Can we at least agree on something?

Andrew Trotman
Nils Pharo
Dylan Jenkinson

At-INEX Experiment

- INEX Workshop in Dagstuhl
  - Lockdown scenario
  - Many XML-IR researchers
  - Nothing to do in evenings (except talk shop)
- IR Experiments
  - Require coordination
  - Require participants
  - Time consuming to perform
- This is a perfect match

Possible At-INEX Experiments

- Interactive
  - Needs substantial number of participants
  - Needs an hour (or two) to run
  - Both requirements are met
- Assessment
  - Has been time consuming (11 hours/topic in 2005)
  - Does not require many participants
- CLEF
  - 2006 Real-Time Question Answering

At-INEX Experiment 2006

- Assessment experiment aims:
  - Reduce the assessment load
  - Increase soundness and completeness of assessments
- Specific questions:
  - Do we really need one assessor per topic?
  - Can the INEX document pool be reduced in size?
  - Are short time frame assessments effective?

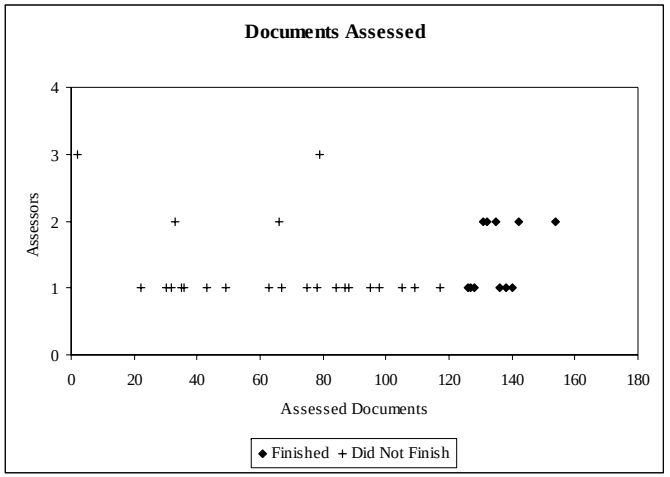
Details

- 41 Participants
- 15 topics (the 2006 double-assessed topics)
- Wikipedia Collection
- No manor to the distribution method
- Time limit of 1:20
- Participants answered short questionnaire on what they did after the experiment
- X-Rai but new pooling software

Pooling

- Reduced pools were used
  - 135 docs per pool on average (~1 hour to assess)
  - Official pool agreements 92%-100% (98% mean)
- New Question
  - Why are the pools different?
    - Official runs are not identified
    - Runs with errors are officially rejected, but later scored!
- ***Proposal:***
  - ***Mark official and rejected runs as so in the XML***
  - ***Either reject or accept runs, no partial accepts***

Assessment Rate

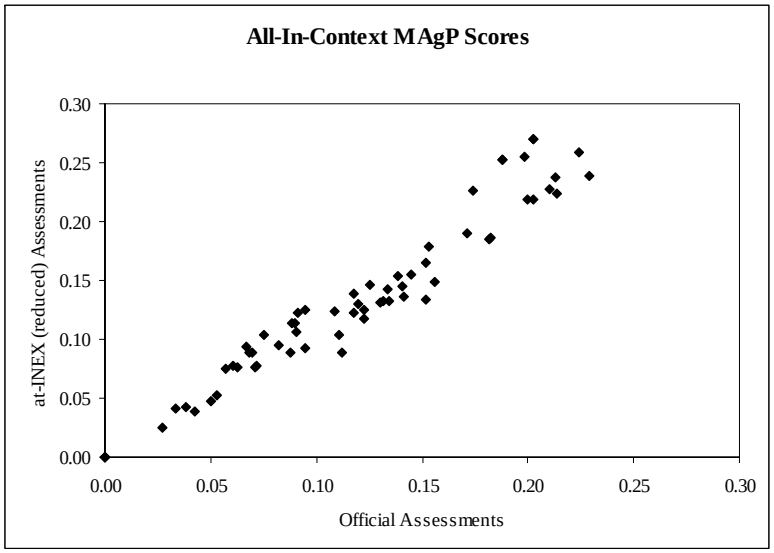


- 16 participants completed the task 29 did not
- 2 to 154 documents per hour (mean 87)

Shallow Pooling

- Are short time frame assessments effective?
- Compare the result (on 15 topics)
  - Official INEX (time:6:15) to at-INEX (time:1:20)
  - Top-n of 500 to top-n of 100
- All-in-Context runs
  - This is the most viable task (in our opinion)
  - MAgP as the metric (as it was the official metric)
- Results
  - Spearman's of 0.97(strong positive correlation)

Shallow Pooling



Top 10 / Top 13

- What about where it matters: The top 10?
  - There are 13 runs in the top 10 of either
  - Spearman's is -0.03 (very weak negative)
- ***Shallow pooling is not good for determining relative performance of the top performers, but can be used for rough overall indicator of performance***

Questionnaire Results

- Factors influencing relevancy of a passage
  - 14 categories including:
    - geographical facet
  - 3 common categories

Factors	Titles	Content	Keywords	Other
Assessors	11	30	12	21

- Barry 1994: factors beyond topical appropriateness influence decisions
- We observe this in XML-IR too
- ***Is this evidence for the All-in-Context task?***

Firmness of Decision

- Did you change your mind during assessment?
  - 17 yes, 22 no, 2 did not respond
  - The main reason was learning about
    - The topic
    - The document types
    - The assessment software
- ***Should we re-assess early documents?***
- ***Should we train the assessors?***

Size of Relevant Passage

- Half the assessors said
  - passages were of a preferred standard size
    - Most of which preferred small or smallish (paragraph)
  - And half said they did not have a preferred size
  - This does not appear to correlate with the topic
- ***Does this mean relevant passages of fixed length is a good retrieval strategy?***

Relationship to Elements

- Most specified one or fewer elements

Elements	**< 1**	**1**	**2**	**3+**
Assessors	25	24	10	15

- ***Evidence for Passage Retrieval?***
- Multiple BEPs
  - 8 yes, 22 no, a few said sometimes

2006 Assessments

- Recently been discussed by INEX organizers
- For X-Rai the document collection is modified

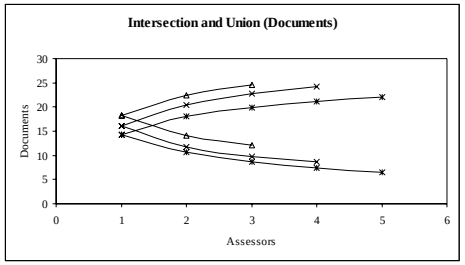
  <a>some.text.with.spaces.after....</c>
- becomes

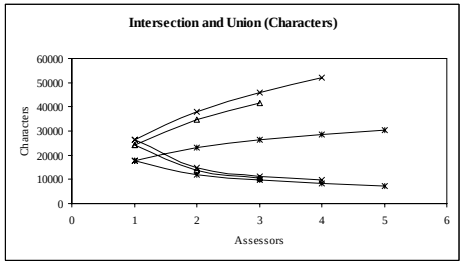
  <a><xrai:s>some.text</xrai:s>.with.spaces.after....</c>
- So, the assessments don't match the collection!
- ***Is the perfect run possible?***

Agreement Levels

- Analysis includes:
  - Official INEX assessments
  - Double-judged assessments
  - Shallow-pools at least half finished
- 15 topics, 60 assessors, 1,471 documents
  - Mean of 98 documents and 4 assessors per topic

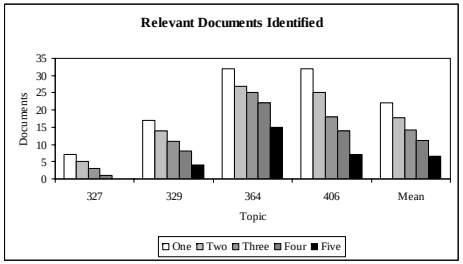
Intersection and Union

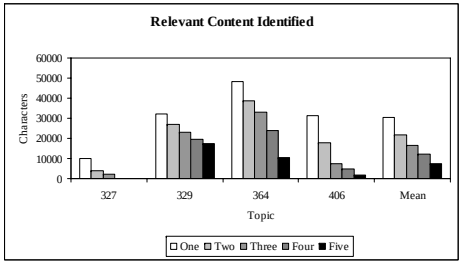




- Extrapolating to 19 assessors there are no documents in common, but 33 documents identified
- Extrapolating to 8 assessors there are no characters in common, but 64,167 characters identified
- ***Relevance is not a universal truth (as we already know)***

Assessor Agreement





- Mean agreement as the number of assessors increases
- Each time a new assessor is added there is new disagreement
- Note topic 327 where they agree on documents but not where within the document (as predicted in previous slide)
- ***Evidence that Passage Retrieval is harder than Document Retrieval?***

Conclusions

- Pooling program
  - We should mark official and rejected runs
  - We should verify all runs at submission time
- Shallow pooling
  - 87 documents in 1:20
  - Not enough for accurate system ranking
- Agreement levels
  - Relevance is in the mind of the assessor
- Questionnaires
  - Little agreement on how assessing is performed