Thomas Mandl

Information Science

Universität Hildesheim

mandl@uni-hildesheim.de

How robust is CLIR? Proposal for a new robust task at CLEF

6th Workshop of the

Cross-Language Evaluation Forum (CLEF)

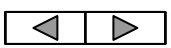
Vienna 23 Sept. 2005

Niedersachsen

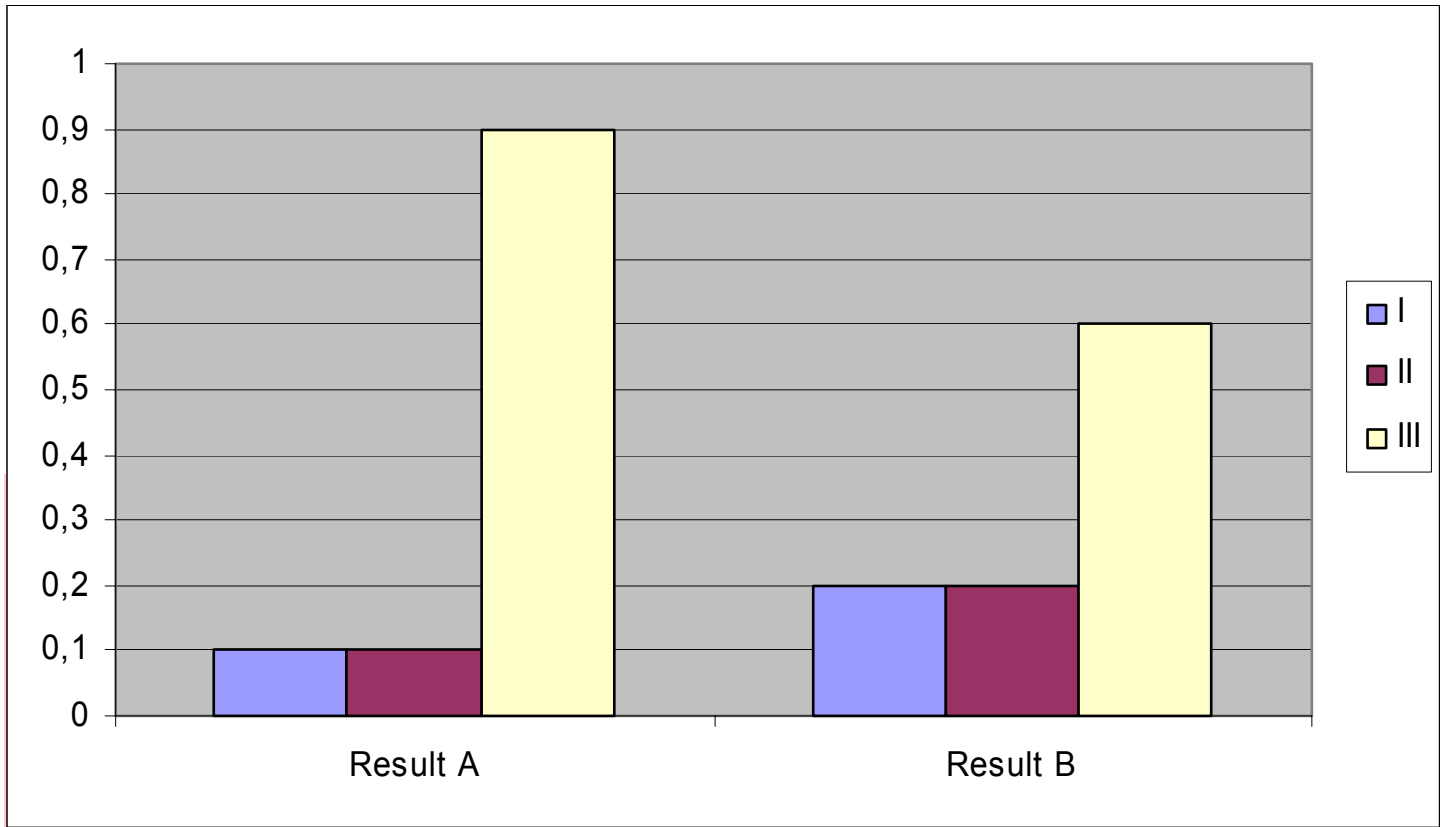
UNIVERSITÄT HILDESHEIM

Robustness?

- Takes great differences over topics into account
  - Performance for difficult topics is emphazised
  - Stable performance over many topics is rewarded
- Mean Average Precision (MAP) is not the main maesure in evaluation
- **Geometric average** (geoAve) is one of the main maesure for evaluation



Example: Which system is better?

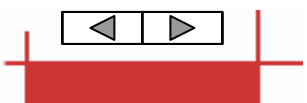


GeoAve	A	0.21	GeoAve	B	**0.29**
MAP	A	**0.37**	MAP	B	0.33

UNIVERSITÄT HILDESHEIM

Why Robustness?

- -> Robustness might be a better approximation of user expectations for many information needs
- -> Robustness might be interesting for practical applications
- It is done at TREC



UNIVERSITÄT HILDESHEIM

Robustness in CLIR

- Robustness in multilingual retrieval could be interpreted in three ways:
  - Stable performance over all topics instead of high average performance (like at TREC)
  - Stable performance over different tasks (like at TREC)
  - Stable performance over different languages (so far at CLEF ?)

Does it make a difference? Robustness of past runs

MAP	Geo Ave
1	9
2	1
3	2
4	4
5	6
6	7
7	5
8	3
9	14
10	1

- Ranking of system changes when geoAve is applied instead of MAP
  - Ranking correlation at between 0.99 and 0.91
  - Top system changes for some tasks
- For example:
  - Bilingual, topic language English in CLEF 2002
  - Top system at MAP drops to 10 at geoAVe

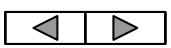
Rankings for Monlingual German task in CLEF 2001

UNIVERSITÄT HILDESHEIM

Cost for Evaluating Robustness?

- Robustness needs at least 250 topics (Voorhees 2005)
  - CLEF has created an ad-hoc topic set of 250 topics in several languages
  - Relevance assessments are available
- Participants may tune their existing systems for robustness

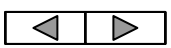
- **Cost is very low**



UNIVERSITÄT HILDESHEIM

Robust CLEF?

It depends on you, the CLEF community!



Cross-Language Evaluation Forum (CLEF)