Research paper at *DESI IV: The ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, June 6, 2011, University of Pittsburgh, Pittsburgh, PA, USA.
http://www.umiacs.umd.edu/~oard/desi4/

Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?

Maura R. Grossman, J.D., Ph.D.
Wachtell, Lipton, Rosen & Katz[1]

Gordon V. Cormack, Ph.D.
University of Waterloo

1 Introduction

In responding to a request for production in civil litigation, the goal is generally to produce, as nearly as practicable, *all* and *only* the non-privileged documents that are *responsive* to the request.[2] *Recall* – the proportion of responsive documents that are produced – and *precision* – the proportion of produced documents that are responsive – quantify how nearly *all* of and o*nly s*uch responsive, non-privileged documents are produced [2, pp 67-68].

The traditional approach to measuring recall and precision consists of constructing a *gold standard* that identifies the set of documents that are responsive to the request. If the gold standard is complete and correct, it is a simple matter to compute recall and precision by comparing the production set to the gold standard. Construction of the gold standard typically relies on human assessment, where a reviewer or team of reviewers examines each document, and codes it as responsive or not [2, pp 73-75].

It is well known that any two reviewers will often disagree as to the responsiveness of particular documents; that is, one will code a document as responsive, while the other will code the same document as non-responsive [1, 3, 5, 8, 9, 10]. Does such disagreement indicate that responsiveness is ill-defined, or does it indicate that reviewers are sometimes mistaken in their assessments? If responsiveness is ill-defined, can there be such a thing as an accurate gold standard, or accurate measurements of recall and precision? Answering this question in the negative might call into question the ability to measure, and thus certify, the accuracy of a response to a production request. If, on the other hand, responsiveness is well-defined, might there be ways to measure and thereby correct for reviewer error, yielding a better gold standard, and therefore, more accurate measurements of recall and precision?

This study provides a qualitative analysis of the cases of disagreement on responsiveness determinations rendered during the course of constructing the gold standard

[1] The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

[2] *See* Fed. R. Civ. P. 26(b) & (g), 34(a), and 37(a)(4).

for the TREC 2009 Legal Track Interactive task ("TREC 2009") [7]. For each disagreement, we examined the document in question, and made our own determination of whether the document was "clearly responsive," "clearly non-responsive," or "arguable," meaning that it could reasonably be construed as either responsive or not, given the production request and operative assessment guidelines.

2 Prediction

Our objective was to test two competing hypotheses:

> **Hypothesis 1:** *Assessor disagreement is largely due to ambiguity or inconsistency in applying the criteria for responsiveness to particular documents.*
>
> **Hypothesis 2:** *Assessor disagreement is largely due to human error.*

Hypothesis 1 and Hypothesis 2 are mutually incompatible; evidence refuting Hypothesis 1 supports Hypothesis 2, and vice versa.

To test the validity of the two hypotheses, we constructed an experiment in which, prior to the experiment, the two hypotheses were used to predict the outcome. An observed result consistent with one hypothesis and inconsistent with the other would provide evidence supporting the former and refuting the latter.

In particular, Hypothesis 1 predicted that if we examined a document about whose responsiveness assessors disagreed, it would generally be difficult to determine whether or not the document was responsive; that is, it would usually be possible to construct a reasonable argument that the document was either responsive or non-responsive. On the other hand, Hypothesis 2 predicted that it would generally be clear whether or not the document was responsive; it would usually be possible to construct a reasonable argument that the document was responsive, or that the document was non-responsive, but not both.

At the outset, we conjectured that the results of our experiment would more likely support Hypothesis 1.

3 TREC Adjudicated Assessments

The TREC 2009 Legal Track Interactive Task used a two-pass adjudicated review process to construct the gold standard [7]. In the first pass, law students or contract attorneys assessed a sample of documents for each of seven production requests – "topics," in TREC parlance – coding each document in the sample as responsive or not. TREC 2009 participants were invited to appeal any of the assessor coding decisions with which they disagreed, and the Topic Authority (or "TA") – a senior lawyer tasked with defining responsiveness – was asked to make a final determination as to whether the appealed document was responsive or not. The gold standard considered a document to be *responsive* if the first-pass assessor coded it as responsive and that decision was not appealed, the first-pass assessor coded it as responsive and that decision was upheld by the Topic Authority, or the first-pass assessor coded it as non-responsive and

Topic	First-Pass Assessment	Assessed	Appealed	Success	% Success
201	Responsive	603	374	363	97%
201	Non-responsive	5,605	123	101	82%
202	Responsive	1,743	167	115	68%
202	Non-responsive	5,462	541	469	86%
203	Responsive	131	74	69	93%
203	Non-responsive	5,296	209	186	88%
204	Responsive	105	59	50	84%
204	Non-responsive	7,024	207	169	81%
205	Responsive	1,631	889	882	99%
205	Non-responsive	4,289	78	50	64%
206	Responsive	235	52	50	96%
206	Non-responsive	6,860	0	0	–
207	Responsive	938	43	23	53%
207	Non-responsive	7,377	154	125	81%
All	Responsive	5,386	1,658	1,552	93%
All	Non-responsive	41,913	1,312	1,100	83%

Table 1: Number of documents assessed, appealed, and the success rates of appeals for the TREC 2009 Legal Track Interactive Task, categorized by topic and first-pass assessment.

that decision was overturned by the Topic Authority. The gold standard considered a document to be *non-responsive* if the first-pass assessor coded it as non-responsive and that decision was not appealed, the first-pass assessor coded it as non-responsive and that decision was upheld by the Topic Authority, or the first-pass assessor coded it as responsive and the decision was overturned by the Topic Authority.

A gold standard was created for each of the seven topics.[3] A total of 49,285 documents – about 7,000 per topic – were assessed for the first-pass review. A total of 2,976 documents (5%) were appealed and therefore adjudicated by the Topic Authority. Of those appeals, 2,652 (89%) were successful; that is, the Topic Authority disagreed with the first-pass assessment 89% of the time. A breakdown of the number of documents appealed per topic, and the outcome of those appeals, appears in Table 1.[4]

4 Post-Hoc Assessment

We performed a qualitative, post-hoc assessment on a sample of the successfully appealed documents from each category represented in Table 1; that is, the documents where the TREC 2009 first-pass assessor and Topic Authority disagreed. Where 50 or more documents were successfully appealed, we selected a random sample of 50.

[3] The gold standard and evaluation tools are available at `http://trec.nist.gov/data/legal09.html`.

[4] The pertinent documents may be identified by comparing files `qrels_doc_pre_all.txt` and `qrels_doc_post_all.txt` in `http://trec.nist.gov/data/legal/09/evalInt09.zip`.

Topic	TA Opinion	TA Correct	Arguable	TA Incorrect
201	Responsive	74%	20%	6%
201	Non-responsive	94%	2%	4%
202	Responsive	96%	2%	2%
202	Non-responsive	96%	0%	4%
203	Responsive	94%	2%	4%
203	Non-responsive	82%	4%	14%
204	Responsive	90%	10%	0%
204	Non-responsive	90%	8%	2%
205	Responsive	100%	0%	0%
205	Non-responsive	82%	4%	14%
206	Responsive	–	–	–
206	Non-responsive	96%	2%	2%
207	Responsive	74%	12%	14%
207	Non-responsive	70%	0%	28%
All	Responsive	88% (84–91%)	8% (5–11%)	4% (2–7%)
All	Non-responsive	89% (85–92%)	3% (2–6%)	8% (5–12%)

Table 2: Post-hoc assessment of documents whose first pass responsiveness assessment was overturned by the Topic Authority in the TREC 2009 Legal Track Interactive Task. The columns indicate the topic number, the TA's assessment, the proportion of documents for which the authors believe the TA was clearly correct, the proportion of documents for which the authors believe the correct assessment is arguable, and the proportion of documents for which the authors believe the TA was clearly incorrect. The final two rows give these proportions over all topics, with 95% binomial confidence intervals.

Doc. Id.	TA Opinion	Post-Hoc Assessment	TA Reconsideration
0.7.47.1151420	Responsive	Arguable	TA Incorrect
0.7.47.1310694	Responsive	Arguable	TA Incorrect
0.7.47.272751	Responsive	TA Incorrect	Arguable
0.7.6.180557	Responsive	Arguable	TA Correct
0.7.6.252211	Responsive	Arguable	TA Incorrect
0.7.47.1082536.1	Non-responsive	Arguable	TA Correct
0.7.47.14687.1	Non-responsive	Arguable	Arguable
0.7.47.758281	Non-responsive	Arguable	TA Correct
0.7.6.707917.2	Non-responsive	Arguable	TA Correct
0.7.6.731168	Non-responsive	Arguable	TA Correct

Table 3: Blind reconsideration of adjudication decisions for Topic 204 by the Topic Authority (Grossman) that were contradicted or deemed arguable by the post-hoc reviewer (Cormack). The columns represent the TREC document identifier for each of the ten documents, the opinion rendered by the TA during the TREC 2009 adjudication process, the opinion rendered by the post-hoc reviewer, and the *de novo* opinion of the same Topic Authority for the purposes of this study.

Where fewer than 50 documents were successfully appealed, we selected all of the appealed documents.

We used the plain-text version of the TREC 2009 Legal Track Interactive Track corpus, downloaded by one of the authors while participating in TREC 2009 [4], and redistributed for use at TREC 2010.[5] One of the authors of this study examined every document, in every sample, and coded each as "responsive," "non-responsive," or "arguable," based on the content of the document, the production request, and the written assessment guidelines composed for TREC 2009 by each Topic Authority. We coded a document as "responsive" if we believed there was no reasonable argument that the document fell outside the definition of responsiveness dictated by the production request and guidelines. Similarly, we coded a document as "non-responsive" if we believed there was no reasonable argument that the document should have been identified as responsive to the production request. Finally, we coded the document as "arguable" if we believed that informed, reasonable people might disagree about whether or not the document met the criteria specified by the production request and guidelines.

Table 2 shows the agreement of our post-hoc assessment with the TREC 2009 Topic Authority's assessment on appeal, categorized by topic and by the TA's assessment of responsiveness. Each row shows the TA opinion (which is necessarily the opposite of the first-pass opinion), the percentage of post-hoc assessments for which we believe that the only reasonable coding was that rendered by the TA, the percentage of post-hoc assessments for which we believe that either coding would be reasonable, and the percentage of post-hoc assessments for which we believe that the only reasonable coding contradicts the one that was made by the TA.

5 Topic Authority Reconsideration

One of the authors (Grossman) was the Topic Authority for Topic 204 at TREC 2009. The other author (Cormack) conducted the post-hoc assessment for Topic 204. The post-hoc assessment clearly disagreed with the Topic Authority in only one case, and was "arguable" in nine other cases. The ten documents were presented to the TA for *de novo* reconsideration, in random order, with no indication as to how they had been previously coded. For this reconsideration effort, the TA used the same three categories as for the post-hoc assessment: "responsive," "non-responsive," or "arguable."[6] Table 3 shows the results of the TA's reconsideration of the ten documents.

6 Document Exemplars

Table 4 lists the production requests for the seven TREC topics. Based on the production request and his or her legal judgement, each Topic Authority prepared a set

[5] Available at `http://plg1.uwaterloo.ca/~gvcormac/treclegal09/`.

[6] Note that when the TA adjudicated documents as part of TREC 2009, she was constrained to the categories of "responsive" and "non-responsive"; there was no category for "arguable" documents. Therefore, we cannot consider a post-hoc determination of "arguable" as necessarily contradicting the TA's original adjudication at TREC 2009.

Topic	Production Request
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
204	All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.
205	All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
206	All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s), or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i) the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv) the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Table 4: Mock production requests ("Topics") composed for the TREC 2009 Legal Track Interactive Task.

```
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

I'm in.  I'll be shredding 'till 11am so I should
haveplenty of time to make it.
```

Figure 1: A clearly responsive document to Topic 204. This document was coded as non-responsive by a contract attorney, although it clearly pertains to document shredding, as specified in the production request.

```
From:   Bass, Eric
Sent:   Thursday, January 17, 2002 11:19 AM
To:     Lenhart, Matthew
Subject:        FFL Dues

You owe $80 for fantasy football.  When can you pay?
```

Figure 2: A clearly responsive document to Topic 207. This document was coded as non-responsive by a contract attorney, although it clearly pertains to fantasy football, as specified in the production request.

of assessment guidelines.[7] We illustrate our post-hoc analysis using exemplar documents that were successfully appealed as responsive to topics 204 and 207. We chose these topics because they were the least technical and, therefore, the most accessible to readers lacking subject-matter expertise.

Figures 1 and 2 provide examples of documents that are clearly responsive to Topics 204 and 207, but were coded as non-responsive by the first-pass assessors. The first document concerns shredding, while the second concerns payment of a Fantasy Football[8] debt. We assert that the only reasonable assessment for both of these documents is "responsive."

Figures 3 and 4, on the other hand, illustrate documents for which the responsiveness to Topics 204 and 207, respectively, is arguable. Reasonable, informed assessors might disagree, or find it difficult to determine, whether or not these documents met the criteria spelled out in the production requests and assessment guidelines.

[7] The guidelines, along with the complaint, production requests, and exemplar documents, may be found at `http://plg1.cs.uwaterloo.ca/trec-assess/`.

[8] "Fantasy football an interactive, virtual competition in which people manage professional football players versus one another." `http://en.wikipedia.org/wiki/Fantasy_football_(American)`.

```
Subject: Original Guarantees
Just a followup note:
We are still unclear as to whether we should continue
to send original  incoming and outgoing guarantees to
Global Contracts (which is what we have  been doing
for about 4 years, since the Corp. Secretary kicked us
out of  using their vault on 48 for originals because
we had too many documents).  I  think it would be
good practice if Legal and Credit sent the originals
to  the same place, so we will be able to find them
when we want them.  So my  question to y'all is, do
you think we should send them to Global Contracts,  to
you, or directly the the 48th floor vault (if they let
us!).
```

Figure 3: A document of arguable responsiveness to Topic 204. This message concerns *where* to store particular documents, not specifically their destruction or retention. Reasonable, informed assessors might disagree as to its responsiveness, based on the TA's conception of relevance.

```
Subject:        RE: How good is Temptation Island 2
They have some cute guy lawyers this year-but I bet you
probably watch that manly Monday night Football.
```

Figure 4: A document of arguable responsiveness to Topic 207. This message mentions football whimsically and in passing, but does not reference a *specific* football team, player, or game. Reasonable, informed assessors might disagree about whether or not it is responsive according to the TA's conception of relevance.

7 Discussion

Our evidence supports the conclusion that responsiveness – at least as characterized by the production requests and assessment guidelines used at TREC 2009 – is fairly well defined, and that disagreements among assessors are largely attributable to human error. As a threshold matter, only 5% of the first-pass assessments were appealed. Since participating teams had the opportunity and incentive to appeal the assessments with which they disagreed, we may assume that, for the most part, they agreed with the first-pass assessments of the documents they chose not to appeal. That is, the first-pass assessments were on the order of 95% accurate. Second, we observe that 89% of the appeals were upheld, suggesting that they had, for the most part, a reasonable basis.

Our study considers only those appealed documents for which the appeals were upheld – about 89% of the appealed documents, or 4.5% of all assessed documents. Are these documents arguably on the borderline of responsiveness, as one might suspect? At the TREC 2009 Workshop, many participants, including the authors, voiced opinions to this effect. An earlier study by the authors preliminarily examined this question and found that, for two topics,[9] the majority of non-responsive assessments that were overturned were the result of human error, rather than questionable responsiveness [6]. The aim of the present study was to further test this hypothesis, by considering the other five topics, and also responsive assessments that were overturned (*i.e.,* adjudicated to be non-responsive). To our surprise, we found that we judged nearly 90% of the overturned documents to be clearly responsive, or clearly non-responsive, in agreement with the Topic Authority. We found another 5% or so of the documents to be clearly responsive or clearly non-responsive, contradicting the Topic Authority. *Only 5% did we find to be arguable*, indicating a borderline or questionable decision. Accordingly, we conclude that the vast majority of disagreements arise due to simple human error; error that can be identified by careful reconsideration of the documents using the production requests and assessment guidelines.

Our results also suggest that the TA assessments, while quite reliable, are not infallible. We confirmed this directly for Topic 204 by having the same TA reconsider ten documents that she had previously assessed as part of TREC 2009. For three of the ten documents, the TA contradicted her earlier assessment; for two of the ten, the TA coded the documents as arguable. For only half of the documents did the TA unequivocally reprise her previous assessment. While we did not have the TAs for the other topics reconsider their assessments, we are confident from our own analysis of the documents that some of their assessments were incorrect.

All in all, the total proportion of documents that are borderline, or for which the adjudication process yielded the wrong result, appears to be quite low. Five percent of the assessed documents were appealed; 90% of those appeals were upheld; and of those, perhaps 10% were borderline – that is, only about 0.45% of the assessed documents were "arguable." It stands to reason that there may be some borderline documents that our study did not consider. In particular, we did not consider documents that the first-pass assessor and the TREC 2009 participants agreed on, and which were therefore not appealed. We also did not consider documents that were appealed, but

[9] Topics 204 and 207, which were chosen because they were the least esoteric of the seven topics.

for which the TA upheld the first-pass assessment. We have little reason to believe that the number of such borderline documents would be large in either case; however, a more extensive study would be necessary to quantify this number. In any event, we are concerned here specifically with the *cause* of assessor disagreement that was observed, and since there is no assessor disagreement on these particular documents, this quantity has no bearing on the hypotheses we were testing.

We characterize our study as qualitative rather than quantitative for several reasons. The documents we examined were not randomly selected from the document collection; they were selected in several phases, each of which identified a disproportionate number of controversial documents:

1. The stratified sampling approach used by TREC 2009 to identify documents for the first-pass assessment emphasized documents for which the participating teams had submitted contradictory results;

2. The appeals process selected from these documents those for which the teams disagreed with the first-pass assessment;

3. For our post-hoc assessment, we considered only appealed documents for which the Topic Authority disagreed with the first-pass assessor; and

4. For our TA reconsideration, we considered only ten percent of the documents from our post-hoc assessment – those for which the post-hoc assessment disagreed with the decision rendered by the TA at TREC 2009.

All of these phases tended to focus on controversial documents, consistent with our purpose of determining whether disagreement arises due to ambiguity concerning responsiveness, or human error. Therefore, it would be inappropriate to use these results to estimate the error rate of either the first-pass assessor or the Topic Authority on the collection as a whole.

Finally, neither of the authors is at arm's length from the TREC 2009 effort; our characterization of responsiveness reflects our informed analysis and as such, is amenable to debate. Accordingly, we invite others in the research community to examine the documents themselves and to let us know their results. Towards this end, we have made publicly available the text rendering of the documents we reviewed for this study.[10]

8 Conclusion

It has been posited by some that it is impossible to derive accurate measures of recall and precision for the results of any document review process because large numbers of documents in the review set are "arguable," meaning that two informed, reasonable reviewers could disagree on whether the documents are responsive or not. The results of our study support the hypothesis that the vast majority of cases of disagreement are a product of human error rather than documents that fall in some "gray area" of responsiveness. Our results also show that while Topic Authorities – like all human

[10] *See* `http://plg1.cs.uwaterloo.ca/~gvcormac/maura1/`.

assessors – make coding errors, adjudication of cases of disagreement in coding using a senior attorney can nonetheless yield a reasonable gold standard that may be improved by systematic correction of the estimated TA error rate.

References

[1] BAILEY, P., CRASWELL, N., SOBOROFF, I., THOMAS, P., DE VRIES, A., AND YILMAZ, E. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), ACM, pp. 667–674.

[2] BÜTTCHER, S., CLARKE, C., AND CORMACK, G. *Information retrieval: Implementing and evaluating search engines*. MIT Press, 2010.

[3] CHU, H. Factors affecting relevance judgment: a report from TREC Legal track. *Journal of Documentation 67*, 2 (2011), 264–278.

[4] CORMACK, G., AND MOJDEH, M. Machine learning for information retrieval: TREC 2009 web, relevance feedback and legal tracks. In *The eighteenth Text REtrieval Conference proceedings (TREC 2009), Gaithersburg, MD* (2009).

[5] EFTHIMIADIS, E., AND HOTCHKISS, M. Legal discovery: Does domain expertise matter? *Proceedings of the American Society for Information Science and Technology 45*, 1 (2008), 1–2.

[6] GROSSMAN, M. R., AND CORMACK, G. V. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology XVII*, 3 (2011).

[7] HEDIN, B., TOMLINSON, S., BARON, J. R., AND OARD, D. W. Overview of the TREC 2009 Legal Track. In *The Eighteenth Text REtrieval Conference (TREC 2009)* (2010). To appear.

[8] ROITBLAT, H. L., KERSHAW, A., AND OOT, P. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology 61* (2010), 70–80.

[9] VOORHEES, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management 36*, 5 (2000), 697–716.

[10] WANG, J., AND SOERGEL, D. A user study of relevance judgments for E-Discovery. *Proceedings of the American Society for Information Science and Technology 47*, 1 (2010), 1–10.