

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

Thomas Mandl¹, Paula Carvalho², Giorgio Di Nunzio³, Fredric Gey⁴
Ray Larson⁴, Diana Santos⁵, Christa Womser-Hacker¹

¹Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de, womser@uni-hildesheim.de

²University of Lisbon, DI, LasiGE, XLDB
Linguatca, Portugal
pqfcarvalho@gmail.com

³Department of Information Engineering, University of Padua, Italy
dinunzio@dei.unipd.it

⁴University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@sims.berkeley.edu

⁵Linguatca, SINTEF ICT, Norway
Diana.Santos@sintef.no

Abstract. GeoCLEF is an evaluation task running under the scope of the Cross Language Evaluation Forum (CLEF). The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR). The GeoCLEF 2008 task presented twenty-five geographically challenging search topics for English, German and Portuguese. Eleven participants submitted 131 runs, based on a variety of approaches, including sample documents, named entity extraction and ontology based retrieval. The evaluation methodology and results are presented in the paper.

1 Introduction

The Internet propelled a variety of geographic services that range from map services to route planning and hotel reservation systems. Many queries for search engines involve some sort of geographic processing and reasoning. Therefore, the development and evaluation of information retrieval systems that optimize the geographically oriented access to information is very important.

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Many documents containing spatial references are important to retrieve, rank and visualize information needs, such as “find me news stories about riots near Paris and their consequences”.

GeoCLEF is the first track of an evaluation campaign dedicated to evaluating geographic information retrieval systems. The aim of GeoCLEF is to provide the

framework for evaluating GIR systems, in both a spatial and a multilingual dimension. Participants were presented with a TREC style ad hoc retrieval task, based on the existing CLEF newspaper collections.

GeoCLEF was a pilot track in 2005 and, since then, it was a regular track. It evaluates document retrieval with an emphasis on geographic information text retrieval. Spatial reasoning is often necessary to solve the search tasks.

Eleven research groups (thirteen in 2007) from different backgrounds and nationalities submitted 131 runs (108 in 2007) to GeoCLEF 2008.

Portuguese, German and English were available as document and topic languages. As in previous editions, there were two Geographic Information Retrieval tasks: monolingual (English to English, German to German and Portuguese to Portuguese) and bilingual (language X to language Y, where X and Y correspond to one of the above mentioned languages).

GeoCLEF developed a standard evaluation collection which supports long-term research. Altogether, 100 topics including relevance assessments have been developed over the last four years (one pilot run and three regular tracks). Additionally, a set of 26 CLEF ad-hoc topics with spatial restrictions has been identified and can be used as a benchmark. Topics and the relevance judgment files will be publicly available on the GeoCLEF website¹.

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
2008	English, German, Portuguese	English, German, Portuguese

Geographic IR is a challenging task, namely because it deals with geographical references which are often vague, ambiguous and multilingually challenging. Multilingual retrieval requires systems matching references to a place from one language to another, which may have different correspondents (e.g. *Athens*, *Athen*, *Atenas*, *Atina*). Spatial reasoning is usually mandatory to solve information needs, such as “demonstrations in cities in *Northern Germany*”, where the geographic term corresponds to a selection of places and locations that are not explicitly specified in the topic.

The GeoCLEF track comprises two sub tasks. The main task is described in the following sections. The GikiP task² which evaluates searches for Wikipedia entries that require some geographical processing, is described in a separate overview paper [5].

¹ <http://www.uni-hildesheim.de/geoclef>

² <http://www.linguatca.pt/GikiP>

2 GeoCLEF 2008 Search Task

The geographic search task is the main task of GeoCLEF and it is developed following the general framework underlying the CLEF ad-hoc task. The following sections describe the test design.

2.1 Document Collections used in GeoCLEF 2008

The document collections used in the third GeoCLEF edition are the same as the ones used in GeoCLEF 2007, and in previous CLEF ad-hoc evaluations [1]. They are newspaper and newswire stories, from 1994 to 1995, covering international and national news and events that mention a wide variety of geographical entities. The English collection contains 169,477 documents, which are made out of stories from the British newspaper *The Glasgow Herald* (1995) and the American newspaper *The Los Angeles Times* (1994). The German collection contains 294,809 documents from the German magazine *Der Spiegel* (1994/95), the German newspaper *Frankfurter Rundschau* (1994) and the Swiss newswire agency *Schweizer Depeschen Agentur* (SDA, 1994/95). The Portuguese collection is made out of two major daily newspapers, namely the Portuguese newspaper *Público* (106,821 documents) and the Brazilian newspaper *Folha de São Paulo* (103,913 documents). The Portuguese collections are distributed by Linguatca as the CHAVE collection³.

Table 2. GeoCLEF 2008 test collection size

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

The documents have a common structure in the three language collections: newspaper-specific information, like date, (optionally) page, issue, special filing numbers and often one or more titles, a by-line and the actual text. Geographic entities were not previously recognized and none semantic location-specific information was added to the documents.

2.2 Generating Search Topics

A total of 25 topics were created for this year's GeoCLEF (GC76 - GC100). Topics express a natural information need that a user of the collection might have. Topic creation was a shared task between the Portuguese and the German groups. The task was supported, by the use of the DIRECT System, provided by the University of Padua. This system includes a search utility for handling the collections.

Topic creation was performed in two stages. First, each group devised a set of candidate topics in their own language, whose appropriateness was checked in the text

³ <http://www.linguatca.pt/CHAVE>

collection available for that language. Topic candidates were subsequently checked for relevant documents in the other collections. Sometimes, it is difficult to find geographically interesting topics below the granularity of a country. Regional events with a wide coverage in one country do not often correspond to many newspaper articles in other countries. As a consequence, some topics needed to be partially modified or refined, by relaxing or tightening the content or the geographic focus.

Other reasons driving this process were the absence of relevant documents in one of the languages, the complexity of topic interpretation and the translation into the other languages. For example, a candidate topic on fish living in the Iberian Peninsula had relevant matches in the Portuguese collection. However, this topic was not mentioned in the other newspapers. Moreover, some of the species described in the “narrative” (e.g. *saramugo*, a species which lives only in Spanish and Portuguese rivers) were difficult to translate into German and English. The spatial parameter (*Iberian Peninsula*) remained in a topic, but the subject was replaced by a matter that potentially interests the international mass media, namely, the state of agriculture in the Iberian Peninsula. In most cases, the changes were not radical. For example, the initial candidate topic “Nobel Prize winners in Physics from Northern European countries” was replaced by a more general one: “Nobel prize winners from Northern European countries”. In other cases, the geographic term was replaced by other(s) involving a more difficult but interesting exercise of geographic reasoning and processing. For example, “Most visited sights in the capital of France” was changed to: “Most visited sights in the capital of France and its vicinity”, which is more challenging from the geographic point of view. The new form involves the processing of relative proximity and neighborhood concepts.

The final topic set was agreed upon after intensive discussion. All missing topics were translated into Portuguese and German and all translations were checked. The next section discusses the creation of topics with spatial parameters for the track.

2.3 Spatial Parameters

One goal of GeoCLEF is the creation of a geographically challenging topic set. Geographic knowledge is necessary to successfully retrieve relevant documents for most documents. While many geographic searches may be reasonably satisfied by keyword approaches, others require geographic reasoning. Most systems, especially keyword based systems, might perform better on average with a realistic topic set, where these difficulties occur less frequently.

To increase the difficulty of the topic set, the following issues were explicitly included in the topics of GeoCLEF 2008:

- imprecise /vague geographic regions (Sub-Saharan Africa, Western Europe)
- geographical relations beyond IN (forest fires on Spanish islands)
- granularity below the country level (fairs in Lower Saxony)
- terms which are not explicitly mentioned in documents (Portuguese communities in other countries)

We tried to create a set of topics representing different kinds of geographic queries.

These queries present different levels of complexity and may require different approaches to process them adequately, and successfully retrieve relevant documents. Instead of privileging specific geographical places, such as a country or city, preference was given to reference geographical regions, comprehending more than one physical or administrative place. Different kinds of regions were, then, considered, which may correspond, for instance, to a delimited geographical area of a given continent (e.g. Sub-Saharan Africa, Southeast Asia, Northern Africa, Western Europe) or country (e.g. Western USA, Lower Saxony, Spanish islands). Other interesting geo-economic-political terms, such as OECD countries, were also considered in the topic creation.

The majority of the GeoCLEF 2008 topics specify complex (multiply defined) geographical relations, a property introduced in the GeoCLEF 2007 [8], kept in this evaluation. Such geographical relations, which can be explicitly or implicitly mentioned in the topic, may represent:

- Proximity (e.g. Most visited sights in the capital of France and its vicinity);
- Inclusion (e.g. Attacks in Japanese subways);
- Exclusion (e.g. Portuguese immigrant communities in the world).

The example illustrating proximity also presents a relation of inclusion (between sights and capital of France), explicitly formalized by the preposition "in". That relation can also be inferred in the phrase "Japanese subways" occurring in the topic illustrating inclusion, which can be paraphrased by the expression "subways in Japan".

Different from the GeoCLEF 2007 topics, which might represent explicit relations of exclusion (e.g. Europe excluding the Alps), such relations were only implicitly represented in the topics of GeoCLEF 2008, as illustrated above. This topic has the particularity of presenting simultaneously a relation of inclusion (communities from Portugal in the world) and exclusion (in this context, world represents any country except Portugal).

Just as in previous GeoCLEF editions, vague geographic designations were introduced for certain topics. For example, in the topic: "Nuclear tests in the South Pacific", the geographical term South Pacific may refer to both Australasia ("an area including Australia, New Zealand, New Guinea and other islands including the eastern part of Indonesia") and Oceania ("a geographical (often geopolitical) region of many countries/territories (mostly islands) in the southern Pacific Ocean"). The interpretation of this geographical term (ambiguous between an ocean and the islands within it) is only possible if the full topic content is considered.

A similar situation is observed in the topic "American troops in the Persian Gulf". In this case, the Persian Gulf does not stand for the gulf itself but for a Southwest Asian region, which is an extension of the Indian Ocean located between Iran and the Arabian Peninsula. Once again, the adequate processing of the information in the topic requires term disambiguation.

Another case of vagueness can be observed in the topic "Environmental pollution in European waters", where the term waters can refer to rivers, lakes or the sea.

2.4 Approaches to Geographic Information Retrieval

The format of GeoCLEF 2008 is identical to that of GeoCLEF 2006 and 2007. Table 3 illustrates the syntax of two different topics, the one on the left hand side in English and the one on the right side in Portuguese. As it can be observed, the topics do not contain any geographic tag.

Table 3: Two examples from the Topics: 10.2452/89-GC and 10.2452/84-GC

<code><num>10.2452/89-GC</num> <title>Trade fairs in Lower Saxony </title> <desc>Documents reporting about industrial or cultural fairs in Lower Saxony. </desc> <narr>Relevant documents should contain information about trade or industrial fairs which take place in the German federal state of Lower Saxony, i.e. name, type and place of the fair. The capital of Lower Saxony is Hanover. Other cities include Braunschweig, Osnabrück, Oldenburg and Göttingen. </narr> </top></code>	<code><num>10.2452/84-GC</num> <title>Atentados à bomba na Irlanda do Norte </title> <desc>Os documentos relevantes mencionam atentados bombistas em localidades da Irlanda do Norte </desc> <narr>Documentos relevantes devem mencionar atentados à bomba na Irlanda do Norte, indicando a localização do atentado. </narr> </top></code>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The table shows, the short topic description, within the title and description tags, is followed by the narrative tag, which contains a detailed description of the geographic requirements and the relevance criteria. In some topics, relevant geographic names are listed in the narrative.

2.5 Approaches to Geographic Information Retrieval

In the last three editions of GeoCLEF, traditional ad-hoc retrieval approaches and specific geographic reasoning systems have been explored in parallel. Successful results have often been achieved by ad-hoc techniques without any specific geographic knowledge or processing. These approaches have sometimes been developed as a baseline for more sophisticated systems. Some of the traditional techniques may have beneficial effects for geographic search tasks. Blind relevance feedback can lead to a geographic term expansion necessary to solve a search problem. For example, a query for riots in German cities does not contain the name of any German city. A query including the term German may lead to documents containing the word German and the names of some cities which can be included in subsequent optimized queries. As a result, geographic term expansion has been achieved without proper geographic knowledge being available to the system. This form of pseudo-geographic processing is not very reliable, but the specific components often have a high error rate or introduce significant noise. In GeoCLEF 2007, some systems tried combinations of both approaches and the dedicated geographic systems have further matured. In 2008, new ideas were introduced. For example, an ontology based approach presented by the DFKI was successful for the most competitive task: monolingual English. The University of Berkeley implemented a system designed like an ad-hoc system without any geographic components.

The participants used varied approaches to the GeoCLEF tasks, ranging from basic IR approaches to deep natural language processing (NLP). The approaches include the use of full documents for ranking the result set, map based techniques and Wikipedia as a knowledge source. For details, the reader can consult the description of the systems in the papers of the participants (in this volume).

2.6 Relevance Assessment

The English assessment was shared by Berkeley and Hildesheim Universities. The German assessment was done by the University of Hildesheim and the Portuguese assessment by Linguateca. The DIRECT System, used for topic development, was also used for relevance assessment. The system provided by the University of Padua allowed the automatic submission of runs by participating groups and supported the GeoCLEF assessment pools by language. All runs were included in the pool. The size of the pool is shown in table 4 and the distribution of relevant documents over topics is given in table 5.

Table 4. GeoCLEF 2008 Size of Pools

Language	# Documents
English	14,528
German	15,081
Portuguese	14,780

Table 5. GeoCLEF 2008 Relevant Documents per Topic

Language	Minimum	Maximum
English	0	109
German	1	146
Portuguese	2	158

During the assessment process, the assessor tried to find the best collection of keywords – namely, based on the detailed information described in the narrative, using the DIRECT system. The following subsections report some of the issues concerning the relevance assessment for each language.

Some topics caused assessment difficulties, especially when the narrative required specific information, not expressed in the text. For example, from the sentence: Bonn ... former chancellor Willy Brandt ... Nobel Peace prize winner... is it possible to infer that Willy Brandt was German?

In assessments, topic drifts typically occur. GeoCLEF 2008 assessment was not an exception. Is a document about kidnapping of a French aid worker in Kenya relevant for "foreign aid in Sub-Saharan Africa"? The kidnapping of an aid worker implies the existence of foreign aid in Kenya, but a kidnapping is not related in any sense to foreign aid.

The assessment usually provides hints on why systems failed. The German topic about “fairs in Lower Saxony” points to inappropriate stemming rules or to high n-gram similarity.. The German word for fairs (*Messe*) was matched against similar words with a different meaning (e.g. *angemessen* -> appropriate, *Messer* -> knife).

The English document pool also led to borderline cases that needed to be discussed among the assessors. One topic required documents on “natural disasters in Western states of the USA”. Some documents only reported the insurance costs caused by natural disaster overall. In such cases, it was decided to consider relevant the documents mentioning a geographically relevant place (for example, *Los Angeles*) even when they did not mention the disaster explicitly and directly.

3 GeoCLEF 2008 Results

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 eleven groups from seven different countries participated in one or more GeoCLEF tasks. A total of 131 experiments (runs) were submitted. Five of these groups participated in GeoCLEF for the first time.

Table 6. GeoCLEF 2008 participants – new groups are indicated by *

Participant	Institution	Country
Alivale*	U.Jaén & U.Politecnica Valencia	Spain
Cheshire	U.C.Berkeley	United States
Csusc	Cal. State U.- San Marcos	United States
dfki*	German Research Center for AI	Germany
Hagen	U.Hagen-Comp.Science	Germany
icl	Imperial College London	United Kingdom
Inaoe*	Lab. Tecnologias del Lenguaje	Mexico
jaen*	U.Jaén	Spain
pittsburgh*	U.Chengdu & U.Pittsburgh,	China & United States
Valencia	U.Politecnica Valencia	Spain
xldb	U.Lisbon	Portugal

Table 7 provides an overview of the experiments submitted per task and participant. Three different topic languages were allowed for the GeoCLEF bilingual experiments. Again, the most popular language for queries was English; German took the second place. The number of bilingual runs by topic language is shown in Table 8.

Table 7. GeoCLEF 2008 experiments by task

Participant	Monolingual Tasks			Bilingual Tasks			TOTAL
	DE	EN	PT	X2DE	X2EN	X2PT	
alivale*		9					9
cheshire	3	3	3	6	6	6	27
csusm	1	1	2	1	1	1	7
Dfki*		5					5
hagen	5			10			15
icl		9					9
inaoe*		12					12
Jaen*		7			6		13
pittsburgh*		4					4
valencia		6					6
xldb		12	12				24
TOTAL	9	68	17	17	13	7	131

Table 8. Bilingual experiments by topic language

Track	Source Language			TOTAL
	DE	EN	PT	
Bilingual X2DE		10	7	17
Bilingual X2EN	4		3	7
Bilingual X2PT	7	6		13
TOTAL	11	16	10	27

3.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, and Portuguese. Figures 1 to 3 show the interpolated recall vs. average precision for the top participants of the monolingual tasks.

The most competitive task was the monolingual English task with half of all GeoCLEF runs. The DFKI submitted the best run based on ontology processing but the results of the other participants are very close. The University of California at Berkeley applied no geographic processing and is not only in the top group for monolingual English but also for the bilingual experiments.

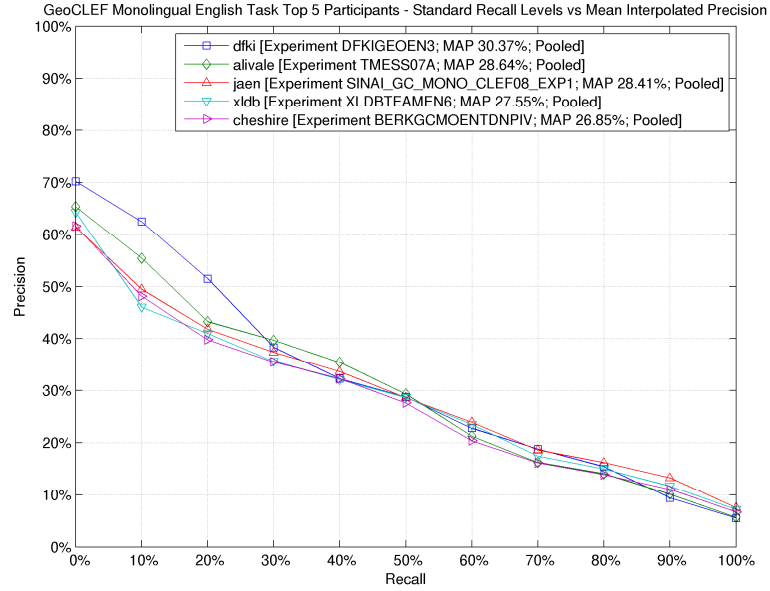


Fig. 1. Monolingual English top participants. Interpolated Recall vs. Average Precision.

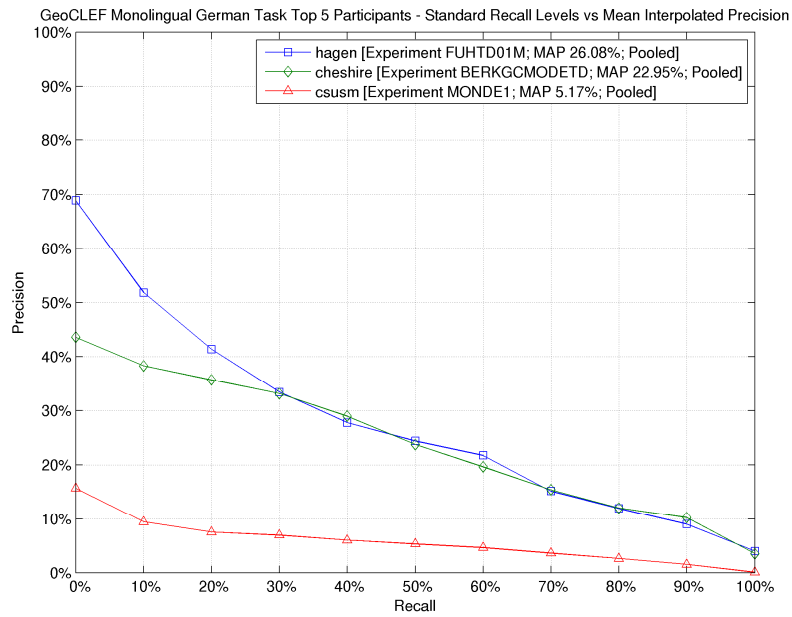


Fig. 2. Monolingual German top participants. Interpolated Recall vs. Average Precision.

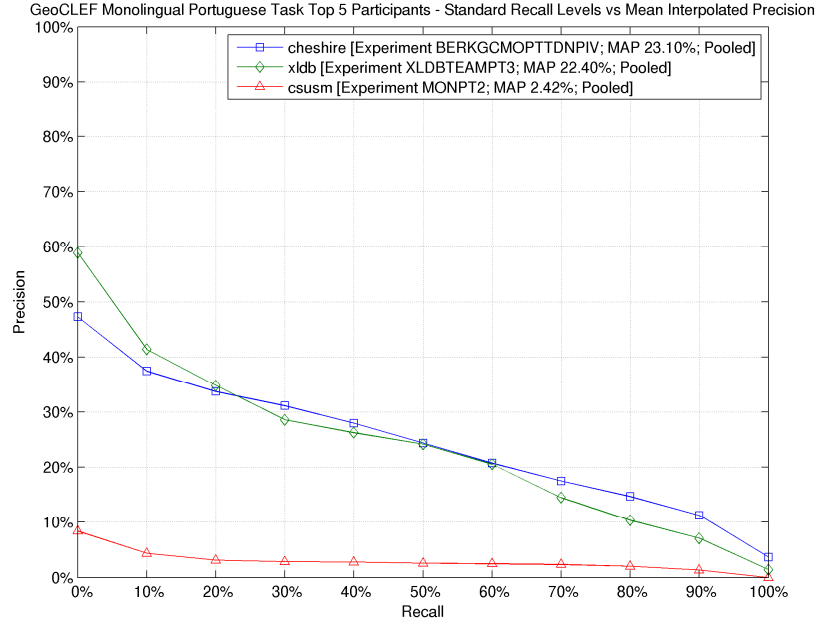


Fig. 3. Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

3.2 Bilingual Experiments

The bilingual task was structured in four subtasks ($X \rightarrow$ DE, EN or PT target collection). The best performing system for each of the three bilingual sub-tasks was presented by the University of California at Berkeley. This system did not use any specific geographic reasoning or knowledge source. Figures 4 to 6 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 increased 25 topics each year. Statistical testing and further analysis were performed to assess the validity of the results obtained.

Statistical testing for retrieval tests is used to determine whether the order of the systems which results from the evaluation reliably measures the quality of the systems in a reliable manner [2]. In most cases, the statistical analysis gives an approximate conservative estimate of the upper level of significance. The MATLAB Statistics Toolbox and the *ANalysis Of VAriance* (ANOVA) test were used for statistical testing. In all the experiments a value of $\alpha = 0.05$ has been used to determine when to accept or reject the null hypothesis.

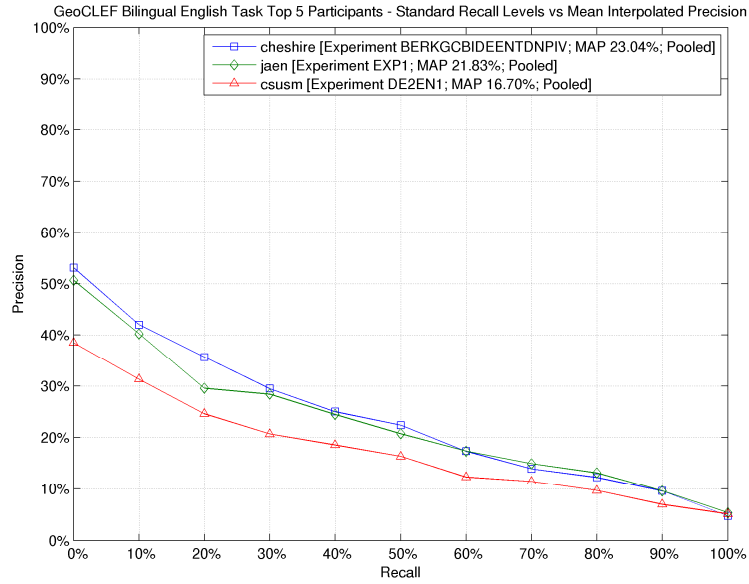


Fig. 4. Bilingual English top participants. Interpolated Recall vs Average Precision.

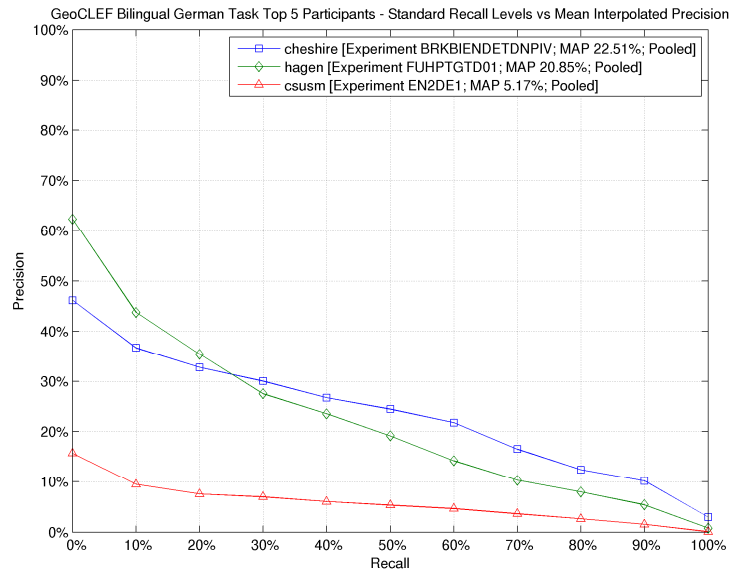


Fig. 5. Bilingual German top participants. Interpolated Recall vs Average Precision.

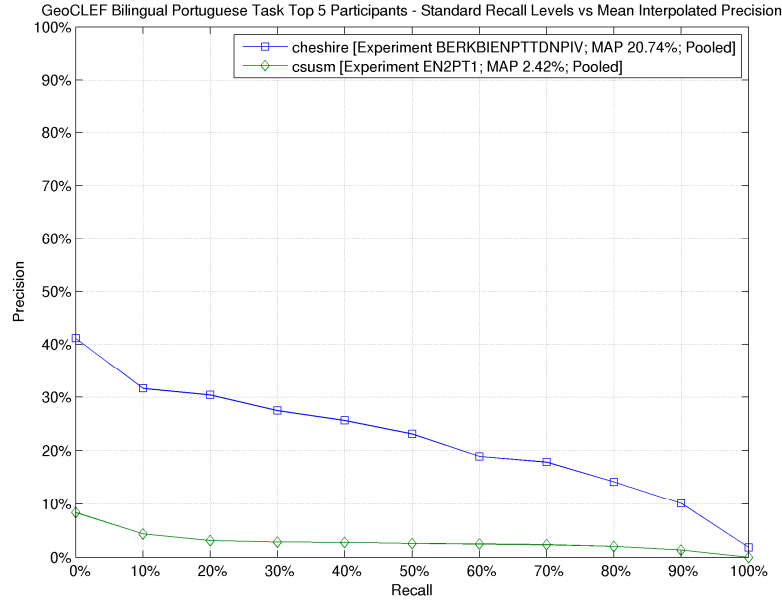


Fig. 6. Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

4.1 Monolingual vs. Bilingual Retrieval

In order to evaluate bilingual retrieval experiments, a common method is to compare results against monolingual baselines, which is comparing the best monolingual experiment vs. the best bilingual experiment and transform the ratio into a percentage:

- $X \rightarrow DE$: 86 % of best monolingual German IR system
- $X \rightarrow EN$: 76 % of best monolingual English IR system
- $X \rightarrow PT$: 90 % of best monolingual Portuguese IR system

Note that there is an almost constant proportion in this result since CLEF 2006: Portuguese is usually the best performer. German and surprisingly English are last, even though there are several geographical and linguistic resources for these languages.

It is possible to run another kind of statistical analysis for a comparison between bilingual and monolingual performance which is not based on the comparison of the single best experiments, but on the average performance of each topic on the monolingual and bilingual task [6]. The results of this analysis are as follows:

- Monolingual German performs better than bilingual German. The mean average precision per topic of the monolingual task is significantly higher than the mean average precision per topic of the bilingual task;

- Monolingual English performs significantly better than bilingual English;
- Monolingual Portuguese performance is not significantly different from bilingual Portuguese.

That means, even though the performance difference between 86% and 90% of the German and Portuguese tasks presented above seems to be small, only for Portuguese, the difference between monolingual and bilingual performance is not significant.

4.2 Grouped Analysis

When the goal is to validate how well results can be expected to hold beyond a particular set of queries, statistical testing can help to determine what differences between runs appear to be real as opposed to differences that are due to sampling issues. For this purpose, a Tukey t-test was performed in order to study the groups of experiments which performed equally or significantly different [7].

There was an interesting result: the performance of all the experiments were statistically not different except for one participant, California State University San Marcos (CSUSM) who performed significantly worse compared to all other experiments. This experiment is an important baseline for comparison with all the approaches because the experiments sent by CSUMS were:

- automatic (no manual processing),
- without any query expansion,
- using only title and description (without narrative),
- without any translation in the bilingual task (no translation module at all),
- without removing diacritic marks in the collection.

This shows that if a cross-lingual system is designed with the basic functionalities, the performances of this system will be significantly worse compared to systems with advanced components. For the other systems, different optimization approaches can lead to optimal performances and no approach can be considered superior yet.

5 Conclusions and Future Work

GeoCLEF developed 100 topics and relevance judgments for geographic information retrieval. Another 26 topics with geographic specification were selected out of previous ad-hoc CLEF topics. This test collection is the first GIR test collection available for the research community and it will be a benchmark for future research.

GIR is receiving increased attention both through the GeoCLEF effort and through scientific workshops on the topic. The wide availability of geographic systems on the Internet will further increase the demand for and the interest in geographic information retrieval.

For GeoCLEF 2009, a new GikIP track is again planned, [5]. A query parsing and classification task is again planned for CLEF 2009. It requires the participants to identify geographic queries within a large set of queries from a search engine log.

Acknowledgments

The organization of GeoCLEF was mainly volunteer work. The English assessment was performed by Fredric Gey and Ray Larson from the University of California at Berkeley and Samaneh Beheshti-Kashi and Wiebke Alscher from the University of Hildesheim. The German assessment was carried out by a group of people from the University of Hildesheim: Lea Drolshagen, Kathrin Stackmann,, Julia Schulz, Nadine Mahrholz, Daniela Wilczek and Ralph Kölle. The Portuguese documents were assessed by Ana Frankenberg, Cláudia Freitas, Cristina Mota, David Cruz, Diana Santos, Hugo Oliveira, Luís Costa, Luís Miguel Cabral, Marcirio Chaves, Paula Carvalho, Paulo Rocha, Pedro Martins, Rosário Silva, Sérgio Matos and Susana Inácio, all from Linguatca (the Linguatca project is 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 checked by Sven Hartrumpf from the University of Hagen (Germany).

References

1. Braschler, M.; Peters, C.: Cross-Language Evaluation Forum: Objectives, Results, Achievements. In *Information Retrieval* 7(1-2) 7-31
2. Buckley, C.; Voorhees, E.: Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge & London: MIT Press. 2005. pp. 53-75.
3. Gey, F.; Larson, R.; Sanderson, M.; Bishoff, K.; Mandl, T.; Womser-Hacker, C.; Santos, D.; Rocha, P.; Di Nunzio, G. M.; Ferro, N.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In *7th Workshop of the Cross-Language Evaluation Forum: CLEF 2006*, Alicante, Spain, Springer [LNCS 4730] 2007. pp. 852-876.
4. Mandl, T.; Gey, F.; Di Nunzio, G. M.; Ferro, N.; Larson, R.; Sanderson, M.; Santos, D.; Womser-Hacker, C.; Xing, X.: GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: *8th Workshop Cross-Language Evaluation Forum. CLEF 2007, Budapest, Hungary*, Springer [LNCS 5152] 2008
5. Santos, D.; Cardoso, N.; Carvalho, P.; Dornescu, I.; Hartrumpf, S.; Leveling, J.; Skalban, Y. GikIP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In *this volume*.
6. Crivellari, F.; Di Nunzio, G. M.; Ferro, N.: A Statistical and Graphical Methodology for Comparing Bilingual to Mono-lingual Cross-Language Information Retrieval. In Agosti, M. (ed.) *Information Access through Search Engines and Digital Libraries*, Springer [Information Retrieval Series 22] 2008. pages 171-188.
7. Gey, F.; Larson, R.; Mandl, T.; Womser-Hacker, C.; Santos, D.; Carvalho, P.; Di Nunzio, G. M.; Ferro, N.: GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In Carol P.; Borri, F. (ed.) *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, Sept. 2008. http://www.clef-campaign.org/2008/working_notes/GeoCLEF-2008-overview-notebook-paperWNfinal.pdf.