Thomas Mandl

8th Workshop of the Cross-Language Evaluation Forum (CLEF) Budapest 19 Sept. 2007

Information Science
Universität Hildesheim
mandl@uni-hildesheim.de

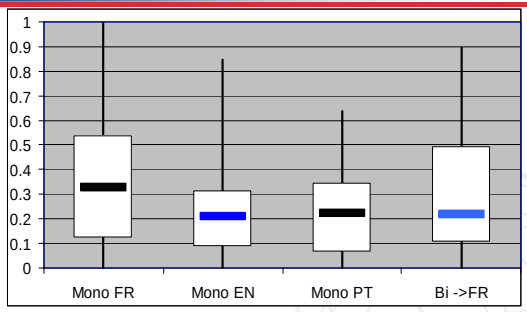
Robust Task - Result Overview and Lessons Learned from Robustness Evaluation

Cross-Language Evaluation Forum (CLEF)

Robustness?

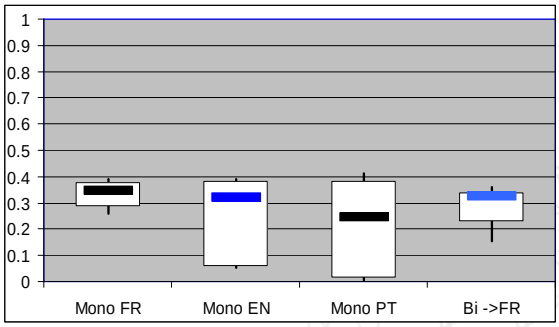
- **Robust …** means … capable of functioning correctly, (or at the very minimum, not failing catastrophically) under a great many conditions. (http://www.reference.com/)
- Robust IR means the capability of an IR system to work well (and reach at least a minimal performance) under a variety of conditions (topics, difficulty, collections, users, languages …)

Variety of conditions …



Variance between topics

System Variance



History of Robust IR Evaluation

- TREC
  - Mono-lingual Retrieval
  - 2003 - 2005
- CLEF
  - Mono-, bi- and Multilingual Retrieval
  - 2006 six languages
  - 2007 three languages

Robust Task 2007

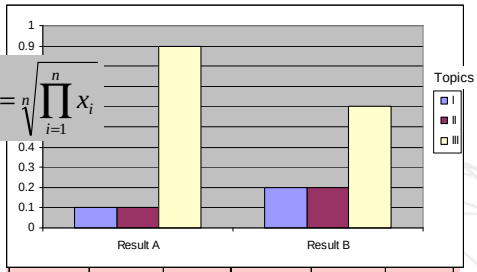
- Again …
  - Use topics and relevance assessment from previous CLEF campaigns
  - Take a different perspective and use a robust evaluation measure (GMAP)
  - Emphasize the difficult (= low performing) topics

Training and Test

- CLEF 2001, 2002 and 2003 for training
- CLEF 2004, 2005 and 2006 for testing

Which system is better?

$$geoAve = \sqrt[n]{\prod_{i=1}^{n} x_i}$$



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

Collections

Language	Target Collection	Training Topics	Test Topics
English	Los Angeles Times 1994	41-200	251-350
French	Le Monde 1994 Swiss News Agency 94	41-140	251-350
Portuguese	Público 1995	-	201-350

Robust Task 2007

- 3 languages (collections and topics)
- 3 mono-lingual tasks
- 1 bi-lingual task (English to French)
- some 300,000 documents
- about 1 gigabyte of text

Participation

- 63 runs submitted by 7 groups
- 2006: 133 runs by 8 groups

Results

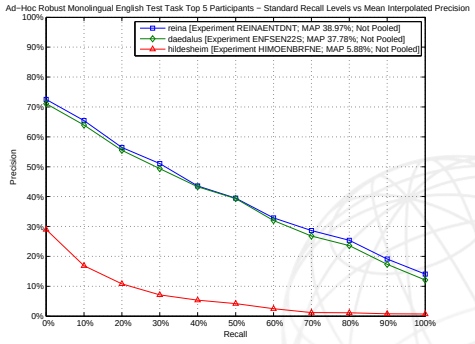
Mono English

Rank	Participant	Experiment	MAP	GMAP
1st	reina	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDNT	38.97%	18.50%
2nd	daedalus	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.DAEDALUS.ENFSEN22S	37.78%	17.72%
3rd	hildesheim	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRFNE	5.88%	0.32%

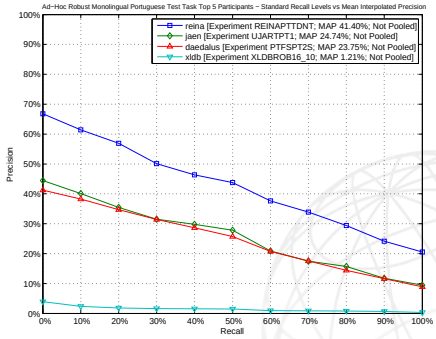
Mono Portuguese

Rank	Participant	Experiment	MAP	GMAP
1st	reina	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDNT	41.40%	12.87%
2nd	jaen	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.JAEN.UJARTPT1	24.74%	0.58%
3rd	daedalus	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.DAEDALUS.PTFSPT2S	23.75%	0.50%
4th	xldb	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB16	1.21%	0.071%

Results Mono English



Results Mono Portuguese



Results

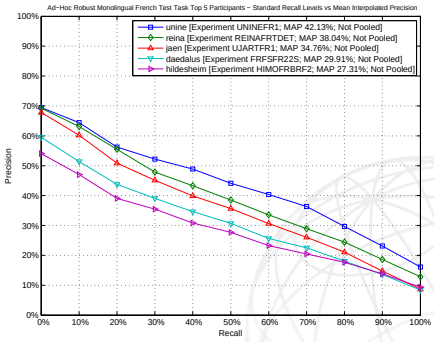
Mono French

Rank	Participant	Experiment	MAP	GMAP
1st	unine	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.UNINE.UNINEFR1	42.13%	14.24%
2nd	reina	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDET	38.04%	12.17%
3rd	jaen	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.JAEN.UJARTFR1	34.76%	10.69%
4th	daedalus	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.DAEDALUS.FRFSFR22S	29.91%	7.43%
5th	hildesheim	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBRF2	27.31%	5.47%

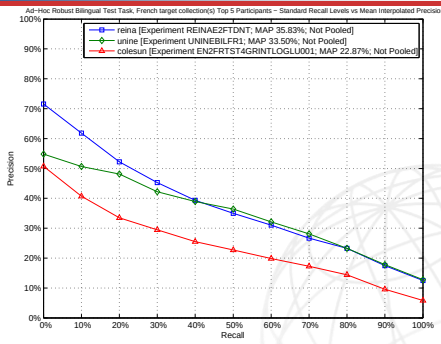
Bi -> French

Rank	Participant	Experiment	MAP	GMAP
1st	reina	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDNT	35.83%	12.28%
2nd	unine	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.UNINE.UNINEBILFR1	33.50%	5.01%
3rd	colesun	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.COLESUN.EN2FRTST4GRINTLOGLU001	22.87%	3.57%

Results Mono French



Results Bi-lingual X -> French



Approaches

- Adoption of traditional and “advanced” CLIR methods
  - BM 25 (*Miracle*)
  - N-gram translation (*CoLesIR*)
  - Weighting, stemming (*Uni NE*)
- Adoption of “robust” heuristics
  - Expansion with an external resource (*SINAI*)

Percentage of *Bad* Topics

- Percentage of Topics which received an MAP below 0.1

	Mono PT	Mono EN	Mono FR	Bi -> FR
Best System	26	17	18	23
Average	32	27	20	25

Topics

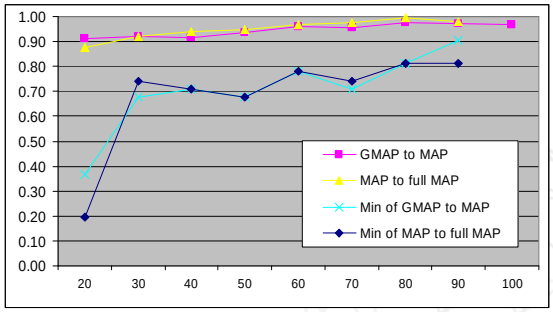
- Large improvements are still possible
- Difficult topics can be solved better

Task	Topic	Average	Best System	System Nr. 1
Mono PT	222	0.0108	0.0478	0.0183
Mono EN	266	0.0217	0.1120	0.0357
Mono FR	192	0.0157	0.0247	0.0160
Bi -> FR	282	0.0342	0.1588	0.1588

Correlation between Measures?

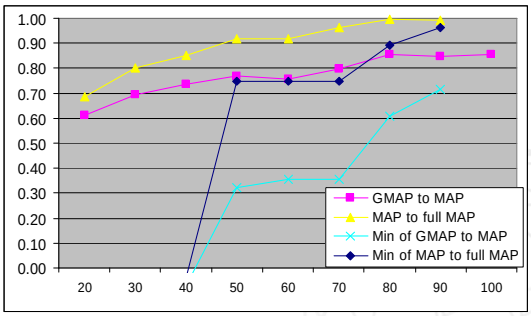
- Often IR measures correlation highly
- For a larger topic set – as used in the robust task – the correlation might be even higher
  - More topics make a test more reliable
- If correlation is high, it makes no sense to use alternative measures

Analysis with Reduced Topic Set



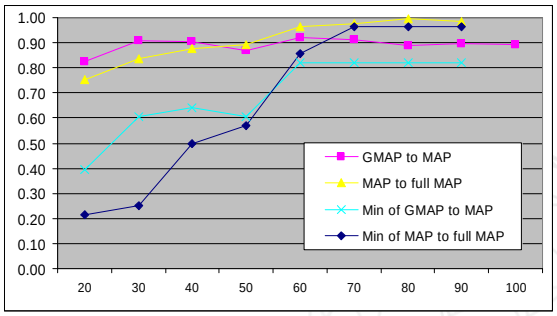
Robust task 2007 Mono-lingual English

Analysis with Reduced Topic Set



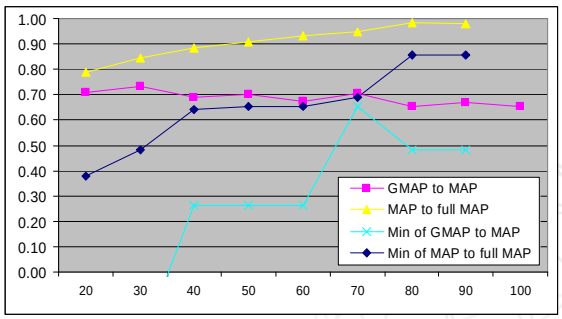
Robust task 2007 Bi-lingual -> FR

Analysis with Reduced Topic Set



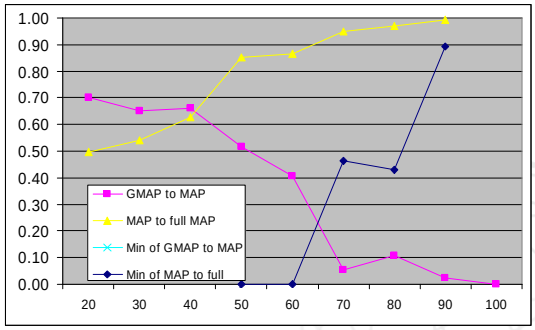
Robust task 2007 Mono-lingual Portuguese

Analysis with Reduced Topic Sets



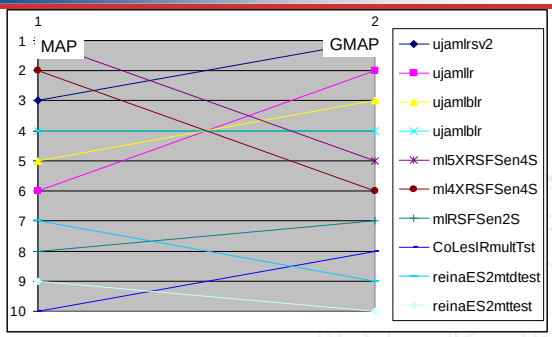
Robust task 2007 Mono-lingual French

Analysis with Reduced Topic Sets



Robust task **2006** Multi-lingual

Changes in Rankings



Robust task **2006** Multi-lingual