9th Workshop of the Cross-Language Evaluation Forum (CLEF) Århus, 18th Sept. 2008

GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview

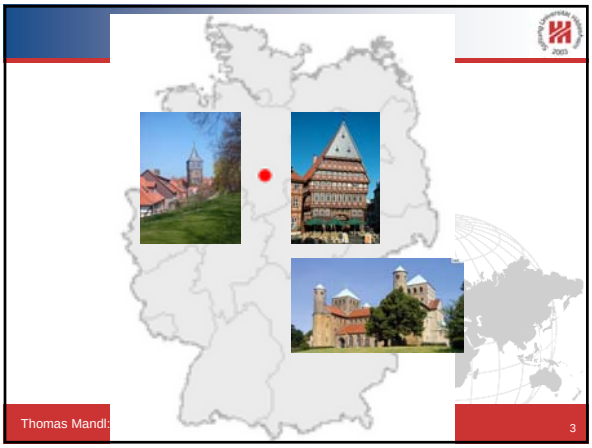
Thomas Mandl[1], Paula Carvalho[2], Giorgio Maria Di Nunzio[3], Nicola Ferro[3]
Fredric Gey[4], Ray Larson[4], Diana Santos[5], Christa Womser-Hacker[1]

[1]Information Science, University of Hildesheim, GERMANY
mandl@uni-hildesheim.de, womser@uni-hildesheim.de
[2]University of Lisbon, DI, LasiGE, XLDB Linguateca, PORTUGAL
pqfcarvalho@gmail.com
[3]Department of Information Engineering, University of Padua, Italy
dinunzio@dei.unipd.it , ferro@dei.unipd.it
[4]University of California, Berkeley, CA, USA
gey@berkeley.edu , ray@sims.berkeley.edu
[5]Linguateca, SINTEF ICT, NORWAY
Diana.Santos@sintef.no

Università degli Studi di Padova

GeoCLEF Administration

- Joint effort of
  - Fredric Gey, Ray Larson (U. California at Berkeley)
  - Diana Santos (Linguateca, SINTEF ICT, Norway)
  - Paula Carvalho (Linguateca, U. Lisbon)
  - Nicola Ferro, Giorgio Di Nunzio (U. Padua)
  - Christa Womser-Hacker (U. Hildesheim)
  - many relevance assessors ….
  - and others



Content

- Introduction
- Geographic Search Task
- Topic Development
- Relevance Assessment
- Results in a Nutshell
- Giorgio Di Nunzio: *Results and Statistical Analysis*
- Diana Santos: *GikiP task*

Initial Aim of GeoCLEF

- Aim: to evaluate retrieval of multilingual documents with an emphasis on geographic search (GIR)
  - Example query:
  - "*find me news stories about riots near Dublin*"

content part | geo part

(Fred Gey @ CLEF Workshop 2005)

Interesting Issues

- Ambiguity
  - Santos, Neustadt, Albertville
  - Galizien, Galicien (Spain, Poland)
  - *Oder* (River but also a stop word in German)
- Different Translations
  - Peking, Beijing
  - Deutschland, Allemagne, Germany
- Name changes
  - Bombay -> Mumbai
  - St. Petersburg ->Leningrad -> St. Petersburg
- multi word groups
  - Rio Grande do Sul, Newcastle upon Tyne

Search Task

- How much and which geo knowledge and reasoning is necessary?
  - spatial reasoning is necessary to solve information needs
    - demonstrations in cities in *Northern Germany*
    - -> *Northern Germany* may not appear in documents
- Often, keyword based systems do well on the task
  - E.g. Blind relevance feedback may lead to expansion with names of cities

Search Task 2008

- Three languages
  - English, Portuguese, German
- 600,000 + docs
- 25 topics
  - so far 100 in four years
  - + 26 geo Topics from prev. CLEF campaigns
- Test collection is available for future use
  - Do experiments with the whole set and publish them

Search Task 2008

- **Monolingual Retrieval**
  - Topic- and document language identical
  - English, Portuguese, German
- **Bilingual Retrieval**
  - Topic- and document language identical
  - {English, Portuguese, German} -> {English, Portuguese, German}

Topic Development

- Topics are meant to express a natural information need which a user of the collection might have
- Goal: creation of a geographically challenging topic set
- Geographic knowledge should be necessary to be successful

Topic Development

- Each group devised a set of candidate topics in their own language, whose appropriateness was checked in the text collection available for that language.
- The candidate topics were subsequently translated into English and checked for relevant documents in the other collections.
- Some candidate topics were modified or refined, due to the absence of relevant documents in one of the languages, the complexity of topic interpretation and/or the translation into the other.
- The final topic set was agreed upon after intensive discussion, and all topics were translated into Portuguese and German
- Final translation and check (Thanks to Sven Hartrumpf)

Topic Development

- Topic development is hard for multilingual collections
- Geo entities below the country level are interesting
- But these geo entities below the country level may not appear in newspapers in other countries
- Relevant documents are required in all three languages

Topic Development

- Several issues were explicitly included:
  - vague geographic regions (Sub-Saharan Africa , Western Europe)
  - geographical relations beyond IN (forest fires on Spanish islands)
  - granularity below the country level (Industrial or cultural fairs in Lower Saxony)
  - terms which do not occur in documents (Portuguese communities in other countries, demonstrations in German cities)

Topic Modifications

Subject modification:

Endangered animal species in Iberian Peninsula

» **Agriculture**

Subject extension:

Nobel Prize winners in Physics from Northern European countries

» **Nobel Prize winners**

Topic refinement:

» **Most visited sights in the capital of France and its vicinity**

Most visited sights in the capital of France

Topic Creation (spatial parameters)

- The majority of the topics specify complex (multiply defined) geographical relations, which may represent:
  - Inclusion (e.g. Attacks **in** Japanese subways);
  - Exclusion (e.g. Portuguese immigrant communities in the world).

    [the generic geographical term world must be interpreted, in this context, as the entire world excluding Portugal]

Example

<num>10.2452/89-GC</num>
<title>Trade fairs in Lower Saxony </title>
<desc>Documents reporting about industrial or cultural fairs in Lower Saxony. </desc>
<narr>Relevant documents should contain information about trade or industrial fairs which take place in the German federal state of Lower Saxony, i.e. name, type and place of the fair. The capital of Lower Saxony is Hanover. Other cities include Braunschweig, Osnabrück, Oldenburg and Göttingen. </narr>
</top>

Reliability?

- 25 topics are sufficient under most circumstances to reliably order systems (*Sanderson & Zobel 2005*)
- Analysis of the Results of GeoCLEF 2007 hint that the results are reliable

Participation Main Task

CLEF Year	2005	2006	2007	2008
Nr. of Participants	11	17	13	**11** **5 newcomer**
Nr. of submitted Experiments	117	149	108	**131**

Approaches

- No geographic components
  - Elaborated weighting (U Berkeley)
- Specific geographic processing
  - Geo filter and gazetteer (Imperial College)
  - GeoWordNet and distance function for geo entities (U Valencia)
  - Expansion by geo coordinates (U Chengdu & U Pittsburgh)
  - NER and disambiguation, fusion by Fuzzy Borda (U Jaén & U Valencia)
  - Ontology based approach (DFKI)
  - Deep natural language processing (U Hagen)

Relevance Assessment

- Different range of meanings
  - Portuguese "monumentos"
  - English "sights"
  - German "Sehenswürdigkeiten"
- Euro Disney might be a *sight*, but it cannot be considered as a *monumento*

Relevance Assessment

- Indirect Information
  - „foreign aid in Sub-Saharan Africa „
    - Is a document on the kidnapping of an aid worker relevant?
  - „natural desasters in the Western USA“
    - Is a document on the insurance costs caused by a natural desaster relevant?

Relevance Assessment

- **Hints for problems of the systems**
  - German word for fails (Messe) was matched against similar words which have a different meaning
    - angemessen -> appropriate
    - Messer -> knife

Relevant Docs per Topic

Language	Min	Max
English	0	109
German	1	146
Portuguese	2	158

Results in a Nutshell

- How much and which geo knowledge and reasoning is necessary?
- Often, keyword based systems do well on the task
- Best system in most competitive task (many runs) uses specific geo reasoning
  - Significant?
- For most other tasks (esp. cross lingual), the best system uses no specific geo components
  - Significant?

More on GeoCLEF

Parallel Session
on Thursday 14:30 – 16:00

Please come to the Breakout Session
on Friday
and help us to form GeoCLEF 2009

GeoCLEF 2009

- Continuation of GikiP
  - Search for Wiki-Entires with geographic constraints
- Query Classification Task
  - Find geo queries in a search engine log

Breakout Session
on Friday

Overview

More on the geographic search task …

- Giorgio Di Nunzio:
  Results and Statistical Analysis
- Diana Santos:
  GikiP Task

http://www.uni-hildesheim.de/geoclef

Acknowledgments

- This work was partly done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.