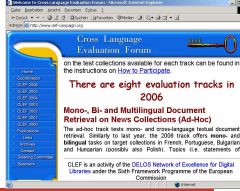


ACM SAC
Fortaleza
18 March 2008

THOMAS MANDL

Information Science
Universität Hildesheim
mandl@uni-hildesheim.de

How Robust are Multilingual Information Retrieval Systems?



There are eight evaluation tracks in 2006


Mono-, Bi- and Multilingual Document Retrieval on News Collections (pl-4ec)

The address track tests mono- and cross-language textual document retrieval. Similarity to last year, the 2006 track offers mono- and bilingual tracks on target collections in French, Portuguese, Bulgarian and Indonesian (available after March). Tracks for statements of...

CLEF is an activity of the DELOS Network of Excellence for Digital Libraries under the Sixth Framework Programme of the European Commission.



Multilingual Information Retrieval

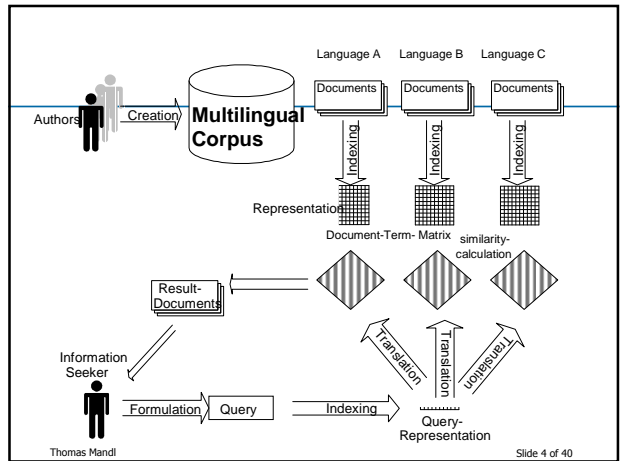
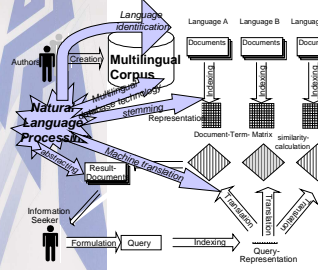


Cross-Language Evaluation Forum

- Query-Language different from Document-Language
- Usage Scenario:
 - User can understand documents in foreign language (passive competence), but cannot formulate a query (active competence)
- Access to documents
 - Dependent of relevance
 - Independent of language

Thomas Mandl

Slide 3 of 40

How can this complex process be evaluated?

Cross-Language Evaluation Forum (CLEF)


Thomas Mandl

Slide 5 of 40

Cross-Language Evaluation Forum (CLEF)

Research on cross- and multi-lingual Information Retrieval Systems

- Test Suites
- System Development
- Benchmarks
- Evaluation Research



EU Funding: DELOS NoE for Digital Libraries

Mandl et al. @ CLEF 2002 - 2007

Thomas Mandl

Slide 6 of 40

Evaluation

- System evaluation:
how good are document rankings?
- User-based evaluation:
how satisfied is the user?

Thomas Mandl

Slide 7 of 40

Why Robust?

The user never sees the perspective of an evaluation
(=MAP)
but only the performance on his/her request(s).

Thomas Mandl

Slide 8 of 40

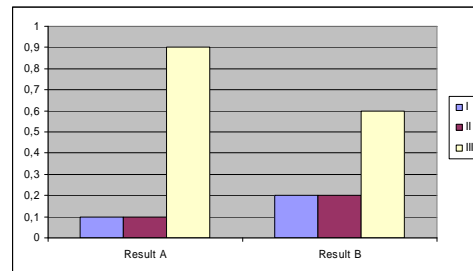
Cross-Language Evaluation Forum (CLEF)

- Create an infrastructure for research and development on cross- und multi-lingual information retrieval
 - Test multilingual information retrieval systems
 - Evaluate systems
 - Create reusable testsuites

Thomas Mandl

Slide 9 of 40

Which system is better?

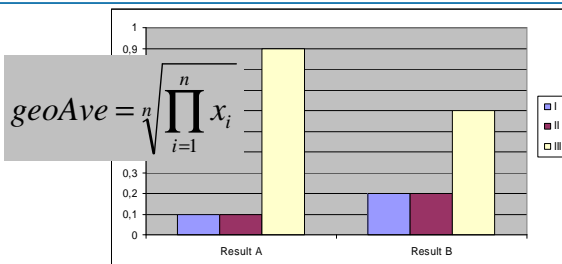


Two Systems, three Requests

Thomas Mandl

Slide 10 of 40

Which system is better?



GeoAve	A	0,21	GeoAve	B	0,29
MAP	A	0,37	MAP	B	0,33

Thomas Mandl

Why Robust?

«The **unhappy customer**, on average, **will tell 27** other people about their experience. With the use of the internet, whether web pages or e-mail, that number can increase to the thousands ...»
«**Dissatisfied customers tell** an average of **ten** other people about their bad experience. Twelve percent tell up to twenty people.»

→ **Bad news travels fast.**

Thomas Mandl

Slide 12 of 40


Why Robust?

On the other hand, satisfied customers will tell an average of n people about their positive experience.



→ **Good news travels somewhat slower**

Your system should produce less bad news!

 improve on *hard* topics

Don't worry too much about the good news

 best topics

Thomas Mandl

Credit to Jacques Savoy for slide
Slide 13 of 40

Purpose: 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

Thomas Mandl

Slide 14 of 40

Ultimate Goal

- „More work needs to be done on customizing methods for each topic“ (Harman 2005)
- Hard, but some work is done in that direction
 - RIA Workshop
 - SIGIR Workshop on Topic Difficulty
 - TREC Robust Track (until 2005)
 - ...

Thomas Mandl

Slide 15 of 40

Robustness

- Robustness in multilingual retrieval
 - Stable performance over all topics instead of high average performance
 - Minimal performance level for all topics and tasks
 - Difficult topics become more important
 - User oriented evaluation measures

Coordination Robust Track CLEF in cooperation with

- National Institute of Standards and Technology (USA)
- City U Dublin
- Université de Neuchâtel
- U Padua

Thomas Mandl

Slide 16 of 40

Data Collection

- Six languages
- 1.35 million documents
- 3.6 gigabytes of text

Language	Collection
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Italian	La Stampa 94, AGZ (SDA) 94/95
Dutch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
German	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanish	EFE 94/95

Thomas Mandl

Slide 17 of 40

Test and Training Set

- Arbitrary split of the 160 topics
 - 60 training topics
 - 100 test topics

Thomas Mandl

Slide 18 of 40

More than 100 Runs ...

Task	Language	nr test runs	nr training runs	nr groups
mono	en	13	7	6
	fr	18	10	7
	nl	7	3	3
	de	7	3	3
	es	11	5	5
bi	it->es	8	2	3
	fr->nl	4	0	1
	en->de	5	1	2
multi	multi	10	3	4

Thomas Mandl

Slide 19 of 40

Results – Mono-lingual

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Dutch	hummingbird	daedalus	colesir			1st vs 3rd
MAP	51.06%	42.39%	41.60%			22.74%
GMAP	25.76%	17.57%	16.40%			57.13%
Run	humNL06Rtd	nlFSnlR2S	CoLesIRnlTst			
English	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	47.63%	43.66%	43.48%	39.69%	37.64%	26.54%
GMAP	11.69%	10.53%	10.11%	8.93%	8.41%	39.00%
Run	humEN06Rtd	reinaENTdtes	dcudesceng1	enFSenR2S	CoLesIRenTst	
French	urine	hummingbird	reina	dcu	colesir	1st vs 5th
MAP	47.57%	45.43%	44.58%	41.08%	39.51%	20.40%
GMAP	15.02%	14.90%	14.32%	12.00%	11.91%	26.11%
Run	UniNEfr1	humFR06Rtd	reinaFRtdtes	dcudescfr120	CoLesIRfrTst	

Thomas Mandl

Slide 20 of 40

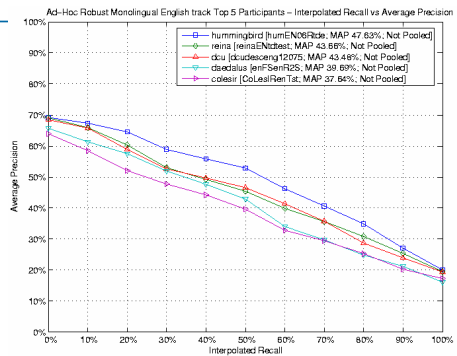
Results – Mono-lingual

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
German	hummingbird	colesir	daedalus			1st vs 3rd
MAP	48.30%	37.21%	34.06%			41.81%
GMAP	22.53%	14.80%	10.61%			112.35%
Run	humDE06Rtd	CoLesIRdeTs	deFSdeR2S			
Italian	hummingbird	reina	daedalus	colesir		1st vs 5th
MAP	41.94%	38.45%	37.73%	35.11%	32.23%	30.13%
GMAP	11.47%	10.55%	9.19%	10.50%	23%	39.37%
Run	humIT06Rtd	reinaITdtes	dcudescit100	itFSitR2S	CoLesIRitTst	
Spanish	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	45.66%	44.01%	42.14%	40.40%	40.17%	13.67%
GMAP	23.61%	22.65%	21.32%	19.64%	18.84%	25.32%
Run	humES06Rtd	reinaESTdtes	dcudescsp12	esFSesR2S	CoLesIResTst	

Thomas Mandl

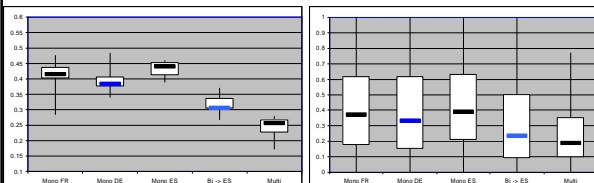
Slide 21 of 40

Results – Mono-lingual English



Thomas Mandl

Slide 22 of 40



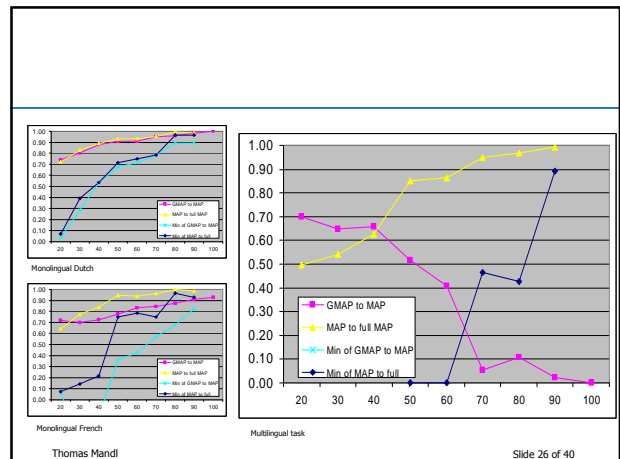
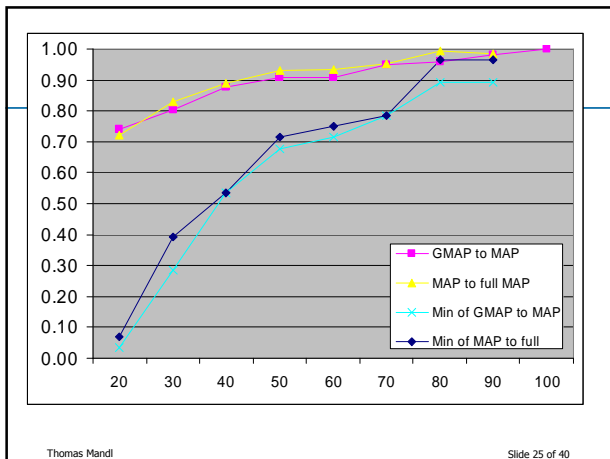
Thomas Mandl

Slide 23 of 40

- Method
- Subset of Topics

Thomas Mandl

Slide 24 of 40



Observations and Conclusions

- Robustness does measure something different in multilingual retrieval
- More „difficult“ topics -> robust measure changes the picture
- CLEF 2008
 - Registration is open
 - <http://www.clef-campaign.org/>
 - Robust Word Sense Disambiguation Task
 - <http://ixa2.si.ehu.es/clirwsd/>

Thomas Mandl

Slide 27 of 40

Acknowledgements

- Giorgio di Nunzio & Nicola Ferro (U Padua)
- Robust Committee
 - Donna Harman (NIST)
 - Carol Peters (ISTI-CNR)
 - Jacques Savoy (U Neuchâtel)
 - Gareth Jones (Dublin City U)
- Ellen Voorhees (NIST)
- Participants

Thomas Mandl

Slide 28 of 40

Download this presentation at:
<http://www.uni-hildesheim.de/~mandl/>



Thomas Mandl

Slide 29 of 40