Better than their Reputation?
On the Reliability of Relevance Assessments with Students

Philipp Schaer
philipp.schaer@gesis.org

CLEF 2012, 2012-09-17

Disagreement in Relevance Assessments

Over the last three years we evaluated three retrieval systems. More than 180 LIS students participated by doing relevance assessments.

- How reliable (and therefore: good) are the relevance assessments of our students?
- Can the quality and reliability be safely quantified and with what methods?
- What effects would data cleaning bring up when we drop unreliable assessments?

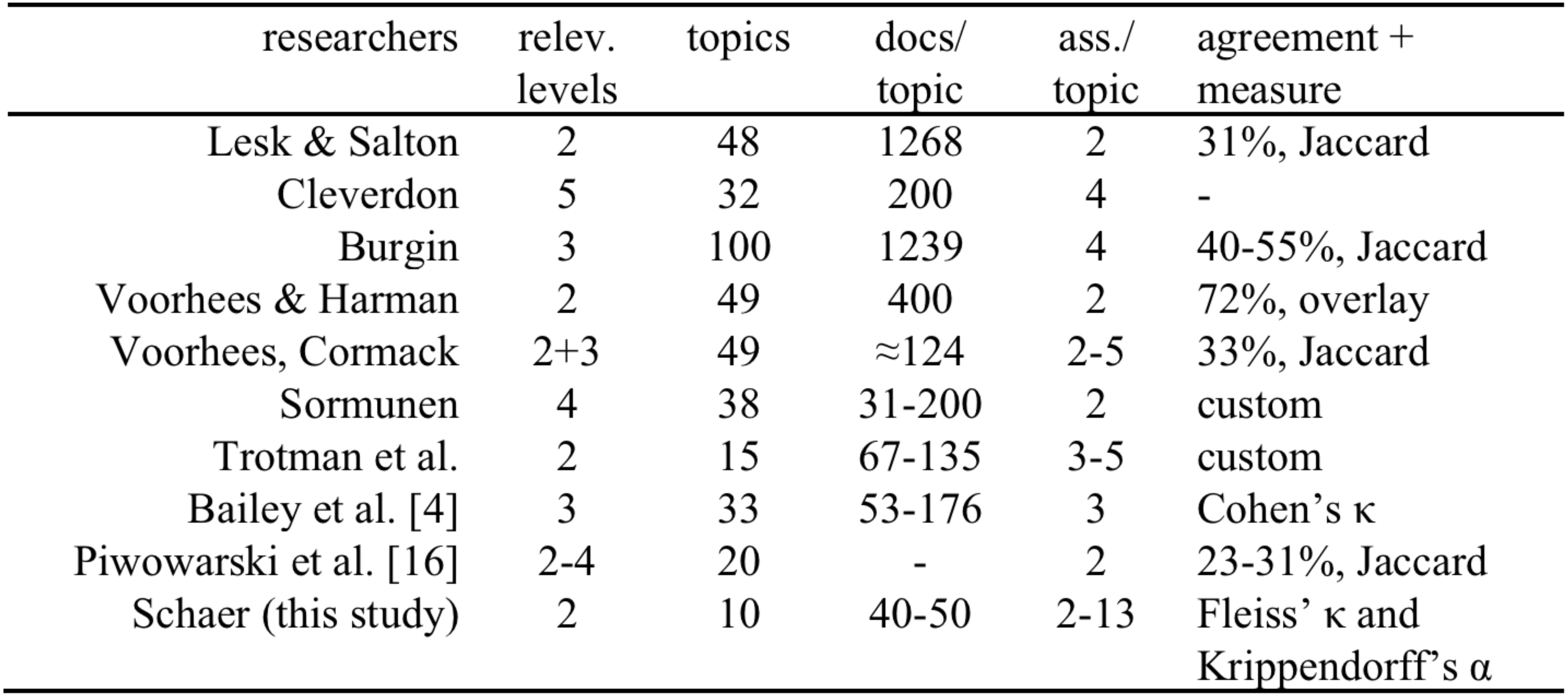
Overall question: What about the bad reputation of relevance assessments studies done with students/colleagues/laymen/turkers … ?

How to measure Inter-Assessor Agreement

- Simple percentage agreement and Jaccard' coefficient (intersection/union)
  - Used in early TREC studies
  - Misleading and unstable to number of topics, documents/topic, assessor/topic …
- Cohen's Kappa, Fleiss's Kappa
  - Described in IR standard literature (Manning et al.), but rarely used in IR
  - Statistical rate of agreement that exceeds random ratings
  - Cohen's Kappa can only compare two assessors, Fleiss's Kappa more than two
- Krippendorff's Alpha
  - Uncommon in IR, but used in opinion retrieval or computational linguistics
  - More robust against imperfect and incomplete data, no. of assessors and values

All approaches return a value (usually between -1, 0, and 1) that is hard to interpret. As Krippendorff (2006) pointed out: "There are no magical numbers".

Literature Review



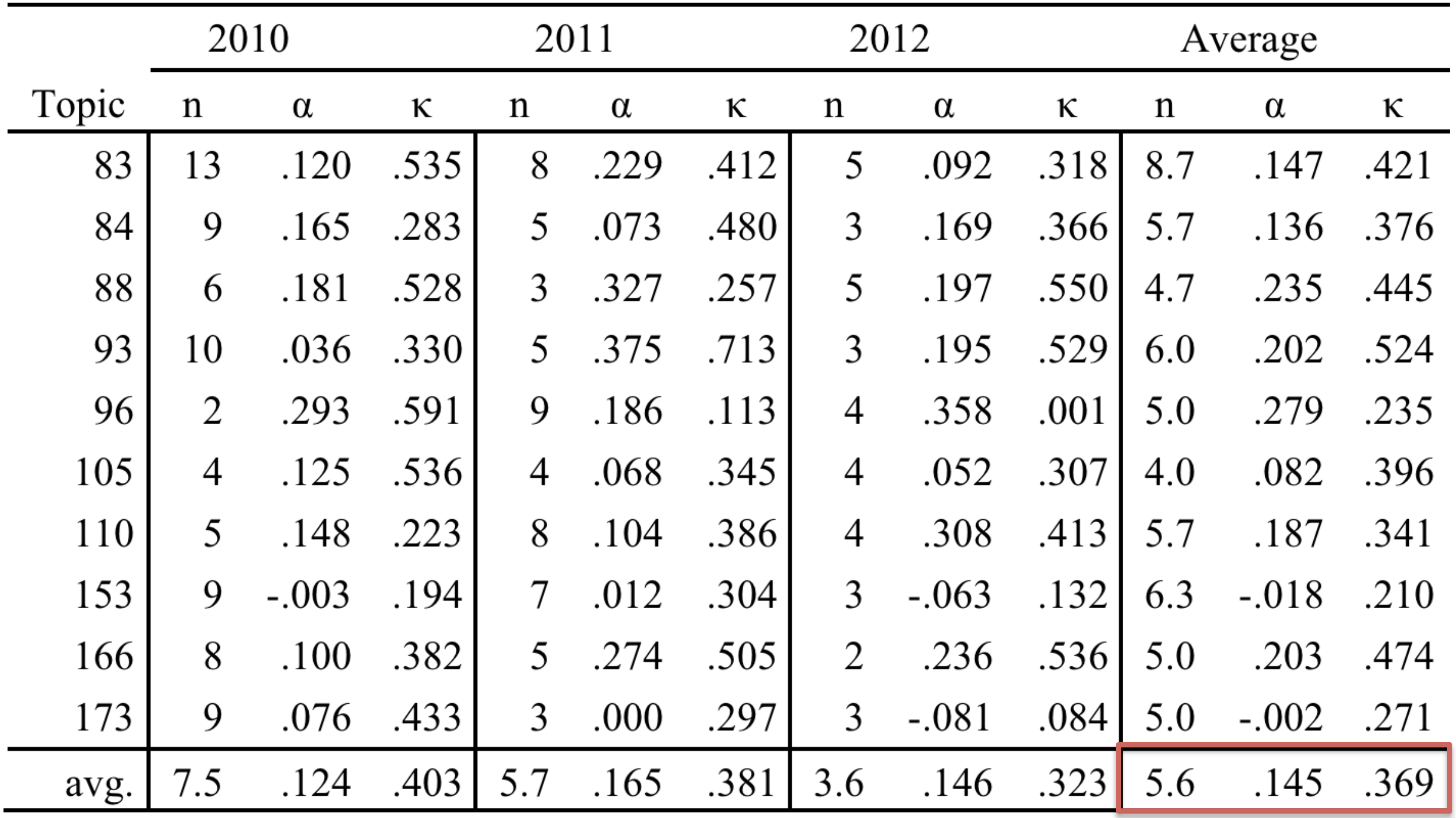
researchers	relev. levels	topics	docs/ topic	ass./ topic	agreement + measure
Lesk & Salton	2	48	1268	2	31%, Jaccard
Cleverdon	5	32	200	4	-
Burgin	3	100	1239	4	40-55%, Jaccard
Voorhees & Harman	2	49	400	2	72%, overlay
Voorhees, Cormack	2+3	49	≈124	2-5	33%, Jaccard
Sormunen	4	38	31-200	2	custom
Trotman et al.	2	15	67-135	3-5	custom
Bailey et al. [4]	3	33	53-176	3	Cohen's κ
Piwowarski et al. [16]	2-4	20	-	2	23-31%, Jaccard
Schaer (this study)	2	10	40-50	2-13	Fleiss' κ and Krippendorff's α

Based on work by Bailey et al. (2008)

Evaluation Setup

- ~370,000 documents from SOLIS (super set of GIRT, used in TREC/CLEF).
- Ten topic from CLEF's domain specific track (83,84, 88, 93, 96, 105, 110, 153, 166, and 173) based on the their ability to be common-sense topics.
- Five different systems
  - SOLR baseline system
  - QE based on thesaurus terms (STR)
  - Re-Ranking with Core Journals (BRAD) and author networks (AUTH)
  - A random ranker (RAND)
- Assessments in Berlin (Vivien Petras) and Darmstadt (Philipp Mayr)
  - 75 participants in 2010 (both), 57 participants in 2011 (both), and 36 in 2012 (only Darmstadt)
  - 168 participants after data cleaning (removed incomplete topic judgements)
  - Binary judgements, 9226 single documents assessments in total

Results: Inter-Assessor Agreement

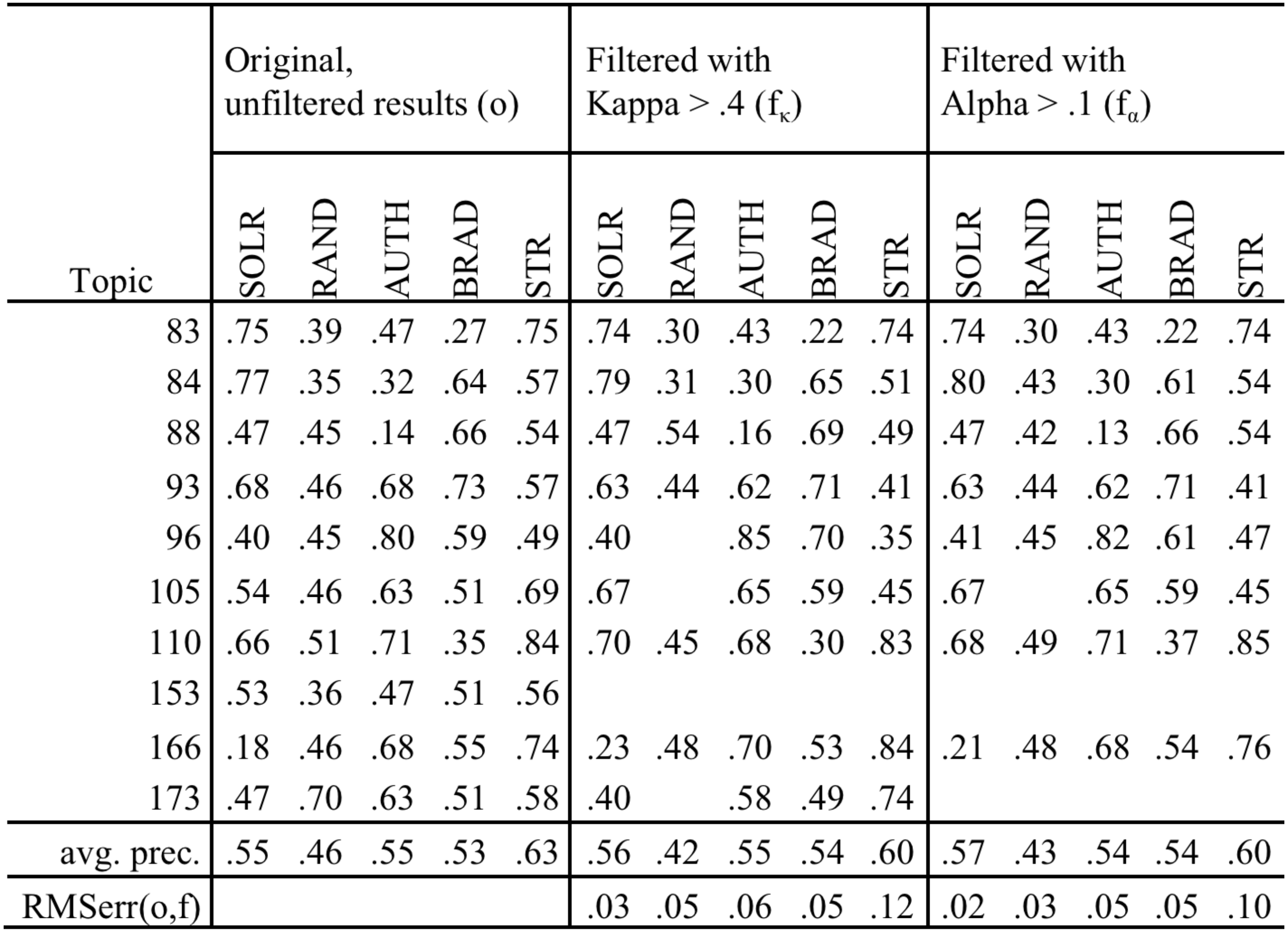


	2010			2011			2012			Average		
Topic	n	α	κ	n	α	κ	n	α	κ	n	α	κ
83	13	.120	.535	8	.229	.412	5	.092	.318	8.7	.147	.421
84	9	.165	.283	5	.073	.480	3	.169	.366	5.7	.136	.376
88	6	.181	.528	3	.327	.257	5	.197	.550	4.7	.235	.445
93	10	.036	.330	5	.375	.713	3	.195	.529	6.0	.202	.524
96	2	.293	.591	9	.186	.113	4	.358	.001	5.0	.279	.235
105	4	.125	.536	4	.068	.345	4	.052	.307	4.0	.082	.396
110	5	.148	.223	8	.104	.386	4	.308	.413	5.7	.187	.341
153	9	-.003	.194	7	.012	.304	3	-.063	.132	6.3	-.018	.210
166	8	.100	.382	5	.274	.505	2	.236	.536	5.0	.203	.474
173	9	.076	.433	3	.000	.297	3	-.081	.084	5.0	-.002	.271
avg.	7.5	.124	.403	5.7	.165	.381	3.6	.146	.323	5.6	.145	.369

Summary: Inter-Assessor Agreement

- The general agreement rate is low
  - Avg. Kappa values between 0.210 and 0.524 → "fair" to "moderate"
  - Avg. Alpha values between -0.018 and 0.279 → away from "acceptable"
  - Alpha values are generally below Kappa values
- Correlation between between Kappa and Alpha (Pearson): 0.447
  - 0.581 in 2010, 0.406 in 2011, and 0.326 in 2012
  - Some outliners like topic 96 in 2012 and topic 83 in 2010
- Large differences between topics
  - Based on number of students per topic and the specific topic
  - In 2010 7.5 students per topic and relatively high correlation between Alpha and Kappa
  - In 2012 fewer students and lower correlation
  - Topic 153 and 173 both got very low Alpha and Kappa values

Results: Dropping Unreliable Assessments



Topic	Original, unfiltered results (o)					Filtered with Kappa > .4 (f_κ)					Filtered with Alpha > .1 (f_α)				
	SOLR	RAND	AUTH	BRAD	STR	SOLR	RAND	AUTH	BRAD	STR	SOLR	RAND	AUTH	BRAD	STR
83	.75	.39	.47	.27	.75	.74	.30	.43	.22	.74	.74	.30	.43	.22	.74
84	.77	.35	.32	.64	.57	.79	.31	.30	.65	.51	.80	.43	.30	.61	.54
88	.47	.45	.14	.66	.54	.47	.54	.16	.69	.49	.47	.42	.13	.66	.54
93	.68	.46	.68	.73	.57	.63	.44	.62	.71	.41	.63	.44	.62	.71	.41
96	.40	.45	.80	.59	.49	.40		.85	.70	.35	.41	.45	.82	.61	.47
105	.54	.46	.63	.51	.69	.67		.65	.59	.45	.67		.65	.59	.45
110	.66	.51	.71	.35	.84	.70	.45	.68	.30	.83	.68	.49	.71	.37	.85
153	.53	.36	.47	.51	.56										
166	.18	.46	.68	.55	.74	.23	.48	.70	.53	.84	.21	.48	.68	.54	.76
173	.47	.70	.63	.51	.58	.40		.58	.49	.74					
avg. prec.	.55	.46	.55	.53	.63	.56	.42	.55	.54	.60	.57	.43	.54	.54	.60
RMSerr(o,f)						.03	.05	.06	.05	.12	.02	.03	.05	.05	.10

Summary: Dropping Unreliable Assessments

- “There are no magical numbers” … but…
  - Applying high thresholds like Alpha and Kappa > 0.8 → no remaining data
  - Moderate/low thresholds of Alpha > 0.1 and Kappa 0.4 lead to a different view
  - A total of 17 out of 30 assessments sets had to be dropped due to Kappa filter and 11 due to Alpha filter
- Large differences between topics
  - No single topic had reliable assessments for all three years
  - Topic 153 and 173 both got very low Alpha and Kappa values, no data remains
- Root mean square (RMS) as an error measure
  - Moderate, but clear differences between 0.05 and 0.12
  - In both cases STR had the highest differences

Discussion and Conclusion

- Student's assessments are inconsistent and contain disagreement!
- We didn't compare to an expert group yet, but n=168 is a large sample group, so somehow reliable results
- But: Many users and agreement don't go hand-in-hand
- And: The effects of throwing away inconsistent assessments is considerable
- Especially true for new evaluation settings like crowd sourcing using Amazon's Mechanical Turk etc.
- Remember: Agreement != reliability, but is gives clues on stability and reproducibility. Not necessarily on accuracy.

Despite "no consistent conclusion on how disagreement affects the reliability of evaluation" (Song et al, 2011), **report on the disagreement and consider data filtering!**

Mini-statistic based on the Lab's overview articles (done yesterday after a 6 hour trip… so please don't take this tooooo serious… :)

Did the organizers report on inter-assessor agreement/no. of assessors etc.?

- **CHiC:** Didn't report (no multiple assessors per topic? Unclear…)
- **CLEF-IP:** Didn't report ("main challenges faced by the organizers were obtaining relevance judgments…")
- **Image-CLEF (Medical Image):** Didn't report, but "Many topics were judged by two or more judges to explore inter-rater agreements and its effects on the robustness of the rankings of the systems".
- **Inex (Social Book):** Didn't report
- **PAN:** Unsure… (reused TREC qrels?!?)
- **QA4MRE:** Didn't report
- **RepLab:** Couldn't download
- **CLEF eHealth:** Didn't report