GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview

Fredric Gey[1], Ray Larson[1], Mark Sanderson[2], Kerstin Bischoff [3], Thomas Mandl [3], Christa Womser-Hacker [3], Diana Santos[4], Paulo Rocha[4], Giorgio M. Di Nunzio[6], and Nicola Ferro[6]

[1] University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@sims.berkeley.edu
[2]Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk
[3] Information Science, University of Hildesheim, Germany
{mandl, womser}@uni-hildesheim.de
[4] Linguateca, SINTEF ICT, Norway
Diana.Santos@sintef.no, Paulo.Rocha@di.uminho.pt
[6]Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

Abstract. After being a pilot track in 2005, GeoCLEF advanced to be a regular track within CLEF 2006. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR): retrieval for topics with a geographic specification. For GeoCLEF 2006, twenty-five search topics were defined by the organizing groups for searching English, German, Portuguese and Spanish document collections. Topics were translated into English, German, Portuguese, Spanish and Japanese. Several topics in 2006 were significantly more geographically challenging than in 2005. Seventeen groups submitted 149 runs (up from eleven groups and 117 runs in GeoCLEF 2005). The groups used a variety of approaches, including geographic bounding boxes, named entity extraction and external knowledge bases (geographic thesauri and ontologies and gazetteers).

1 Introduction

Existing evaluation campaigns such as TREC and CLEF have not, prior to 2005, explicitly evaluated geographical relevance. The aim of GeoCLEF is to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants are offered a TREC style ad hoc retrieval task based on existing CLEF collections. GeoCLEF 2005 was run as a pilot track to evaluate retrieval of multilingual documents with an emphasis on geographic search on English and German document collections. Results were promising, but it was felt that more work needed to be done to identify the research and evaluation issues surrounding geographic information retrieval from text. Thus 2006 was the second year in which GeoCLEF was run as a track within CLEF. For 2006, two additional document languages were added to GeoCLEF, Portuguese and Spanish. GeoCLEF was a

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 852 – 876, 2007.

collaborative effort by research groups at the University of California, Berkeley (USA) , the University of Sheffield (UK), University of Hildesheim (Germany), Linguateca (Norway and Portugal), and University of Alicante (Spain). Seventeen research groups (increased from eleven in 2005) from a variety of backgrounds and nationalities submitted 149 runs (up from 117 in 2005) to GeoCLEF.

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Given that many documents contain some kind of spatial reference, there are examples where geographical references (geo-references) may be important for IR. For example, to retrieve, re-rank and visualize search results based on a spatial dimension (e.g. "find me news stories about riots near Dublin City"). In addition to this, many documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. For example, the city of Cologne (English) is also Köln (German), Colónia in Portuguese from Portugal, Colônia in Brazilian Portuguese, and Colonia (Spanish). Queries with names such as this may require an additional translation step to enable successful retrieval.

For 2006, Spanish and Portuguese, in addition to German and English, were added as document languages, while topics were developed in all four languages with topic translations provided for the other languages. In addition the National Institute of Informatics of Tokyo, Japan translated the English version of the topics to Japanese. There were two Geographic Information Retrieval tasks: monolingual (English to English, German to German, Portuguese to Portuguese and Spanish to Spanish) and bilingual (language X to language Y, where X or Y was one of English, German, Portuguese or Spanish and additionally X could be Japanese).

2 Document Collections Used in GeoCLEF

The document collections for this year's GeoCLEF experiments are all newswire stories from the years 1994 and 1995 used in previous CLEF competitions. Both the 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 consists of 169,477 documents and was 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 news agency SDA (1994/95). Although there are more documents in the German collection, the average document length (in terms of words in the actual text) is much larger for the English collection. In both 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 or contained any other location-specific information. For Portuguese, GeoCLEF 2006 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. The collections are also distributed for IR and NLP research by Linguateca as the CHAVE collection (www.linguateca.pt/CHAVE/, see URL for DTD and document examples). The Spanish collection was composed of Spanish newspapers EFE 1994-1995 distributed by the Spanish Agency EFE (http://www.efe.es/). EFE 1994 are made up of 215,738 documents and EFE 1995 of 238,307 documents.

3 Generating Search Topics

A total of 25 topics were generated for this year's GeoCLEF. Topic creation was shared among the four organizing groups, each group creating initial versions of their proposed topics in their language, with subsequent translation into English. In order to support topic development, Ray Larson indexed all collections with his Cheshire II document management system and this was made available to all organizing groups for interactive exploration of potential topics. While the aim had been to prepare an equal number of topics in each language, ultimately only two topics (GC026 and GC027) were developed in English. Other original language numbers were German, 8 topics (GC028 to GC035), Spanish, 5 topics (GC036 to GC040) and Portuguese, 10 topics (GC041 to GC050). This section will discuss the creation of the spatially-aware topics for the track.

3.1 Topic Generation

In GeoCLEF 2005 some criticism arose about the lack of geographical challenges of the topics (favouring keyword-based approaches) and the German task was inherently more difficult because several topics had no relevant documents in the German collections. Therefore geographical and cross-lingual challenge and equal distribution across language collections was considered central during topic generation. Topics should vary according to the granularity and kind of geographic entity and should require adequate handling of named entities within the process of translation (e.g. regarding decompounding, transliteration or translation).

For English topic generation, Fred Gey simply took two topics he had considered in the past (Wine regions around rivers in Europe and Cities within 100 kilometers of Frankfurt, Germany) and developed them. The latter topic (GC027) evolved into an exact specification of the latitude and longitude of Frankfurt am Main (to distinguish it from Frankfurt an der Oder) in the narrative section. Interactive exploration verified that documents could be found which satisfied these criteria on the basis of geographic knowledge by the proposer (i.e. the Rhine and Moselle valleys of Germany and cities Heidelberg, Koblenz, Mainz, and Mannheim near Frankfurt).

The German group at Hildesheim started with brain storming on interesting geographical notions and looking for potential events via the Cheshire II Interface, we unfortunately had to abandon all smaller geographic regions soon. Even if a suitable number of relevant documents could be found in one collection, most times there were few or no respective documents in the other language collections. This may not be surprising, because within the domain of news criteria like (inter)national relevance, prominence, elite nation or elite person besides proximity, conflict/negativism

and continuity etc. (for an overview see Eidlers[2]) are assumed to affect what will become a news article. Thus, the snow conditions or danger of avalanches in Grisons (canton in Switzerland) may be reported frequently by the Swiss news agency *SDA* or even German newspapers, whereas the British or American newspapers may not see the relevance for their audience. In addition, the geographically interesting issue of tourism in general is not well represented in the German collection. As a result well known places and larger regions as well as international relevant or dramatic concepts had to be focused on, although this may not reflect all user needs for GIR systems (see also Kluck & Womser-Hacker[7]). In order not to favor systems relying purely on keywords we concentrated on more difficult geographic entities like historical or political names used to refer geographically to a certain region and imprecise regions like *the Ruhr* or *the Middle East*. Moreover some topics should require the use of external geographic knowledge e.g. to identify cities onshore of the *Sea of Japan* or *the Tropics.* The former examples introduce ambiguity or translation challenges as well. *Ruhr* could be the river or the area in Germany and *the Middle East* may be translated to German *Mittlerer Osten*, which is nowadays often used, but would denote a slightly different region. The naming of the *Sea of Japan* is difficult as it depends on the Japanese and Western perspective, whereas in Korea it would be named *East Sea (of Korea)*. After checking such topic candidates for relevant documents in other collections we proposed eight topics, which we thought would contribute to a topic set varying in thematic content and system requirements.

The GeoCLEF topics proposed by the Portuguese group (a total of 10) were discussed between Paulo Rocha and Diana Santos, according to an initial typology of possible geographical topics (see below for a refined one) and after having scrutinized the frequency list of proper names in both collections, manually identifying possible places of interest. Candidate topics were then checked in the collections, using the Web interface to the AC/DC project [8], to investigate whether they were well represented. We included some interesting topics from a Portuguese (language) standpoint, including "ill-defined" or at least little known regions in an international context, such as *norte de Portugal* (North of Portugal) or *Nordeste brasileiro* (Brazilian Northeast). Basically, they are very familiar and frequently used concepts in Portuguese, but have not a purely geographical explanation. Rather, they have a strongly cultural and historical motivation. We also inserted a temporally dependent topic (outdated Champion's Cup, now Champion's League – and already in 1994-1995 as well, but names continue their independent life in newspapers and in folk's stock of words). This topic is particularly interesting, since it in addition concerns "European" football, where one of the partners is (non-geographically-European) Israel.

We also strove to find topics which included more geographical relations than mere "in" (homogeneous region), as well as different location types as far as grain and topology are concerned. As to the first concern, note that although "shipwrecks in the Atlantic Ocean" seem to display an ordinary "in"-relation, shipwrecks are often near the coasts, and the same is still more applicable about topics such as "fishing in Newfoundland", where it is presupposed that you fish on the sea near the place (or that you are concerned with the impact of fishing to Newfoundland). Likewise, anyone who knows what ETA stands for would at once expect that "ETA's activities in France" would be mostly located in the French Basque country (and not anywhere in France).

For the second concern, that of providing different granularity and/or topology, note that the geographical span of forest fires is clearly different from that of lunar of solar eclipses (a topic suggested by the German team). As to form of the region, the "New England universities" topic circumscribes the "geographical region" to a set of smaller conceptual "regions", each represented by a university. Incidentally, this topic displays another complication, because it involves a multiword named entity: not only "New England" is made up of two different words but both are very common and have a specific meaning on its own (in English and Portuguese alike). This case is further interesting because it would be as natural to say *New England* in Portuguese as *Nova Inglaterra*, given that the name is not originally Portuguese.

We should also report interesting problems caused by translation into Portuguese from topics originally stated in other languages (they are not necessarily translation problems, but were spotted because we had to look into the particular cases of those places or expressions). For example, Middle East can be equally translated by *Próximo Oriente* and *Médio Oriente*, and it is not politically neutral how precisely in that area some places are described. For example, we chose to use the word *Palestina* (together with *Israel*) and leave out Gaza Strip (which, depending on political views, might be considered a part of both). What is interesting here is that the political details are absolutely irrelevant for the topic in question (which deals with archaeological findings), but the GeoCLEF organizers (in common) decided to specify a lower level, or a higher precision description, of every location/area mentioned, in the narrative, so that a list of Middle East countries and regions had to be supplied, and agreed upon.

3.2 Format of Topic Description

The format of GeoCLEF 2006 differed from that of 2005. No explicit geographic structure was used this time, although such a structure was discussed by the organizing groups. Two example topics are shown in Figure 1.

```
<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>
<top>
<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive vaccination </EN-desc>
<EN-narr> Relevant documents state cases of malaria in tropical regions and possible
preventive measures like chances to vaccinate against the disease. Outbreaks must be of
epidemic scope. Tropics are defined as the region between the Tropic of Capricorn,
latitude 23.5 degrees South and the Tropic of Cancer, latitude 23.5 degrees North.  Not
relevant are documents about a single person's infection.</EN-narr>
</top>
```

Fig. 1. Topics GC027: Cities within 100 Kilometers of Frankfurt and GC034: Malaria in the Tropics

As can be seen, after the brief descriptions within the title and description tags, the narrative tag contains detailed description of the geographic detail sought and the relevance criteria.

3.3 Several Kinds of Geographical Topics

We came up with a tentative classification of topics according to the way they depend on place (in other words, according to the way they can be considered "geographic"), which we believe to be one of the most interesting results of our participation in the choice and topic formulation for GeoCLEF. Basically, this classification was done as an answer to the overall too simplistic assumption of first GeoCLEF[3], namely the separation between subject and location as if the two were independent and therefore separable pieces of information. (Other comments to the unsuitability of the format used in GeoCLEF can be found in Santos and Cardoso [9], and will not be repeated here.)

While it is obvious that in some (simple) cases geographical topics can be modeled that way, there's much more to place and to the place of place in the meaning of a topic than just that, as we hope this categorization can help making clear:

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?*)

Note that we here are not even dealing with the obviously equally relevant interdependence of the temporal dimension, already mentioned above, and which was actually extremely conspicuous in the preliminary discussions among this year's organizing teams, concerning the denotation of "former Eastern bloc countries" and "former Yugoslavia" now (that is, in 1994-1995). In a way, as argued in Santos and Chaves[10], which countries or regions to accept as relevant depends ultimately on the user intention (and need). Therefore, pinning down the meaning of a topic depends on geographical, temporal, cultural, and even personal constraints, that are intertwined in

a complex way, and more often than not do not allow a clear separation. To be able to make sense of these complicated interactions and arrive at something relevant for a user by employing geographical reasoning seems one of the challenges that lies ahead in future GeoCLEF tracks.

4 Approaches to Geographic Information Retrieval

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 NLP processing to extract place and topological clues from the texts and queries. Specific techniques used included:

- Ad-hoc techniques (blind feedback, German word decompounding, manual query expansion)
- Gazetteer construction (GNIS, World Gazetteer)
- Gazetteer-based query expansion
- Question-answering modules utilizing passage retrieval
- Geographic Named Entity Extraction
- Term expansion using Wordnet
- Use of geographic thesauri (both manually and automatically constructed)
- Resolution of geographic ambiguity
- NLP – part-of-speech tagging

5 Relevance Assessment

English assessment was shared by Berkeley and Sheffield Universities. German assessment was done by the University of Hildesheim, Portuguese assessment by Linguateca, and Spanish assessment by University of Alicante. All organizing groups utilized the DIRECT System provided by the University of Padua. The Padua system allowed for automatic submission of runs by participating groups and for automatic assembling of the GeoCLEF assessment pools by language.

5.1 English Relevance Assessment

The English document pool extracted from 73 monolingual and 12 bilingual (language X to) English runs consisted of 17,964 documents to be reviewed and judged by our 5 assessors or about 3,600 documents per assessor. In order to judge topic GC027 (Cities within 100km of Frankfurt), Ray Larson used data from the GeoNames Information System along with the Cheshire II geographic distance calculation function, 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. If a document in the pool contained the name of a German city or town, it was checked against the spreadsheet to see if it was within 100km of Frankfurt. Thus documents with well-known names (Mannheim, Heidelberg) were easily recognized, but Mecklenberg (where the German Grand Prix auto race is held) was not so easily

recognized. In reading the documents in the pool, we were surprised to find many Los Angeles Times documents about secondary school sports events and scores in the pool. A closer examination revealed that these documents contained the references to American students who had the same *family name* as German cities and towns. It is clear that geographic named entity disambiguation from text still needs some improvement.

5.2 German Relevance Assessment

For the pool of German monolingual and bilingual runs X2German 14.094 documents from the newspaper *Frankfurter Rundschau*, the Swiss news agency *SDA* and the news magazine *Spiegel* had to be assessed. Every assessor had to judge a number of assigned topics. Decisions on dubious cases were left open and then discussed within the group and/or the other language co-ordinators. Since many topics had clear, predefined criteria as specified in title, description and narrative, searching first the key concepts and their synonyms within the documents and then identifying their geographical reference led to rejecting the bulk of documents as irrelevant. Depending on the geographic entity asked for, manual expansion, e.g., the country names of the Middle East and their capitals, was done to query the DIRECT System provided by the University of Padua. Of course, such a list could never be complete and available resources would not be comprehensive enough to capture all possible expansions (e.g. we could not verify the river Code on the island of Java). Thus skimming over the text was often necessary to capture the documents main topic and geographical scope.

While judging relevance was generally easier for the short news agency articles of *SDA* with their headlines, keywords and restriction to one issue, *Spiegel* articles took rather long to judge, because of their length and essay-like stories often covering multiple events etc. without a specific narrow focus. Many borderline cases for relevance resulted from uncertainties about how broad/narrow a concept term should be interpreted and how explicit the concept must be stated in the document (e.g. do parked cars destroyed by a bomb correspond to a car bombing? Are attacks on foreign journalists and the Turkish invasion air attacks to be considered relevant as fulfilling the concept of combat?). Often it seems that for a recurring news issue it is assumed that facts are already known, so they are not explicitly cited. To keep the influence of order effects minimal is critical here.

Similarly, assessing relevance regarding the geographical criterion brought up a discussion on specificity *wrt* implicit inclusion. In all cases, reference to the required geographic entity had to be explicitly made, i.e., a document reporting about Fishing in the Norwegian Sea or the Greenland Sea without mentioning e.g. a certain coastal city in Greenland or Newfoundland was not considered relevant. Moreover, the borders of oceans and its minor seas are often hard to define (e.g. does Havana, Cuba border the Atlantic Ocean?). Figuring out the location referred to was frequently difficult, when the city mentioned first in an article could have been the domicile of the news agency or/and the city some event occurred in. This was especially true for GC040 *active volcanoes* and for GC027 *cities within 100km from Frankfurt*, with Frankfurt being the domicile of the *Frankfurter Rundschau*, which formed part of the collection. Problems with fuzzy spatial relations or imprecise regions on the other hand did not figure very prominently as they were defined in the extended narratives

(e.g. “near” Madrid includes only Madrid and its outskirts) and the documents to be judged did not contain critical cases. However, one my have argued against the decision to exclude all districts of Frankfurt as they do not form own cities, but have a common administration.

The topic on *cities around Frankfurt* (GC027) was together with GC050 about *cities along the Danube and the Rhine* the most difficult one to judge. Although a list of relevant cities containing more than 4000 names was provided by Ray Larson, this could not be used efficiently for relevance assessment to query the DIRECT system. Moreover, the notion of an event or a description made assessment even more time-consuming. We queried about 40 or 50 prominent relevant cities and actually read every document except tabular listings of sports results or public announcements in tabular form. Since the *Frankfurter Rundschau* is also a regional newspaper, articles on nearby cities, towns and villages are frequent. Would one consider the selling of parking meters to another town an event? Or a public invitation to fruit picking or the announcement of a new vocational training as nurse? As the other assessors did not face such a problem, we decided to be rather strict, i.e. an event must be something popular, public and have a certain scope or importance (not only for a single person or a certain group) like concerts, strikes, flooding or sports. In a similar manner, we agreed on a narrower interpretation of the concept of description for GC026 and for GC050 as something unique or characteristic to a city like statistical figures, historical reviews or landmarks. What would be usually considered a description was not often found due to the kind of collection, likewise relevant documents for GC045 *tourism in Northeast Brazil* were also few. While the *SDA* news agency articles will not treat traveling or tourism, such articles may sometimes be found in *Frankfurter Rundschau* or *Spiegel*, but there is no special section on that issue.

Finally, for topic GC027 errors within the documents from the *Frankfurter Rundschau* will have influenced retrieval results: some articles have duplicates (sometimes even up to four versions), different articles thrown together in one document (e.g. one about *Frankfurt* and one about *Wiesbaden*), sentences or passages of articles are missing. Thus a keyword approach may have found many relevant documents, because *Frankfurt* was mentioned somewhere in the document.

5.3 Portuguese Relevance Assessment

Details of Portuguese group’s assessment are as follows: The assessor tried to find the best collection of keywords – based on the detailed information in the narrative and his/her knowledge of the geographical concepts and subjects involved – and queried the DIRECT system. Often there was manual refinement of the query after finding new spellings in previous hits (note that our collections are written in two different varieties of Portuguese). For example, for topic GC050, "cities along the Danube and the Rhine", the following (final) query was used: *Danúbio Reno Ulm Ingolstadt Regensburg Passau Linz Krems Viena Bratislava Budapeste Vukovar Novi Sad Belgrado Drobeta-Turnu Severin Vidin Ruse Brăila Galaţi Tulcea Braila Galati Basel Basiléia Basileia Estrasburgo Strasbourg Karlsruhe Carlsruhe Mannheim Ludwigshafen Wiesbaden Mainz Koblenz Coblença Bona Bonn Colónia Colônia Cologne*

Düsseldorf Dusseldorf Dusseldórfia Neuss Krefeld Duisburg Duisburgo Arnhem Nederrijn Arnhemia Nijmegen Waal Noviomago Utrecht Kromme Rijn Utreque Rotterdam Roterdão. A similar strategy was used for cities within 100 km from Frankfurt am Main (GC027), where both particular cities were mentioned, as well as words like *cidade* (city), Frankfurt, *distância* (distance), and so on. Obviously, the significant passages for all hits were read, to assess whether the document actually mentioned cities near Frankfurt.

5.4 Spanish Relevance Assessment

For the evaluation of the Spanish documents in the GeoCLEF task, the research group of Language and Information Systems at the University of Alicante, Spain followed the following procedure. The returned documents for each topic from the GeoCLEF collection have been assigned to a member of the group. Each member had to read the question and to identify the relevant keywords. Afterwards, these keywords have been searched in the document together with the geographic names that appeared in the documents. The names were queried in the GeoNames database in order to determine whether this location corresponded to the necessary latitude and magnitude. In addition, the assessors read the title, the narrative and the content of the document, and on the basis of this information decided whether the answer is relevant with the presented topic. The total number of assessors who took part in the evaluation procedure was 36.

5.5 Challenges to Relevance Assessment

One of the major challenges facing the assessors in GeoCLEF 2006 was the substantial increase in the level of geographic knowledge required to assess particular topics. As discussed above, topic GC027 (cities around Frankfurt) was assessed after creating a custom list of cities within 100 km of Frankfurt from the NGA gazetteer. However, for GC050 (cities along the Danube and the Rhine rivers), no equivalent database was created, and the creation of such a database would have posed major challenges. The details of this challenge are described in a separate paper presented at the workshop on Evaluation of Information Access in Tokyo in May 2007 [4].

6 GeoCLEF Performance

6.1 Participants and Experiments

As shown in Table 1, a total of 17 groups from 8 different countries submitted results for one or more of the GeoCLEF tasks - an increase on the 13 participants of last year. A total of 149 experiments were submitted, which is an increase on the 117 experiments of 2005. There is almost no variation in the average number of submitted runs per participant: from 9 runs/participant of 2005 to 8.7 runs/participant of this year.

Table 1. GeoCLEF 2006 participants – new groups are indicated by *

Participant	Institution	Country
alicante	University of Alicante	Spain
berkeley	University of California, Berkeley	United States
daedalus*	Daedalus Consortium	Spain
hagen	University of Hagen	Germany
hildesheim*	University of Hildesheim	Germany
imp-coll*	Imperial College London (imp-coll)*	United Kingdom
jaen*	University of Jaen	Spain
ms-china*	Microsoft China – Web Search and Mining Group	China
nicta	NICTA, University of Melbourne	Australia
rfia-upv	Universidad Politècnica de Valencia	Spain
sanmarcos	California State University, San Marcos	United States
talp	TALP – Universitat Politècnica de Catalunya	Spain
u.buffalo*	SUNY at University of Buffalo	United States
u.groningen*	University of Groningen	The Netherlands
u.twente*	University of Twente	The Netherlands
unsw*	University of New S. Wales	Australia
xldb	Grupo XLDB – Universidade de Lisboa	Portugal

Table 2 reports the number of participants by their country of origin.

Table 2. GeoCLEF 2006 participants by country

Country	# Participants
Australia	2
China	1
Germany	2
Portugal	1
Spain	5
The Netherlands	2
United Kingdom	1
United States	3
TOTAL	**17**

Table 3 provides a breakdown of the experiments submitted by each participant for each of the offered tasks. With respect to last year there is an increase in the number of runs for the monolingual English task (73 runs in 2006 *wrt* 53 runs of 2005) and a decrease in the monolingual German (16 runs in 2006 *wrt* 25 runs in 2005); on the other hand, there is a decrease for both bilingual English (12 runs in 2006 *wrt* 22 runs in 2005) and bilingual German (11 runs in 2006 *wrt* 17 runs in 2005). Note that the Spanish and the Portuguese collections have been introduced this year.

Table 3. GeoCLEF 2006 experiments by task – new collections are indicated by*

Participant	Monolingual Tasks				Bilingual Tasks				TOTAL
	DE	EN	ES*	PT*	X2DE	X2EN	X2ES*	X2PT*	
alicante		4	3						**7**
berkeley	2	4	2	4	2		2	2	**18**
daedalus	5	5	5						**15**
hagen	5				5				**10**
hildesheim	4	5			4	5			**18**
imp-coll		2							**2**
jaen		5				5			**10**
ms-china		5							**5**
nicta		5							**5**
rfia-upv		4							**4**
sanmarcos		5	5	4		2	3	2	**21**
talp		5							**5**
u.buffalo		4							**4**
u.groningen		5							**5**
u.twente		5							**5**
unsw		5							**5**
xldb		5		5					**10**
TOTAL	**16**	**73**	**15**	**13**	**11**	**12**	**5**	**4**	**149**

Four different topic languages were used for GeoCLEF bilingual experiments. As always, the most popular language for queries was English; German and Spanish tied for the second place. Note that Spanish is a new collection added this year. The number of bilingual runs by topic language is shown in Table 4.

Table 4. Bilingual experiments by topic language

Track	Source Language				TOTAL
	DE	EN	ES	PT	
Bilingual X2DE		11			**11**
Bilingual X2EN	7		5		**12**
Bilingual X2ES		3		2	**5**
Bilingual X2PT		2	2		**4**
TOTAL	**7**	**16**	**7**	**2**	**32**

6.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, Portuguese, and Spanish. As can be seen from Table 3, the number of participants and runs for each language was quite similar, with the exception of English, which has the greatest participation. Table 5 shows the top five groups for each target collection, ordered by mean average precision. Note that only the best run is selected

Table 5. Best entries for the monolingual track (title+description topic fields only). Performance difference between the best and the last (up to 5) group is given (in terms of average precision) – new groups indicated by *.

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Monolingual English	Part.	xldb	alicante	sanmarcos	unsw*	jaen*	
	Run	XLDBGeo ManualEN not pooled	enTD pooled	SMGeoEN4 not pooled	unswTitle-Baseline pooled	sinaiE-nEnExp4 not pooled	
	Avg. Prec.	30.34%	27.23%	26.37%	26.22%	26.11%	16.20%
Monolingual German	Part.	hagen	berkeley	hildesheim*	daedalus*		
	Run	FUHddGY YYTD pooled	BKGeoD1 pooled	HIGeoded-erun4 pooled	GCdeNtLg pooled		
	Avg. Prec.	22.29%	21.51%	15.58%	10.01%		122.68%
Monolingual Portuguese	Part.	xldb	berkeley	sanmarcos			
	Run	XLDBGeo ManualPT pooled	BKGeoP3 pooled	SMGeoPT2 pooled			
	Avg. Prec.	30.12%	16.92%	13.44%			124,11%
Monolingual Spanish	Part.	alicante	berkeley	daedalus*	Sanmarcos		
	Run	esTD pooled	BKGeoS1 pooled	GCesNtLg pooled	SMGeoES1 pooled		
	Avg. Prec.	35.08%	31.82%	16.12%	14.71%		138,48%

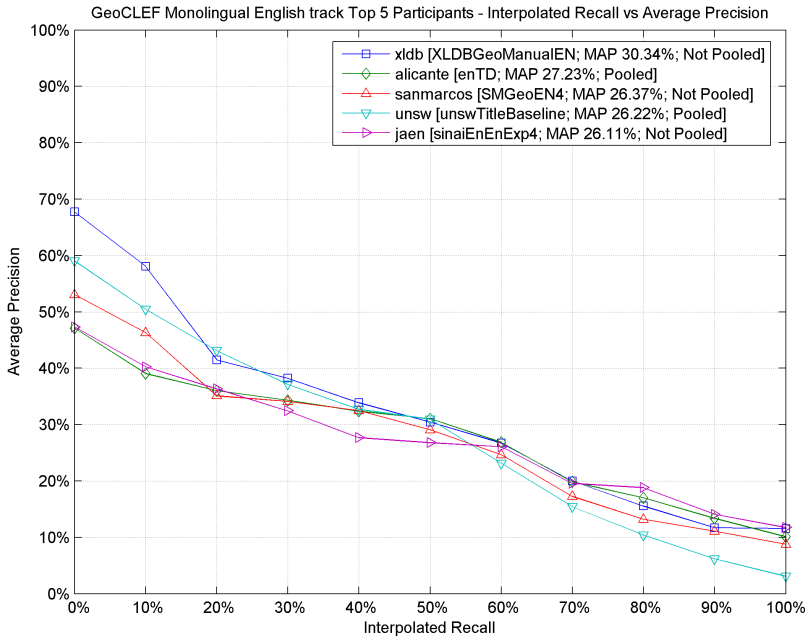


Fig. 2. Monolingual English top participants. Interpolated Recall vs. Average Precision.

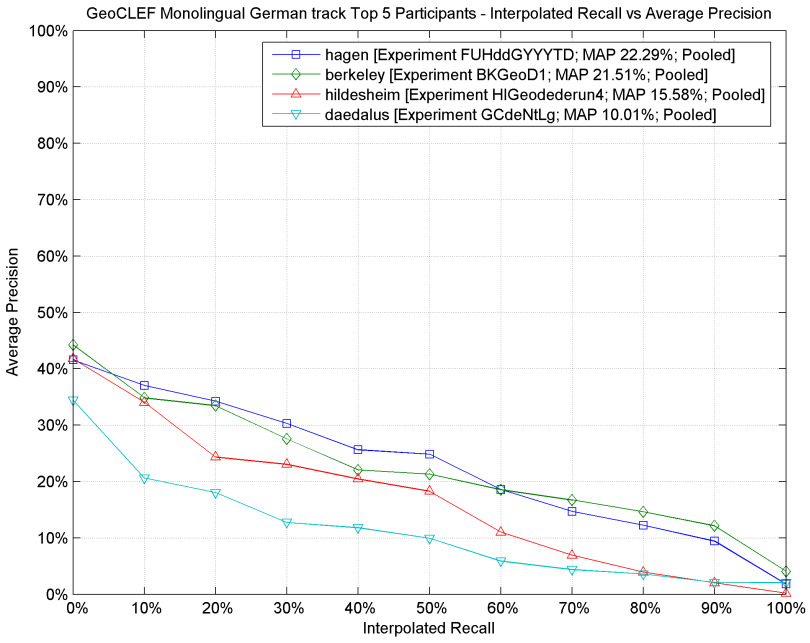


Fig. 3. Monolingual German top participants. Interpolated Recall vs. Average Precision.

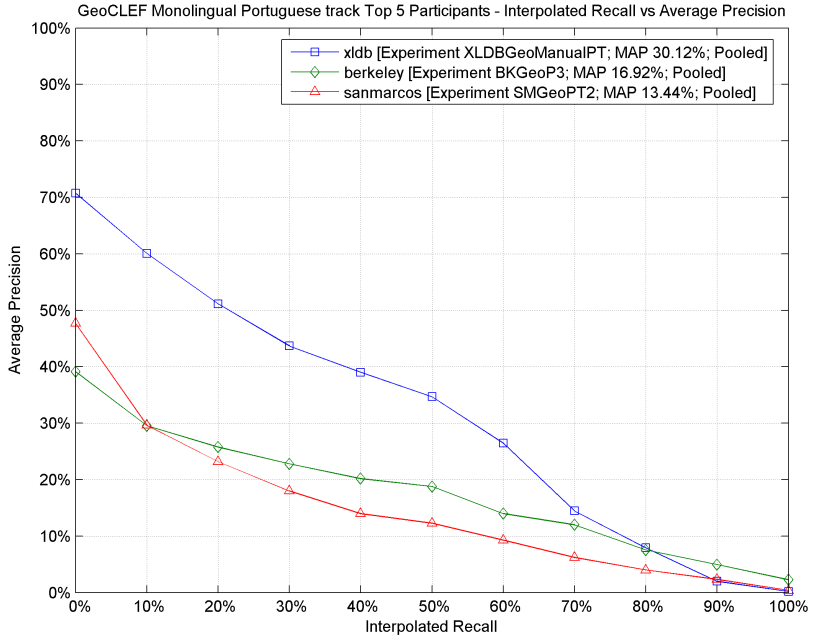


Fig. 4. Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

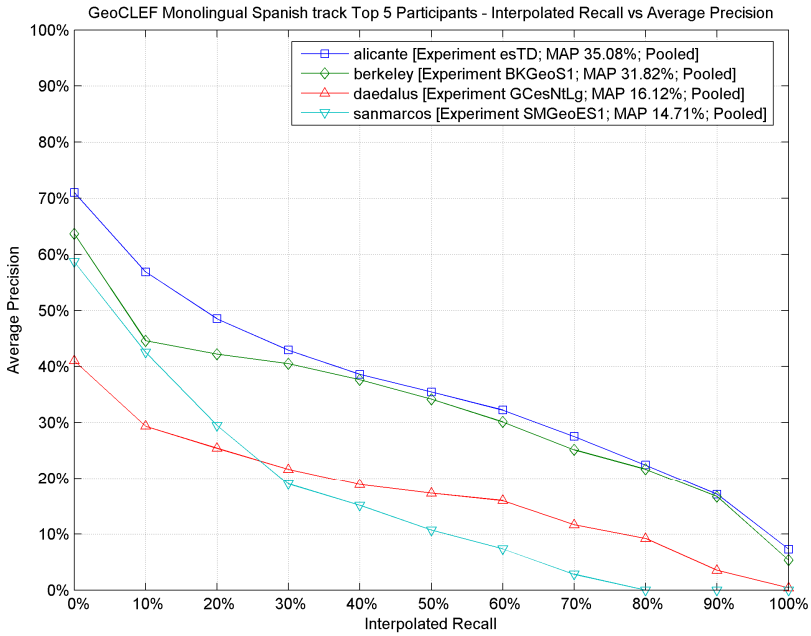


Fig. 5. Monolingual Spanish top participants. Interpolated Recall vs. Average Precision.

for each group, even if the group may have more than one top run. The table reports: the short name of the participating group; the run identifier, specifying whether the run has participated in the pool or not; the mean average precision achieved by the run; and the performance difference between the first and the last participant. Table 5 regards runs using title + description fields only.

Note that the top five participants contain both "newcomer" groups (i.e. groups that had not previously participated in GeoCLEF) and "veteran" groups (i.e. groups that had participated in previous editions of GeoCLEF), with the exception of monolingual Portuguese where only "veteran" groups were subscribed. Both pooled and not pooled runs are in the best entries for each track.

Figures 2 to 5 show the interpolated recall vs. average precision for top participants of the monolingual tasks.

6.3 Bilingual Experiments

The bilingual task was structured in four subtasks (X → DE, EN, ES or PT target collection). Table 6 shows the best results for this task with the same logic of Table 5. Note that the top five participants contain both "newcomer" groups and "veteran" groups, with the exception of monolingual Portuguese and Spanish where only "veteran" groups were subscribed.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines:

- X → DE: 70% of best monolingual German IR system
- X → EN: 74% of best monolingual English IR system

- X → ES: 73% of best monolingual Spanish IR system
- X → PT: 47% of best monolingual Portuguese IR system

Note that the apparently different result for Portuguese may be explained by the fact that the best group in the monolingual experiments did not submit runs for bilingual experiments. If one compares the results per groups, sanmarcos's run of Spanish to Portuguese had even better results than their monolingual Portuguese run, while berkeley's English to Portuguese achieved a similar performance degradation as the one reported for the other bilingual experiments (74%).

Table 6. Best entries for the bilingual task (title+description topic fields only). The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision) – new groups are indicated by *

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Bilingual English	Part.	jaen*	sanmarcos	Hildesheim*			
	Run	sinai-ESENEXP2 pooled	SMGeoE-SEN2 pooled	HIGeodeen-run12 pooled			
	Avg. Prec.	22.56%	22.46%	16.03%			40.74%
Bilingual German	Part.	berkeley	hagen	Hildesheim*			
	Run	BKGeoED1 pooled	FU-HedGY-YYTD pooled	HI-Geoenderun21 pooled			
	Avg. Prec.	15.61%	12.80%	11.86%			31.62%
Bilingual Portuguese	Part.	sanmarcos	berkeley				
	Run	SMGeoESPT2 pooled	BKGeoEP1 pooled				
	Avg. Prec.	14.16%	12.60%				12,38%
Bilingual Spanish	Part.	berkeley	sanmarcos				
	Run	BKGeoES1 pooled	SMGeoENES1 pooled				
	Avg. Prec.	25.71%	12.82%				100.55%

Figure 6 to 9 show the interpolated recall vs. average precision graph for the top participants of the different bilingual tasks.

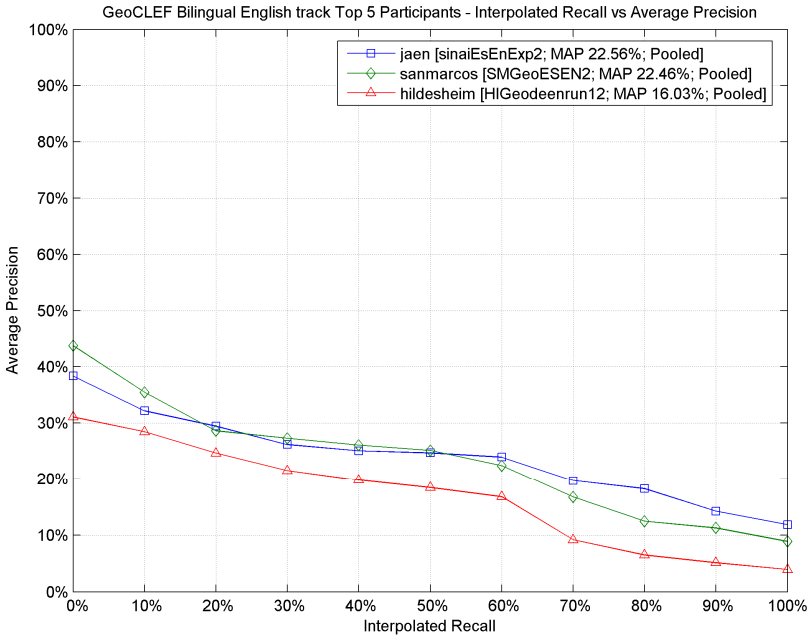


Fig. 6. Bilingual English top participants. Interpolated Recall vs Average Precision.

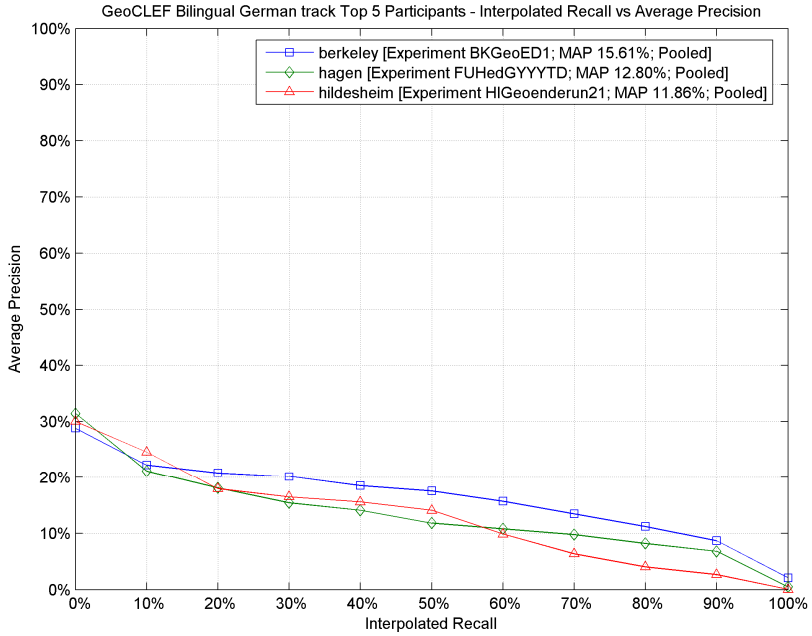


Fig. 7. Bilingual German top participants. Interpolated Recall vs Average Precision.

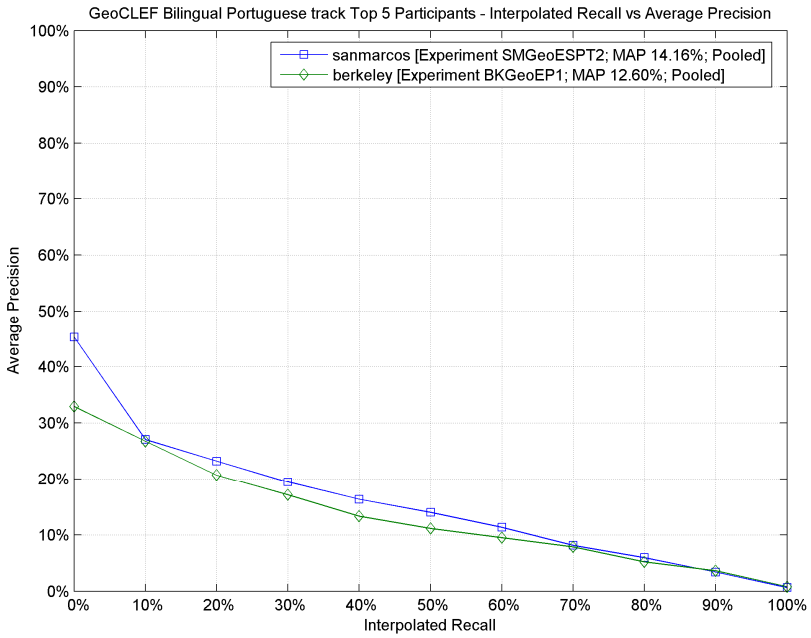


Fig. 8. Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

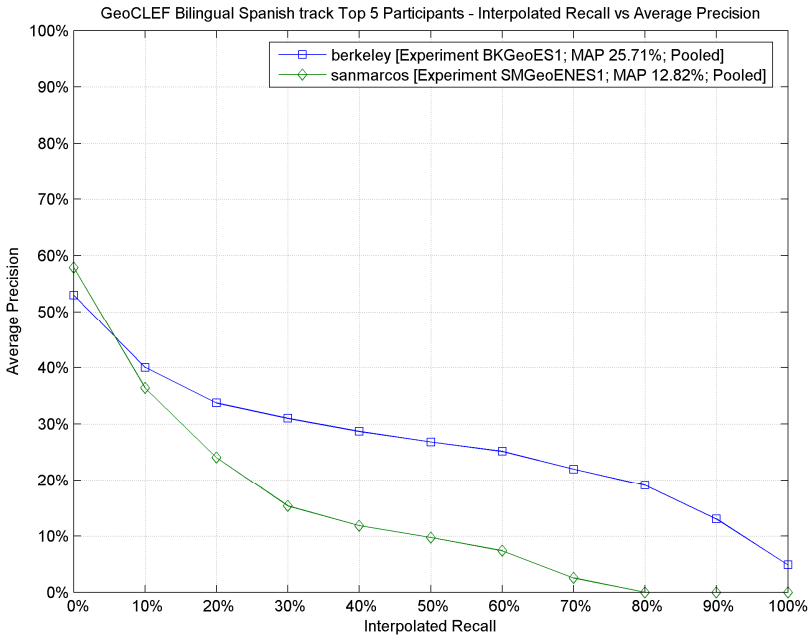


Fig. 9. Bilingual Spanish top participants. Interpolated Recall vs Average Precision.

6.4 Statistical Testing

We used the MATLAB Statistics Toolbox, which provides the necessary functionality plus some additional functions and utilities. We use the *ANalysis Of VAriance* (ANOVA) test. ANOVA makes some assumptions concerning the data be checked. Hull [5] provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. Two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test [1] and the Jarque-Bera test [6]. In the case of the GeoCLEF tasks under analysis, both tests indicate that the assumption of normality is violated for most of the data samples (in this case the runs for each participant).

In such cases, a transformation of data should be performed. The transformation for measures that range from 0 to 1 is the arcsin-root transformation:

$$\arcsin\left(\sqrt{x}\right)$$

which Tague-Sutcliffe [11] recommends for use with precision/recall measures.

Table 7. Lilliefors test for each track with (LL) and without Tague-Sutcliffe arcsin transformation (LL & TS). Jarque-Bera test for each track with (JB) and without Tague-Sutcliffe arcsin transformation (JB & TS).

Track	LL	LL & TS	JB	JB & TS
Monolingual English	2	42	32	54
Monolingual German	0	3	3	11
Monolingual Portuguese	4	5	3	13
Monolingual Spanish	2	12	4	11
Bilingual English	0	4	2	6
Bilingual German	0	1	0	4
Bilingual Portuguese	0	3	0	4
Bilingual Spanish	0	2	2	5

Table 7 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. After the transformation the analysis of the normality of samples distribution improves significantly, with the exception of the bilingual Bulgarian. Each entry shows the number of experiments whose performance distribution can be considered drawn from a Gaussian distribution, with respect to the total number of experiment of the track. The value of alpha for this test was set to 5%. The same table shows also the same analysis with respect to the Jarque-Bera test. The value of alpha for this test was set to 5%. The difficulty to transform the data into normally distributed samples derives from the original distribution of run performances which tend towards zero within the interval [0,1].

The following tables, from Table 8 to Table 13, summarize the results of this test. All experiments, regardless the topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of

absolute performance. Each table shows the overall results where all the runs that are included in the same group do not have a significantly different performance. All runs scoring below a certain group performs significantly worse than at least the top entry of the group. Likewise all the runs scoring above a certain group perform significantly better than at least the bottom entry in that group. Each table contains also a graph which shows participants' runs (y axis) and performance obtained (x axis). The circle

Table 8. Monolingual English: experiment groups according to the Tukey T Test

Run ID	Groups							
XLDBGeoManualEN	X							
sinaiEnEnExp1	X							
enTDN	X	X						
SMGeoEN3	X	X						
unswNarrBaseline	X	X						
BKGeoE4	X	X						
BKGeoE3	X	X	X					
SMGeoEN1	X	X	X					
SMGeoEN4	X	X	X					
unswTitleBaseline	X	X	X					
sinaiEnEnExp4	X	X	X	X				
enTD	X	X	X	X				
rfiaUPV02	X	X	X	X				
BKGeoE2	X	X	X	X	X			
rfiaUPV04	X	X	X	X	X			
sinaiEnEnExp2	X	X	X	X	X			
BKGeoE1	X	X	X	X	X			
MuTdnTxt	X	X	X	X	X	X		
sinaiEnEnExp5	X	X	X	X	X	X		
rfiaUPV01	X	X	X	X	X	X		
UBManual2	X	X	X	X	X	X		
SMGeoEN5	X	X	X	X	X	X		
MuTdTxt	X	X	X	X	X	X		
UBGTDrf1	X	X	X	X	X	X		
UBGTDrf2	X	X	X	X	X	X		
msramanual	X	X	X	X	X	X		
MuTdnManQexpGeo	X	X	X	X	X	X		
sinaiEnEnExp3	X	X	X	X	X	X		
UBGManual1	X	X	X	X	X	X		
MuTdRedn	X	X	X	X	X	X		
rfiaUPV03	X	X	X	X	X	X		
unswTitleF46	X	X	X	X	X	X		
UAUJAUPVenenExp1	X	X	X	X	X	X		
XLDBGeoENAut05	X	X	X	X	X	X		
MuTdQexpPrb	X	X	X	X	X	X		
CLCGGeoEE11	X	X	X	X	X	X	X	
CLCGGeoEE2	X	X	X	X	X	X	X	X
XLDBGeoENAut03_2	X	X	X	X	X	X	X	X
msrawhitelist	X	X	X	X	X	X	X	X
ICgeoMLtdn	X	X	X	X	X	X	X	X
msralocal	X	X	X	X	X	X	X	X
msratext	X	X	X	X	X	X	X	X
CLCGGeoEE1	X	X	X	X	X	X	X	X
utGeoTIBm	X	X	X	X	X	X	X	X
utGeoTdnIBm	X	X	X	X	X	X	X	X
XLDBGeoENAut03	X	X	X	X	X	X	X	X
CLCGGeoEE5	X	X	X	X	X	X	X	X
utGeoTIB	X	X	X	X	X	X	X	X
CLCGGeoEE10	X	X	X	X	X	X	X	X
XLDBGeoENAut02	X	X	X	X	X	X	X	X
HIGeoenenrun3	X	X	X	X	X	X	X	X
ICgeoMLtd	X	X	X	X	X	X	X	X
HIGeoenenrun1n	X	X	X	X	X	X	X	X
msraexpansion	X	X	X	X	X	X	X	X
HIGeoenenrun1	X	X	X	X	X	X	X	X
GCenAtLg	X	X	X	X	X	X	X	X
TALPGeoIRTD1		X	X	X	X	X	X	X
TALPGeoIRTDN1		X	X	X	X	X	X	X
GCenAA		X	X	X	X	X	X	X
HIGeoenenrun2n			X	X	X	X	X	X
utGeoTdnIB			X	X	X	X	X	X
enTDNGeoNames			X	X	X	X	X	X
GCenNtLg			X	X	X	X	X	X
TALPGeoIRTDN3				X	X	X	X	X
HIGeoenenrun2					X	X	X	X
GCenAO					X	X	X	X
utGeoTdIB						X	X	X
GCenNA						X	X	X
TALPGeoIRTDN2							X	X
unswNarrF41							X	X
TALPGeoIRTD2							X	X
unswNarrMap								X

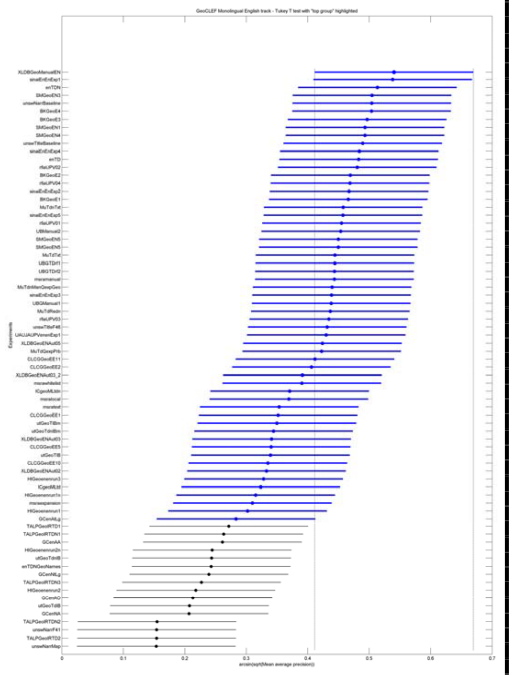


Table 9. Monolingual German: experiment groups according to the Tukey T Test

Run ID	Groups		
BKGeoD1	X		
FUHddGYYYTD	X		
FUHddGYYYTDN	X	X	
FUHddGYYYMTDN	X	X	X
BKGeoD2	X	X	X
FUHddGNNNTD	X	X	X
HIGeodederun4n	X	X	X
HIGeodederun4	X	X	X
FUHddGNNNTDN	X	X	X
GCdeNtLg	X	X	X
HIGeodederun6	X	X	X
HIGeodederun6n	X	X	X
GCdeNA	X	X	X
GCdeAtLg	X	X	X
GCdeAA		X	X
GCdeAO			X

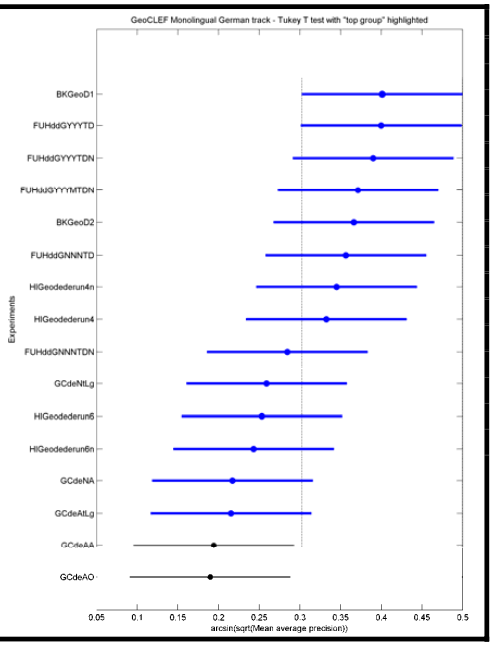


Table 10. Monolingual Portuguese: experiment groups according to the Tukey T Test

Run ID	Groups		
XLDBGeoManualPT	X		
XLDBGeoPTAut05	X		
XLDBGeoPTAut02	X	X	
XLDBGeoPTAut03	X	X	X
BKGeoP3		X	X
BKGeoP4		X	X
BKGeoP1		X	X
BKGeoP2		X	X
XLDBGeoPTAut03_2		X	X
SMGeoPT3			X
SMGeoPT1			X
SMGeoPT4			X

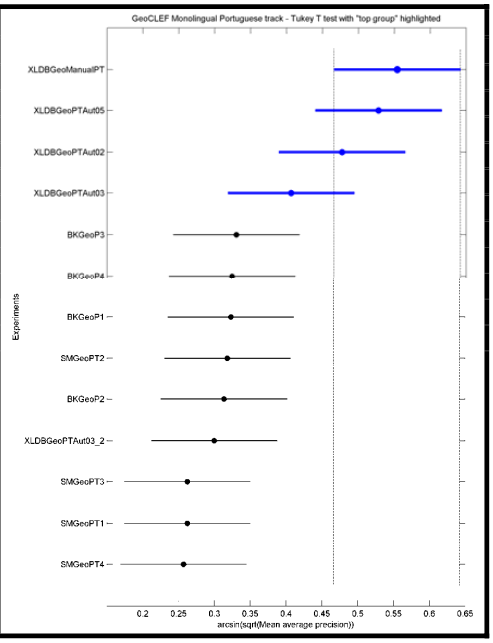


Table 11. Monolingual Spanish: experiment groups according to the Tukey T Test

Run ID	Groups			
esTD	X			
esTDN	X			
BKGeoS1	X	X		
BKGeoS2	X	X	X	
GCesNtLg		X	X	X
SMGeoES2		X	X	X
SMGeoES5		X	X	X
SMGeoES1		X	X	X
SMGeoES3		X	X	X
GCesAtLg			X	X
SMGeoES4			X	X
GCesAA				X
GCesAO				X
esTDNGeoNames				X
GCesNA				X

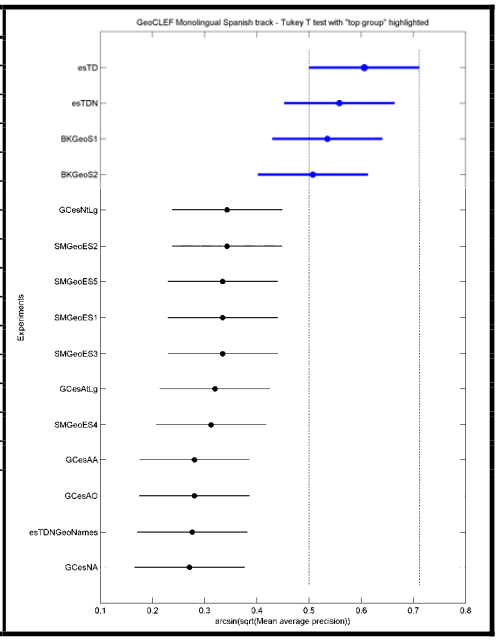


Table 12. Bilingual English: experiment groups according to the Tukey T Test

Run ID	Groups	
SMGeoESEN1	X	
sinaiEsEnExp1	X	
sinaiEsEnExp2	X	X
sinaiEsEnExp3	X	X
sinaiDeEnExp2	X	X
SMGeoESEN2	X	X
sinaiDeEnExp1	X	X
HIGeodeenrun11n	X	X
HIGeodeenrun12	X	X
HIGeodeenrun13n	X	X
HIGeodeenrun11	X	X
HIGeodeenrun13		X

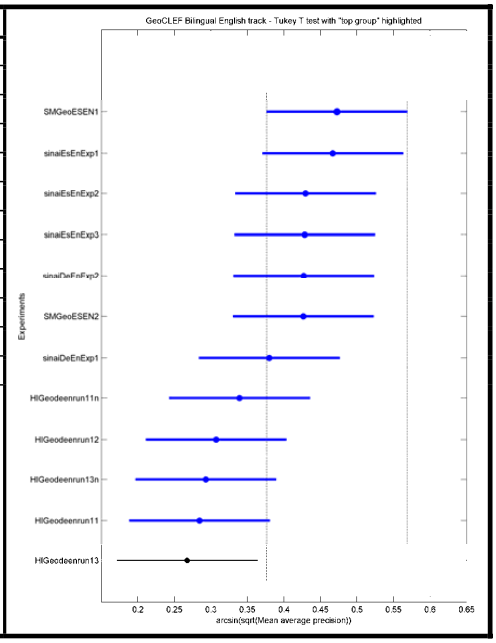
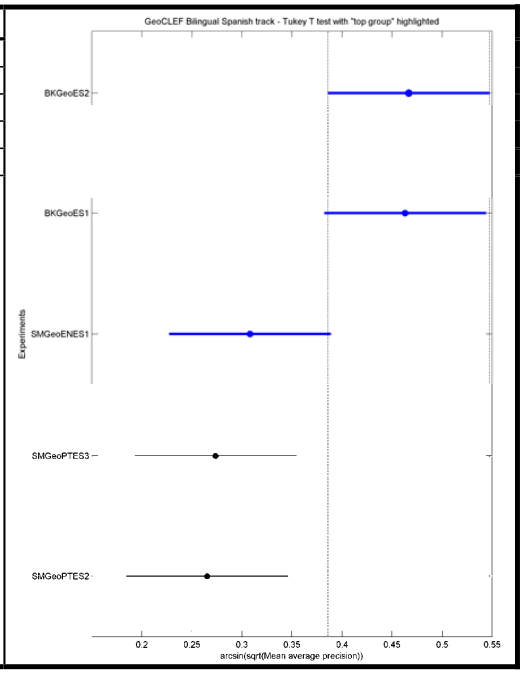


Table 13. Bilingual Spanish: experiment groups according to the Tukey T Test

Run ID	Groups	
BKGeoES2	X	
BKGeoES1	X	
SMGeoENES1	X	X
SMGeoPTES2		X
SMGeoPTES3		X



indicates the average performance while the segment shows the interval in which the difference in performance is not statistically significant; for each graph the best group is highlighted.

Note that there are no tables for Bilingual German and Bilingual Portuguese since, according to the Tukey T, all the experiments of these tasks belong to the same group.

6.5 Limitations to the Non-english Experiments, Particularly Portuguese

One must be cautious about drawing conclusions about system performance where only a limited number of groups submitted experimental runs. For the non-English collections this is especially true for GeoCLEF 2006. Six groups submitted German language runs, four groups submitted Spanish runs, and three groups submitted Portuguese runs. The Portuguese results are dominated by the XLDB group (the only participating group with native Portuguese language speakers). XLDB's overall results are nearly twice that of Berkeley's and more than twice that of San Marcos. XLDB contributed uniquely nearly 40% of the relevant Portuguese documents found, most of them through their manual run. Thus the Portuguese results of the other groups would have been higher without the XLDB participation.

7 Conclusions and Future Work

GeoCLEF 2006 increased the geographic challenge contained within the topics over 2005. Several topics could not be fully exploited without external gazetteer resources. Such resources were utilized by both the participants and the organizing groups doing relevance assessment. The test collection developed for GeoCLEF is the first GIR test collection available to the GIR research community. GIR is receiving increased notice both through the GeoCLEF effort as well as due to the GIR workshops held annually since 2004 in conjunction with SIGIR or CIKM.

At the GIR06 workshop held in August 2006, in conjunction with SIGIR 2006 in Seattle, 14 groups participated and 16 full papers were presented, as well as a keynote address by John Frank of MetaCarta, and a summary of GeoCLEF 2005 presented by Ray Larson. Six of the groups also were participants in GeoCLEF 2005 or 2006. Six of the full papers presented at the GIR workshop used GeoCLEF collections. Most of these papers used the 2005 collection and queries, although one group used queries from this year's collection with their own take on relevance judgments. Of particular interest to the organizers of GeoCLEF are the 8 groups working in the area of GIR who are not yet participants in GeoCLEF. All attendees at the GIR06 workshop were invited to participate in GeoCLEF for 2007.

Acknowledgments

The English assessment by the GeoCLEF organizers was volunteer labor – none of us has funding for GeoCLEF work. Assessment was performed by Hans Barnum, Nils Bailey, Fredric Gey, Ray Larson, and Mark Sanderson. German assessment was done by Claudia Bachmann, Kerstin Bischoff, Thomas Mandl, Jens Plattfaut, Inga Rill and Christa Womser-Hacker of University of Hildesheim. Portuguese assessment was done by Paulo Rocha, Luís Costa, Luís Cabral, Susana Inácio, Ana Sofia Pinto, António Silva and Rui Vilela, all of Linguateca, and Spanish by Andrés Montoyo Guijarro, Oscar Fernandez, Zornitsa Kozareva, Antonio Toral of University of Alicante. The Linguateca 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. The future direction and scope of GeoCLEF will be heavily influenced by funding and the amount of volunteer effort available.

References

[1] Conover, W.J.: Practical Nonparametric Statistics. John Wiley and Sons, New York, USA (1971)

[2] Eilders, C.: The role of news factors in media use. Technical report, Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB). Forschungsschwerpunkt Sozialer Wandel, Institutionen und Vermittlungsprozesse des Wissenschaftszentrums Berlin für Sozialforschung. Berlin (1996)

[3] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
[4] Gey, F., et al.: Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF. In: Proceedings of the 1st International Workshop on Evaluation of Information Access, Tokyo Japan, May 15, 2007, pp. 74–77 (2007)
[5] Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Korfhage, R., Rasmussen, E., Willett, P. (eds.) Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1993), pp. 329–338. ACM Press, New York (1993)
[6] Judge, G.G., Hill, R.C., Griffiths, W.E., Lutkepohl, H., Lee, T.C.: Introduction to the Theory and Practice of Econometrics, 2nd edn. John Wiley and Sons, New York, USA (1988)
[7] Kluck, M., Womser-Hacker, C.: Inside the evaluation process of the cross-language evaluation forum (CLEF): Issues of multilingual topic creation and multilingual relevance assessment. In: Proceedings of the third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas, Spain, pp. 573–576 (2002)
[8] Santos, D., Bick, E.: Providing internet access to portuguese corpora: the ac/dc project. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, Athens, 31 May-2 June 2000, pp. 205–210 (2000)
[9] Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and challenges. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
[10] Santos, D., Chaves, M.: The place of place in geographical information retrieval. In: Jones, C., Purves, R. (eds.) Workshop on Geographic Information Retrieval (GIR06), SIGIR06, Seattle, August 10, 2006, pp. 5–8 (2006)
[11] Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Sparck, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 205–216. Morgan Kaufmann Publishers, San Francisco, California (1997)