



# Do detailed instructions improve assessor agreement?

William Webber

College of Information Studies  
The University of Maryland

University of North Carolina, March 23rd, 2012





# Outline

## Assessor disagreement

Measuring

Observed

Impact on evaluation

## E-discovery

Introducing e-discovery

TREC Legal Track

## Our experiment

Experimental setup

Experimental results

Conclusions



## Assessor disagreement on relevance

- Two assessors independently assess a document for relevance to a topic.
- Each must classify the document as either relevant or irrelevant to the topic.
- How often do they disagree?



## Agreement matrix

		Assessor B		Total
		1	0	
Assessor A	1	$n_{11}$	$n_{10}$	$n_{1.}$
	0	$n_{01}$	$n_{00}$	$n_{0.}$
Total		$n_{.1}$	$n_{.0}$	$n$

- In fact, there are two types of agreement, and two types of disagreement
- making a 2 by 2 table which I'm going to call an agreement matrix (sometimes called a confusion matrix or a contingency table)
- Several point measures of agreement can be derived from this

## Positive agreement or Mutual F1

$$MF1 = \frac{2 * n_{11}}{n_{1.} + n_{.1}} = \frac{2 * n_{11}}{2 * n_{11} + n_{10} + n_{01}} \quad (1)$$

- Positive agreement
- Same as mutual F1:<sup>1</sup>
  - Make one assessor authoritative
  - Measure F1 score of other assessors “retrieval”
  - Note that this is symmetric (one assessor’s recall is the other’s precision)
- Measures agreement in terms of (upper bound on) retrieval performance

**Not** the same as positive overlap (but monotonically equivalent)

---

<sup>1</sup>Harmonic mean of precision and recall

## Cohen's $\kappa$

$$\Pr(a) = \frac{n_{11} + n_{00}}{n}$$

$$\Pr(e) = \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n} + \frac{n_{0.}}{n} \cdot \frac{n_{.0}}{n}$$

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- Many agreement measures are affected by inherent prevalence of one class or another.
- Cohen's  $\kappa$  measures chance-corrected agreement
- Score of 0 means “agreement expected by chance (given marginal prevalence of classifications)”
- Less immediately interpretable than Mutual F1, but more statistically stable.



## Disagreement among TREC assessors

Assessors	MF1	$\kappa$
Primary & A	0.59	0.45
Primary & B	0.66	0.51
A & B	0.60	0.47

Voorhees (2000)<sup>2</sup>:

- Threefold assessment of documents for TREC 4 AdHoc
- First by primary assessor (topic author); then by two other TREC assessors (authors of other topics)

<sup>2</sup>“Variations in relevance judgments and the measure of retrieval effectiveness”, IPM

## Disagreement among closely collaborating assessors

Assessors	MF1	$\kappa$
Voorhees (2002) P & A	0.59	0.45
P & B	0.66	0.51
A & B	0.60	0.47
<b>Sormunen (2002)</b>	<b>0.83</b>	<b>0.69</b>

Sormunen (2002)<sup>3</sup>:

- 2772 TREC documents rejudged by six Masters students in information science
- Judgment performed over 6 months, with initial trial set and corrections, and regular meetings
- Four-grade relevance assessments; folded to binary for above figures.

<sup>3</sup>“Liberal Relevance Criteria of TREC – Counting on Negligible Documents?”





## Disagreement amongst legally trained assessors

Assessors	MF1	$\kappa$
Voorhees (2002) P & A	0.59	0.45
P & B	0.66	0.51
A & B	0.60	0.47
Sormunen (2002)	0.83	0.69
Roitblat et al. (2010) P vs. A	0.36	0.16
P vs. B	0.35	0.15
A vs. B	0.47	0.24

Roitblat et al. (2010)<sup>4</sup>:

- Original review by team of lawyers in real case
- Re-review performed by two other teams of lawyers at same professional review firm

<sup>4</sup>“Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review”



## Disagreement in relative evaluation

What about comparative evaluation between systems?

- Voorhees (2002) found Kendall's  $\tau$  of 0.94 between system AP scores on different assessment sets.
  - that is, system comparisons are very robust to assessor disagreement
- Not surprising if we expect assessor differences to be uncorrelated with system differences
- . . . though beware assessors who just look for keywords





## Disagreement in absolute evaluation

Assessor disagreement an issue in absolute evaluation:

- Often, assessor is not same as querier
  - Web search engines (use to?) do assessment by sampling queries, have raters recreate intent
- Even if assessor is querier, human agreement may set realistic upper bound to absolute automatic performance
- Conservativeness, liberality (coherence?) of assessor can affect absolute scores





# Outline

## Assessor disagreement

Measuring

Observed

Impact on evaluation

## E-discovery

Introducing e-discovery

TREC Legal Track

## Our experiment

Experimental setup

Experimental results

Conclusions



# E-discovery

## E-discovery:

- Retrieval of relevant documents in civil litigation
- ... in response to production request from (or negotiated with) other side
- ... with documents produced to other side

Strong emphasis upon (demonstrated) comprehensiveness of production.

- We'd like reliable absolute measures of performance





## The topic authority

- Responding side performs production under supervision of senior attorney, who certifies production to the court
- This senior attorney's conception of relevance is authoritative; hence, call them the **topic authority** (TA)
- Disagreement with topic authority is not merely assessor disagreement; it is **assessor error**



## Manual review

Established standard is manual review:

- Documents reviewed for relevance by team of junior attorneys, working under TA's directions
- . . . often after a filtering Boolean query

Disagreement here means not just inaccurate effectiveness evaluation, but producing the wrong documents!





## Disagreement amongst legally trained assessors

Assessors	MF1	$\kappa$
Voorhees (2002) P & A	0.59	0.45
P & B	0.66	0.51
A & B	0.60	0.47
Sormunen (2002)	0.83	0.69
Roitblat et al. (2010) P vs. A	0.36	0.16
P vs. B	0.35	0.15
A vs. B	0.47	0.24

But we know from Roitblat et al. that assessor disagreement in e-discovery can be alarmingly high.







# The TREC Legal Track

## TREC Legal Track:

- Set up to examine e-discovery
- Running since 2006
- Being quoted in precedent-establishing court cases



# Interactive task

Interactive task of the Legal Track:

- Has senior lawyer playing TA role
- Participants interact with TA in developing their runs
- TA instructs assessors through detailed written guidelines





## The appeal process

- Participants appeal erroneous assessments to TA for adjudication
- Post-adjudication assessments are the authoritative ones
- Evidence that, for certain topics in 2009, the appeal process was *reasonably* thorough in finding clear errors





# Outline

## Assessor disagreement

Measuring

Observed

Impact on evaluation

## E-discovery

Introducing e-discovery

TREC Legal Track

## Our experiment

Experimental setup

Experimental results

Conclusions



## Experiment question

Do more detailed instructions lead to higher levels of assessor agreement?

- Between two assessors
- Between an assessor and the conception of relevance of the person writing the instructions (the topic authority)



# Data set

TREC Legal Track, interactive task.

- General instructions taken from topic statement
- Detailed instructions taken from the assessment guidelines
- TA's conception of relevance embodied in post-adjudication relevance assessments.





## Detailed guidelines

### Detailed guideline document:

- Written by TA with extensive experience in e-discovery
- Written after dozens of hours of interacting with teams in developing their runs
- 5 pages in length



# Topic statement

## Topic 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.







## Detailed guidelines: criteria

- 3.1. Relevant Subject Matter. Documents that discuss, or are evidence of, the following activities or subject matter are to be considered relevant for the purposes of this exercise.
  - 3.1.1. Non-routine alteration of documents or evidence
    - 3.1.1.1. Non-routine editing of documents or evidence, in particular, with the purpose of eliminating information.
    - 3.1.1.2. Non-routine removal of document content



## Detailed guidelines: instances

Guidelines also includes instances, examples of relevant documents:

### 4.4. Retaining documents or evidence.

#### 4.4.1. Examples of **Responsive** Content

- *I kept all my files on the share drive and have backed them up on an external drive.*
- *You need to talk to him about the records management systems.*
- *Did we ever look at that document storage facility up near Sacramento?*
- *Subject: Preservation of records*





## Experimental subjects

- Two final-year high school students working as interns
- Worked with browser-based review system
- Documents presented as TIFF images, as with official assessments



## Trial experiment

Performed trial experiment on different topic (Topic 301, from 2010).

- To iron out issues with experimental setup
- To determine sample size
  - Yes, we tried to estimate statistical power!
  - Please be relatively impressed
- Third treatment of consultation between assessors on their conception of relevance. (Not done in full experiment because insufficient relevant documents.)
- Sample size of 40 messages (c. 80 documents) per treatment.



## Trial experiment results

Assessors		$\kappa$ for treatment		
		General	Detailed	Consult
Marjorie	Bryan	0.229	0.275	0.325
Marjorie	Official	* 0.557	0.439	* 0.220
Bryan	Official	0.417	0.294	0.325

- More detailed instructions appear to decrease agreement
- Consultation between assessors doesn't help much
- But no result is significant (though \* is borderline,  $p = 0.053$ )



## Sample size conclusions

Statistical power: probability of finding significance for a given true  $\delta$  (here,  $\delta_{\kappa}$ ).

- We nominated  $\delta_{\kappa} = 0.23$ 
  - change from second tercile to first tercile agreement between 2009 assessors and TA
  - Happens to be roughly the gap between Roitblat et al. (2010) ( $\bar{\kappa} \approx 0.39$ ) and Voorhees (2002) ( $\bar{\kappa} \approx 0.62$ ); and again between Voorhees and Sormunen (2002) ( $\bar{\kappa} = 0.83$ )
- Assuming a variance-minimizing prevalence of 0.5 (which we can enforce in selection of documents)
- ... we need a sample size of 215 documents per treatment needed to achieve power  $1 - \beta = 0.8$  for significance level  $\alpha = 0.05$ .





## Full experiment setup

- 160 messages, roughly 234 documents per treatment.
- Stratified sample, 50/50 relevant/irrelevant, mostly appealed documents
- No consultation stage (not enough relevant documents)
- Instead, joint-rereview of batches of first two treatments at end.



## Full experiment results

Assessors		$\kappa$ for treatment			
		General	Detailed	Joint G.	Joint D.
Marjorie	Bryan	0.519	0.528	<i>0.992</i>	<i>0.950</i>
Marjorie	Official	0.454 <sup>ab</sup>	0.555	0.677 <sup>a</sup>	0.665 <sup>b</sup>
Bryan	Official	0.710	0.637	0.686	0.674

- No significant increase (or even clear positive trend) in agreement with detailed instructions
- Joint assessment did lead to significant improvement ( $p < 0.01$ ) for one assessor ...
- but that may be because that assessor dragged to other's conception







## High school students vs lawyers

- Across all documents, the official assessors (professional reviewers with legal training) achieved  $\kappa = 0.320$  with TA.
- Looking only at experimental documents (80% appealed, presumably difficult to assess), our high school students achieved  $\kappa = 0.555$  and  $\kappa = 0.637$  with TA.
- Prejudice in legal community that manual reviewer only performable with legal training confounded.



## Caveats

- While we can be fairly confident that more detailed instructions did not lead to major improvement in agreement
  - on this particular topic
  - with these particular instructions
  - and these particular assessors
- . . . we don't know how well this generalizes to other topics, other assessors



# Conclusions

Nevertheless:

- Experiment confounded the common-sense expectation (and our hypothesis) that greater details lead to better agreement
- Why?
  - Inability to specify a conception of relevance in writing?
  - Incapacity of human mind to hold too many instructions?



# Impact

- Manual review being challenged in market by automated methods (basically text classification)
- Automated methods just this month achieved court recognition
- Our results strengthen belief that delegated manual review is irreparably unreliable
  - ... though we haven't considered active monitoring
- A human may be better able to communicate conception of relevance to an algorithm by training examples, than to another human by instructions

Done!

