Thomas Mandl

Robust Task at CLEF 2007?

Break-Out Session

7th Workshop of the

Cross-Language Evaluation Forum (CLEF)

Alicante 22 Sept. 2006

Purpose of Robust 2006: Robustness in CLIR

- Robustness in multilingual retrieval
  - Stable performance over all topics instead of high average performance (like at TREC, just for other languages)
  - Stable performance over all topics for multi-lingual retrieval
  - Stable performance over different languages (so far at CLEF ?)

Ultimate Goal

- „More work needs to be done on customizing methods for each topic“ (Harman 2005)

Positive Aspects

- Many runs were submitted
  - by merely 8 groups
- Large topic set has been assembled
  - e.g. sensitivity analysis can be done with a larger topic set
- Cheap
  - no relevance assessments were necessary

Negative Aspects

- Probably too many sub tasks
- Not too much specific work on robustness
- Inconsistency in data has been criticized
- Robustness over languages has not been tackled
  - probably unpractical
  - so this seems to be no option for another robust task

Results

- High correlation between MAP and GMAP (higher than at TREC)
  - Is robust analysis necessary?
- Problems finding difficult topics in multilingual retrieval
  - Topics are not inherently difficult, but in combination with a collection

Future of Robust CLEF task?

- **No more robust task**
- **Robustness between different collections**
- **Enforce Topic specific treatment**
- Repeat a similar robust task
- Robustness over different user models
- Participants need to determine Topic difficulty
- In depth failure analysis for individual topics (RIA)
- Robustness over different tasks / Team up with another track
- New Measures

Proved hard at TREC

hard to organize in a CLEF style

hard to organize

Robustness over different user models?

- high recall vs. high precision

Robustness between different collections

- Similar to human-assisted runs at last robust track at TREC
- Use relevance assessments from one collection to find documents in another
- Routing
- Can easily be done cross-lingual
- Is it about robustness?

Enforce Topic specific treatment

- Force participants to submit runs
  - where the topic set is split into two or more subsets
  - which are treated with specific methods

Repeat a similar robust task?

- People want to work on a task at least a second year
- Less sub-tasks
  - No bi-lingual
  - Drop German (inconsistency)
- New topic split?

Thanks for your Attention

I am looking forward to the Discussion