Gauging the Quality of Relevance Assessments using Inter-Rater Agreement

Tadele T. Damessie
RMIT University
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Thao P. Nghiem
RMIT University
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

1 INTRODUCTION
Gauging relevance judgments using humans is a key component in building Information Retrieval test collections. However, human interpretation of “relevance” is an inherently subjective process [11]. According to Tang and Solomon [16], judging relevance is a dynamic, multidimensional process likely to vary between assessors, and sometimes even with a single assessor at different stages of the process. For example, Scholer et al. [13] found that 19% of duplicate document pairings were judged inconsistently in the TREC-7 and TREC-8 test collections. Understanding the factors that lead to such variation in relevance assessments is crucial to reliable test collection development.

To address this issue, Bailey et al. [3] proposed three classes of judges — gold, silver and bronze — based on the expertise of the assessor. Gold judges are topic originators as well as subject experts; whereas silver judges are subject experts but not topic originators. Bronze judges are neither topic originators nor subject experts. But are all judges in a single class really the same? Secondary assessors who are neither topic creators nor experts are all bronze assessors, but there are in fact many different types of assessors who fall into this class. As assessment at the bronze level is now becoming a common practice in IR, in particular with the growing popularity of crowdsourcing, we set up an experiment to investigate the homogeneity of assessment quality using three different variants of bronze judges. The classes used in this study are:

- Lab: This group of assessors carried out a relevance assessment task in a monitored lab environment, with a requirement to assess a pre-determined number of 30 documents in relation to a single search topic.
- CF-Topic: This group of assessors are an exact replica of the Lab group task except that the task was administered using the CrowdFlower crowdsourcing platform.
- CF-Document: This group of assessors performed the task using CrowdFlower just as the CF-Topic group, but unlike the other two groups, each participant could judge as few (minimum 1) or as many (maximum 30) documents as they liked for a topic.

Our main research question can formally be stated as:

Research Question: Are there differences in the quality of relevance judgments gathered from different sub-classes of bronze-level judges?
Number of documents judged

| Assessor Reliability. The pairwise overall average reliability score of the Lab, CF-Topic and CF-Document groups, measured using Krippendorff’s α, Cohen’s κ] is [0.687, 0.581], [0.407, 0.236] and [0.561, 0.522] respectively. The κ scores are calculated on binary foldings of the 4-level graded relevance levels – non-relevant (0), marginally relevant (1), relevant (2) and highly relevant (3). The marginally relevant (1) and non-relevant (0) judgments are binary, and the others as relevant as recommended by Scholer and Turpin [12].

The results in Table 1 indicate Lab and CF-Document assessors are more reliable than CF-Topic assessors. The statistical significance of the differences is evaluated using an unpaired two-tailed t-test across the individual pairwise agreement scores, and reported in Table 2. For both α and κ, the overall pattern from highest to lowest reliability score measured using the Sormunen judgments.

4 RESULTS AND DISCUSSION
Assessor reliability – measured by the mean pairwise agreement between each assessor and the Sormunen gold standard assessments – is used to assess the quality of the assessments from each experimental group. This analysis is then compared with a measure of assessment quality using only inter-rater agreement, in the absence of any ground truth.

Assessor Reliability. The pairwise overall average reliability score of the Lab, CF-Topic and CF-Document groups, measured using Krippendorff’s α, Cohen’s κ] is [0.687, 0.581], [0.407, 0.236] and [0.561, 0.522] respectively. The κ scores are calculated on binary foldings of the 4-level graded relevance levels – non-relevant (0), marginally relevant (1), relevant (2) and highly relevant (3). The marginally relevant (1) and non-relevant (0) judgments are binary, and the others as relevant as recommended by Scholer and Turpin [12].

The results in Table 1 indicate Lab and CF-Document assessors are more reliable than CF-Topic assessors. The statistical significance of the differences is evaluated using an unpaired two-tailed t-test across the individual pairwise agreement scores, and reported in Table 2. For both α and κ, the overall pattern from highest to lowest reliability score measured using the Sormunen judgments.
as a baseline is: \textit{Lab, CF-Document} and \textit{CF-Topic} respectively. One explanation for this trend might be that the \textit{Lab} study is a more directed environment, and assessors know that they are being closely monitored the entire time. This could contribute to longer periods of focus, resulting in a higher overall agreement with the gold standard, and therefore a presumed higher overall quality of obtained judgments.

When comparing only the two crowdsourcing groups, the \textit{CF-Document} assessors show higher reliability. This is somewhat surprising result, since the judges assess fewer documents and therefore spend less time overall forming a notion of relevance for a particular topic. However, this lack of "domain knowledge" might be counteracted by task completion time: an assessor in \textit{CF-Topic} had to judge all 30 documents to get paid, and when an assessor encounters long or difficult documents at the tail of an assessment list, the likely outcome is that the assessor becomes less motivated to get any single judgment exactly right. Fatigue and motivation are known to influence relevance judgment outcomes [9, 19], and perhaps contribute to the drop in quality. In contrast, \textit{CF-Document} assessors may perceive that less effort is required on their behalf to judge a single topic-document pair before getting paid. These "micro" transactions could very well be a strong motivator for crowdsourced assessors, despite having an implicit startup cost in understanding the task at hand that is amortized when judging multiple documents for the same topic. We plan to study this phenomenon in more detail in future work.

Figure 2 and Figure 3 give further insight on the reliability levels (agreement with the gold standard) of individual \textit{CF-Topic} and \textit{CF-Document} assessors, respectively. Results for the \textit{Lab} group were omitted due to space limitations; the reliability score for this group was consistently well above \( \alpha > 0.2 \), with no negative scores for any assessor. A number of assessors in \textit{CF-Topic} showed lower levels of agreement with the gold standard than expected by chance alone for 2 of the topics as shown in Figure 2. Reliability for the other 2 topics in this group is similar to the trend observed for the \textit{Lab} assessors. Only one assessor’s relative performance in the \textit{CF-Document} setup deviated significantly from the others, as shown in Figure 3. We plan to further investigate the reasons for why such low reliability scores were observed for some individual assessors in these groups in followup work. Note that all of these assessors passed manual sanity control measures, and appeared to be performing judgments in good faith.

**Agreement**. As can be seen in Table 3, overall agreement is higher in \textit{Lab}, followed by \textit{CF-Document} and \textit{CF-Topic}, which are in the same relative order as the reliability scores when comparing against a gold standard, suggesting that inter-rater reliability is a reasonable proxy for the quality of judgments.

To further establish our belief of assessor reliability, we computed the median of the multiple assessments made for each document in each experimental group, and computed the Krippendorff’s Alpha (\( \alpha \)) agreement between individual assessors and this score, shown in Table 3 (right). The overall trend is again consistent with the findings of Table 1.

**Table 1**: Average pairwise agreement between judges and Sormunen gold standard judgments, measured across All and individual topics using Krippendorff’s Alpha (\( \alpha \)) on a 4-levels of ordinal scale and Cohen’s Kappa(\( \kappa \)) on a binary scale.

<table>
<thead>
<tr>
<th></th>
<th>Krippendorff’s Alpha (( \alpha ))</th>
<th>Cohen’s Kappa(( \kappa ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab</td>
<td>CF-Topic</td>
</tr>
<tr>
<td>All</td>
<td>0.687</td>
<td>0.407</td>
</tr>
<tr>
<td>el nino</td>
<td>0.843</td>
<td>0.531</td>
</tr>
<tr>
<td>schengen agreement</td>
<td>0.622</td>
<td>0.057</td>
</tr>
<tr>
<td>euro opposition</td>
<td>0.665</td>
<td>0.437</td>
</tr>
<tr>
<td>ship losses</td>
<td>0.617</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**Table 2**: Statistical significance of Table 1 results, evaluated using an unpaired two-tailed t-test for all bronze assessors. Results for Krippendorff’s Alpha (\( \alpha \)) are shown below the diagonal line with ratings on a 4-level ordinal scale, while results for Cohen’s Kappa (\( \kappa \)) are shown above the diagonal line with ratings on a binary scale, flattening 0 and 1 to 0; and 2 and 3 to 1.
Table 3: Inter-rater agreement (left) and majority vote (right) measured between assessors in the Lab, CF-Topic and CF-Document groups using Krippendorff’s alpha (α) across All and individual topics with ratings on a 4-level ordinal scale. The number of assessors for inter-rater agreement is shown in parenthesis next to each α value.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Inter-rater agreement</th>
<th>Majority vote</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab</td>
<td>CF-Topic</td>
</tr>
<tr>
<td>All</td>
<td>0.657 (32)</td>
<td>0.426 (35)</td>
</tr>
<tr>
<td>el nino</td>
<td>0.845 (8)</td>
<td>0.394 (8)</td>
</tr>
<tr>
<td>schengen agreement</td>
<td>0.634 (8)</td>
<td>0.170 (8)</td>
</tr>
<tr>
<td>euro opposition</td>
<td>0.565 (8)</td>
<td>0.464 (9)</td>
</tr>
<tr>
<td>ship losses</td>
<td>0.558 (8)</td>
<td>0.377 (10)</td>
</tr>
</tbody>
</table>

Figure 2: Reliability of CF-Topic assessors when compared with the Sormunen judgments using Krippendorff’s Alpha (α) for the topics: (a) El nino; and (b) Schengen agreement.

Figure 3: Reliability of CF-Document assessors when compared to the Sormunen judgments using Krippendorff’s Alpha (α).

Getting gold standard relevance labels is rarely possible in a live judging scenario, but it is possible to compute inter-rater agreement between assessors, and use this to establish the quality of assessments. Our experiments confirm that using agreement between judges to gauge the quality of relevance judgments collected is indeed one possible approach to controlling the quality of judgments gathered by bronze level assessors.

5 CONCLUSION

This study analyzed the quality of relevance judgments generated in three (of many possible) different sub-classes of bronze assessors, using Krippendorff’s Alpha (α) and Cohen’s Kappa (κ). The results of both metrics confirm the existence of assessment quality differences among the three sub-classes of bronze assessors, warranting further study. Nevertheless, inter-rater agreement can be a reliable tool to benchmark the quality of relevance judgments when gold standard judgments are not readily available.

Acknowledgment. This work was supported by the Australian Research Council’s Discovery Projects Scheme (DP170102231 and DP140101587).

REFERENCES