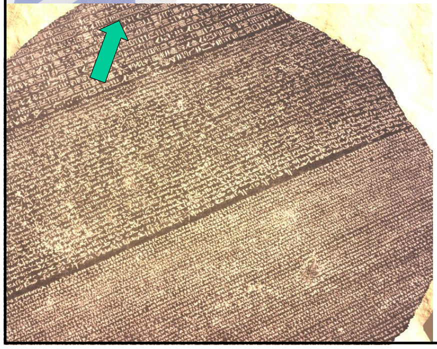
Download this presentation at:
http://www.uni-hildesheim.de/~mandl/

The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation

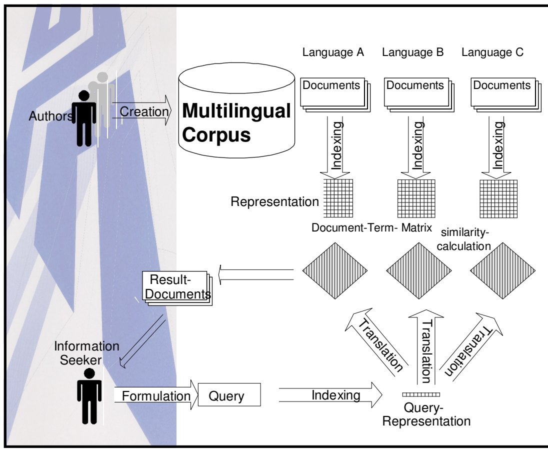
Thomas Mandl, Christa Womser-Hacker
Information Science
University of Hildesheim
Germany
{mandl,womser}@uni-hildesheim.de

Named Entitites

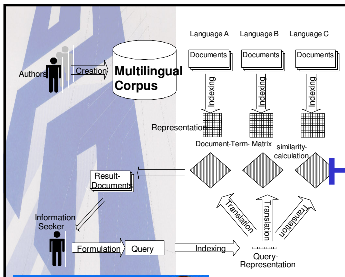


Overview

- What is CLIR and CLEF?
- Named entities in multilingual retrieval
- Relation between system performance and named entities
- System improvement
- Current research



Language A
Language B
Language C
Documents
Documents
Documents
Authors
Creation
Multilingual Corpus
Indexing
Indexing
Indexing
Representation
Document-Term- Matrix
similarity-calculation
Result-Documents
Information Seeker
Translation
Translation
Translation
Formulation
Query
Indexing
Query-Representation



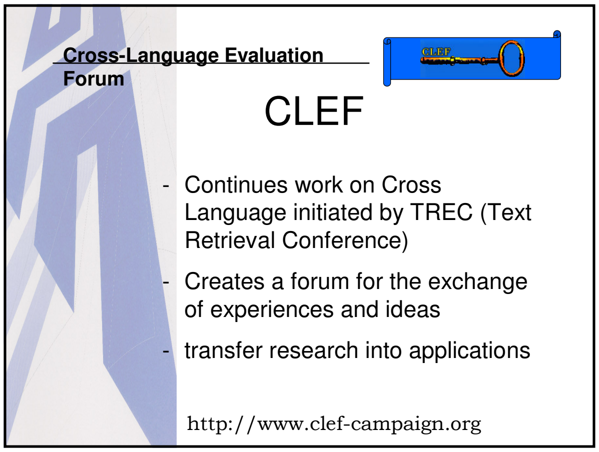
How can this complex process be evaluated?

Cross-Language Evaluation Forum (CLEF)

Cross-Language Evaluation Forum

Goals of 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

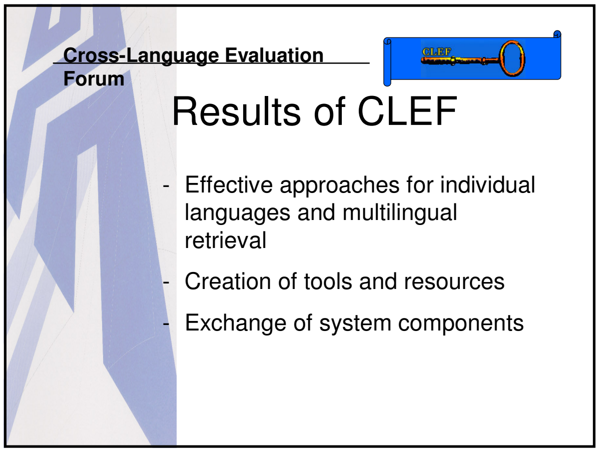


Cross-Language Evaluation Forum

CLEF

- Continues work on Cross Language initiated by TREC (Text Retrieval Conference)
- Creates a forum for the exchange of experiences and ideas
- transfer research into applications

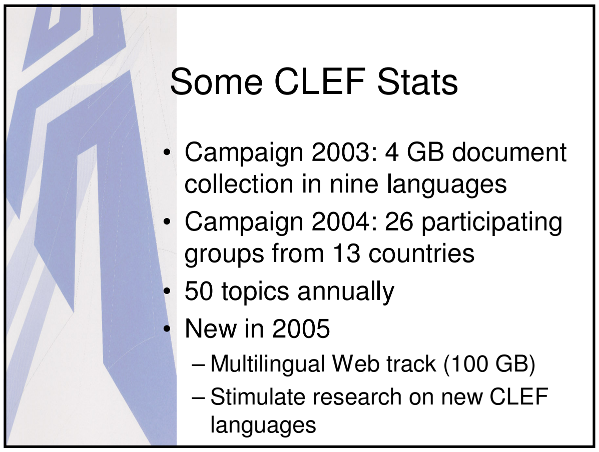
http://www.clef-campaign.org



Cross-Language Evaluation Forum

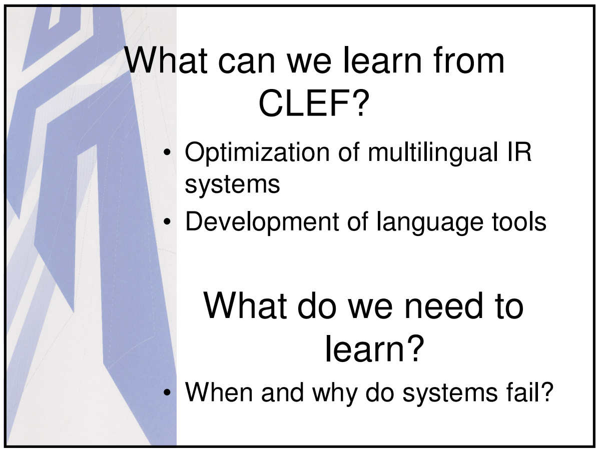
Results of CLEF

- Effective approaches for individual languages and multilingual retrieval
- Creation of tools and resources
- Exchange of system components



Some CLEF Stats

- Campaign 2003: 4 GB document collection in nine languages
- Campaign 2004: 26 participating groups from 13 countries
- 50 topics annually
- New in 2005
  - Multilingual Web track (100 GB)
  - Stimulate research on new CLEF languages

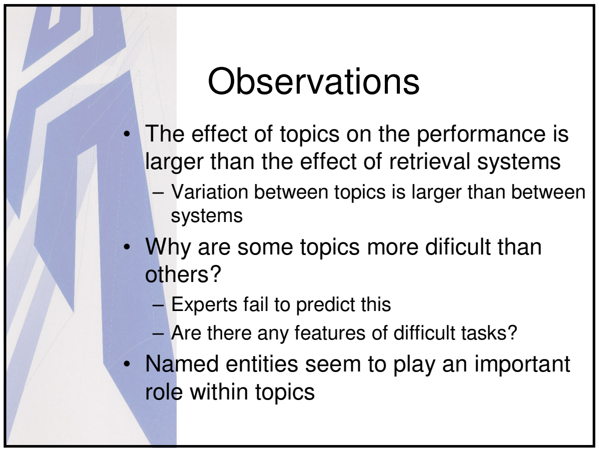


What can we learn from CLEF?

- Optimization of multilingual IR systems
- Development of language tools

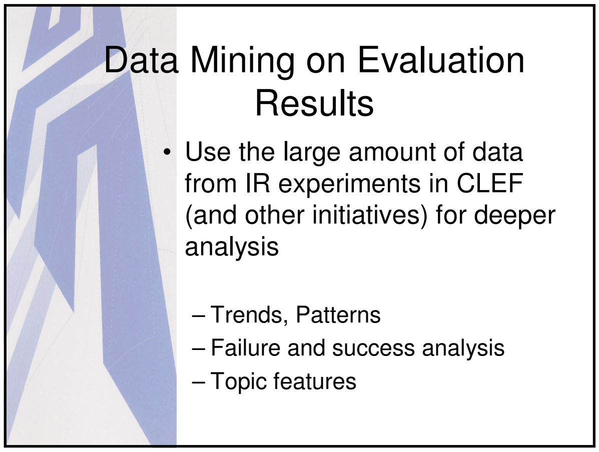
What do we need to learn?

- When and why do systems fail?



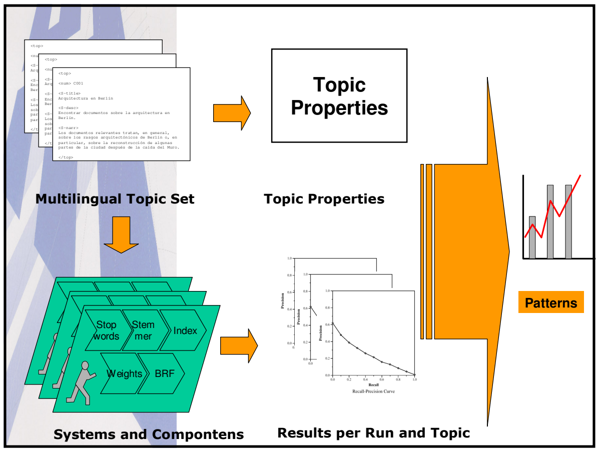
Observations

- The effect of topics on the performance is larger than the effect of retrieval systems
  - Variation between topics is larger than between systems
- Why are some topics more dificult than others?
  - Experts fail to predict this
  - Are there any features of difficult tasks?
- Named entities seem to play an important role within topics



Data Mining on Evaluation Results

- Use the large amount of data from IR experiments in CLEF (and other initiatives) for deeper analysis
  - Trends, Patterns
  - Failure and success analysis
  - Topic features



Topic Properties

Multilingual Topic Set

Topic Properties

Stop words

Stemmer

Index

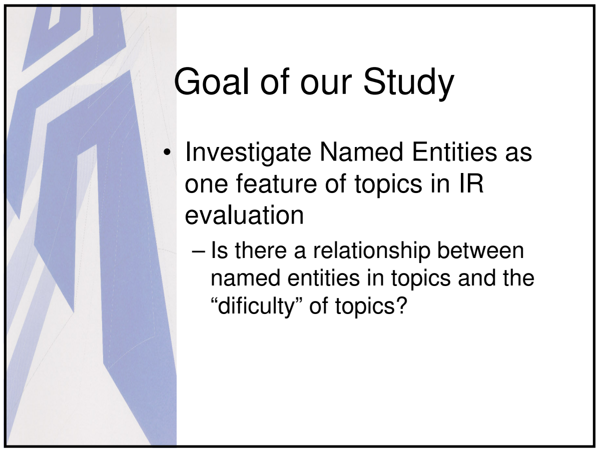
Weights

BRF

Patterns

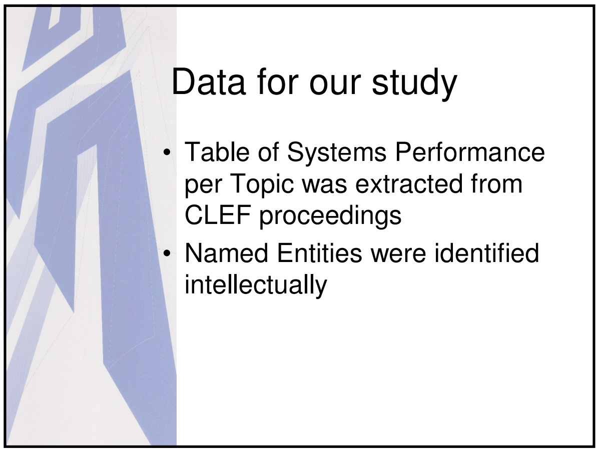
Systems and Compontens

Results per Run and Topic



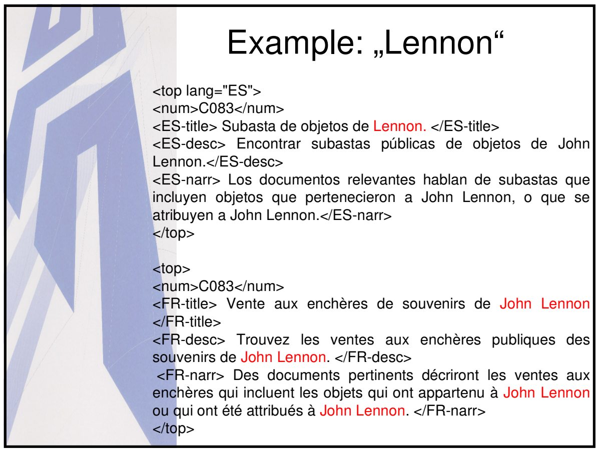
Goal of our Study

- Investigate Named Entities as one feature of topics in IR evaluation
  – Is there a relationship between named entities in topics and the “dificulty” of topics?



Data for our study

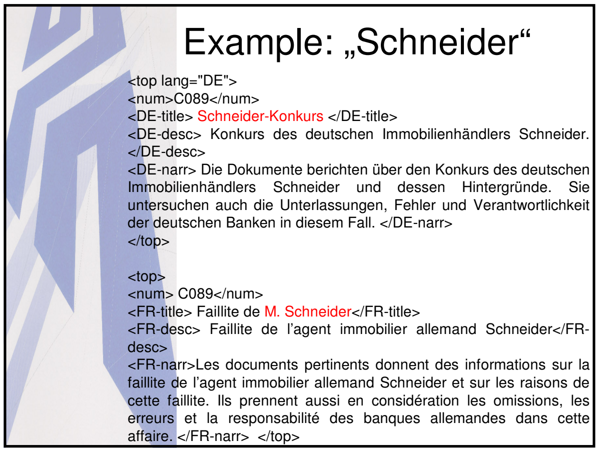
- Table of Systems Performance per Topic was extracted from CLEF proceedings
- Named Entities were identified intellectually



Example: „Lennon“

<top lang="ES">
<num>C083</num>
<ES-title> Subasta de objetos de Lennon. </ES-title>
<ES-desc> Encontrar subastas públicas de objetos de John Lennon.</ES-desc>
<ES-narr> Los documentos relevantes hablan de subastas que incluyen objetos que pertenecieron a John Lennon, o que se atribuyen a John Lennon.</ES-narr>
</top>

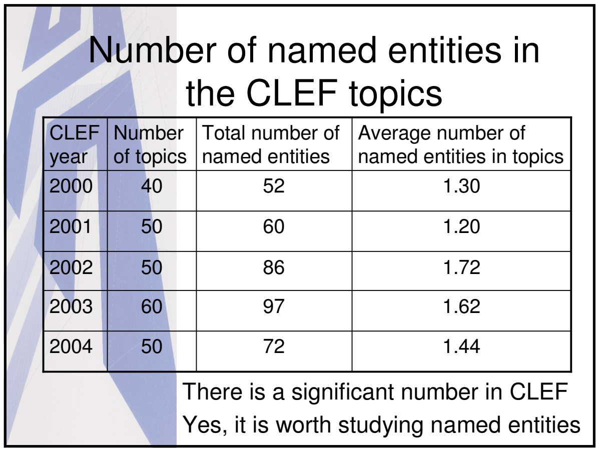
<top>
<num>C083</num>
<FR-title> Vente aux enchères de souvenirs de John Lennon </FR-title>
<FR-desc> Trouvez les ventes aux enchères publiques des souvenirs de John Lennon. </FR-desc>
<FR-narr> Des documents pertinents décriront les ventes aux enchères qui incluent les objets qui ont appartenu à John Lennon ou qui ont été attribués à John Lennon. </FR-narr>
</top>



Example: „Schneider“

<top lang="DE">
<num>C089</num>
<DE-title> Schneider-Konkurs </DE-title>
<DE-desc> Konkurs des deutschen Immobilienhändlers Schneider. </DE-desc>
<DE-narr> Die Dokumente berichten über den Konkurs des deutschen Immobilienhändlers Schneider und dessen Hintergründe. Sie untersuchen auch die Unterlassungen, Fehler und Verantwortlichkeit der deutschen Banken in diesem Fall. </DE-narr>
</top>

<top>
<num> C089</num>
<FR-title> Faillite de M. Schneider</FR-title>
<FR-desc> Faillite de l’agent immobilier allemand Schneider</FR-desc>
<FR-narr>Les documents pertinents donnent des informations sur la faillite de l’agent immobilier allemand Schneider et sur les raisons de cette faillite. Ils prennent aussi en considération les omissions, les erreurs et la responsabilité des banques allemandes dans cette affaire. </FR-narr> </top>



Number of named entities in the CLEF topics

CLEF year	Number of topics	Total number of named entities	Average number of named entities in topics
2000	40	52	1.30
2001	50	60	1.20
2002	50	86	1.72
2003	60	97	1.62
2004	50	72	1.44

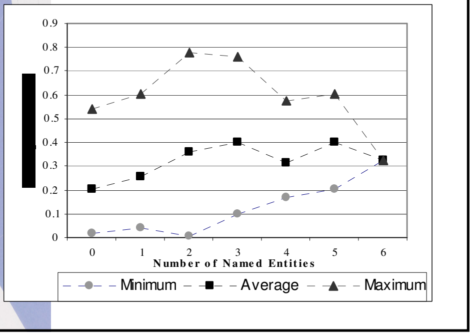
There is a significant number in CLEF

Yes, it is worth studying named entities

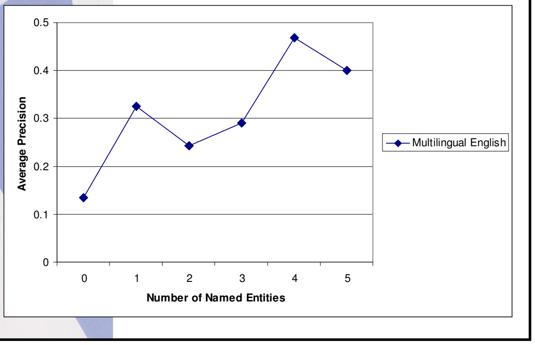
Analyzed Runs (examples)

CLEF year	Task	Topic language	Nr. runs	Topics without amed entities	Topics with one or two named entities	Topics with more than three named entities
2001	Bi	German	9	16	24	7
2001	Multi	German	5	16	24	7
2001	Bi	English	3	16	24	7
2001	Multi	English	17	17	26	7
2002	Bi	German	4	14	21	15

Relation between system performance and named entities



Latest Results: CLEF 2004



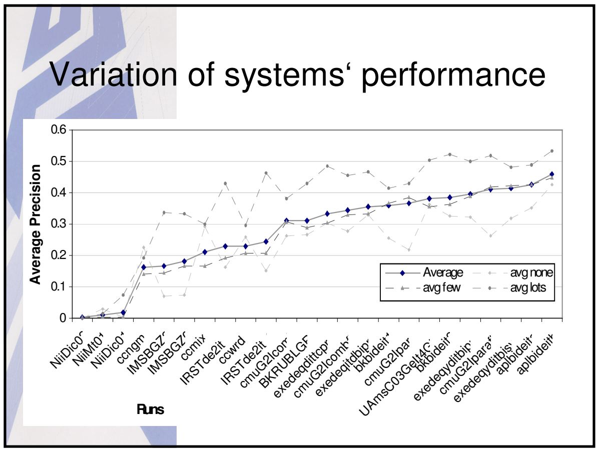
Results

- Named entities make retrieval „easier“ for systems
  - This effect is statistically significant for several tracks, but not for all
- Topic creation in CLEF needs to pay special attention to named entities
- The results can be confirmed by a correlation analysis

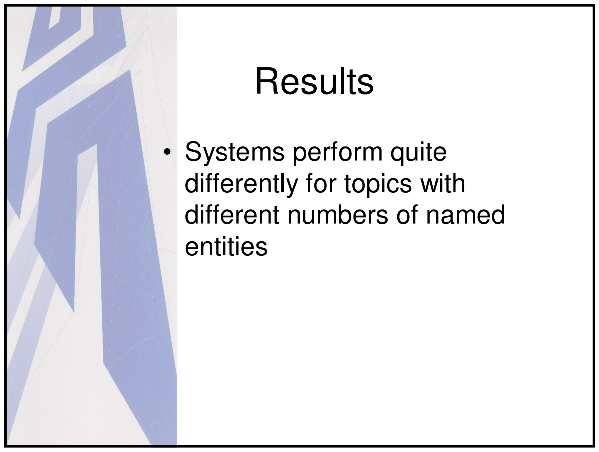
CLEF year	Run type	Topic language	Num-ber of runs	Correlation of average precision per topic to number of named entities	Level of statistical significance (t-distribution) for last column	Correlation of maximum precision per topic to number of named entities
2001	Bilingual	German	9	0.44	99%	0.32
2001	Multilingual	German	5	0.19	-	0.24
2001	Bilingual	English	3	0.20	-	0.13
2001	Multilingual	English	17	-0.34	95%	-0.36
2002	Bilingual	German	4	0.33	-	0.25
2002	Multilingual	German	4	0.43	-	0.41
2002	Bilingual	English	51	0.40	99%	0.36

Variation of systems‘ performance

- Are there systems which perform e.g. especially well for difficult topics?

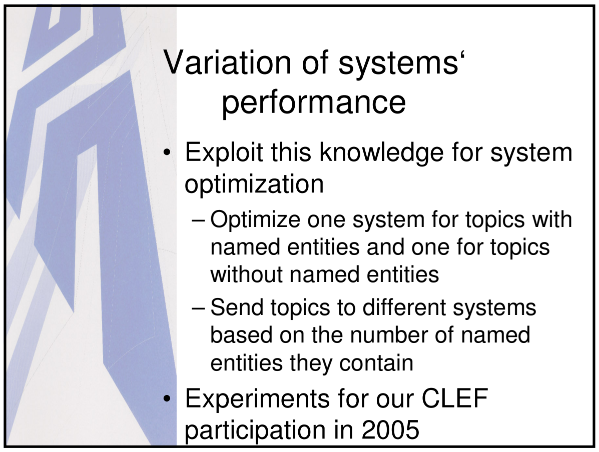


Variation of systems‘ performance



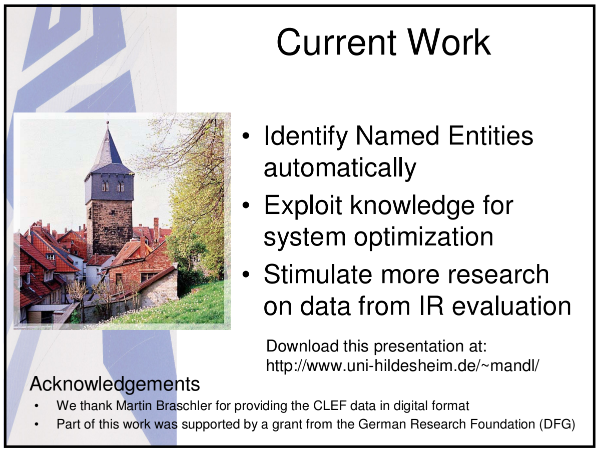
Results

- Systems perform quite differently for topics with different numbers of named entities



Variation of systems‘ performance

- Exploit this knowledge for system optimization
  - Optimize one system for topics with named entities and one for topics without named entities
  - Send topics to different systems based on the number of named entities they contain
- Experiments for our CLEF participation in 2005



Current Work

- Identify Named Entities automatically
- Exploit knowledge for system optimization
- Stimulate more research on data from IR evaluation

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

Acknowledgements

- We thank Martin Braschler for providing the CLEF data in digital format
- Part of this work was supported by a grant from the German Research Foundation (DFG)