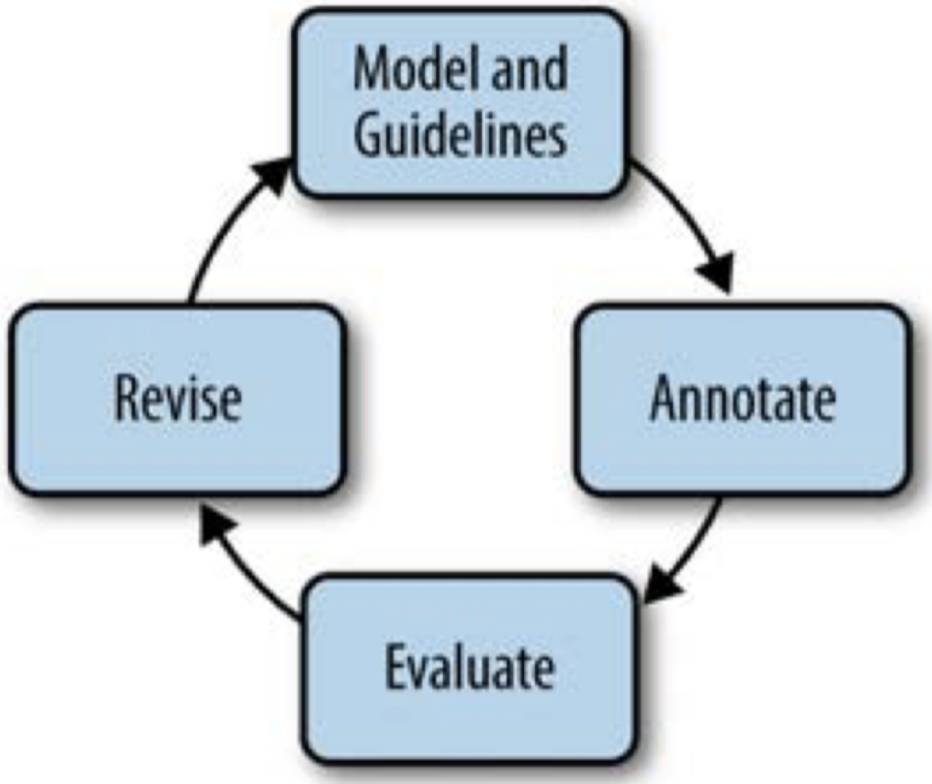
Overview

NLP and Human Annotations

NLP and Human Annotations

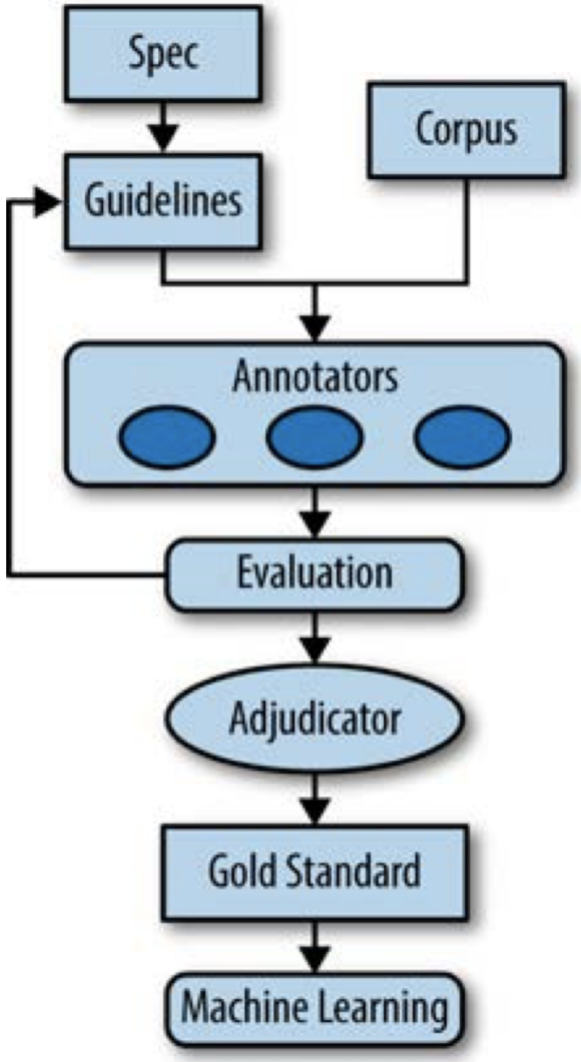
- NLP (and ML in general) is driven by human-annotated corpora.
- `http://nlpprogress.com`.
- Annotation is **difficult** and **expensive**.

Annotation pipeline



Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Annotation pipeline



Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Specs and guidelines

- Goal: given our problem, *how can we formalize our description of the annotation process for multiple annotators to provide the same judgment?*
  - What is the goal of the project?
  - How will the annotation be created? (For example, which tags or documents to annotate first, how to use the annotation tools, etc.)
  - What is each tag called and how is it used? (provide examples and discuss problematic choices.)
  - What parts of the text do you want annotated?
- Note: annotation is usually boring and time-consuming, and cannot be done for 8 hours straight. Annotators also get better over time: early annotations might be discarded.

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Adjudication

- **Adjudication** is the process of deciding on a single annotation for a piece of text, using information from all independent annotators.
- Yes, it is only possible when multiple annotators independently annotate (at least some) of the corpus. *This is a very good procedure to follow, and the only one which will allow to evaluate results*.
- It can be as time-consuming (or more so) as a primary annotation.
- It does not need to be identical with a primary annotation (all annotators can be wrong by chance), but unlikely so.

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Evaluation

Interannotator agreement

annotator A

annotator B

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

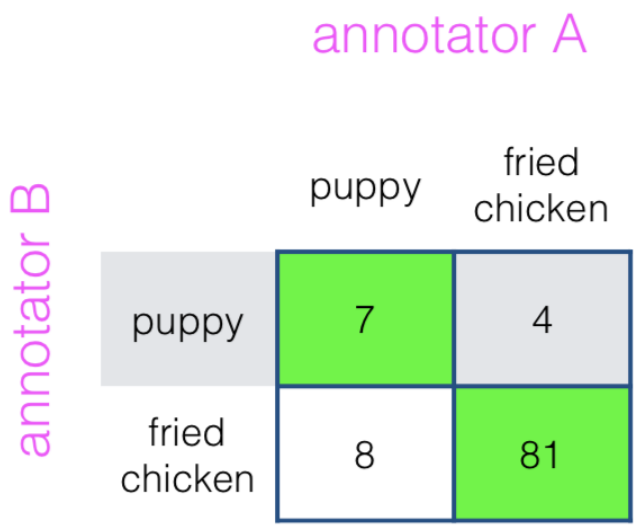
Credit: David Bamman, UCBerkeley.

Cohen's kappa

- Similar idea to mutual information: observed minus expected agreement.
- Cohen's kappa is defined for two annotators over the same set of annotation tasks:

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

Where p_o is the observed correct agreement and p_e the expected correct agreement.

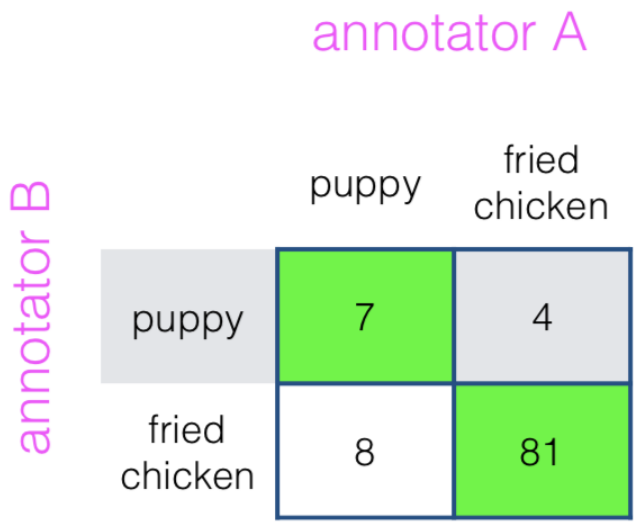


Credit: David Bamman, UCBerkeley.

Cohen's kappa example

- $p_o = 0.88$
- $p_e = P(A = puppy)P(B = puppy) + P(A = chicken)P(B = chicken)$
-

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773} = 0.471$$



Credit: David Bamman, UCBerkeley.

Cohen's kappa scores

Note: these are rules of thumb.

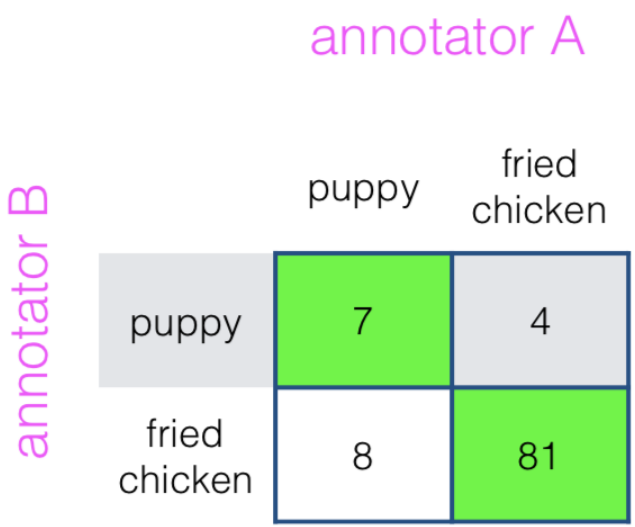
0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement

Exercise: try to calculate fringe cases. E.g., 50/50 puppet/chicken all in agreement, 0/100 puppet/chicken all in agreement, 50/50 wrong puppet/chicken all in agreement.
Credit: David Bamman, UCBerkeley.

Fleiss' kappa

- Extension to multiple annotators (> 2).
- Defined as Cohen's kappa but comparing pairs of annotators:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$



Credit: David Bamman, UCBerkeley.

Fleiss' kappa

- Number of annotators who assign category j to item i: n_{ij}.
- For item i with n annotations, how many annotators agree among all $n(n-1)$ possible pairs:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

- Note that N is the number of items, and K the available annotation categories. Average agreement among all items:

$$P_o = \frac{1}{N} \sum_{i=1}^{N} P_i$$

Fleiss' kappa, continued

- Probability of category j:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

- Expected agreement by chance:

$$P_e = \sum_{j=1}^{K} p_j^2$$

- Back to original formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Platforms and shared tasks

Annotation tools

- Brat `http://brat.nlplab.org`
- Inception `https://inception-project.github.io`
- Prodigy `https://prodi.gy`

Annotation platforms

- Supervisely `https://supervise.ly`
- Dataturks `https://dataturks.com`
- Amazon Mechanical Turk `https://www.mturk.com`
- Figure Eight `https://www.figure-eight.com`
- Alcrowd `https://www.aicrowd.com`