GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview

Thomas Mandl[1], Fredric Gey[2], Giorgio Di Nunzio[3], Nicola Ferro[3], Ray Larson[2], Mark Sanderson[4], Diana Santos[5], Christa Womser-Hacker[1], and Xing Xie[6]

[1] Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de, womser@uni-hildesheim.de
[2] University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@sims.berkeley.edu
[3] Department of Information Engineering, University of Padua, Italy
dinunzio@dei.unipd.it, ferro@dei.unipd.it
[4] Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk
[5] Linguateca, SINTEF ICT, Norway
Diana.Santos@sintef.no
[6] Microsoft Research Asia, Beijing, China
Xingx@microsoft.com

Abstract. GeoCLEF ran as a regular track for the second time within the Cross Language Evaluation Forum (CLEF) 2007. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR): retrieval for topics with a geographic specification. GeoCLEF 2007 consisted of two sub tasks. A search task ran for the third time and a query classification task was organized for the first. For the GeoCLEF 2007 search task, twenty-five search topics were defined by the organizing groups for searching English, German, Portuguese and Spanish document collections. All topics were translated into English, Indonesian, Portuguese, Spanish and German. Several topics in 2007 were geographically challenging. Thirteen groups submitted 108 runs. The groups used a variety of approaches. For the classification task, a query log from a search engine was provided and the groups needed to identify the queries with a geographic scope and the geographic components within the local queries.

1 Introduction

GeoCLEF[1] is the first track in an evaluation campaign dedicated to evaluating geographic information retrieval systems. 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 newspaper collections. GeoCLEF 2005 was run as a pilot track and in 2006, GeoCLEF was a regular CLEF track. GeoCLEF has continued

[1] http://www.uni-hildesheim.de/geoclef/

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 745–772, 2008.

to evaluate 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.

GeoCLEF was a collaborative effort by research groups at the University of California, Berkeley (USA), the University of Sheffield (UK), the University of Hildesheim (Germany) and Linguateca (Norway and Portugal). Thirteen research groups (17 in 2006) from a variety of backgrounds and nationalities submitted 108 runs (149 in 2006) to GeoCLEF.

For 2007, Portuguese, German and English were available as document and topic languages. There were two Geographic Information Retrieval tasks: monolingual (English, German and Portuguese) where both topics and queries were in a single language and bilingual (topics in language X to documents in language Y, where X or Y was one of English, German or Portuguese, and X could in addition be Spanish or Indonesian).

In the three editions of GeoCLEF so far, 75 topics with relevance assessments have been developed. Thus, GeoCLEF has developed a standard evaluation collection which supports long-term research.

Table 1. GeoCLEF test collection – collection and topic languages

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

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). 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). Queries with names may require an additional translation step to enable successful retrieval. Depending on the language and the culture, translation may not helpful in some cases. For example, the word *new* within *New York* is often translated in Spanish (*Nueva York*) and Portuguese (*Nova Iorque*), but never in German. On some occasions, names may be changed and a recent modification may not be well reflected within a foreign collection. E.g. there were still references to the German city *Karl-Marx-Stadt* in Spain after it had been renamed to *Chemnitz* in 1990. Geographical references are often ambiguous (e.g. there is a *St. Petersburg* also in Florida and Pennsylvania in the United States).

The query parsing (and classification) task was offered for the first time at GeoCLEF 2007. This task was dedicated to identifying geographic queries within a log file from the MSN search engine. A log of real queries was provided. Some were labeled as training data and some as test data. The task required participants to find geographic queries within the set and to further mark the geographic entities within the query. The task is briefly described in section 5.

2 GeoCLEF 2007 Search Task

Search is the main task of GeoCLEF. The following sections describe the test design adopted by GeoCLEF.

2.1 Document Collections Used in GeoCLEF 2007

The document collections for this year's GeoCLEF experiments consists of newspaper and newswire stories from the years 1994 and 1995 used in previous CLEF ad-hoc evaluations [1]. 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 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 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[2] collection [2].

Table 2. GeoCLEF 2007 test collection size

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

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.

2.2 Generating Search Topics

A total of 25 topics were generated for this year’s GeoCLEF (GC51 - GC75). Topic creation was shared among the three organizing groups, who all utilized the DIRECT

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

System provided by the University of Padua [3]. A search utility for the collections was provided within DIRECT to facilitate the interactive exploration of potential topics. Each group 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 [4]. 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 relevant documents in the German and Portuguese collections. Relevant documents were marked within the DIRECT system. After intensive discussion, a decision was made about the final set of 25 topics. Finally, all missing topics were translated into Portuguese and German and all translations were checked. The following section will discuss the creation of topics with spatial parameters for the track.

The organizers continued the efforts of GeoCLEF 2006 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 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 that, several difficulties were explicitly included into the topics of GeoCLEF 2006 and 2007:

- ambiguity (*St. Paul's 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*)

However, it was difficult to develop topics which fulfilled all 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.

For English topic generation, topics were initially generated by Mark Sanderson and tested on the DIRECT system. Additional consultation was conducted with other members of the GeoCLEF team to determine if the topics had at least some relevant documents in the German and Portuguese collections. Those found to have few such documents were altered in order ensure that at least some relevant documents existed for each topic.

The German group at Hildesheim started with brain storming on interesting geographical notions. Challenging geographic notions below the country granularity were procured. We came up with German speaking part of Switzerland, which is a vaguely defined region. A check in the collection showed that there were sport events, but not enough to specify a sport discipline. Another challenge was introduced with Nagorno-Karabakh which has many spelling variants.

The Portuguese topics were chosen in a way similar to the one suggested for the choice of ad-hoc topics in previous years [2]. The tripartite division among international, European and national, however, was reduced to national vs. international because we did not consider European as a relevant category (given that neither Portuguese nor English language newspaper collections used in CLEF are totally based in Europe): so, we chose some culturally-bound topics (Senna, crime in Grande Lisboa), some purely international or global (sharks and floods) and some related to specific regions (because of the geographic relevance to GeoCLEF).

In all cases, but especially for those focusing on a particular region (inside or outside the national borders covered by any newspaper collection), we tried to come up with a sensible user model: either a prospective tourist (St. Paul's or Northern Italy) or a cub reporter (Myanmar human rights violation or casualties in the Himalaya). In some cases, we managed to create topics whose general relevance could be either, although naturally the choices would be different for the different kind of users – consider the case of navigation in the Portuguese islands, both relevant for a tourist and for a journalist discussing the subject.

We were also intent on trying some specifically known geographically ambiguous topics, such as St. Paul's or topics where the geographical names were ambiguous with non geographic concepts, such as Madeira (means wood in Portuguese and can also mean a kind of wine).

All the topics were then tried out in the CHAVE collection, encoded in CQP [5] and available for Web search through the AC/DC[3] project [6] at in order to estimate the number of possible hits. In general, there were very few hits for all topics, as can be appreciated by the number of relevant documents per topic found in the Portuguese pool (see Table 5).

The translation of the topics leads to new challenges. One of the English topics about the Scottish town, St. Andrews, was judged to be challenging as it was more ambiguous than in English, because Santo André also denotes a village in Portugal and a city in Brazil. So this is a case where depending on the language the kind of results expected is different. While we are not defending a user model where this particular case would be relevant, we are showing that a mere topic translation (as might be effected by a cross lingual system) would not be enough if one were interested in the Scottish St. Andrews alone.

Another interesting remark is the use of the word "continent", which is very much context dependent and again therefore cannot be translated simply from "continent" to "continente", because depending on your spatial basis the continent is different. Again this requires some clever processing and/or processing for the translation.

Finally, it appears that *perto de X* (near X, or close to X) carries in Portuguese the presupposition that X is not included, and this made us consider that we would have translated better "airports near to London" by "que servem Londres" (i.e., that are used to reach London). (Although we also used the phrase aeroportos londrinos which may also include airports inside London). On the other hand, airplane crashes close to Russian cities seemed more naturally translated by "na proximidade" and not included. We used *perto* for both, but this might have been a translation weakness.

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

2.3 Format of Topic Description

The format of GeoCLEF 2007 was the same of the one of 2006 [7], in that no markup of geographic entities in the topics was provided as had been the case in 2005 [8]. Systems were expected to extract the necessary geographic information from the topic. Two examples of full topics are shown in Figure 1.

<num>10.2452/58-GC</num> <title>Travel problems at major airports near to London</title> <desc>To be relevant, documents must describe travel problems at one of the major airports close to London.</desc> <narr>Major airports to be listed include Heathrow, Gatwick, Luton, Stanstead and London City airport.</narr> </top>	<num>10.2452/75-GC</num> <title>Violation of human rights in Burma</title> <desc>Documents are relevant if they mention actual violation of human rights in Myanmar, previously named Burma.</desc> <narr>This includes all reported violations of human rights in Burma, no matter when (not only by the present government). Declarations (accusations or denials) about the matter only, are not relevant.</narr> </top>

Fig. 1. Topics GC058 and GC075

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. In some topics, lists of relevant regions or places were given.

2.4 Several Kinds of Geographical Topics

A tentative classification for geographical topics was suggested at GIR 2006 [9] and applied at GeoCLEF2006 [7]:

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 a 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?*)

This year we kept topics of both kinds 1 and 2 as last year. The major innovation and diversity introduced in GeoCLEF 2007 were more complicated geographic restriction than at previous GeoCLEF editions. The following three difficulties were introduced:

1. by specifying complex (multiply defined) geographic relations: East Coast of Scotland; Europe excluding the Alps, main roads north of Perth, Mediterranean coast, Portuguese islands, and "the region between the UK and the Continent";
2. by insisting on as politically defined regions, both smaller than countries, such as French speaking part of Switzerland, the Bosporus, Northern Italy, Grande Lisboa, or larger than countries: East European countries, Africa and north western Europe;
3. by having finer geographic subjects, such as lakes, airports, F1 circuits, and even one cathedral as place.

2.5 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 natural language processing (NLP) processing to extract place and topological clues from the texts and queries. Specific techniques used included:

- 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 (LingPipe, GATE, etc.)
- Geographic disambiguation
- Geographic scope and relevance models
- Geographic relation analysis
- Geographic entity type analysis
- Term expansion using WordNet
- Part-of-speech tagging.

2.6 Relevance Assessment

English assessment was shared by Berkeley and Sheffield Universities. German assessment was done by the University of Hildesheim and Portuguese assessment by Linguateca. The DIRECT System [3] was utilized for assessment. The system provided by the University of Padua allowed the automatic submission of runs by participating groups and supported assembling the GeoCLEF assessment pools by language.

2.6.1 English Relevance Assessment

English relevance assessment was conducted primarily by a group of ten paid volunteers from the University of Sheffield, who were paid a small sum of money for each topic assessed. The English document pool extracted from 53 monolingual and 13 bilingual (language X to) English runs consisted of 15,637 documents to be reviewed and judged by our 13 assessors or about 1,200 documents per assessor.

Table 3. GeoCLEF English 2007 Pool

Pool Size	15,637 documents • 14,987 not relevant • 650 relevant 25 topics • about 625 documents per topic
Pooled Experiments	27 out of 66 submitted experiments • monolingual: 21 out of 53 submitted experiments • bilingual: 6 out of 13 submitted experiments
Assessors	13 assessors • about 1,200 documents per assessor

The box plot of figure 2 shows the distribution of different types of documents across the topics of the English pool. In particular, the upper box shows the distribution of the number of pooled documents across the topics; as it can be noted, the distribution is a little bit asymmetric towards topics with a higher number of pooled documents and does not present outliers. The middle box shows the distribution of the number of not relevant documents across the topics; as it can be noted, the distribution is a little bit asymmetric towards topics with a lower number of not relevant documents and does not present outliers. Finally, the lower box shows the distribution of the number of relevant documents across the topics; as it can be noted, the distribution is almost symmetric; with a median number of relevant documents around 20 per topic, but it present some outliers, which are topics with a large number of relevant documents.

2.6.2 German Relevance Assessment

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. One topic required systems to find documents which report shark attacks. Documents telling the reader that a certain area is “full of sharks” were not judged as relevant.

For other topics, implicit information in the document was used for the decision. For example, the topic sport events in German speaking Switzerland led to documents where the place of a soccer game was not mentioned, but the result was included in a standardized form which indicates that the game was played in the first city

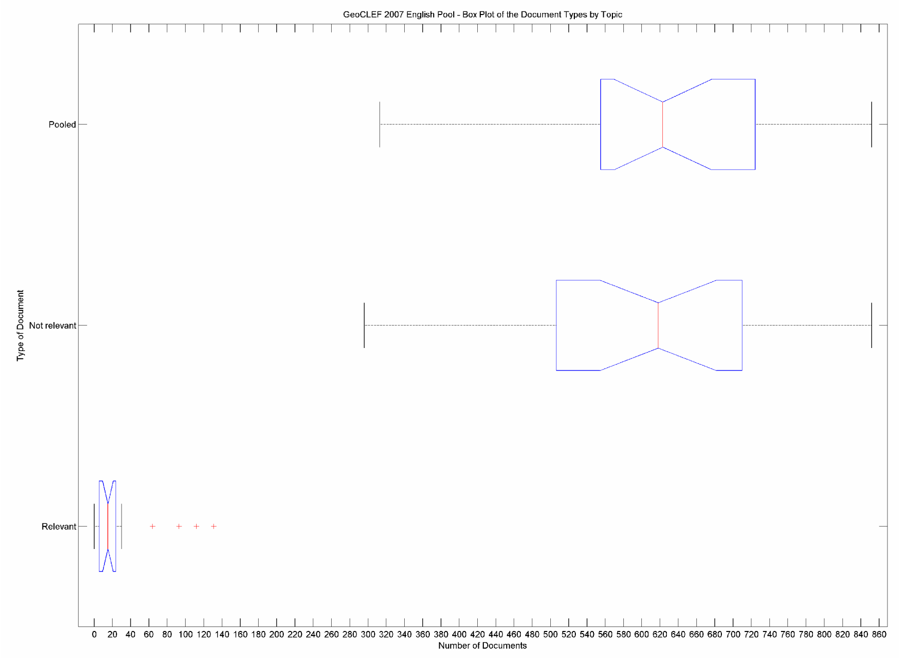


Fig. 2. GeoCLEF English 2007 Pool: distribution of the different document types

Table 4. GeoCLEF German 2007 Pool

Pool Size	15,488 documents • 14,584 not relevant • 904 relevant 25 topics • about 620 documents per topic
Pooled Experiments	24 out of 24 submitted experiments • monolingual: 16 experiments • bilingual: 8 experiments
Assessors	8 assessors • about 1,900 documents per assessor

mentioned (e.g. Lausanne - Genf 0:2, has most usually been played in Lausanne). It was also assumed that documents which report that hikers are missing in the Himalayas are relevant for the topic casualties in the Himalayas.

Many documents are at first identified as borderline cases and need to be discussed further. One topic requested topics on travel delays at London airports. One document mentioned that air travel had been delayed and some flight had to be directed to Gatwick. Because a delay at Gatwick is not explicitly mentioned, the document was regarded as not relevant.

The box plot of figure 3 shows the distribution of different types of documents across the topics of the German pool. It shows for the three sets of pooled, relevant and non relevant documents how they are distributed over the topics. This graph shows that the medium number of non relevant documents for a topic is 640. There is one topic with 300 non relevant documents which represents the minimum of the distribution. The maximum is a topic with 850 documents. The number of the topics is not given in this graph.

As it can be noted, the distribution of the pooled documents is almost symmetrical with no outliers. On the other hand, the distribution of non relevant documents is asymmetrical with a tail towards topics with a lower number of not relevant documents and does not present outliers; finally, also the distribution of the relevant documents is asymmetrical but towards topics with a greater number of relevant documents and presents outliers, which are topics with a great number of relevant documents

2.6.3 Portuguese Relevance Assessment

In addition to the problem (already reported before) that some if the news articles included in the CHAVE collection are in fact a list of "last news" which concern several different subjects (and have therefore to be read in their entirety, making it especially tiresome), we had some general problems assessing topics, which we illustrate here in detail for the "free elections in Africa" subject:

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 more time...

In fact, how far in time is information relevant? For example, does mention to the murder of the first democratically elected president in Ruanda qualify as text about free elections in Africa? And if elections took place and were subsequently annulated as in Argelia, do they count as elections or not? Also, how much indirectly conveyed information can be considered relevant? A text about the return of Portuguese citizens to Portugal after the (free) South African elections is about free elections in South Africa?

The decision on whether the elections were free or not might by arbitrary when this fact is not mentioned in the text. Should the juror assume anything? As in the case of a text about Uganda mentioning "voltou à Presidência no fim de 1980, pela via eleitoral" (X came back to presidency through the electoral path). Are either our knowledge or our opinions going to play a role on the relevance assessment, or we are supposed to just look at the document and not bring our own bias?

Finally, how much difference of opinions is relevant to a topic? Consider the following piece of news "Savimbi considera ilegais as eleições consideradas livres e justas pela ONU..." (Savimbi considers illegal the elections considered free and just by UN). Are we to stand with UN or with Savimbi, as far as the elections in Angola are concerned? (In our opinion, this text is very relevant to the subject, anyway, since it mentions, and discusses, precisely the issue of "free elections in an African country".)

Due to this (acknowledged) difficulty of assessing relevance for some topics, it would have been beneficial to have a pool of judges assessing the same documents and produce a relevance cline. Although this is currently not possible with the DIRECT system, it might make sense in the future, especially for more evaluative topics that involve complex issues.

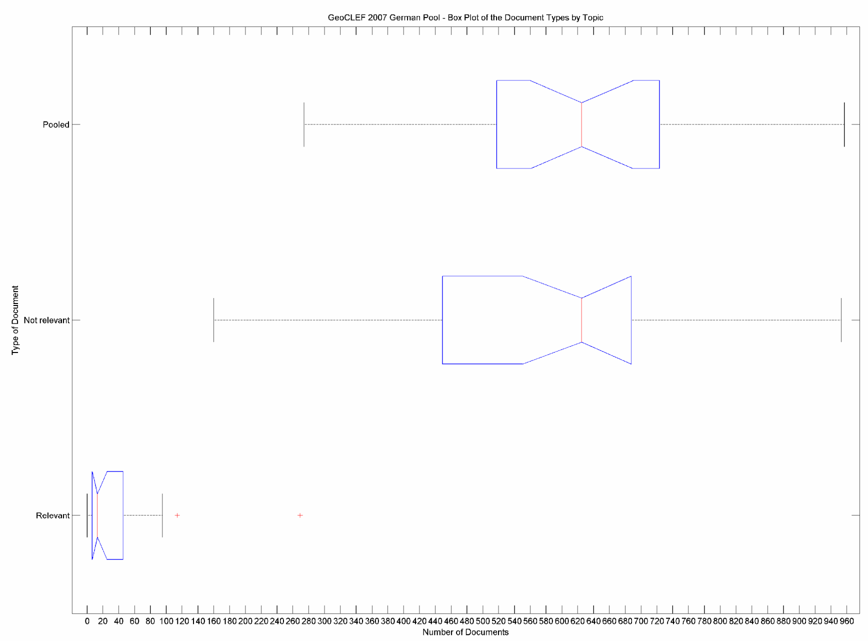


Fig. 3. GeoCLEF German 2007 Pool: distribution of the different document types

Table 5. GeoCLEF Portuguese 2007 Pool

Pool Size	15,572 documents • 14,810 not relevant • 762 relevant 25 topics • about 623 documents per topic
Pooled Experiments	18 out of 18 submitted experiments • monolingual: 11 experiments • bilingual: 7 experiments
Assessors	6 assessors • about 2,600 documents per assessor

The box plot of figure 4 shows the distribution of different types of documents across the topics of the Portuguese pool. As it can be noted the distribution of the pooled documents is a little bit asymmetrical towards topics with a lower number of pooled document and presents both upper and lower outliers, i.e. topics with many or few pooled documents; on the other hand, the distribution of not relevant documents is almost symmetrical with an outlier, which is a topic with few not relevant documents; finally, also the distribution of the relevant documents is asymmetrical towards topics with a greater number of relevant documents and presents outliers, which are topics with a great number of relevant documents.

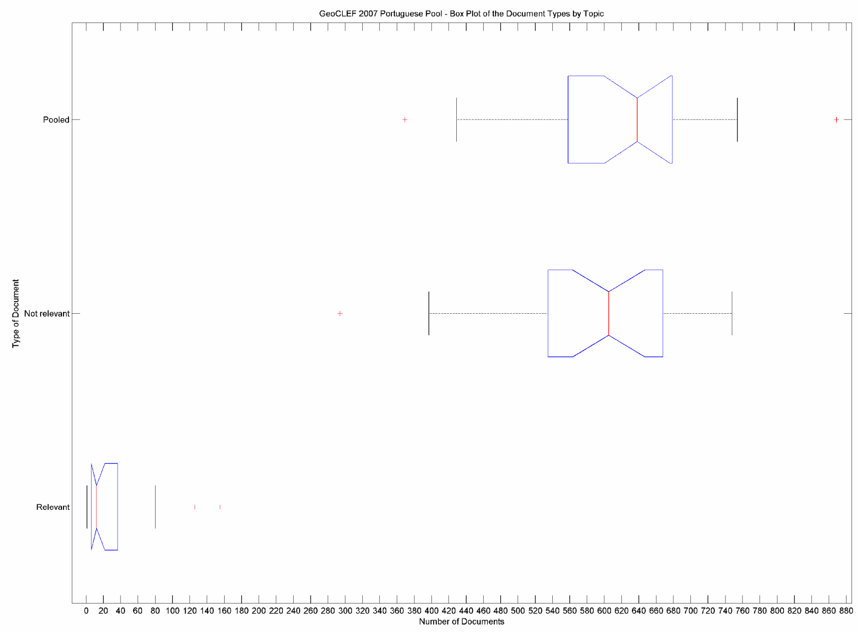


Fig. 4. GeoCLEF Portuguese 2007 Pool: distribution of the different document types

3 Results of the GeoCLEF 2007 Search Task

The results of the participating groups are reported in the following sections.

3.1 Participants and Experiments

As shown in Table 6, a total of 13 groups from 9 different countries submitted results for one or more of the GeoCLEF tasks. A total of 108 experiments were submitted.

Table 6. GeoCLEF 2007 participants – new groups are indicated by *

Participant	Institution	Country
catalunya	U.Politecnica Catalunya	Spain
cheshire	U.C.Berkeley	United States
csusm	Cal State U.- San Marcos	United States
depok*	U. Indonesia	Indonesia
groningen	U. Groningen	The Netherlands
hagen	U. Hagen-Comp.Sci	Germany
hildesheim	U. Hildesheim	Germany
icl	Imperial College London - Computing	United Kingdom
linguit*	Linguit Ltd	United Kingdom
moscow*	Moscow State U.	Russia
msasia	Microsoft Asia	China
valencia	U.Politecnica Valencia	Spain
xldb	U.Lisbon	Portugal

Five different topic languages were used for GeoCLEF bilingual experiments: German, English, Indonesian, Portuguese, and Spanish. Differently from usual, the most popular language for queries was Spanish (11 experiments out of 28 bilingual experiments); English (7 experiments) and Indonesian (6 experiments) almost tied for the second place; German (2 experiments) and Portuguese (2 experiments) tied for the third place. The number of bilingual runs by topic language is shown in Table 9.

Table 7 reports the number of participants by their country of origin.

Table 7. GeoCLEF 2007 participants by country

Country	# Participants
China	1
Germany	2
Indonesia	1
Portugal	1
Russia	1
Spain	2
The Netherlands	1
United Kingdom	2
United States	2
TOTAL	**13**

Table 8 provides a breakdown of the experiments submitted by each participant for each of the offered tasks.

Table 8. GeoCLEF 2007 experiments by task

Participant	Monolingual Tasks			Bilingual Tasks			TOTAL
	DE	EN	PT	X2DE	X2EN	X2PT	
catalunya		5					**5**
cheshire	1	1	1	3	3	3	**12**
csusm	6	6	5		4	4	**25**
depok*					6		**6**
groningen		5					**5**
hagen	5			5			**10**
hildesheim	4	4					**8**
icl		4					**4**
linguit*		4					**4**
moscow*		2					**2**
msasia		5					**5**
valencia		12					**12**
xldb		5	5				**10**
TOTAL	**16**	**53**	**11**	**8**	**13**	**7**	**108**

3.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, and Portuguese. Table 10 shows the top five groups for each target collection,

Table 9. Bilingual experiments by topic language

Track	Source Language					TOTAL
	DE	EN	ES	ID	PT	
Bilingual X2DE		6	1		1	**8**
Bilingual X2EN	1		5	6	1	**13**
Bilingual X2PT	1	1	5			**7**
TOTAL	**2**	**7**	**11**	**6**	**2**	**28**

ordered by mean average precision. Note that only the best run is selected 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 experiment Digital Object Identifier (DOI); the mean average precision achieved by the experiment; and the performance difference between the first and the last participant.

Due to an error, the XLDB group submitted the wrong run files for monolingual Portuguese. Because of the low number of participants, this run appears among the top runs. This explains the large difference between the second and the third run in Table 10.

Figures 5 to 7 show the interpolated recall vs. average precision for the top participants of the monolingual tasks.

Table 10. Best entries for the monolingual track. Additionally, the performance difference between the best and the last (up to 5) placed group is given (in terms of mean average precision) – new groups are indicated by *.

Track	Rnk	Partner	Experiment DOI	MAP
Mono-lingual English	1st	catalunya	10.2415/GC-MONO-EN-CLEF2007.CATALUNYA.TALPGEOIRTD2	28.5%
	2nd	cheshire	10.2415/GC-MONO-EN-CLEF2007.CHESHIRE.BERKMOENBASE	26.4%
	3rd	valencia	10.2415/GC-MONO-EN-CLEF2007.VALENCIA.RFIAUPV06	26.4%
	4th	groningen	10.2415/GC-MONO-EN-CLEF2007.GRONINGEN.CLCGGEOEETD00	25.2%
	5th	csusm	10.2415/GC-MONO-EN-CLEF2007.CSUSM.GEOMOEN5	21.3%
	Δ			**33.7%**
Mono-lingual German	1st	hagen	10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTDN5DE	25.8%
	2nd	csusm	10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE4	21.4%
	3rd	hildesheim	10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODENE2NA	20.7%
	4th	cheshire	10.2415/GC-MONO-DE-CLEF2007.CHESHIRE.BERKMODEBASE	13.9%
	Δ			**85.1%**
Mono-lingual Portuguese	1st	csusm	10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT3	17.8%
	2nd	cheshire	10.2415/GC-MONO-PT-CLEF2007.CHESHIRE.BERKMOPTBASE	17.4%
	3rd	xldb	10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_1	3.3%
	Δ			**442 %**

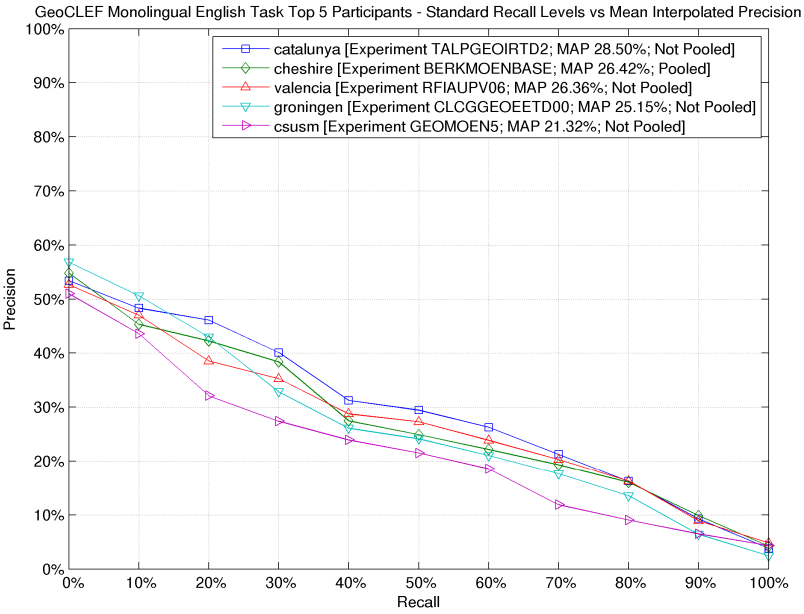


Fig. 5. Monolingual English top participants. Interpolated Recall vs. Average Precision.

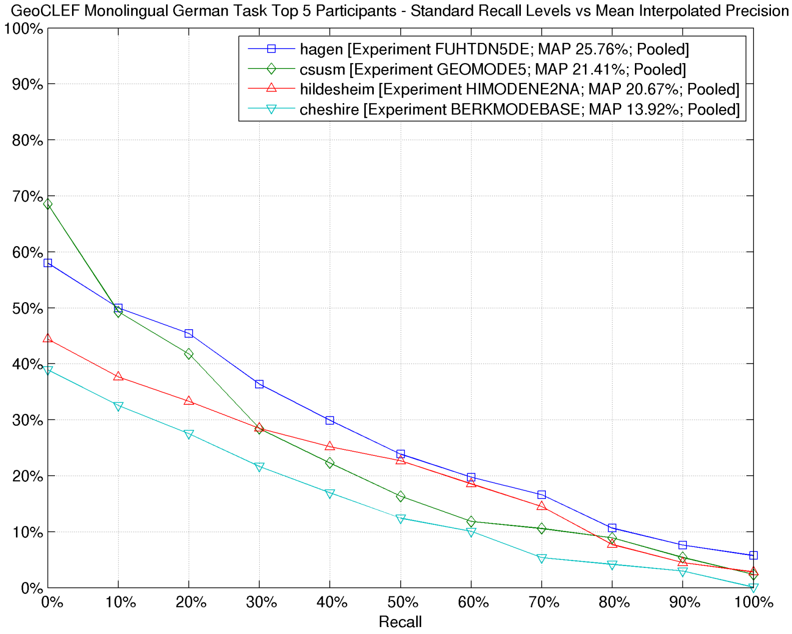


Fig. 6. Monolingual German top participants. Interpolated Recall vs. Average Precision.

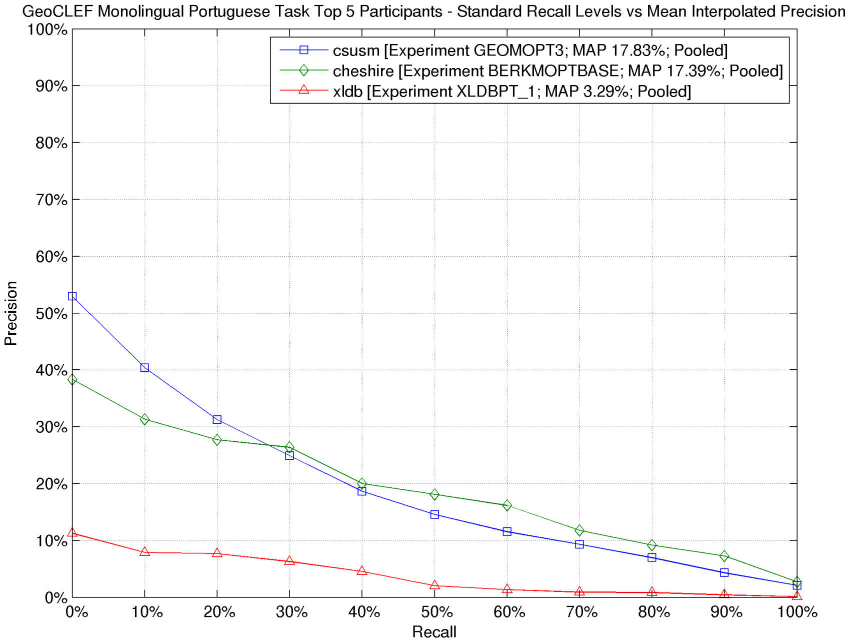


Fig. 7. Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

3.3 Bilingual Experiments

The bilingual task was structured in three subtasks (X → DE, EN, or PT target collection). Table 11 shows the best results for this task with the same logic of Table 7. Note that the top five participants contain both "newcomer" groups and "veteran" groups.

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

- X → DE: 81.1% of best monolingual German IR system
- X → EN: 77.4% of best monolingual English IR system
- X → PT: 112.9% of best monolingual Portuguese IR system

Note that there is a significant improvement for Bilingual German since CLEF 2006, when it was 70% of the best monolingual system; Bilingual English shows a small improvement, with respect to the 74% of the best monolingual system in CLEF 2006; finally, Bilingual Portuguese is quite surprising since it outperforms the monolingual and it represents a complete overturn with respect to the 47% of CLEF 2006. Figures 8 to 10 show the interpolated recall vs. average precision graph for the top participants of the different bilingual tasks.

4 Result Analysis

The test collection of GeoCLEF grew of 25 topics each year. This is usually considered the minimal test collection size to produce reliable results. Therefore, statistical testing and further reliability analysis are performed to assess the validity of the results obtained. The range of difficulties in the topics might have led to topics more difficult and more diverse than in traditional ad-hoc evaluations. To gain some insight on this issue, a topic performance analysis was also conducted.

Table 11. Best entries for the bilingual task. The performance difference between the best and the last (up to 5) placed group is given (in terms of mean average precision) – new groups are indicated by *.

Track	Rnk.	Partner	Experiment DOI	MAP
Bilingual English	**1st**	cheshire	10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDEENBASE	22.1%
	2nd	depok*	10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGP	21.0%
	3rd	csusm	10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN2	19.6%
	Diff.			**12.5%**
Bilingual German	**1st**	hagen	10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN4EN	20.9%
	2nd	cheshire	10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIPTDEBASE	11.1%
	Diff.			**88.6%**
Bilingual Portuguese	**1st**	cheshire	10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIENPTBASE	20.1%
	2nd	csusm	10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT4	5.3%
	Diff.			**277.5%**

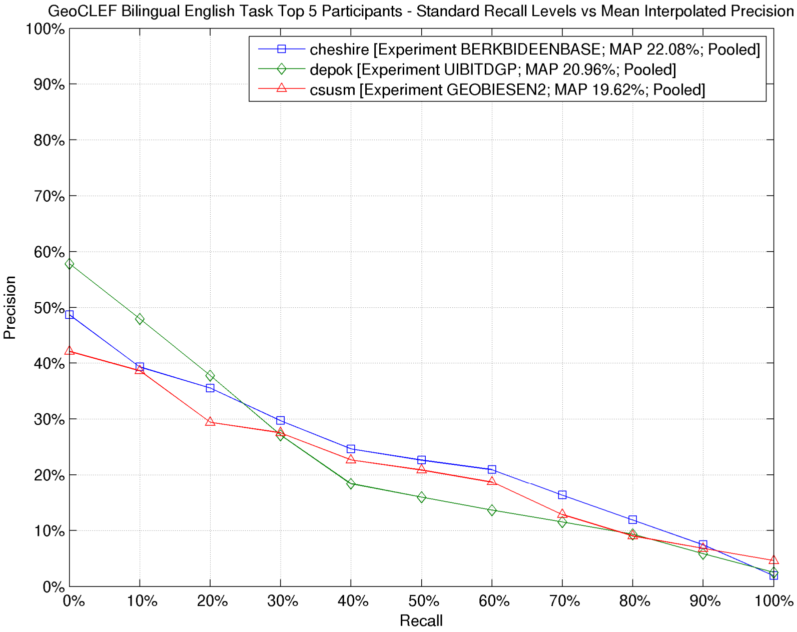


Fig. 8. Bilingual English top participants. Interpolated Recall vs Average Precision.

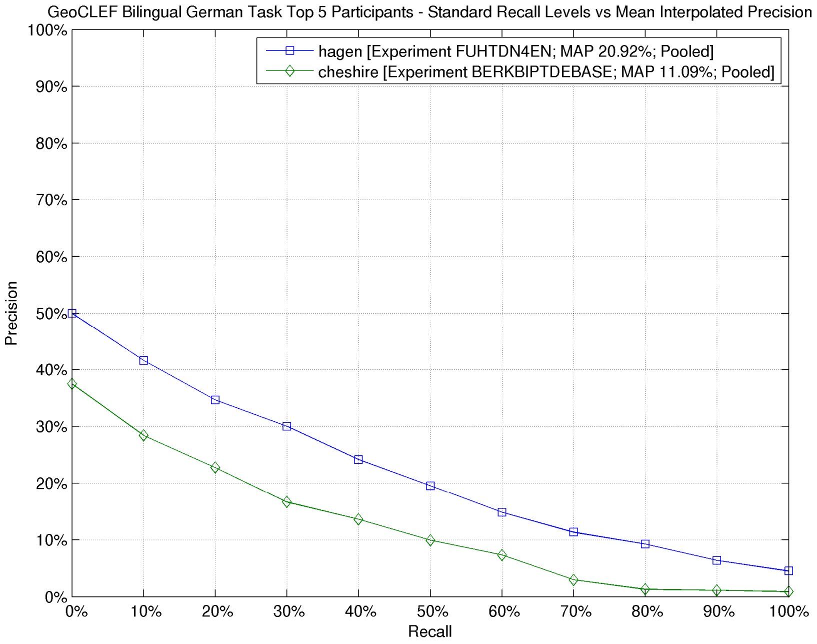


Fig. 9. Bilingual German top participants. Interpolated Recall vs Average Precision.

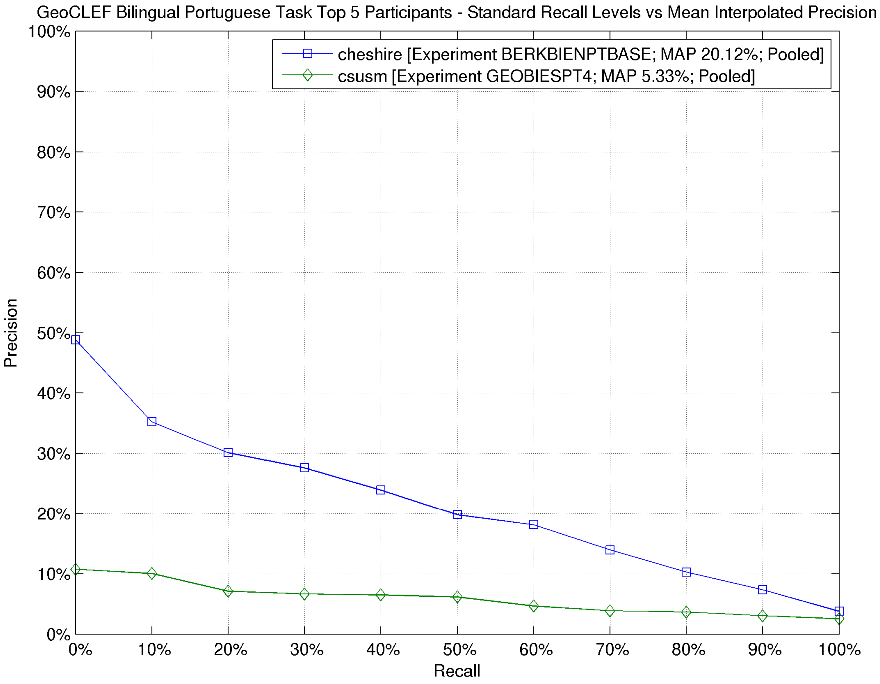


Fig. 10. Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

Table 12. 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	10	39	27	45
Monolingual German	0	13	8	14
Monolingual Portuguese	2	5	5	8
Bilingual English	1	7	10	13
Bilingual German	1	4	3	7
Bilingual Portuguese	0	2	2	3

4.1 Statistical Testing

Statistical testing for retrieval tests is intended to determine whether the order of the systems which results from the evaluation reliably measures the quality of the systems [10]. In most cases, the statistical analysis gives a conservative estimate of the upper level of significance [11]. 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.

Table 12 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. The results of the statistical analysis are shown in tables 13-18. Again, it is necessary to point out that among the few runs for monolingual Portuguese, one group were submitted with errors.

4.2 Stability Analysis

As for many other information retrieval evaluations, the variance between topics is much larger than between the systems. This fact has led 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 [12]. 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. The rankings have been compared by a rank correlation coefficient. It can be observed that the system ranking remains stable even until topic sets of size 11 which is less than half of the original topic set. The correlation remains above 80% and even 90% depending on the sub task. This stability is surprisingly high and shows that the GeoCLEF results are considerably reliable.

Table 13. Monolingual German: experiment groups according to the Tukey T Test

Experiment DOI	Grps.	
10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTDN5DE	X	
10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTDN4DE	X	X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE4	X	X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE5	X	X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE1	X	X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE6	X	X
10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODENE2NA	X	X
10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTD6DE	X	X
10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODEBASE	X	X
10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTD3DE	X	X
10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODENE2	X	X
10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTD2DE	X	X
10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODENE3	X	X
10.2415/GC-MONO-DE-CLEF2007.CHESHIRE.BERKMODEBASE	X	X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE2		X
10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE3		X

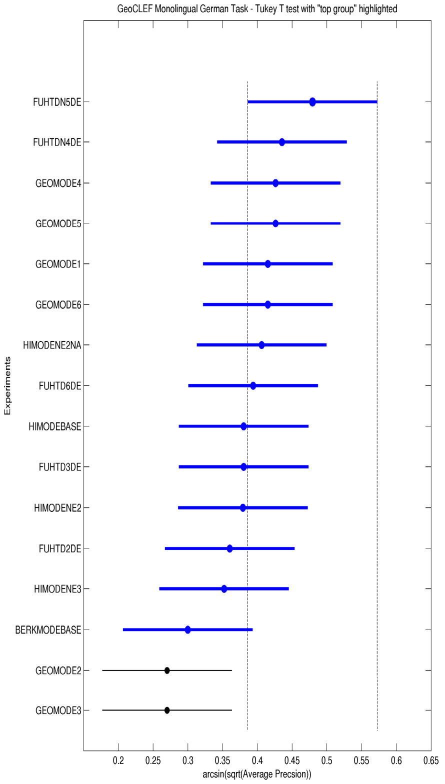


Table 14. Monolingual English: experiment groups according to the Tukey T Test. Experiment DOI is proceeded by *10.2415/GC-MONO-EN-CLEF2007.*

Experiment DOI	Groups											
CATALUNYA.TALPGEOIRTD2	x											
CATALUNYA.TALPGEOIRTDN3	x											
CATALUNYA.TALPGEOIRTDN2	x											
CATALUNYA.TALPGEOIRTD1	x											
CHESHIRE.BERKMOENBASE	x	x										
GRONINGEN.CLCGGEOEETD00	x	x										
VALENCIA.RFIAUPV06	x	x										
VALENCIA.RFIAUPV04	x	x										
CATALUNYA.TALPGEOIRTDN1	x	x										
VALENCIA.RFIAUPV02	x	x										
VALENCIA.RFIAUPV11	x	x	x									
VALENCIA.RFIAUPV09	x	x	x									
VALENCIA.RFIAUPV05	x	x	x									
VALENCIA.RFIAUPV10	x	x	x									
VALENCIA.RFIAUPV12	x	x	x	x								
VALENCIA.RFIAUPV01	x	x	x	x								
VALENCIA.RFIAUPV03	x	x	x	x	x							
VALENCIA.RFIAUPV08	x	x	x	x	x	x						
VALENCIA.RFIAUPV07	x	x	x	x	x	x						
GRONINGEN.CLCGGEOEETDN01	x	x	x	x	x	x						
GRONINGEN.CLCGGEOEET00	x	x	x	x	x	x						
GRONINGEN.CLCGGEOEETDN00	x	x	x	x	x	x						
CSUSM.GEOMOEN5	x	x	x	x	x	x	x					
GRONINGEN.CLCGGEOEETDN01B	x	x	x	x	x	x	x					
CSUSM.GEOMOEN6	x	x	x	x	x	x	x					
ICL.IMPCOLTEXTONLY	x	x	x	x	x	x	x	x				
CSUSM.GEOMOEN4	x	x	x	x	x	x	x	x				
HILDESHEIM.HIMOENNE	x	x	x	x	x	x	x	x				
HILDESHEIM.HIMOENNANE	x	x	x	x	x	x	x	x				
MOSCOW.CIRGEOEN07_RUN1	x	x	x	x	x	x	x	x				
HILDESHEIM.HIMOENBASE	x	x	x	x	x	x	x	x				
MOSCOW.CIRGEOEN07_RUN2	x	x	x	x	x	x	x	x				
MSASIA.MSRATEXT	x	x	x	x	x	x	x	x	x			
CSUSM.GEOMOEN1	x	x	x	x	x	x	x	x	x			
HILDESHEIM.HIMOENNE2	x	x	x	x	x	x	x	x	x			
LINGUIT.LTITLEDESC2007	x	x	x	x	x	x	x	x	x			
LINGUIT.LTITLEEXPMANUAL2007	x	x	x	x	x	x	x	x	x			
LINGUIT.LTITLE2007	x	x	x	x	x	x	x	x	x	x		
CSUSM.GEOMOEN3	x	x	x	x	x	x	x	x	x	x	x	
CSUSM.GEOMOEN2		x	x	x	x	x	x	x	x	x	x	x
ICL.IMPCOLCOMBINATION			x	x	x	x	x	x	x	x	x	x
ICL.IMPCOLNOGEO				x	x	x	x	x	x	x	x	x
MSASIA.MSRALDA					x	x	x	x	x	x	x	x
LINGUIT.LTITLEGEORERANK2007						x	x	x	x	x	x	x
MSASIA.MSRALOCATION							x	x	x	x	x	x
MSASIA.MSRAWHITELIST							x	x	x	x	x	x
MSASIA.MSRAEXPANSION								x	x	x	x	x
XLDB.XLDBEN_2									x	x	x	x
XLDB.XLDBEN_3									x	x	x	x
XLDB.XLDBEN_5										x	x	x
XLDB.XLDBEN_4										x	x	x
XLDB.XLDBEN_1											x	x
ICL.IMPCOLGEOONLY												x

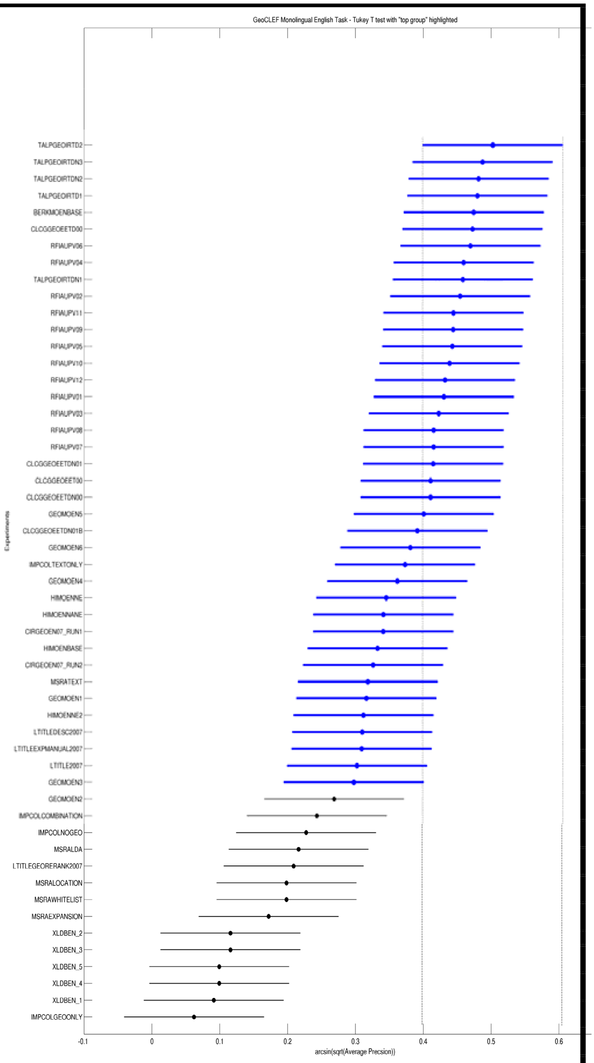


Table 15. Monolingual Portuguese: experiment groups according to the Tukey T Test

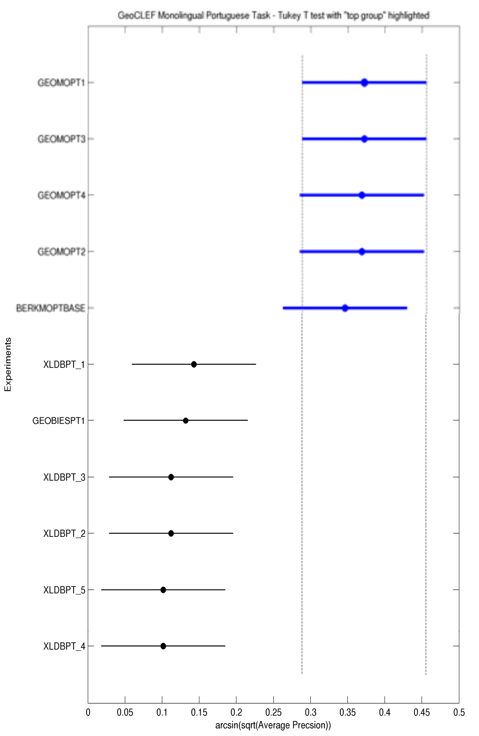
Experiment DOI	Grps.		
10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT1	X		
10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT3	X		
10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT4	X		
10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT2	X		
10.2415/GC-MONO-PT-CLEF2007.CHESHIRE.BERKMOPTBASE	X		
10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_1		X	
10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOBIESPT1		X	
10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_3		X	
10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_2		X	
10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_5		X	
10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_4		X	

Table 16. Bilingual English: experiment groups according to the Tukey T Test

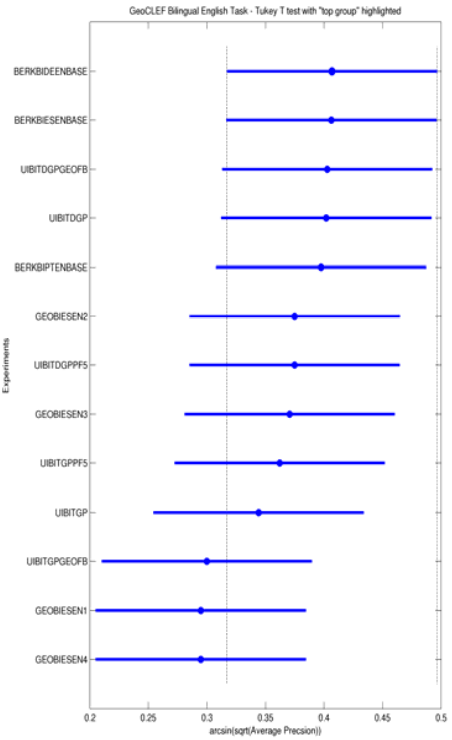
Experiment DOI	Grps	
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDEENBASE	X	
10.2415/GC-BILI-X2EN-CLEF 2007.CHESHIRE.BERKBIESENBASE	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPGEOFB	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGP	X	
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIPTENBASE	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN2	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPPF5	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN3	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGPPF5	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGP	X	
10.2415/GC-BILI-X2EN-CLEF 2007.DEPOK.UIBITGPGEOFB	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN1	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN4	X	

Table 17. Bilingual English: experiment groups according to the Tukey T Test

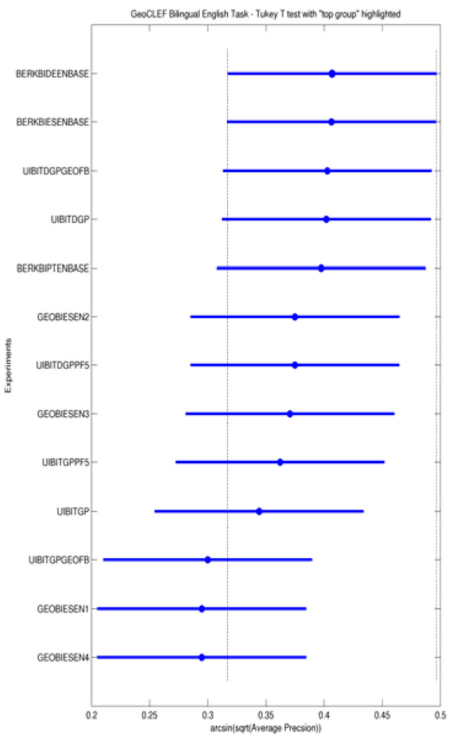
Experiment DOI	Grps	
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDEENBASE	X	
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIESENBASE	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPGEOFB	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGP	X	
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIPTENBASE	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN2	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPPF5	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN3	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGPPF5	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGP	X	
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGPGEOFB	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN1	X	
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN4	X	

Table 18. Bilingual German: experiment groups according to the Tukey T Test

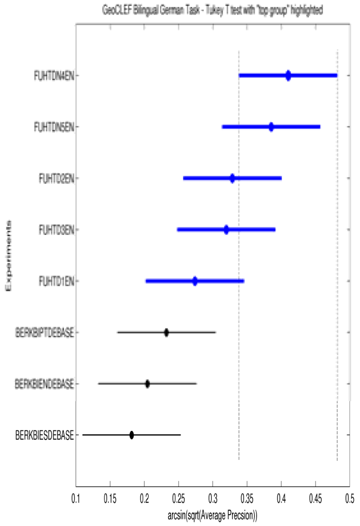
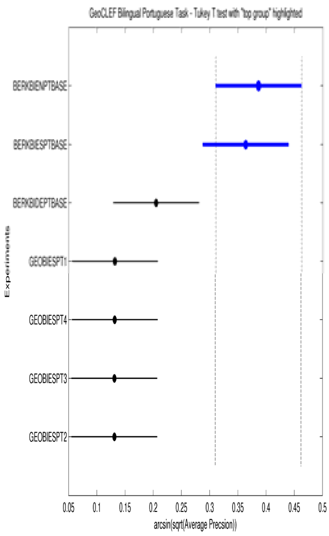
Experiment DOI	Grps			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN4EN	X			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN5EN	X			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD2EN	X	X		
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD3EN	X	X	X	
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD1EN	X	X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIPTDEBASE		X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIENDEBASE		X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIESDEBASE			X	

Table 19. Bilingual Portuguese: experiment groups according to the Tukey T Test

Experiment DOI	Grps		
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIENPTBASE	X		
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIESPTBASE	X		
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIDEPTBASE		X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT1		X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT4		X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT3		X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT2		X	

5 Query Classification Task

The query parsing and classification task was offered for the first time at GeoCLEF 2007. It was dedicated to identifying geographic queries within a log file from the MSN search engine. This task has been organized by Xie Xing from Microsoft Research Asia. The task is of high practical relevance to GeoCLEF and the real log data is of great value for research.

The task required participants to find the geographic entity, the relation type and the non geographic topic of the query. In details, the systems needed to find the queries with a geographic scope, extract the geographic component (where), extract the type of the geographic relation (e.g. in, north of) and extract the topic of the query (what component). In addition, the systems were required to classify the query type. The classes defined were information, yellow page and map. For a query Lottery in Florida, for example, the systems were required to respond that this is a geographic query of the type information, return Florida as the where-component, lottery as the what component and extract in as the geographic relation. There were 27 geographic relations given.

For this task, a log of 800,000 real queries was provided. Out of these, 100 were labeled as training data and 500 were assessed as test data. The labeling was carried out by three Microsoft employees. They reached a consensus on each decision. In the randomly chosen and manually cleansed set, there were 36% non local queries. The geographic queries comprised 16% map queries, 29% yellow page type queries and 19% information (ad-hoc type) queries.

The results were analyzed by calculating the recall, the precision and a combined F-Score for the classification task. The task attracted six participating groups. The performance for classifying whether a query was local or not were used as a primary evaluation measure. The results are shown in Table 19.

Table 19. Results of the Query Classification Task

Team	Recall	Precision	F1-Score
Ask.com	**0.625**	0.258	0.365
csusm	0.201	0.197	0.199
linguit	0.112	0.038	0.057
miracle (DAEDALUS)	0.428	**0.566**	**0.488**
catalunya	0.222	0.249	0.235
xldb	0.096	0.08	0.088

The overall results are quite low. This shows that further research is necessary. Most participants used approaches which combined heuristic rules and lists and gazetteers of geographic named entities. More details on the task design, the data, participation and evaluation results are provided in an overview paper [13].

6 Conclusions and Future Work

GeoCLEF 2007 has continued to create an evaluation resource or geographic information retrieval. Spatially challenging topics have been developed and interesting experiments have been submitted. 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 the SIGIR or CIKM conferences. All participants of GeoCLEF 2007 are invited to actively contribute to the discussion of the future of GeoCLEF.

Acknowledgments

German assessment was based entirely on volunteer effort; it was carried out by Julia Jürgens, Theresa Märtl, Ben Heuwing, Frauke Zurek, Robert Stoldt, Jens Plattfaut, and Rafael Hellmann of the University of Hildesheim. The English assessment was performed by Fredric Gey, Ray Larson, Mark Sanderson and the Sheffield student team (paid for by the EU 6th Framework project Tripod). The Portuguese documents were assessed by Paulo Rocha, Luís Costa, Luís Miguel Cabral, Susana Inácio, Maria Cláudia Freitas and Diana Santos, in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC. The topics were thoroughly checked by Sven Hartrumpf from the University of Hagen (Germany). The query classification task was carried out by Microsoft Research Asia without any external financial support.

References

1. Braschler, M., Peters, C.: Cross-Language Evaluation Forum: Objectives, Results, Achievements. Information Retrieval 7(1-2), 7–31
2. Santos, D., Rocha, P.: The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 821–832. Springer, Heidelberg (2005)
3. Di Nunzio, G.M., Ferro, N.: DIRECT: A System for Evaluating Information Access Components of Digital Libraries. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 483–484. Springer, Heidelberg (2005)
4. 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)
5. Evert, S.: The CQP Query Language Tutorial (CWB version 2.2.b90) University of Stuttgart (July 10, 2005), `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/`

6. Santos, D., Eckhard, B.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavriladou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, pp. 205–210 (2000)
7. Gey, F., Larson, R., Sanderson, M., Bishoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Di Nunzio, G., Ferro, N.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 852–876. Springer, Heidelberg (2007)
8. 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, pp. 908–919. Springer, Heidelberg (2006)
9. Santos, D., Chaves, M.S.: The place of place in geographical IR. In: GIR 2006, the 3rd Workshop on Geographic Information Retrieval, SIGIR 2006, Seattle, 10 August 2006 (presentation at, 2006), `http://www.linguateca.pt/Diana/download/acetSantosChavesGIR2006.pdf`
10. Buckley, C., Voorhees, E.: Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval, pp. 53–75. MIT Press, Cambridge (2005)
11. Sanderson, M., Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: 28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR) 2005, pp. 162–169 (2005)
12. Zobel, J.: 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 1998), Melbourne, Australia, pp. 307–314. ACM Press, New York (1998)
13. Li, Z., Wang, C., Xing, X., Ma, W.-Y.: Query Parsing Task for GeoCLEF 2007 Report. In: Nardi, A., Peters, C. (eds.) Working Notes of the Cross Language Evaluation Forum (CLEF) (2007)