Estimating precision and recall

The estimation requires four numbers that are easily obtained from the results of the matching algorithm. Let the total number of event-article pairs that we compared within a given time window be denoted as N, which we split into pairs where the news is published before the event N_{before} and news published after the event N_{after}. After applying a similarity threshold, we then count the discrete matches M, which we also split into matches before the event M_{before} and matches after the event M_{after}.

Our first step is to calculate the false positive rate (FPR), which is the number of false positives divided by the number of cases that is actually negative (i.e. event-article pairs where the article did not cover the event).

$$FPR = \frac{\sum false\, positive}{\sum actual\, negative}$$

An article cannot cover an event that has not yet happened. Therefore, all N_{before} are certain to be *actual negative*, and all M_{before} are certain to be $false\, positive$. This means that we can calculate the FPR for event-article pairs where the article was published before the event.

$$FPR_{before} = \frac{M_{before}}{N_{before}}$$

Here we introduce the assumption that the FPR is approximately the same for event-article pairs where the article was published after the event. Our reasoning is that in the case of false positives the article does not cover the event, so whether the article was published before or after the event is irrelevant.

$$FPR \approx FPR_{before}$$

Given the false positive rate and the total number of pairs N_{after}, we can then estimate the number of false positives $\widehat{FP}$. This is an overestimation, because N_{after} can include *actual positives*. If *actual positives* are extremely rare, as in our data, this is neglectable.

$$\widehat{FP} = FPR \times N_{after}$$

Then, given the number of matches M_{after} we can subtract $\widehat{FP}$ to get the estimated number of true positives $\widehat{TP}$.

$$\widehat{TP} = M_{after} - \widehat{FP}$$

And now we can calculate the estimated precision.

$$\widehat{P} = \frac{\widehat{TP}}{\widehat{TP} + \widehat{FP}}$$

To estimate the recall R we rely on the accuracy of estimating $\widehat{TP}$ for data with a low similarity threshold. The recall is the number of true positives divided by the number of cases that is actually positive (i.e. event-article pairs where the article did cover the event).

$$R = \frac{\sum true\,positive}{\sum actual\,positive}$$

If recall is 1 (100%), then $\sum actual\,positive$ has to be identical to the number of true positives. Accordingly, to get the $\sum actual\,positive$, we can use our estimation of $\widehat{TP}$ for data of which we are confident that the recall is close to 1.

$$R = \frac{\widehat{TP}}{\widehat{TP}_{recall \approx 1}}$$

In our case, this means using a very low similarity threshold. If the threshold is zero (all events connected to all news articles) the recall is certain to be 1, but then our estimation of $\widehat{TP}$ would not work. In practice, we thus need to use a low threshold, such as used in the first bar chart in Figure 2. We used the lowest threshold for which our estimated precision is at least 5%.[1] For our data this was 1.28.

[1] We verified that using values between 1% to 20% made little difference, and did not affect our findings.