

Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF

Fredric Gey¹, Ray Larson¹, Mark Sanderson²,
Kerstin Bischoff³, Thomas Mandl³, Christa Womser-Hacker³,
Diana Santos⁴, Paulo Rocha⁴, Andres Montoyo⁵

¹University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@simms.berkeley.edu

²Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk

³Information Science, University of Hildesheim, GERMANY
{mandl|womser}@uni-hildesheim.de

⁴Linguateca, SINTEF ICT, NORWAY
Diana.Santos@sintef.no, Paulo.Rocha@di.uminho.pt

⁵University of Alicante, SPAIN

WITH

Giorgio M. Di Nunzio⁶, Nicola Ferro⁶

⁶Department of Information Engineering, University of Padua, Italy
{dinunzio|ferro}@dei.unipd.it

Abstract

This is the third year of the evaluation of geographic information retrieval (GeoCLEF) within the Cross-Language Evaluation Forum (CLEF). GeoCLEF 2006 presented topics and documents in four languages (English, German, Portuguese and Spanish). After two years of evaluation we are beginning to understand the challenges to both Geographic Information Retrieval from text and of evaluation of the results of geographic information retrieval. This poster enumerates some of these challenges to evaluation and comments on the limitations encountered in the first two evaluations.

Keywords: *Geographic Information Retrieval, Cross-Language Information Retrieval*

1 Introduction

For the past several years researchers in Information Retrieval have been taking a closer look at the opportunities and challenges to Geographic Information Retrieval. A series of three workshops on the special nature of Geographic Information Retrieval have been held jointly with SIGIR 2004, CIKM 2006, SIGIR 2006 [4, 8, 9] with a current call for papers associated with CIKM 2007.

Geographic Information Retrieval (GIR) is concerned with the locational aspect of searching text collections. More will be said about this in subsequent sections.

The first three authors above, together with Paul Clough of Sheffield and Hideo Joho (now of Glasgow) proposed in 2004 at the Cross Language Evaluation Forum (CLEF) that a pilot track of Geographic Information Retrieval be undertaken within CLEF 2005. After approval by the CLEF steering committee, the organizers from UC Berkeley and University of Sheffield choose two languages, English and German, to be the document languages and named the track GeoCLEF. Eleven groups participated in GeoCLEF 2005 and seventeen groups in 2006. Papers and results of GeoCLEF 2005 have been published by Springer [7]. As we will see, the evaluation taken in GeoCLEF has thus far been one of manual judging of pooled results using measures of recall/precision for performance measurement. Other approaches have included evaluation of components such as Named Entity Recognition for place names [6] and spatial relevance judgment [1].

2. Collections and Approach

The GeoCLEF document collections have been those news collections traditionally used for ad-hoc retrieval evaluation within CLEF. For both GeoCLEF 2005 and 2006, the English document collection consisted of 169,477 documents from The Glasgow Herald (1995) and the Los Angeles Times (1994). The German collection consisted of 294,809 documents from Der Spiegel, Frankfurter Rundschau (1994) and the Swiss news agency

SDA (1994/95). For Portuguese, GeoCLEF 2006 used newspaper collections, covering 1994-1995, for the Portuguese newspaper *Público* (106,821 documents) and the Brazilian *Folha de São Paulo* (103,913 documents). The Spanish collection of GeoCLEF 2006 was from the Spanish newspapers EFE 1994-1995. This collections are distributed by the Spanish Agency EFE (www.efe.es). EFE 1994 are made up of 215,738 documents and EFE 1995 of 238,307 documents.

3 GeoCLEF Topics

Twenty-five topics were evaluated in GeoCLEF 2005 and 2006. In 2005, 15 topics were chosen from prior CLEF ad-hoc topics modified for geography. A typical 2005 topic was;

```
<top> <num>GC001</num>
  <orignum>C084</orignum>
  <EN-title>Shark Attacks off Australia and
  California</EN-title>
  <EN-desc>Documents will report any
  information relating to shark attacks on
  humans.</EN-desc>
  <EN-concept>Shark Attacks</EN-concept>
  <EN-spatialrelation>near</EN-
  spatialrelation>
  <EN-location>Australia</EN-location>
  <EN-location>California</EN-location>
</top>
```

Note that spatial operators were included in the topic. Operators were abandoned in 2006.

One of the problems with GeoCLEF 2005 was that ordinary ac-hoc techniques of blind feedback (without the use of external resources) performed better than a host of geographically oriented methods. Thus, in GeoCLEF 2006, the geography became pre-eminent and the challenge was substantially increased, as in the following topic:

```
<top><num>GC027</num>
  <EN-title>Cities within 100km of
  Frankfurt</EN-title>
  <EN-desc>Documents about cities within 100
  kilometers of the city of Frankfurt in Western
  Germany</EN-desc>
  <EN-narr>Relevant documents discuss cities
  within 100 kilometers of Frankfurt am Main
  Germany, latitude 50.11222, longitude
  8.68194. To be relevant the document must
  describe the city or an event in that city.
  Stories about Frankfurt itself are not
  relevant</EN-narr>
</top>
```

In evaluating this topic we encounter two problems:

1) Ambiguity of specification of the term ‘city.’ Is a city a major city in Germany as defined by the World Gazetteer¹? Is it size-related (the USA Census Bureau tabulates cities of population greater than 30,000 in their City Data Book)? Or is it the technical definition of ‘populated place’ found in the NGA Geographic Names Server (GNS)²?

2) Given a choice of definition, how to develop a reference set of ‘cities.’ We took the liberal assumption and queried the GNS to “to extract and prepare a spreadsheet of populated places whose latitude and longitude was within a distance of 100 km of the latitude and longitude of Frankfurt. This spreadsheet contained 5342 names and was made available to all groups doing assessment.” [2]

4 GeoCLEF topic classification

The overview of GeoCLEF 2006 has presented the following classification of geographic topics to be applied to search against natural language free-text documents.

- 1 non-geographic subject restricted to a place (music festivals in Germany) [only kind of topic in GeoCLEF 2005]
- 2 geographic subject with non-geographic restriction (rivers with vineyards) [new kind of topic added in GeoCLEF 2006]
- 3 geographic subject restricted to a place (cities in Germany)
- 4 non-geographic subject associated to a place (independence, concern, economic handlings to favour/harm that region, etc.) Examples: *independence of Quebec, love for Peru* (as often remarked, this is frequently, but not necessarily, associated to the metonymical use of place names)
- 5 non-geographic subject that is a complex function of place (for example, place is a function of topic) (*European football cup matches, winners of Eurovision Song Contest*)
- 6 geographical relations among places (*how are the Himalayas related to Nepal? Are they inside? Do the Himalaya mountains cross Nepal's borders? etc.*)
- 7 geographical relations among (places associated to) events (*Did Waterloo occur more north than the battle of X? Were the findings of Lucy more to the south than those of the Cromagnon in Spain?*)
- 8 relations between events which require their precise localization (*was it the same river that flooded last year and in which killings occurred in the XVth century?*)

¹ <http://world-gazetteer.com/>

² <http://earth-info.nga.mil/gns/html/index.html>

Not all of these types of topics of user interest have yet been included within GeoCLEF.

5 Evaluation Challenges

We divide the challenges to evaluation within GeoCLEF into two areas – evaluation problems which occur within the general area of ad-hoc retrieval of multilingual collections and evaluation problems specifically arising out of the particular character of Geographic Information Retrieval. Problems of the first character include:

- Insufficient judgment pools – for GeoCLEF 2006 only three groups participated using the Portuguese language (monolingual), and 4 groups participated, respectively, in Spanish and German, while 16 groups did English GIR. By contrast in GeoCLEF 2005 five groups participated in the German GIR task, while ten groups tried English GIR.

We expect to present results at the EVIA workshop of the effect of insufficient judgment pools.

Problems of the second character include:

- Finding the (consistent) basis for relevance assessment.

For example topic 50:

<top><num>GC050</num>

<EN-title>Cities along the Danube and the Rhine</EN-title>

<EN-desc>Documents describe cities in the shadow of the Danube or the Rhine</EN-desc>

<EN-narr>Relevant documents should contain at least a short description of cities through which the rivers Danube and Rhine pass, providing evidence for it. The Danube flows through nine countries (Germany, Austria, Slovakia, Hungary, Croatia, Serbia, Bulgaria, Romania, and Ukraine). Countries along the Rhine are Liechtenstein, Austria, Germany, France, the Netherlands and Switzerland</EN-narr> </top>

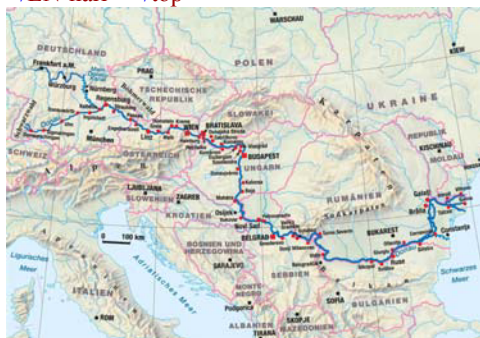


Figure 1: Danube river map

To our knowledge, all organizing groups found this topic to be difficult to assess. The groups utilized printed atlases of Europe in order to obtain the basis for judgments (see the map above for the Danube³).

However, to be consistent with assessment of Topic 26 above, a complex geospatial query should have been applied to the GNS by taking the digital lat-long coordinates the Rhine and Danube Rivers and computing a geospatial cover with a perpendicular line to each line segment specifying the river, together with circles covering the join points on the where the polygonal curve changes to a new line segment. The cover could be computed for some threshold distance (1, 3, 10km) from the river. Then any city or town whose latitude and longitude lay within that cover would become a candidate upon which to do an assessment search.

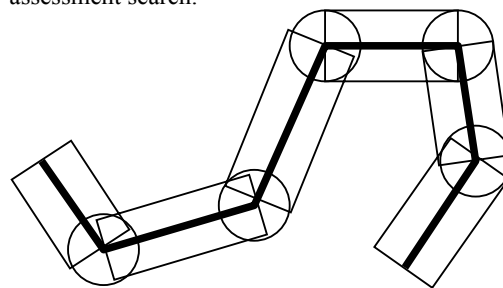


Figure 2: Geographic Cover for River Coordinates

A more complex approximation would run quadratic splines through the lat/lon points specifying the river's coordinates and then compute a cover perpendicular to each parabolic spline point.

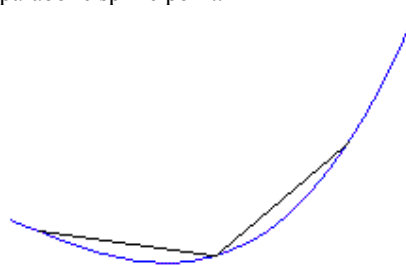


Figure 3: Example quadratic spline

Larson and Frontiera have evaluated spatial ranking methods for digital library search based upon minimum bounding rectangles and convex hulls for geographic areas [5]

³ Map adapted from http://www.danube-river.org/en_fluss-karte.html

6 Conclusions and future directions

This poster paper has begun to outline the particular problems with evaluation of Geographic Information Retrieval from text. The insights we have obtained so far come from our experience in two years of running the GeoCLEF evaluation within the CLEF workshop. Our insights are still preliminary and we are conducting further investigations.

7 Acknowledgment

The initial work of evaluation of the German language for GeoCLEF 2005 depended upon relevance judgments made by Vivien Petris (now with GESIS in Bonn, Germany). The work on Portuguese was supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI. The GeoCLEF website is maintained by University of Sheffield, England, United Kingdom.

8 References

- [1] P Clough, H Joho, R Purves: Judging the Spatial Relevance of Documents for GIR. *Proceedings of ECIR 2006*: 548-552
- [2] F Gey, et al: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview, Springer LNCS, forthcoming 2007.
- [3] F Gey, et al: GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview, *Accessing*

Multilingual Information Repositories, Springer LNCS 4022, May 2006.

- [4] C B Jones, R Purves (Eds.): *Proceedings of the 2005 Workshop On Geographic Information Retrieval, GIR 2005*, Bremen, Germany, November 4, 2005. ACM 2005, ISBN 1-59593-165-1
- [5] R Larson and P Frontiera, Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries, in *Proceedings of ECDL 2004, Research and Advanced Technology for Digital Libraries*, Bath, UK, September 2004, 45-56.
- [6] Martins, B., Silva, M. J., and Chaves, M. S. 2005. Challenges and resources for evaluating geographical IR. In *Proceedings of the 2005 Workshop on Geographic information Retrieval* (Bremen, Germany, November 04 - 04, 2005). GIR '05. ACM Press, New York, NY, 65-69.
- [7] C Peters, F Gey, J Gonzalo, H Müller, G J. F. Jones, M Kluck, B Magnini, M de Rijke (Eds.): *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, Vienna, Austria, 21-23 September, 2005, Springer LNCS 4022, 2006
- [8] R Purves, C B Jones: Workshop on geographic information retrieval held at SIGIR 2006. SIGIR Forum 40(2): (December 2006)
- [9] R Purves, C B Jones: Workshop on geographic information retrieval, SIGIR 2004. SIGIR Forum 38(2): 53-56 (2004)