An Evaluation Resource for Geographic Information Retrieval

Thomas Mandl[1], Fredric Gey[2], Giorgio Di Nunzio[3], Nicola Ferro[3], Mark Sanderson[4], Diana Santos[5], Christa Womser-Hacker[1]
[1]University of Hildesheim, Germany, [2]University of California, Berkeley, CA, USA
[3]University of Padua, Italy , [4]University of Sheffield, Sheffield, UK, [5]Linguateca, SINTEF ICT, Norway
E-mail: mandl@uni-hildesheim.de

Abstract

In this paper we present an evaluation resource for geographic information retrieval developed within the Cross Language Evaluation Forum (CLEF). The GeoCLEF track is dedicated to the evaluation of geographic information retrieval systems. The resource encompasses more than 600,000 documents, 75 topics so far, and more than 100,000 relevance judgments for these topics. Geographic information retrieval requires an evaluation resource which represents realistic information needs and which is geographically challenging. Some experimental results and analysis are reported.

1. Geographic Information Retrieval Evaluation

The Cross Language Evaluation Forum[1] (CLEF) is a large European evaluation initiative dedicated to cross-language retrieval for European languages [Peters et al. 2004]. CLEF was implemented as a consequence to the rising need for cross- and multi-lingual retrieval research and applications. CLEF provides a multi-lingual testbed for retrieval experiments. The evaluation campaign of CLEF comprises several components: the evaluation methodology, the evaluation software packages, the data collections, the topics, the overall results of the participants, the assessed results of the participants, and the calculated statistical results.

GeoCLEF[2] was the first track at an evaluation campaign dedicated to evaluating geographic information retrieval (GIR) systems ever. The aim of GeoCLEF is the provision of the necessary framework for the evaluation of GIR systems for search tasks involving both spatial and multilingual aspects. Participants are offered a TREC style ad-hoc retrieval task based on newspaper collections. GeoCLEF started as a pilot track in 2005 [Gey et al. 2006] and was a regular CLEF track since then [Gey et al. 2007, Mandl et al. 2008].

GeoCLEF evaluates the retrieval of documents with an emphasis on geographic information retrieval from text. Geographic search requires the combination of spatial and content based relevance into one result. Many research and evaluation issues surrounding geographic mono- and bilingual search have been addressed in GeoCLEF. It is still an open research question how to best combine semantic knowledge on geographic relations with vague document representations [Chaves et al 2005] as well as how to encode place knowledge in NLP [Santos & Chaves 2006]. Especially the multilingual aspect of geographic retrieval is not trivial [Gey & Carl 2004].

[1] http://www.clef-campaign.org
[2] http://www.uni-hildesheim.de/geoclef

2. Evaluation Resources

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Many documents contain some kind of spatial reference which may be important for IR. For example, to retrieve, rank and visualize search results based on a spatial dimension (e.g. “find me news stories about bush fires near Sidney”). Many challenges of geographic IR involve geographical references (geo-references) which systems need to recognized and treated properly. Documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. For example, the city *Cape Town* (English) is also *Kapstadt* (German), *Cidade do Cabo* in Portuguese and *Ciudad del Cabo* (Spanish).

For 2007, Portuguese, German and English were available as document and topic languages. There were two Geographic Information Retrieval tasks: monolingual (English to English, German to German and Portuguese to Portuguese) and bilingual (language X to language Y, where X or Y was one of English, German or Portuguese). In the first three editions of GeoCLEF, 75 topics with relevance assessments have been developed. Thus, GeoCLEF has developed a standard evaluation collection which supports long-term research.

Topic creation is a collaborative activity of the three organizing groups, who all utilize the DIRECT System provided by the University of Padua [Agosti et al. 2007]. DIRECT has been designed to extend the current IR methodology in order to provide an integrated vision of the scientific data involved in an international evaluation campaign. It offers tools to support tasks related to different areas such as, for example, the creation of the topics and the management of relevance assessments. A search utility for the collections is provided to facilitate the interactive exploration of potential topics.

Each group initially created initial versions of nine proposed topics in their language, with subsequent translation into English. Topics are meant to express a natural information need which a user of the collection might have. These candidates were subsequently checked for relevant documents in the other collections. In many cases, topics needed to be refined. For example, the topic candidate *honorary doctorate degrees at Scottish universities* was expanded to topic GC53 *scientific research at Scottish universities* due to an initial lack of documents in the German and Portuguese collections. After the translation, all topics were thoroughly checked. An example of a topic in the three languages is shown below:

```
<top lang="en">
<num>10.2452/63-GC</num>
 <title>Water quality along coastlines of the Mediterranean
      Sea</title>
  <desc>Find documents on the water quality at the coast of the
      Mediterranean Sea</desc>
  <narr>Relevant documents report on the water quality along
      the coast and coastlines of the Mediterranean Sea. The
      coasts must be specified by their names.</narr>
</top>

<top lang="pt">
  <num>10.2452/63-GC</num>
  <title>Qualidade da água na costa
      mediterrânica</title>
  <desc>Os documentos devem referir a qualidade da água nas
      praias ou costas do Mediterrâneo. </desc>
  <narr>As zonas a que se refere essa qualidade têm
      de figurar no documento. </narr>
</top>

<top lang="de">
  <num>10.2452/63-GC</num>
  <title>Wasserqualität an der Küste des Mittelmeers</title>
   <desc>Dokumente über die Wasserqualität an Küsten im
      Mittelmeer</desc>
     <narr>Relevante Dokumente berichten von der
      Wasserqualität im Mittelmeer in Zusammenhang mit den
      Namen der Küsten und Küstenabschnitte, an denen die
      Verschmutzungen aufgetreten sind.</narr>
</top>
```

The organizers aimed at creating a geographically challenging topic set. This means that explicit geographic knowledge should be necessary in order for the participants to successfully retrieve relevant documents. Keyword-based approaches only should not be favored by the topics. While many geographic searches may be well served by keyword approaches, others require a profound geographic reasoning. We speculate that, for a realistic topic set where these difficulties might be less common, most systems could perform better.

In order to achieve a geographically challenging topic set, several difficulties were explicitly included in the topics of GeoCLEF 2006 and 2007:

- Ambiguity (a church called *St. Pauls Cathedral*, exists in London and São Paulo)
- Vague geographic regions (*Near East*)
- Geographical relations beyond IN (*near Russian cities*, *along Mediterranean Coast*)
- Cross-lingual issues (Greater *Lisbon* , Portuguese: *Grande Lisboa* , German: *Großraum Lissabon*)
- Granularity below the country level (*French speaking part of Switzerland, Northern Italy*)
- Complex region shapes (*along the rivers Danube and Rhine*)
- Differences between local and national newspapers (local events are not often mentioned in national newspapers of other countries)

However, it was often difficult to develop multilingual topics which fulfilled these criteria. For example, local events which allow queries on a level of granularity below the country often do not lead to newspaper articles outside the national press. This makes the development of cross-lingual topics difficult.

The topics are used by the systems to produce results which are then joined in a document pool which is evaluated by human assessors. The spatial dimension is an additional factor in this relevance judgment process. Documents need to be relevant and geographically adequate.

The participants used a wide variety of approaches to the GeoCLEF tasks, ranging from basic IR approaches (with no attempts at spatial or geographic reasoning or indexing) to deep natural language processing (NLP) processing to extract place and topological clues from the texts and queries. Specific techniques used included (see more details in the overview paper Mandl et al. 2008):

- Ad-hoc techniques (weighting, probabilistic retrieval, language model, blind relevance feedback)
- Semantic analysis (annotation and inference)
- Geographic knowledge bases (gazetteers, thesauri, ontologies)
- Text mining
- Query expansion techniques (e.g. geographic feedback)
- Geographic Named Entity Extraction
- Geographic disambiguation
- Geographic scope and relevance models
- Geographic relation analysis
- Geographic entity type analysis
- Term expansion using Wordnet
- Part-of-speech tagging

The relevance judgments posed several problems, illustrated here in detail for the "free elections in Africa" topic: What is part of an election (or presupposed by it)? In other words, which parts are necessary or sufficient to consider that a text talks about elections: campaign, direct results, who were the winners, "tomada de posse", speeches when receiving the power, cabinet constitution, balance after one month, after a longer period?.

3. GeoCLEF Collection

The document collections for 2007 GeoCLEF experiments consisted of newspaper and newswire stories from the years 1994 and 1995 used in previous CLEF ad-hoc evaluations. The Portuguese, English and German collections contain stories covering international and national news events, therefore representing a wide variety of geographical regions and places. The English document collection contains 169,477 documents and is composed of stories from the British newspaper The *Glasgow Herald* (1995) and the American newspaper *The Los Angeles Times* (1994). The German document collection consists of 294,809 documents from the German news magazine *Der Spiegel* (1994/95), the German newspaper *Frankfurter Rundschau* (1994) and the Swiss newswire agency *Schweizer Depeschen Agentur* (SDA, 1994/95). For Portuguese, GeoCLEF 2007 utilized two newspaper collections, spanning over 1994-1995, for respectively the Portuguese and Brazilian newspapers *Público* (106,821 documents) and *Folha de São Paulo* (103,913 documents). Both are major daily newspapers in their countries. Not all material published by the two newspapers is included in the collections (mainly for copyright reasons), but every day is represented with documents. The Portuguese collections are also distributed for IR and NLP research by Linguateca as the CHAVE collection[3], recently distributed with automatic syntactic annotation as well. The English and German collections are available in a CLEF package from ELDA/ELRA.

GeoCLEF Year	Collection Languages	Topic Languages
2005 (pilot)	English, German	English, German
2006	English, German, Portuguese, Spanish	English, German, Portuguese, Spanish, Japanese
2007	English, German, Portuguese	English, German, Portuguese, Spanish, Indonesian
2008 (planned)	English, German, Portuguese	English, German, Portuguese

Table 1: GeoCLEF 2007 test collection size.

In all collections, the documents have a common structure: newspaper-specific information like date, page, issue, special filing numbers and usually one or more titles, a byline and the actual text. The document collections were not geographically tagged and contained no semantic location-specific information.

[3] http://www.linguateca.pt/CHAVE/

Language	English	German	Portuguese
Number of documents	169,477	294,809	210,734

Table 2: GeoCLEF 2007 test collection size.

A query classification task has also been conducted. The challenge for systems was the identification of the geographic queries within a real search engine query log and the recognition of the geographic and the thematic parts (Li et al. 2008). Training and test data labeled by humans was created as the test environment.

4. Results

GeoCLEF 2007 attracted 13 participating groups from nine countries. They developed or modified their systems and ran experiments with the benchmark data. All groups together submitted 108 runs for all sub tasks.

The detailed results for all sub tasks are provided in the overview paper (Mandl et al., 2008). As an example, the systems for two sub tasks of GeoCLEF 2007 are displayed in figure 1 and 2. It can be observed that the systems perform quite similarly. Furthermore, it can be seen that the performance of systems for bilingual retrieval remains weaker than for monolingual. The results show that the topics are indeed challenging and the performance of the systems lags behind typical ad-hoc topics without geographical parameters (e.g. di Nunzio et al., 2008).

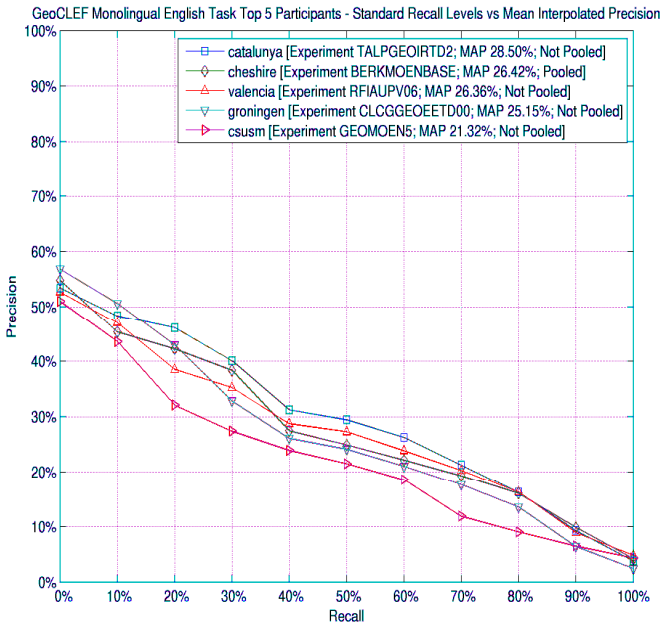

Figure 1: Results of GeoCLEF 2007: Monolingual English (Mandl et al. 2008)

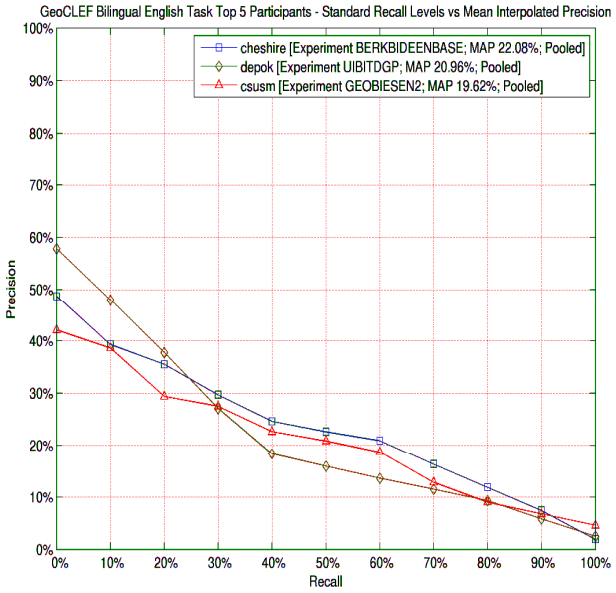

Figure 2: Results of GeoCLEF 2007: Bilingual English (Mandl et al. 2008)

The two best systems for the monolingual English task used quite different approaches although both are based on a state-of-the-art text retrieval engine. Ceshire relied solely on probabilistic text retrieval based on logistic regression and incorporating blind relevance feedback and did not include geographic reasoning (Larson 2008). On the other hand, TALP implemented several steps for incorporating geographical knowledge bases. Topic analysis and a geographic document retrieval phase complemented the text retrieval (Ferrés & Rodríguez 2008).

5. Analysis

As a first step toward the analysis, the variance between topics as well as between systems was calculated for all sub tasks for GeoCLEF 2006 and GeoCLEF 2007. These values are shown in Box-and-Whiskers diagrams which visualize the distribution of the data. For all sub tasks, we calculated the average for all systems for one topic to get the average performance for that topic. The same is done for the systems. The average performance of one system is calculated as the average of its performance for all topics. This can also be interpreted as the mean average precision (MAP) usually given as result for retrieval tests (di Nunzio et al., 2008). The distribution of all average topic and system performances is illustrated in the figures 3a through 3d and 4a though 4d. The dark line in the middle of the box shows the median and the box represents the interval with 50% of the data points. The end points of the antennas show the minimum and maximum.

No dramatic differences between the distribution of GeoCLEF 2006 and GeoCLEF 2007 occur. Overall, the maximal performance for topics lies lower especially for the bilingual tasks. Nevertheless, the median performance for topics varies more between languages than between the two GeoCLEF editions.

As for many other information retrieval evaluations, the variance is much larger for the topics than for the systems. This has also been shown by test theoretic analysis (Bodoff & Li, 2007). This fact has led to ideas for topic specific optimization approaches (Mandl & Womser-Hacker, 2005, Savoy, 2007). Moreover, it has led to serious doubts about the validity and reliability of tests in information retrieval. Since the variance between topics is so large, the results can depend much on the arbitrary choice of topics.

To measure this effect, a method which uses simulations with sub sets of the original topic set has been established (Zobel, 1998). The simulation uses smaller sets of topics and compares the resulting ranking of the systems to the ranking obtained when using all topics. If the systems are ranked very differently when only slightly smaller sets are used, the reliability is considered as small. The rankings can be compared by counting the number of position changes in the system ranking (swap rate). For GeoCLEF, such a simulation has been carried out as well (Mandl, 2008). The rankings have been compared by a rank correlation coefficient. A result is shown in figure 5. it can be observed that the system ranking remain stable even until topic sets of size 11 which is less than half of the original topic set. This stability is surprisingly high and shows that the GeoCLEF results are considerably reliable.

The variance between systems has also led to optimization efforts. In order to illustrate how much one could achieve by combining systems effectively for the topics for which they are most appropriate, an analysis on the most difficult and the easiest topics for GeoCLEF 2006 was carried out. Tables 3a through 3c show these topics and gives the average performance for all systems for them and the performance of the system with the best result for that topic. It can be seen that there is large room for improvement.

The tables also make it obvious that the level of the difficulty of a topics is dependent on the language. For the hard topics, only one topics appears twice in the list. For the easy topics, three topics (30, 32, 46) rank among the best performing topics for English and Portuguese. However, there is no overlap with German. This proves how much the difficulty or the success of the systems depends on the collection behind the system and the pecularities of the language.

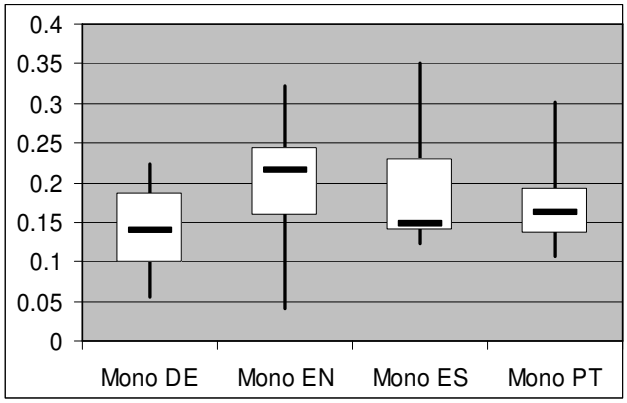

Figure 3a: GeoCLEF 2006. Monolingual Systems

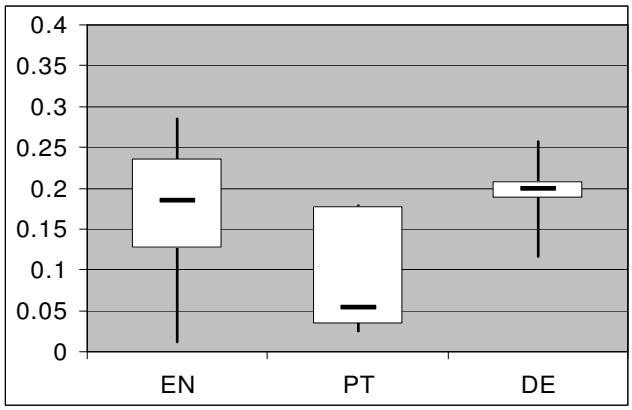

Figure 4a: GeoCLEF 2007. Monolingual Systems

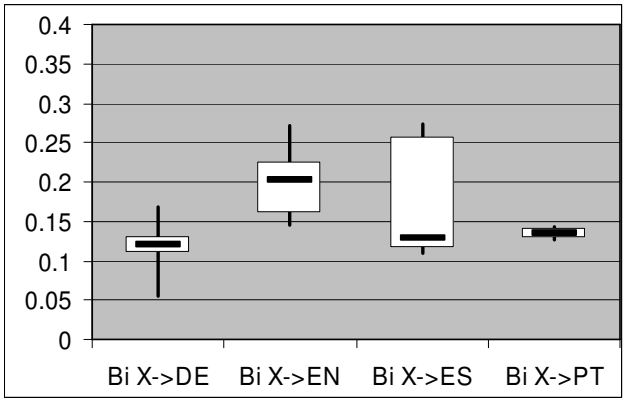

Figure 3b: GeoCLEF 2006. Bilingual Systems

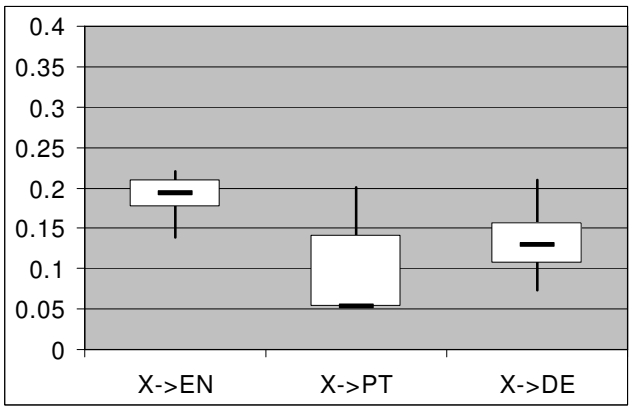

Figure 4b: GeoCLEF 2007. Bilingual Systems

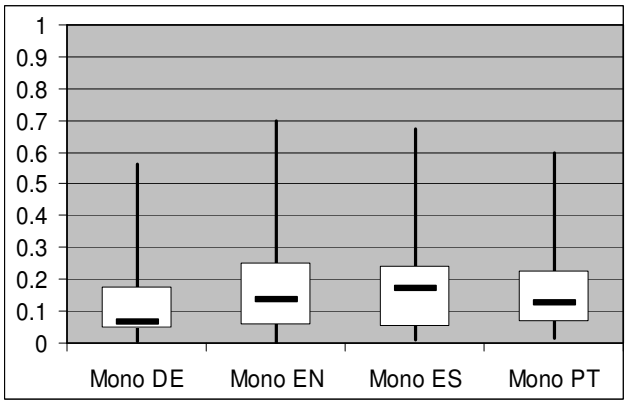

Figure 3c: GeoCLEF 2006. Monolingual Topics

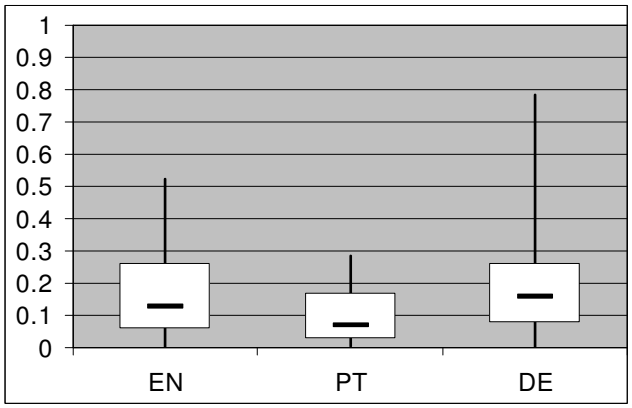

Figure 4c: GeoCLEF 2007. Monolingual Topics

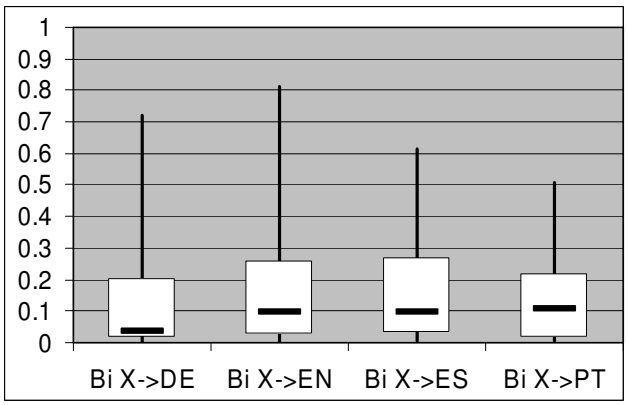

Figure 3d: GeoCLEF 2006. Bilingual Topics

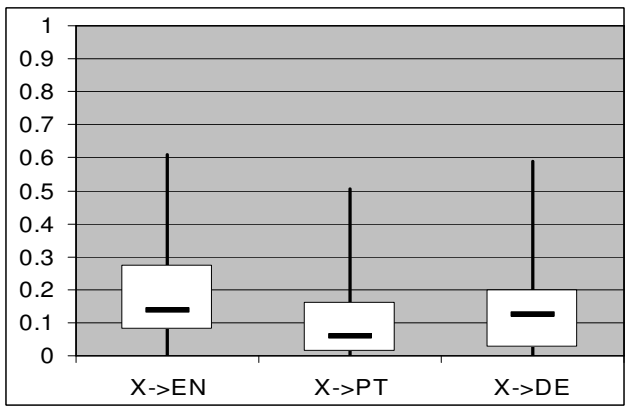

Figure 4d: GeoCLEF 2007. Bilingual Systems

Topic	AP	Max AP
48) Fishing in Newfoundland and Greenland	0.5646	0.9161
30) Car bombings near Madrid	0.53	0.7862
32) Independence movement in Quebec	0.4625	0.7861
34) Malaria in the tropics	0.3122	0.6704
40 Cities near active volcanoes	0.2285	0.4016
35) Credits to the former Eastern Bloc	0.0377	0.1231
50) Cities along the Danube and the Rhine	0.0352	0.0755
43) Scientific research in New England Universities	0.0239	0.0617
27) Cities within 100km of Frankfurt	0.0132	0.0359
26) Wine regions around rivers in Europe	0.0034	0.0172

Table 3a: GeoCLEF 2006: Hardest and easiest topics for mono-lingual German.

Topic	AP	Max AP
36) Automotive industry around the Sea of Japan	0	0
41) Shipwrecks in the Atlantic Ocean	0.0123	0.25
43) Scientific research in New England Universities	0.0290	0.3115
27) Cities within 100km of Frankfurt	0.0373	0.1257
47) Champions Cup games near the Mediterranian	0.0450	0.1914
49) ETA in France	0.2953	0.6429
46) Forest fires in Northern Portugal	0.5205	1
30) Car bombings near Madrid	0.5443	1
48)Fishing in Newfoundland and Greenland	0.6182	0.9086
32) Independence movement in Quebec	0.6988	0.9631

Table 3b: GeoCLEF 2006: Hardest and easiest topics for mono-lingual English

6. Outlook

GeoCLEF has created an important evaluation resource for geographic information retrieval. Spatially challenging topics have been developed and interesting experiments have been submitted. The search task based on newspaper collections will continue to run at CLEF 2008. The test collection developed for GeoCLEF is the first GIR test collection available to the GIR research community.

Topic	AP	Max AP
43) Pesquisa científica em universidades da Nova Inglaterra	0.0174	0.0809
36) Indústria automóvel no Mar do Japão	0.0177	0.0952
35) Empréstimos ao antigo Bloco de Leste	0.0202	0.1505
33) Competições desportivas internacionais no Ruhr	0.0232	0.0988
26) Regiões vinícolas à beira de rios na Europa	0.0239	0.1959
42) Eleições regionais no norte da Alemanha	0.3060	0.524
46) Fogos florestais no norte de Portugal	0.3510	0.5987
30) Carros armadilhados em Madrid e arredores	0.3965	0.6566
32) Movimento para a independência do Québec	0.5878	0.8587
48) Pescas na Terra Nova e na Gronelândia	0.5979	0.9241

Table 3c: GeoCLEF 2006: Hardest and easiest topics for mono-lingual Portuguese

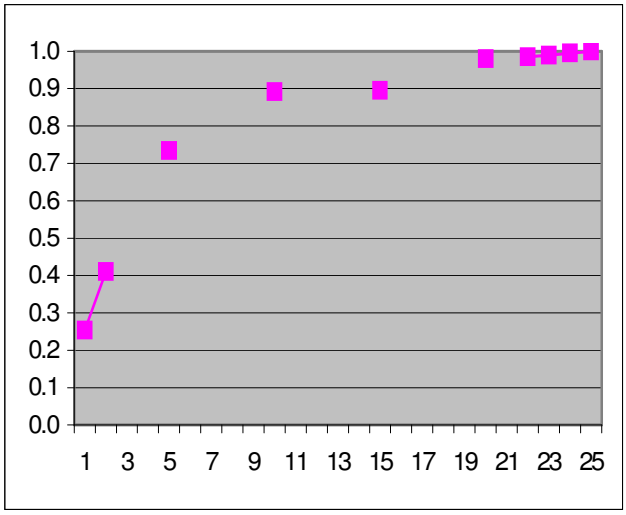

Figure 5: Correlation of Topic Subsets with final Result GeoCLEF 2007, Monolingual German (Mandl 2008)

For future GeoCLEF campaigns, both an image and a question answering task are envisioned to investigate geographic issues in a wider variety of retrieval applications.

7. Acknowledgements

The work for Portuguese was done in the scope of Linguateca, contract no. 339/1.3/C/NAC, a project jointly funded by the Portuguese Government and the European Union.

8. References

Agosti, M., Di Nunzio, G. M., Ferro, N. (2007). A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In Sakay, T., Sanderson, M. (Eds.): *Proc. 1st International Workshop on Evaluating Information Access* (EVIA 2007), National Institute of Informatics, Tokyo, Japan, pp. 62--73.

Bodoff, D., Li, P. (2007). Test Theory for Assessing IR Test Collections. In: *30th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR) Amsterdam. pp. 367--374.

Chaves, M., Martins, B., Silva, M.J. (2005). Challenges and Resources for Evaluating Geographical IR. In: *Proceedings of the 2nd International Workshop on Geographic Information Retrieval, CKIM*. Bremen, Germany. pp. 65--69.

Gey, F., R. Larson, M. Sanderson, H. Joho, P. Clough, Petras, V. (2006). GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In: *6th Workshop of the Cross-Language Evaluation Forum: CLEF 2005*. Springer [LNCS 4022]

Gey, F., Larson, R., Sanderson, M., Bishoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Di Nunzio, G., Ferro, N. (2007). GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In *7th Workshop of the Cross-Language Evaluation Forum: CLEF 2006*, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer [LNCS 4730] 2007. pp. 852--876.

Gey, F.C., Carl, K. (2004). Geotemporal Access to Multi-lingual Documents. In: *Workshop on Geographic Information Retrieval* (GIR at SIGIR 2004). http://www.geo.unizh.ch/~rsp/gir/abstracts/gey.pdf

Larson, R. (2008). Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines. In: Peters, C. et al. (Eds.): *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear. Preprint: http://www.clef-campaign.org/

Ferrés, D., Rodríguez, H. (2008). TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering. Peters, C. et al. (Eds.): *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear. Preprint: http://www.clef-campaign.org/

Li, Z., Wang, C., Xie, X., Ma, W.-Y. (2008). Query Parsing Task for GeoCLEF2007 Report. In: Peters, C. et al. (Eds.): *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear. Preprint: http://www.clef-campaign.org/2007/working_notes/LI_OverviewCLEF2007.pdf

Mandl, T. (2008). Die Reliabilität der Evaluierung von Information Retrieval Systemen am Beispiel von GeoCLEF. In: *Datenbank-Spektrum: Zeitschrift für Datenbanktechnologie und Information Retrieval*. Heft 24. pp. 40--47.

Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xing, X. (2008). GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol et al. (Eds.). 8th *Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear

Mandl, T., Womser-Hacker, C. (2005). The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: *Proceedings of 2005 ACM SAC Symposium on Applied Computing* (SAC). Santa Fe, New Mexico, USA. March 13.-17. pp. 1059--1064.

Di Nunzio, G., Ferro, N., Mandl, T., Peters, C. (2008). CLEF 2007: Ad Hoc Track Overview. In: *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [Lecture Notes in Computer Science] to appear. Preprint: http://www.clef-campaign.org

Santos, D., Chaves, M. (2006). The place of place in geographical IR. In: *Proceedings of GIR06, the 3rd Workshop on Geographic Information Retrieval* (SIGIR 2006) Seattle, pp. 5--8.

Savoy, J. (2007). Why do successful search systems fail for some topics. In: *Proceedings of the ACM Symposium on Applied Computing* (Seoul, Korea, March 11-15). SAC '07. ACM Press, pp. 872--877.

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '98) Melbourne, Australia, ACM Press, New York. pp. 307--314.