13

Bias decreases in proportion to the number of annotators

Ron Artstein and Massimo Poesio †

Abstract

The effect of the individual biases of corpus annotators on the value of reliability coefficients is inversely proportional to the number of annotators (less one). As the number of annotators increases, the effect of their individual preferences becomes more similar to random noise. This suggests using multiple annotators as a means to control individual biases.

Keywords corpus annotation, reliability, kappa

13.1 Introduction

One of the problems of creating an annotated corpus is inter-annotator reliability—the extent to which different annotators "do the same thing" when annotating the corpus. Among the factors that may affect reliability is what we will call the individual annotator bias, informally thought of as the differences between the individual preferences of the various annotators. Methods to control bias include the development of clear annotation schemes, detailed and explicit manuals, and extensive training. Nevertheless, some individual differences in the interpretation of such schemes and manuals will always remain. We suggest another means to control for bias—increasing the number of annotators. We give a proof that the effect of individual annotator bias on standard measures of reliability decreases in proportion to the number of anno-

†This work was supported in part by EPSRC project GR/S76434/01, ARRAU. We wish to thank Tony Sanford, Patrick Sturt, Ruth Filik, Harald Clahsen, Sonja Eisenbeiss, and Claudia Felser.

FG-MoL 2005.
James Rogers (ed.).

tators (or, to be pedantic, in proportion to the number of annotators less one).

In order to test inter-annotator reliability, two or more annotators annotate the same text, and their annotations are compared using some statistical measure. Since the publication of Carletta (1996) it has been common in computational linguistics to use a family of related but distinct agreement coefficients often subsumed under the name "kappa". Recently, Di Eugenio and Glass (2004) have pointed out that different members of this family make different assumptions about, among other things, individual annotator bias: some coefficients treat this bias as noise in the data (e.g. π, Scott, 1955), while others treat it as a genuine source of disagreement (e.g. κ, Cohen, 1960). Di Eugenio and Glass demonstrate, using examples with two annotators, that the choice of agreement coefficient can affect the reliability values.

In this paper we use the difference between the two classes of coefficients in order to quantify individual annotator bias. We then show that this measure decreases in proportion to the number of annotators. Of course, multiple annotators may still vary in their individual preferences. However, as the number of annotators grows, the effect of this variation as a source of disagreement decreases, and it becomes more similar to random noise.

While the results of this study are purely mathematical, they have also been tested in the field: we conducted a study of the reliability of coreference annotation using 18 subjects (the largest such study we know of), and we found that the differences between biased and unbiased agreement coefficients were orders of magnitude smaller than any of the other variables that affected reliability values. This shows that using many annotators is one way to overcome individual biases in corpus annotation.

13.2 Agreement among two coders: pi and kappa

We start with a simple case, of two annotators who have to classify a set of items into two categories. As a concrete example, we will call our annotators Alice and Bill, call the categories "yes" and "no", and assume they classified ten items with the following results.

Alice:	Y Y N Y N Y N N Y Y
Bill:	Y Y N N Y Y Y N Y Y

Since Alice and Bill agree on the classification of seven of the ten items, we say that their observed agreement is 7/10 or 0.7. Generally, when two annotators classify a set of items into any number of distinct and mutually exclusive categories, their observed agreement is simply the

proportion of items on whose classification they agree.

Observed agreement in itself is a poor measure of inter-annotator reliability, because a certain amount of agreement is expected purely by chance; this amount varies depending on the number of categories and the distribution of items among categories. For this reason it is customary to report an agreement coefficient in which the observed agreement A_o is discounted by the amount of agreement expected by chance A_e. Two such coefficients, suitable for judging agreement between just two annotators, are π (Scott, 1955) and κ (Cohen, 1960); both are calculated according to the following formula.

$$\pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

The difference between π and κ is in the way the expected agreement is calculated. Both coefficients define expected agreement as the probability that the two annotators will classify an arbitrary item into the same category. But while π assumes that this probability is governed by a single distribution, κ assumes that each annotator has a separate probability distribution.

Let's see what this means in our toy example. According to π, we calculate a single probability distribution by looking at the totality of judgments: there are 13 "yes" judgments and 7 "no" judgments, so the probability of a "yes" judgment is 0.65 while that of a "no" judgment is 0.35; overall, the probability that the two annotators will classify an arbitrary item into the same category is $0.65^2 + 0.35^2 = 0.545$. According to κ, we calculate a separate probability distribution for each coder: for Alice the probability of a "yes" judgment is 0.6 and that of a "no" judgment is 0.4, while for Bill the probability of a "yes" judgment is 0.7 and that of a "no" judgment is 0.3; the overall probability that the two annotators will classify an arbitrary item into the same category is $0.6 \cdot 0.7 + 0.4 \cdot 0.3 = 0.54$, slightly lower than the probability calculated by π. This, in turn, makes the value of κ slightly higher than π.

$$\pi = \frac{0.7 - 0.545}{1 - 0.545} \approx 0.341 \qquad \kappa = \frac{0.7 - 0.54}{1 - 0.54} \approx 0.348$$

More generally, for π we use $P(k)$, the overall probability of assigning an item to category k, which is the total number of such assignments by both coders $\mathbf{n}_k$ divided by the overall number of assignments, which is twice the number of items $\mathbf{i}$. For κ we use $P(k|c)$, the probability of assigning an item to category k by coder c, which is the number of such assignments $\mathbf{n}_{ck}$ divided by the number of items $\mathbf{i}$.

$$P(k) = \frac{1}{2\mathbf{i}}\,\mathbf{n}_k \qquad P(k|c) = \frac{1}{\mathbf{i}}\,\mathbf{n}_{ck}$$

According to π, the probability that both coders assign an item to a particular category $k \in K$ is $\mathrm{P}(k)^2$, so the expected agreement is the sum of $\mathrm{P}(k)^2$ over all categories $k \in K$. As for κ, the probability that the two coders c_1 and c_2 assign an item to a particular category $k \in K$ is $\mathrm{P}(k|c_1)\mathrm{P}(k|c_2)$, so the expected agreement is the sum of $\mathrm{P}(k|c_1)\mathrm{P}(k|c_2)$ over all categories $k \in K$.

$$\mathrm{A}_\mathrm{e}^\pi = \sum_{k \in K} \mathrm{P}(k)^2 \qquad \mathrm{A}_\mathrm{e}^\kappa = \sum_{k \in K} \mathrm{P}(k|c_1)\mathrm{P}(k|c_2)$$

Since $\mathrm{P}(k)$ is the mean of $\mathrm{P}(k|c_1)$ and $\mathrm{P}(k|c_2)$ for each category $k \in K$, it follows that for any set of coding data, $\mathrm{A}_\mathrm{e}^\pi \geq \mathrm{A}_\mathrm{e}^\kappa$, and consequently $\pi \leq \kappa$, with the limiting case obtaining when the distributions of the two coders are identical.

13.3 Measuring the bias

Di Eugenio and Glass (2004) point out that π and κ reflect two different conceptualizations of the reliability problem (they refer to π and κ by the names $\kappa_{\mathrm{S\&C}}$ and κ_{Co}, respectively). For π, differences between the coders in the observed distributions of judgments are considered to be noise in the data, whereas for κ they reflect the relative biases of the individual coders, which is one of the sources of disagreement (Cohen, 1960, 40–41). Here we will show how this difference can be quantified and related to an independent measure—the variance of the individual coders' distributions.

We should note that a single coder's bias cannot be measured in and of itself—it can only be measured by comparing the coder's distribution of judgments to some other distribution. Our agreement coefficients do not include reference to any source external to the coding data (such as information about the distribution of categories in the real world), and therefore we cannot measure the bias of an individual coder, but only the bias of the coders with respect to each other.

We are aware of several proposals in the literature for measuring individual coder bias. Zwick (1988) proposes a modified χ^2 test (Stuart, 1955), and Byrt et al. (1993) define a "Bias Index" which is the difference between the individual coders' proportions for one category label (this only applies when there are exactly two categories). Since we are interested in the effect of individual coder bias on the agreement coefficients, we define B, the overall bias in a particular set of coding data, as the difference between the expected agreement according to π

and the expected agreement according to κ.

$$\begin{aligned}
\mathrm{B} = \mathrm{A}_\mathrm{e}^\pi - \mathrm{A}_\mathrm{e}^\kappa &= \sum_{k \in K} \mathrm{P}(k)^2 - \sum_{k \in K} \mathrm{P}(k|c_1)\mathrm{P}(k|c_2) \\
&= \sum_{k \in K} \left(\frac{\mathrm{P}(k|c_1) + \mathrm{P}(k|c_2)}{2} \right)^2 - \mathrm{P}(k|c_1)\mathrm{P}(k|c_2) \\
&= \sum_{k \in K} \left(\frac{\mathrm{P}(k|c_1) - \mathrm{P}(k|c_2)}{2} \right)^2
\end{aligned}$$

The bias is a measure of variance. Take c to be a random variable, with equal probabilities for each of the two coders: $\mathrm{P}(c_1) = \mathrm{P}(c_2) = 0.5$. For each category $k \in K$, we calculate the mean μ and variance σ^2 of $\mathrm{P}(k|c)$.

$$\begin{aligned}
\mu_{\mathrm{P}(k|c)} &= \frac{\mathrm{P}(k|c_1) + \mathrm{P}(k|c_2)}{2} \\
\sigma^2_{\mathrm{P}(k|c)} &= \frac{(\mathrm{P}(k|c_1) - \mu_{\mathrm{P}(k|c)})^2 + (\mathrm{P}(k|c_2) - \mu_{\mathrm{P}(k|c)})^2}{2} \\
&= \left(\frac{\mathrm{P}(k|c_1) - \mathrm{P}(k|c_2)}{2} \right)^2
\end{aligned}$$

We find that the bias B is the sum of the variances of $\mathrm{P}(k|c)$ for all categories $k \in K$.

$$\mathrm{B} = \sum_{k \in K} \sigma^2_{\mathrm{P}(k|c)}$$

This is a convenient way to quantify the relative bias of two coders. In the next section we generalize π and κ to apply to multiple coders, and see that the bias drops in proportion to the number of coders.

13.4 Agreement among multiple coders

We now provide generalizations of π and κ which are applicable when the number of coders $\mathbf{c}$ is greater than two. The generalization of π is the same as the coefficient which is called, quite confusingly, κ by Fleiss (1971). We will call it π because it treats individual coder bias as noise in the data and is thus better thought of as a generalization of Scott's π, reserving the name κ for a proper generalization of Cohen's κ which takes bias as a source of disagreement. As far as we are aware, ours is the first generalization of κ to multiple coders—other sources which claim to give a generalization of κ actually report Fleiss's coefficient (e.g. Bartko and Carpenter, 1976, Siegel and Castellan, 1988, Di Eugenio

and Glass, 2004).

With more than two coders we can no longer define the observed agreement as the percentage of items on which there is agreement, since there will inevitably be items on which some coders agree amongst themselves while others disagree. The amount of agreement on a particular item is therefore defined as the proportion of agreeing judgment pairs out of the total number of judgment pairs for the item. Let $\mathbf{n}_{ik}$ stand for the number of times an item i is classified in category k (i.e. the number of coders that make such a judgment). Each category k contributes $\binom{\mathbf{n}_{ik}}{2}$ pairs of agreeing judgments for item i; the amount of agreement agr_i for item i is the sum of $\binom{\mathbf{n}_{ik}}{2}$ over all categories $k \in K$, divided by $\binom{\mathbf{c}}{2}$, the total number of judgment pairs per item.

$$\mathrm{agr}_i = \frac{1}{\binom{\mathbf{c}}{2}} \sum_{k \in K} \binom{\mathbf{n}_{ik}}{2} = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

The overall observed agreement is the mean of agr_i for all items $i \in I$.

$$\mathrm{A_o} = \frac{1}{\mathbf{i}} \sum_{i \in I} \mathrm{agr}_i = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_{i \in I} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

Since agreement is measured as the proportion of agreeing judgment pairs, the agreement expected by chance is the probability that any given pair of judgments for the same item would agree; this, in turn, is equivalent to the probability that two arbitrary coders would make the same judgment for a particular item by chance. For π we use $\mathrm{P}(k)$, the overall probability of assigning an item to category k, which is the total number of such assignments by all coders $\mathbf{n}_k$ divided by the overall number of assignments, which is the number of items $\mathbf{i}$ multiplied by the number of coders $\mathbf{c}$. For κ we use $\mathrm{P}(k|c)$, the probability of assigning an item to category k by coder c, which is the number of such assignments $\mathbf{n}_{ck}$ divided by the number of items $\mathbf{i}$.

$$\mathrm{P}(k) = \frac{1}{\mathbf{ic}} \mathbf{n}_k \qquad \mathrm{P}(k|c) = \frac{1}{\mathbf{i}} \mathbf{n}_{ck}$$

According to π, the probability that two arbitrary coders assign an item to a particular category $k \in K$ is $\mathrm{P}(k)^2$, so the expected agreement is the sum of $\mathrm{P}(k)^2$ over all categories $k \in K$. As for κ, the probability that two particular coders c_m and c_n assign an item to category $k \in K$ is $\mathrm{P}(k|c_m)\mathrm{P}(k|c_n)$; since all coders judge all items, the probability that an arbitrary pair of coders assign an item to category k is the arithmetic mean of $\mathrm{P}(k|c_m)\mathrm{P}(k|c_n)$ over all coder pairs c_m, c_n, and the expected

agreement is the sum of this probability over all categories $k \in K$.

$$A_e^{\pi} = \sum_{k \in K} P(k)^2 \qquad A_e^{\kappa} = \sum_{k \in K} \frac{1}{\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} P(k|c_m)P(k|c_n)$$

It is easy to see that A_e^{κ} for multiple coders is the mean of the two-coder A_e^{κ} values from section 13.2 for all coder pairs.

We start with a numerical example. Instead of two annotators we now have four; furthermore, it so happens that Claire gives exactly the same judgments as Alice, and Dave gives exactly the same judgments as Bill.

Alice, Claire:	Y Y N Y N Y N N Y Y
Bill, Dave:	Y Y N N Y Y Y N Y Y

The expected agreement according to π remains 0.545 as in the case of just Alice and Bill, since the overall proportion of "yes" judgments is still 0.65 and that of "no" judgments is still 0.35. But for the calculation of expected agreement according to κ we also have to take into account the expected agreement between Alice and Claire and the expected agreement between Bill and Dave. Overall, the probability that two arbitrary annotators will classify an item into the same category is $\frac{1}{6}[0.6^2 + 4 \cdot 0.6 \cdot 0.7 + 0.7^2] + \frac{1}{6}[0.4^2 + 4 \cdot 0.4 \cdot 0.3 + 0.3^2] = 0.54333\ldots$; this value is still lower than the probability calculated by π, but higher than it was for two annotators. If we add a fifth annotator with the same judgments as Alice and Claire and a sixth with the judgment pattern of Bill and Dave, expected agreement according to π remains 0.545 while expected agreement according to κ rises to 0.544. It appears, then, that as the number of annotators increases, the value of A_e^{κ} approaches that of A_e^{π}. We now turn to the formal proof.

We start by taking the formulas for expected agreement above and putting them into a form that is more useful for comparison with one another.

$$\begin{aligned}
A_e^{\pi} &= \sum_{k \in K} P(k)^2 = \sum_{k \in K} \left(\frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} P(k|c_m) \right)^2 \\
&= \sum_{k \in K} \frac{1}{\mathbf{c}^2} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} P(k|c_m)P(k|c_n) \\
A_e^{\kappa} &= \sum_{k \in K} \frac{1}{\binom{\mathbf{c}}{2}} \sum_{m=1}^{\mathbf{c}-1} \sum_{n=m+1}^{\mathbf{c}} P(k|c_m)P(k|c_n) \\
&= \sum_{k \in K} \frac{1}{\mathbf{c}(\mathbf{c}-1)} \left(\sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} P(k|c_m)P(k|c_n) - \sum_{m=1}^{\mathbf{c}} P(k|c_m)^2 \right)
\end{aligned}$$

The overall bias is the difference between the expected agreement according to π and the expected agreement according to κ.

$$\begin{aligned} \mathrm{B} &= \mathrm{A}_\mathrm{e}^\pi - \mathrm{A}_\mathrm{e}^\kappa \\ &= \frac{1}{\mathbf{c}-1} \sum_{k \in K} \frac{1}{\mathbf{c}^2} \left(\mathbf{c} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m)^2 - \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \mathrm{P}(k|c_m)\mathrm{P}(k|c_n) \right) \end{aligned}$$

We now calculate the mean μ and variance σ^2 of $\mathrm{P}(k|c)$, taking c to be a random variable with equal probabilities for all of the coders: $\mathrm{P}(c) = \frac{1}{\mathbf{c}}$ for all coders $c \in C$.

$$\mu_{\mathrm{P}(k|c)} = \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m)$$

$$\begin{aligned} \sigma^2_{\mathrm{P}(k|c)} &= \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} (\mathrm{P}(k|c_m) - \mu_{\mathrm{P}(k|c)})^2 \\ &= \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m)^2 - 2\mu_{\mathrm{P}(k|c)} \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m) + \mu^2_{\mathrm{P}(k|c)} \frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} 1 \\ &= \left(\frac{1}{\mathbf{c}} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m)^2 \right) - \mu^2_{\mathrm{P}(k|c)} \\ &= \frac{1}{\mathbf{c}^2} \left(\mathbf{c} \sum_{m=1}^{\mathbf{c}} \mathrm{P}(k|c_m)^2 - \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \mathrm{P}(k|c_m)\mathrm{P}(k|c_n) \right) \end{aligned}$$

The bias B is thus the sum of the variances of $\mathrm{P}(k|c)$ for all categories $k \in K$, divided by the number of coders less one.

$$\mathrm{B} = \frac{1}{\mathbf{c}-1} \sum_{k \in K} \sigma^2_{\mathrm{P}(k|c)}$$

Since the variance does not increase in proportion to the number of coders, we find that the more coders we have, the lower the bias; at the limit, κ approaches π as the number of coders approaches infinity.

13.5 Conclusion

We have seen that one source of disagreement among annotators, individual bias, decreases as the number of annotators increases. This does not mean that reliability increases with the number of annotators, but rather that the individual coders' preferences become more similar to random noise. This suggests using multiple annotators as a means for controlling bias.

There is a further class of agreement coefficients which allow for gradient disagreements between annotators, for example weighted kappa κ_w (Cohen, 1968) and α (Krippendorff, 1980). Passonneau (2004), for example, uses α to measure reliability of coreference annotation, where different annotators may partially agree on the identity of an anaphoric chain. We cannot treat these coefficients here due to space limitations, but the same result holds for gradient coefficients—bias decreases in proportion to the number of annotators. We performed an experiment testing the reliability of coreference annotation among 18 naive subjects, using α and related measures (Poesio and Artstein, 2005); we found that the effect of bias on the agreement coefficients was substantially lower than any of the other variables that affected reliability.

References

Bartko, John J. and William T. Carpenter, Jr. 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease* 163(5):307–317.

Byrt, Ted, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46(5):423–429.

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

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4):213–220.

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics* 30(1):95–101.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*, chap. 12, pages 129–154. Beverly Hills: Sage.

Passonneau, Rebecca J. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*. Lisbon.

Poesio, Massimo and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*. Ann Arbor.

Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19(3):321–325.

Siegel, Sidney and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, chap. 9.8, pages 284–291. New York: McGraw-Hill, 2nd edn.

Stuart, Alan. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42(3/4):412–416.

Zwick, Rebecca. 1988. Another look at interrater agreement. *Psychological Bulletin* 103(3):374–378.