The CLEF Relevance Assessments in Practice

Djoerd Hiemstra

University of Twente

Overview

- Evaluation and the blessings of TREC
- Special cross-language issues
- The judgements
- Conclusion

Retrieval system evaluation

Evaluation of a retrieval system is concerned with how well the system is satisfying users, not just in individual cases, but collectively, for all actual and potential users in the community

Jean Tague-Sutcliffe, 1996

Retrieval system evaluation

- Real (potential) users
- Controlled lab environment
  - clear procedure: which documents selected?
  - clear instruction: unambiguous decisions, consistent per topic
- “Double blind”
  - subjects and supervisors do not know which system(s) produced the results
- Clear - common - evaluation measures

The blessings of TREC / CLEF

- NIST / CLEF takes care of all of the above
  - (on the previous slide)
- We, IR system developers, use the data!
  - The collection can be used without consulting the users again
- But...

Controlled lab environment for cross-language retrieval

- Need good judgements for each language
  - each language a different (native) assessor

- Need a good pool for each language
  - each language a different pool (?)
  - ironically, monolingual runs are of the utmost importance for CLIR system evaluation!

Conflicts of interest

- A multilingual run adds a small, unbalanced, number of documents to each pool
  - maybe larger pool for multilingual task?
  - maybe treat a multilingual run as merged run of 5 bilingual runs?
- A consistent pool depth for each run may result in pools of very different sizes
  - maybe different pool depth per language?

Two less ambitious goals

- At least:
  - Consistency within each subcollection / language
  - Consistency within each experiment / task
- Possibly:
  - Similar approaches for each subcollection / language
  - Similar approaches for each experiment / task:

The judgements in practice

- The pools

– English:	25085 docs (502 per topic)
– French:	12613 docs (252 per topic)
– German:	16872 docs (337 per topic)
– Italian:	11505 docs (230 per topic)
– Spanish:	14549 docs (291 per topic)
– Dutch:	16774 docs (335 per topic)

The judgements in practice

	dutch	english	french	german	italian	spanish
nr. of assessors	10	6	5	2	3	4
experienced as user?	yes	some	yes	one	some	no
experienced as assesors?	no	yes	no	one	two	no
written/oral instruction	oral	both	oral	both	both	oral
native topics by assesors	no	yes	yes	one	one	no
translated by assessors	no	no	no	one	one	no
transl. from source lang.?	yes	yes	yes	no	mostly	no
discussion possible?	some	some	yes	yes	yes	yes
single/group opinion?	single	single	?	group	group	group
supervisors involved?	no	no	yes	yes	yes	no
post-assessm. narrative?	no	yes	no	sketchy	sketchy	no

The judgements in practice

- Quality: Agreement between judges
  - all docs, average over 10 topics: 0.934
  - only rel. docs of user 1, average: 0.569
  - only rel. docs of user 2, average: 0.635
  - overlap: 0.405 (TREC-4 ad-hoc: 0.426)
- Completeness: Removing runs from pool
  - mean absolute difference: 0.0013 avg. prec.
  - max absolute difference: 0.0059 avg. prec.

The judgements in practice

- What did we learn?
  - CLEF judgements as good as TREC ad-hoc and better than the early CLIR task in TREC
  - Some variety in organisation
  - Often group opinion on problematic cases!
    - two-stage assessments
    - is this really impossible for the entire collection?
  - Post-assessment narratives
    - good idea, next year for all languages!

Conclusion

- Diverse pool depths (per language / topic)?
- Per language pool for multilingual task, or at least a larger pool?
- Can we do a consistent group opinion for the multilingual collection?