Annotation Analysis Exercise
Methods in Computational Linguistics

Franziska Weeber

November 24, 2025

This exercise is based on material from Eva Maria Vecchi.

Introduction

The goal of this exercise is for you to evaluate annotation agreement and understand how separate annotations can be combined to produce a gold standard.

The main issues for the first part of this exercise are discreticizing the linguistic phenomenon you will annotate, creating the annotation manual, and annotating a small corpus fragment.

What to do before the exercise:

- Read the exercise file
- Go through the relevant parts of the annotation lecture
- Go through the example calculations in the exercise file
- Download the python file for the exercise and make sure you can run the code in it (you can, but do not have to, transfer it to a jupyter notebook)
  If you have any issues with executing the code, you can come to class early, I will be there from 9.15 on!
- Implement the code for the exercise

What we will do in class:

- We will go over one (of multiple) possible solution
- You also have time to ask questions

The participation in this lab session is voluntary: you do not have to submit any result. If you want to get feedback, please ask questions during the session.

Data

For analyzing annotations, we will use dummy data to exemplify the process. Let's use the following annotation matrix $A^{N \times |J|}$, showing the category assignment of $N = 3$ human judges, to a set of 15

instances ($|J| = 15$), with 3 categories ($|K| = 3$). The i-th row shows the categories assignments made by annotator i.

$$A = \begin{bmatrix} 1 & 1 & 1 & 3 & 3 & 2 & 3 & 3 & 1 & 2 & 2 & 2 & 1 & 3 & 1 \\ 1 & 2 & 1 & 2 & 3 & 2 & 3 & 3 & 1 & 1 & 2 & 2 & 1 & 3 & 1 \\ 2 & 1 & 1 & 3 & 3 & 2 & 3 & 3 & 2 & 2 & 2 & 2 & 2 & 3 & 1 \end{bmatrix}$$

Annotation Agreement

Compute Cohen's Kappa and Fleiss' Kappa on the assembled annotations. Use python to assemble annotations from the provided dataset, and to compute the scores.

Cohen's Kappa

For **each pair** of annotators that have annotated the same instances, compute Cohen's Kappa as follows:

- Compute the agreement score A_{xy} between annotators $i = x$ and $i = y$:

$$A_{xy} = \frac{|\{j \in J | a_{xj} = a_{yj}\}|}{|J|}$$

$$\left(= \frac{\text{the number of instances for which annotators x and y assigned the same category}}{\text{the number of instances annotated by both x and y}} \right)$$

  where J is the set of instances annotated by annotators $i = x$ and $i = y$, a_{ij} is the category assigned by annotator i to instance j.

  Note that here, A_{xy} does not refer to matrix A but is the agreement score. a_{xj} and a_{yj} refer to annotations in matrix A.

- Compute the chance agreement score E_{xy} between annotators $i = x$ and $i = y$ (the probability that the two annotators will assign the same category):

$$E_{xy} = \sum_{k \in K} \frac{|\{a_{xj} | j \in J, a_{xj} = k\}|}{|J|} \frac{|\{a_{yj} | j \in J, a_{yj} = k\}|}{|J|}$$

- Compute Cohen's kappa for annotators x and y:

$$\kappa_{xy} = \frac{A_{xy} - E_{xy}}{1 - E_{xy}}$$

Example Considering the first $i = 0$ and second $i = 1$ annotator from the example annotation matrix A, let's compute Cohen's kappa:

- A_{01} is the agreement score between the two judges. It is the number of instances on which they agree (they assigned the same category), divided by the total number of instances:

$$A_{01} = \frac{12}{15} = 0.8$$

- E_{01} is the chance agreement score, so the probability that the two annotators assign the same category to an instance, which is the probability that they both assign category 1 + the probability they both assign category 2 + the probability they both assign category 3. The probability that judges $i = 0$ and $i = 1$ assign the same category k to an instance is the probability that judge 0 assigns category k × the probability that judge 1 assigns category k: $p_{01}(k) = p_0(k)p_1(k) = \frac{n_0(k)}{|J|}\frac{n_1(k)}{|J|}$, where $n_i(k)$ is the number of times judge i assigned category k to instances in set J: $n_i(k) = |\{a_{ij} = k | k \in K, j \in J\}|$:

$$\begin{aligned} E_{12} &= \sum_{k=1}^{3} p_{12}(k) \\ &= p_{12}(1) + p_{12}(2) + p_{12}(3) \\ &= p_1(1)p_2(1) + p_1(2)p_2(2) + p_1(3)p_2(3) \\ &= \frac{6}{15}\frac{6}{15} + \frac{4}{15}\frac{5}{15} + \frac{5}{15}\frac{4}{15} \\ &\approx 0.337 \end{aligned}$$

therefore,

$$\kappa_{12} = \frac{A_{12} - E_{12}}{1 - E_{12}} = \frac{0.8 - 0.337}{1 - 0.337} \approx 0.698$$

Fleiss' Kappa

For **the whole group** of annotators that have annotated the same instances, compute Fleiss' kappa as follows:

Hint: When calculating Fleiss' kappa, it can make sense to create a matrix $N^{|}K| \times |J|$ from matrix A, where n_{kj} is the number of annotators who assigned category k to instance j. However, there are also valid solutions without this. Also note that n and its variations are used in many contexts here to indicate the total number of something, so pay attention to what n actually counts!

- Compute the agreement score $\bar{P}$ for all annotators:

$$\bar{P} = \frac{1}{|J|}\sum_{j=1}^{|J|} P_j$$

P_j the extent to which annotators agree for the j-th instance, which has n annotations (i. e., compute how many rater pairs are in agreement, relative to the number of all possible rater pairs).

The number of all possible rater pairs when we have N annotators is N choose 2:

$$\binom{N}{2} = \frac{N!}{2!(N-2)!} = \frac{N(N-1)}{2}$$

The number of pairs that are in agreement for instance j and category k, when n_{kj} is the number of judges that have assigned the same category k to instance j is n_{kj} choose 2:

$$\binom{n_{kj}}{2} = \frac{n_{kj}!}{2!(n_{kj}-2)!} = \frac{n_{kj}(n_{kj}-1)}{2}$$

The agreement score for instance j is then:

$$\begin{aligned} P_j &= \frac{1}{\frac{n(n-1)}{2}} \sum_{k=1}^{|K|} \frac{n_{kj}(n_{kj}-1)}{2} \\ &= \frac{1}{n(n-1)} \sum_{k=1}^{|K|} n_{kj}(n_{kj}-1) \\ &= \frac{1}{n(n-1)} \left(\sum_{k=1}^{|K|} n_{kj}^2 - n \right) \end{aligned}$$

- Compute the chance agreement score $\bar{P}_e$ for all annotators:

$$\bar{P}_e = \sum_{k=1}^{|K|} p_k^2 \qquad p_k = \frac{1}{|J|n} \sum_{j=1}^{|J|} n_{kj}, \qquad 1 = \sum_{k=1}^{|K|} p_k$$

where n – number of ratings per instance; n_{kj} – number of annotators who assigned the k-th category to instance j

- Compute Fleiss' kappa:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Example Use the same annotation matrix A from above.

- Compute the agreement score $\bar{P}$:
  - For each instance j, compute how much judges agree on annotations for this instance (how many rater pairs are in agreement, relative to the total number of annotator pairs). We use the P_j formula:

$$P_j = \frac{1}{n(n-1)} \left(\sum_{k=1}^{|K|} n_{kj}^2 - n \right)$$

for each instance j, where $n = 3$ is the number of ratings for each instance, n_{kj} is the number of times category k was assigned to instance j. For instance 1, we have:

$n_{1\ 1} = 2$, $n_{1\ 2} = 1$, $n_{1\ 3} = 0$. We compute then P_1:

$$\begin{aligned} P_1 &= \frac{1}{3(3-1)} \left(\sum_{k=1}^{|K|} n_{1\ k}^2 - 3 \right) \\ &= \frac{1}{6}(n_{1\ 1}^2 n_{1\ 2}^2 + n_{1\ 3}^2 - 3) \\ &= \frac{1}{6}(2^2 + 1^2 + 0^2 - 3) \\ &= \frac{1}{6}2 = 0.33 \end{aligned}$$

We compute similarly P_2 ... P_{15}

Then, we compute $\bar{P}$:

$$\begin{aligned} \bar{P} &= \frac{1}{|J|} \sum_{j=1}^{|J|} P_j \\ &= \frac{1}{15}(P_1 + P_2 + ... + P_{15}) \end{aligned}$$

- Compute the chance agreement score $\bar{P}_e$:
  - Compute p_k – the agreement for each category k, where $n = 3$ the number of annotations for each instance, and n_{kj} is the number of judges that assigned category k to instance j (that we computed already!). We have 15 instances. For category $k = 1$ we have:

$$\begin{aligned} p_1 &= \frac{1}{|J|n} \sum_{j=1}^{|J|} n_{j\ 1} \\ &= \frac{1}{15 \times 3}(n_{1\ 1} + n_{2\ 1} + ... + n_{15\ 1}) \end{aligned}$$

  Compute $\bar{P}_e = (p_1^2 + p_2^2 + p_3^2)$

- Compute Fleiss' kappa:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \approx 0.599$$

Annotation aggregation

After verifying annotator agreement, we will start assembling the gold standard dataset by combining the individual judgments. We will use the bias-correcting rule. We compute the weights w_{ik} that capture how much weight to give to each annotator i's choice of category k, following the steps:

- First compute the relative frequency with which each annotator assigns category k:

$$Freq_i(k) = \frac{|\{a_{i*} = k\}|}{|\{a_{i*}\}|}$$

- Compute the overall relative frequency of each category k as the ratio between the number of times k was assigned, out of the total number of annotations.

$$Freq(k) = \frac{|\{a_{**} = k\}|}{|\{a_{**} \in K\}|}$$

- Compute $w_{i\ k}$, using one of the following formulas for the bias-correcting rule (BCR):

Diff	difference-based BCR	$w_{i\ k} = 1 + Freq(k) - Freq_i(k)$
Rat	ratio-based BCR	$w_{i\ k} = Freq(k)/Freq_i(k)$
Com	complement-based BCR	$w_{i\ k} = 1 + 1/\|K\| - Freq_i(k)$
Inv	inverse-based BCR	$w_{i\ k} = 1/Freq_i(k)$

After computing all the weights w_{ik}, for each annotated instance j in the dataset, we compute its category assignment as the category that has the highest weight, according to the following weighted aggregator function:

$$F_w(A)_j = argmax_{k \in K} \sum_{i \in N_j, a_{ij} = k} w_{ik}$$

where K is the set of categories, a_{ij} is the category assigned by annotator i to instance j, N_j is the number of annotators that have annotated instance j.

If there are ties – there are several categories with the same score – **include all options in the final dataset**.
Hint: You can achieve this by not finding the index of the maximum value (which would give you the first occurrence only), but by finding the max value and then iterating through all values to see whether they are equal, storing the indices of those which are.

Example Based on the annotation matrix A on page 1:

- Compute the category frequencies for each annotator:
  - annotator 1: $Freq_1(1) = 6/15 = 0.4, Freq_1(2) = 4/15 \approx 0.267, Freq_1(3) = 5/15 \approx 0.333$
  - ...and so on for the other annotators
- Compute the overall category frequencies:
  - $Freq(1) = 15/45 \approx 0.333, Freq(2) = 16/45 \approx 0.356, Freq(3) = 14/45 \approx 0.311$
- Choose one of the formulas for w_{ik} and compute the weights. If we choose the ratio-based formula, for example, we compute the weights:
  - $w_{1\ 1} = Freq(1)/Freq_1(1) = 0.333/0.4 = 0.832$
  - ...and so on for all the other weights

Compute the category assignment for each instance using the weighted aggregator as shown above.