Geographic Information Retrieval (GIR): Algorithms and Approaches

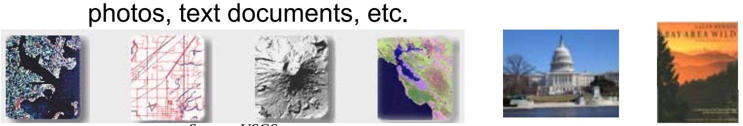
Ray R. Larson
University of California, Berkeley
School of Information

Overview

- What is GIR?
- Spatial Approaches to GIR
- A Logistic Regression Approach to GIR
  - Model
  - Testing and Results
  - Example using Google Earth as an interface
- GIR Evaluation Tests
  - GeoCLEF
  - GikiCLEF
  - NTCIR GeoTime

Geographic Information Retrieval (GIR)

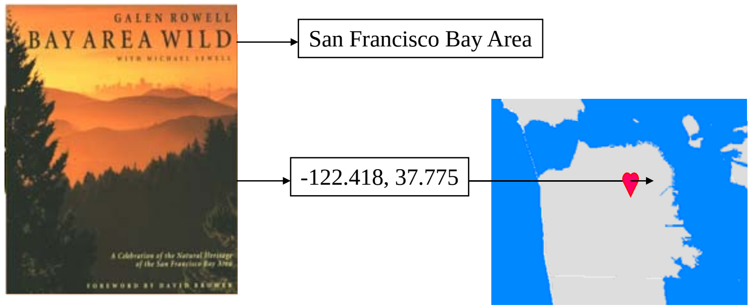
- Geographic information retrieval (GIR) is concerned with spatial approaches to the retrieval of geographically referenced, or georeferenced, information objects (GIOs)
  - about specific regions or features on or near the surface of the Earth.
  - Geospatial data are a special type of GIO that encodes a specific geographic feature or set of features along with associated attributes
    - maps, air photos, satellite imagery, digital geographic data, photos, text documents, etc.



Source: USGS

Georeferencing and GIR

- Within a GIR system, e.g., a geographic digital library, information objects can be georeferenced by place names or by geographic coordinates (i.e. longitude & latitude)



GIR is not GIS

- **GIS** is concerned with spatial representations, relationships, and analysis at the level of the individual spatial object or field

- **GIR** is concerned with the *retrieval* of geographic information resources (and geographic information objects at the set level) that may be relevant to a geographic query region

Spatial Approaches to GIR

- A spatial approach to geographic information retrieval is one based on the integrated use of spatial representations, and spatial relationships.
- A spatial approach to GIR can be qualitative or quantitative
  - **Quantitative:** based on the geometric spatial properties of a geographic information object
  - **Qualitative:** based on the non-geometric spatial properties.

Spatial Matching and Ranking

- Spatial similarity can be considered as a indicator of relevance: documents whose spatial content is more similar to the spatial content of query will be considered more relevant to the information need represented by the query.
- Need to consider both:
  - Qualitative, non-geometric spatial attributes
  - Quantitative, geometric spatial attributes
    - Topological relationships and metric details
- We focus on the latter…

Spatial Similarity Measures and Spatial Ranking

- Three basic approaches to spatial similarity measures and ranking
- Method 1: Simple Overlap
- Method 2: Topological Overlap
- Method 3: Degree of Overlap:

Method 1: Simple Overlap

- Candidate geographic information objects (GIOs) that have any overlap with the query region are retrieved.
- Included in the result set are any GIOs that are contained within, overlap, or contain the query region.
- The spatial score for all GIOs is either relevant (1) or not relevant (0).
- The result set cannot be ranked
  - topological relationship only, no metric refinement

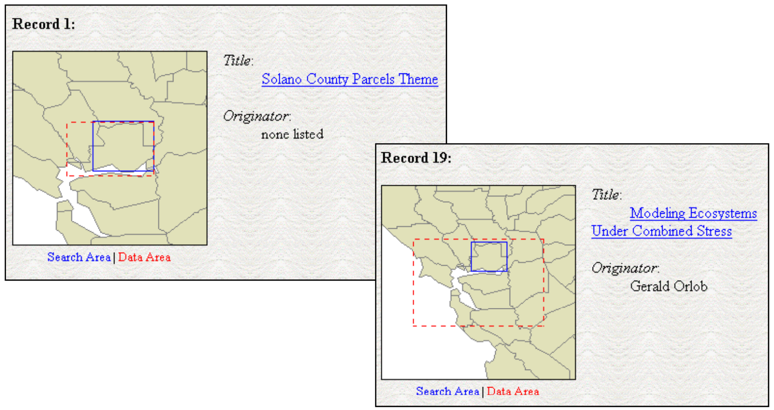
Method 2: Topological Overlap

- Spatial searches are constrained to only those candidate GIOs that either:
  - are completely contained within the query region,
  - overlap with the query region,
  - or, contain the query region.
- Each category is exclusive and all retrieved items are considered relevant.
- The result set cannot be ranked
  - categorized topological relationship only,
  - no metric refinement

Method 3: Degree of Overlap

- Candidate geographic information objects (GIOs) that have any overlap with the query region are retrieved.
- A spatial similarity score is determined based on the degree to which the candidate GIO overlaps with the query region.
- The greater the overlap with respect to the query region, the higher the spatial similarity score.
- This method provides a score by which the result set can be ranked
  - topological relationship: overlap
  - metric refinement: area of overlap

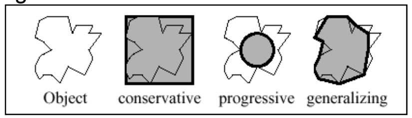
Example: Results display from CheshireGeo:



http://calsip.regis.berkeley.edu/pattyf/mapserver/cheshire2/cheshire_init.html

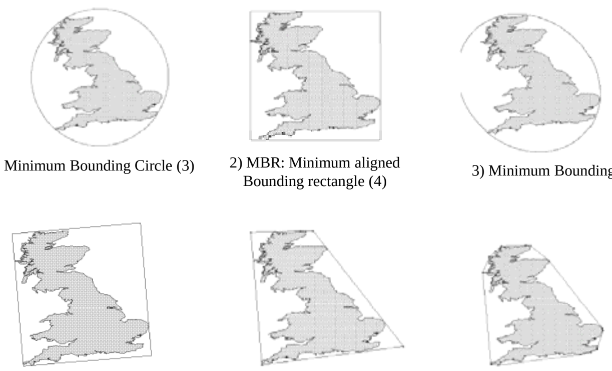
Geometric Approximations

- The decomposition of spatial objects into approximate representations is a common approach to simplifying complex and often multi-part coordinate representations
- Types of Geometric Approximations
  - Conservative: superset
  - Progressive: subset
  - Generalizing: could be either



  - Concave or Convex
    - Geometric operations on convex polygons much faster

Other convex, conservative Approximations



1) Minimum Bounding Circle (3)
2) MBR: Minimum aligned Bounding rectangle (4)
3) Minimum Bounding Ellipse (5)
4) Rotated minimum bounding rectangle (5)
5) 4-corner convex polygon (8)
6) Convex hull (varies)

After Brinkhoff et al, 1993b

Presented in order of increasing quality. Number in parentheses denotes number of parameters needed to store representation

Our Research Questions

- Spatial Ranking
  - How effectively can the spatial similarity between a query region and a document region be evaluated and ranked based on the overlap of the geometric approximations for these regions?
- Geometric Approximations & Spatial Ranking:
  - How do different geometric approximations affect the rankings?
    - MBRs: the most popular approximation
    - Convex hulls: the highest quality convex approximation

Spatial Ranking:
Methods for computing spatial similarity

Reference	Formula
Hill, 1990[10]	$Range = 2\frac{O}{Q+C}$
Walker et al, 1992[19]	$Range = MIN\left(\frac{O}{Q}, \frac{O}{C}\right)$
Beard and Sharma, 1997[3]	Case 1: Q contains C $Range = \frac{C}{Q}$ Case 2: Q and C overlap $Range = \frac{O\%}{(1-O\%)+100}$ Case 3: Q contained in C $Range = \frac{Q}{C}$
Where: Q = area of query region C = area of candidate GIO O = area of overlap for G, C	Range (for all): 0 = no similarity 1 = identical

Proposed Ranking Method

- Probabilistic Spatial Ranking using Logistic Inference
- Probabilistic Models
  - Rigorous formal model attempts to predict the probability that a given document will be relevant to a given query
  - Ranks retrieved documents according to this probability of relevance (Probability Ranking Principle)
  - Rely on accurate estimates of probabilities

Logistic Regression

Probability of relevance is based on Logistic regression from a sample set of documents to determine values of the coefficients.
At retrieval the probability estimate is obtained by:

$$P(R \mid Q, D) = c_0 + \sum_{i=1}^{m} c_i X_i$$

For the *m X* attribute measures (on the following page)

Probabilistic Models: Logistic Regression attributes

- X_1 = area of overlap(query region, candidate GIO) / area of query region
- X_2 = area of overlap(query region, candidate GIO) / area of candidate GIO
- X_3 = 1 – abs(fraction of overlap region that is onshore fraction of candidate GIO that is onshore)
- *Where*:

  Range for all variables is 0 (not similar) to 1 (same)

Probabilistic Models

Advantages

- Strong theoretical basis
- In principle should supply the best predictions of relevance given available information
- Computationally efficient, straight-forward implementation (if based on LR)

Disadvantages

- Relevance information is required -- or is "guestimated"
- Important indicators of relevance may not be captured by the model
- Optimally requires on-going collection of relevance information

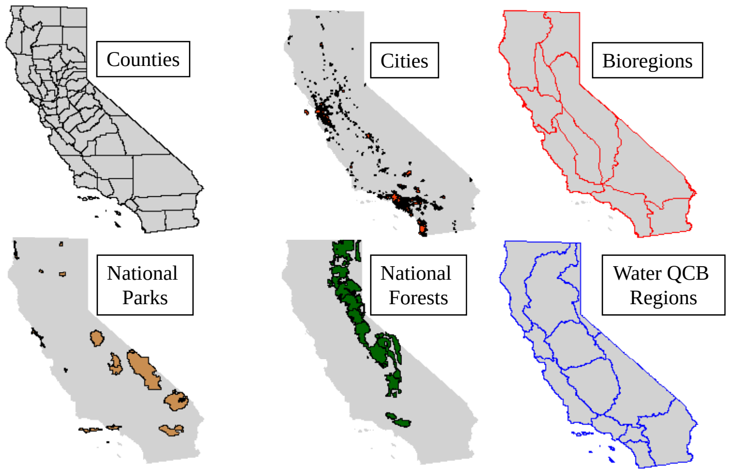
Test Collection

- California Environmental Information Catalog (CEIC)
- http://ceres.ca.gov/catalog.
- Approximately 2500 records selected from collection (Aug 2003) of ~ 4000.

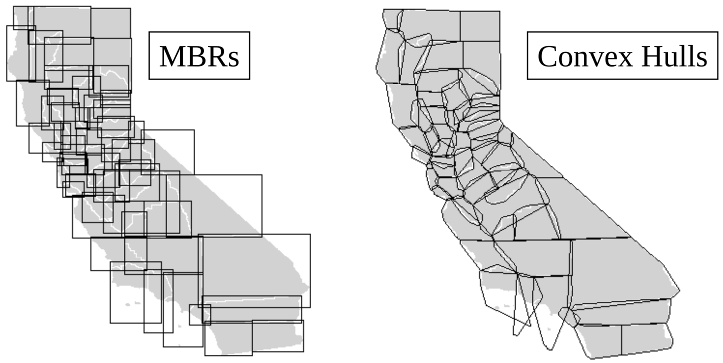
Test Collection Overview

- 2554 metadata records indexed by 322 unique geographic regions (represented as MBRs) and associated place names.
  - 2072 records (81%) indexed by 141 unique CA place names
    - 881 records indexed by 42 unique counties (out of a total of 46 unique counties indexed in CEIC collection)
    - 427 records indexed by 76 cities (of 120)
    - 179 records by 8 bioregions (of 9)
    - 3 records by 2 national parks (of 5)
    - 309 records by 11 national forests (of 11)
    - 3 record by 1 regional water quality control board region (of 1)
    - 270 records by 1 state (CA)
  - 482 records (19%) indexed by 179 unique user defined areas (approx 240) for regions within or overlapping CA
    - 12% represent onshore regions (within the CA mainland)
    - 88% (158 of 179) offshore or coastal regions

CA Named Places in the Test Collection – complex polygons

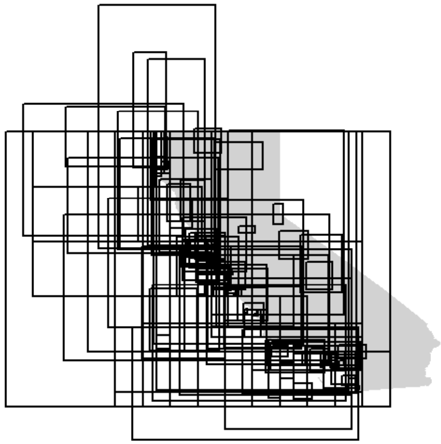


CA Counties – Geometric Approximations

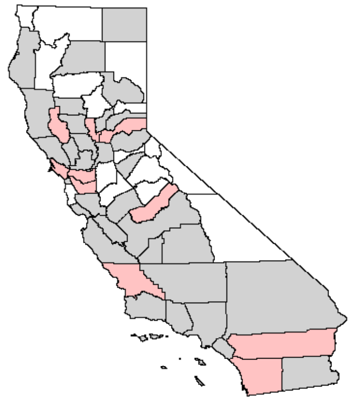


Ave. False Area of Approximation:
MBRs: 94.61% Convex Hulls: 26.73%

CA User Defined Areas (UDAs) in the Test Collection



Test Collection Query Regions: CA Counties



42 of 58 counties referenced in the test collection metadata

- 10 counties randomly selected as query regions to train LR model
- 32 counties used as query regions to test model

Test Collection Relevance Judgements

- Determine the reference set of candidate GIO regions relevant to each county query region:
- Complex polygon data was used to select all CA place named regions (i.e. counties, cities, bioregions, national parks, national forests, and state regional water quality control boards) that overlap each county query region.
- All overlapping regions were reviewed (semi-automatically) to remove sliver matches, i.e. those regions that only overlap due to differences in the resolution of the 6 data sets.
  - Automated review: overlaps where overlap area/GIO area > .00025 considered relevant, else not relevant.
  - Cases manually reviewed: overlap area/query area < .001 and overlap area/GIO area < .02
- The MBRs and metadata for all information objects referenced by UDAs (user-defined areas) were manually reviewed to determine their relevance to each query region. This process could not be automated because, unlike the CA place named regions, there are no complex polygon representations that delineate the UDAs.
- This process resulted in a master file of CA place named regions and UDAs relevant to each of the 42 CA county query regions.

LR model

- X_1 = area of overlap(query region, candidate GIO) / area of query region
- X_2 = area of overlap(query region, candidate GIO) / area of candidate GIO
- *Where*:
  Range for all variables is 0 (not similar) to 1 (same)

Approximation	Logistic Regression Model Fitted on the Training Data
MBR	$LogO(R\|Q,C) = -5.0402 + (6.5154 * X_1) + (5.7729 * X_2)$
Convex Hull	$LogO(R\|Q,C) = -3.4767 + (7.4536 * X_1) + (5.7569 * X_2)$

Some of our Results

Mean Average Query Precision: the average precision values after each new relevant document is observed in a ranked list.

For metadata indexed by CA named place regions:

Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.7193	0.8097
Walker et al., 1992	0.7025	0.8006
Beard & Sharma, 1997	0.7094	0.8116
Logistic Regression	0.9389	0.9973

For all metadata in the test collection:

Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.6722	0.7936
Walker et al., 1992	0.6509	0.7810
Beard & Sharma, 1997	0.6523	0.7778
Logistic Regression	0.8141	0.9099

These results suggest:
- Convex Hulls perform better than MBRs
  - Expected result given that the CH is a higher quality approximation
- A probabilistic ranking based on MBRs can perform as well if not better than a non-probabiliistic ranking method based on Convex Hulls
  - Interesting
  - Since any approximation other than the MBR requires great expense, this suggests that the exploration of new ranking methods based on the MBR are a good way to go.

Some of our Results

Mean Average Query Precision: the average precision values after each new relevant document is observed in a ranked list.

For metadata indexed by CA named place regions:

Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.7193	0.8097
Walker et al., 1992	0.7025	0.8006
Beard & Sharma, 1997	0.7094	0.8116
Logistic Regression	0.9389	0.9973

For all metadata in the test collection:

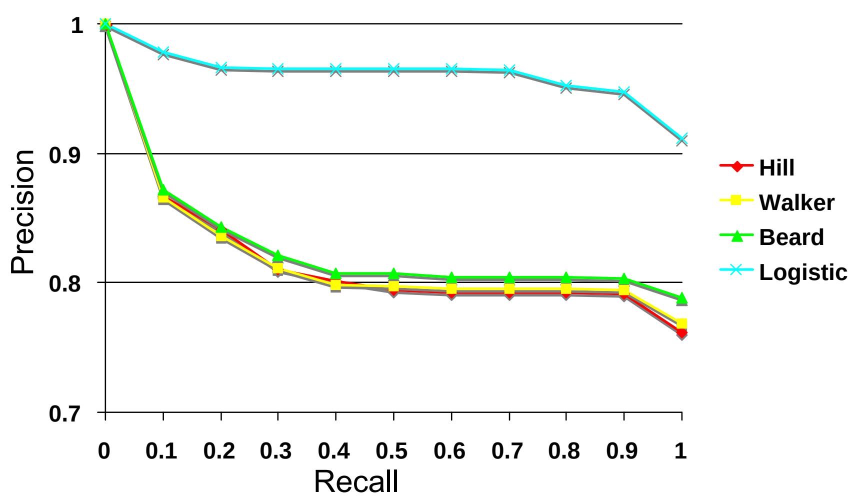
Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.6722	0.7936
Walker et al., 1992	0.6509	0.7810
Beard & Sharma, 1997	0.6523	0.7778
Logistic Regression	0.8141	0.9099

BUT:

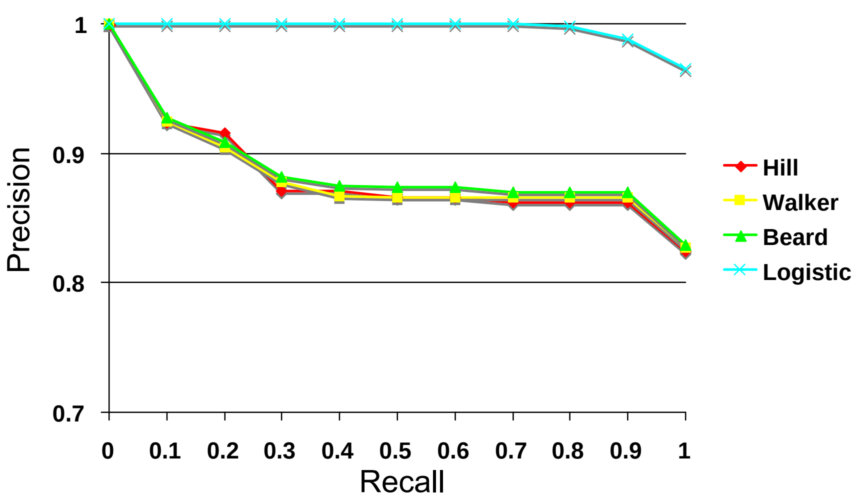
The inclusion of UDA indexed metadata reduces precision.

This is because coarse approximations of onshore or coastal geographic regions will necessarily include much irrelevant offshore area, and vice versa

Results for MBR - Named data



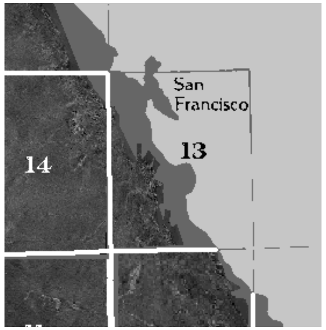
Results for Convex Hulls -Named

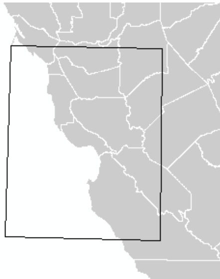


Offshore / Coastal Problem

California EEZ Sonar Imagery Map – GLORIA Quad 13

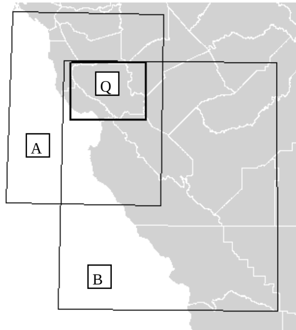
• **PROBLEM: the MBR for GLORIA Quad 13 overlaps with several counties that area completely inland.**





Adding Shorefactor Feature Variable

Shorefactor = 1 – abs(fraction of query region approximation that is onshore – fraction of candidate GIO approximation that is onshore)



Onshore Areas

Candidate GIO MBRs
A) GLORIA Quad 13: fraction onshore = .55
B) WATER Project Area: fraction onshore = .74

Query Region MBR
Q) Santa Clara County: fraction onshore = .95

Computing Shorefactor:
Q – A Shorefactor: 1 – abs(.95 - .55) = .60
Q – B Shorefactor: 1 – abs(.95 - .74) = .79

Even though A & B have the same area of overlap with the query region, B has a higher shorefactor, which would weight this GIO's similarity score higher than A's.

Note: geographic content of A is completely offshore, that of B is completely onshore.

About the Shorefactor Variable

- Characterizes the relationship between the query and candidate GIO regions based on the extent to which their approximations overlap with onshore areas (or offshore areas).
- Assumption: a candidate region is more likely to be relevant to the query region if the extent to which its approximation is onshore (or offshore) is similar to that of the query region's approximation.

About the Shorefactor Variable

- The use of the shorefactor variable is presented as an example of how geographic context can be integrated into the spatial ranking process.
- Performance: Onshore fraction for each GIO approximation can be pre-indexed. Thus, for each query only the onshore fraction of the query region needs to be calculated using a geometric operation. The computational complexity of this type of operation is dependent on the complexity of the coordinate representations of the query region (we used the MBR and Convex hull approximations) and the onshore region (we used a very generalized concave polygon w/ only 154 pts).

Shorefactor Model

- X1 = area of overlap(query region, candidate GIO) / area of query region
- X2 = area of overlap(query region, candidate GIO) / area of candidate GIO
- X3 = 1 – abs(fraction of query region approximation that is onshore – fraction of candidate GIO approximation that is onshore)
  - Where: Range for all variables is 0 (not similar) to 1 (same)

Approximation	Logistic Regression Model Fitted on the Training Data
MBR	1. LogO(R|Q,C) = $-1.6747 + (1.9871 * X_1) + (3.2970 * X_2)$
	2. LogO(R|Q,C) = $-2.1303 + (1.9138 * X_1) + (3.2157 * X_2) + (0.7451 * X_3)$
Convex Hull	1. LogO(R|Q,C) = $-1.2124 + (1.4471 * X_1) + (5.4585 * X_2)$
	2. LogO(R|Q,C) = $-1.2825 + (1.4341 * X_1) + (5.4096 * X_2) + (0.1267 * X_3)$

Some of our Results, with Shorefactor

For all metadata in the test collection:

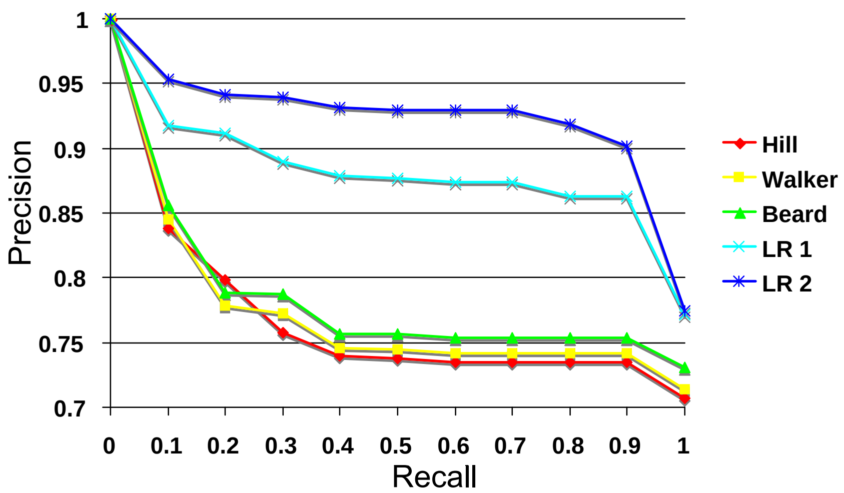
Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.6722	0.7936
Walker et al., 1992	0.6509	0.7810
Beard & Sharma, 1997	0.6523	0.7778
Logistic Regression 1	0.8141	0.9099
Logistic Regression 2	0.8819	0.9238

Mean Average Query Precision: the average precision values after each new relevant document is observed in a ranked list.

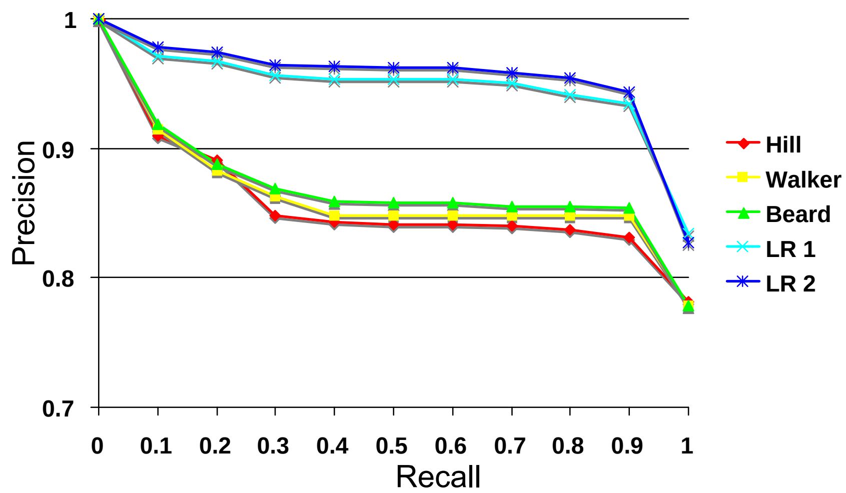
These results suggest:

- Addition of Shorefactor variable improves the model (LR 2), especially for MBRs
- Improvement not so dramatic for convex hull approximations – b/c the problem that shorefactor addresses is not that significant when areas are represented by convex hulls.

Results for All Data - MBRs

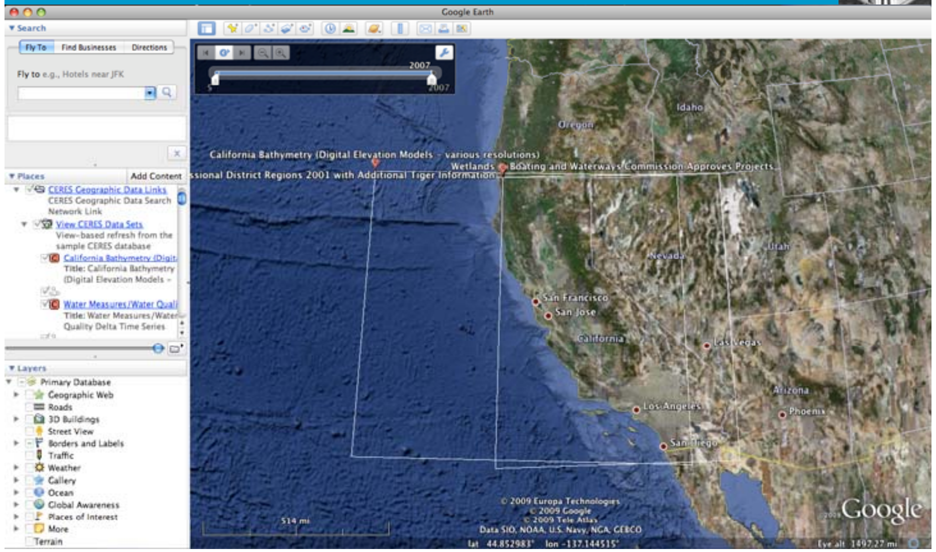


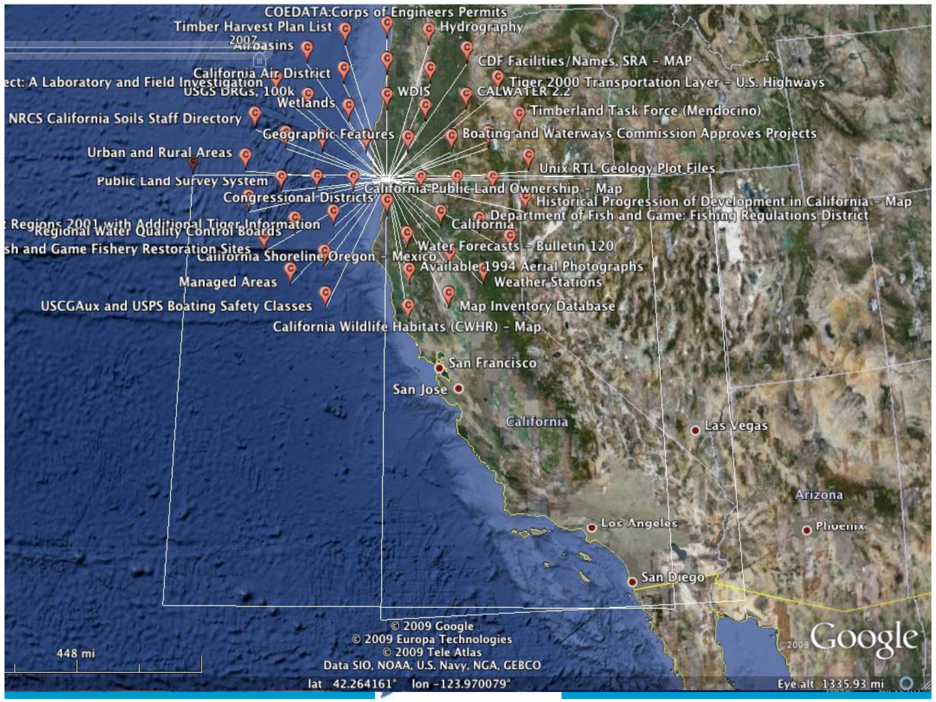
Results for All Data - Convex Hull

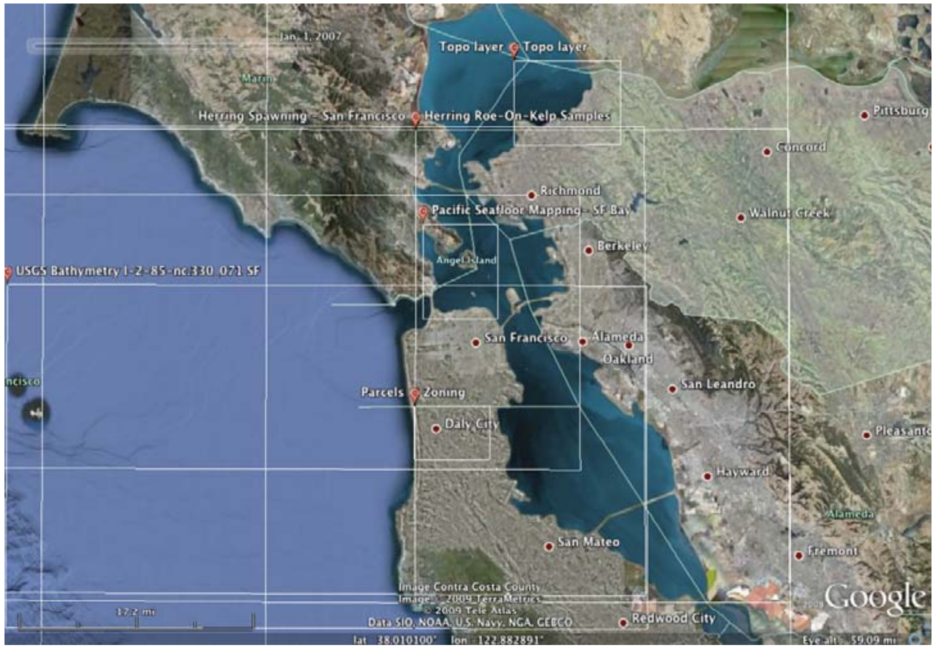


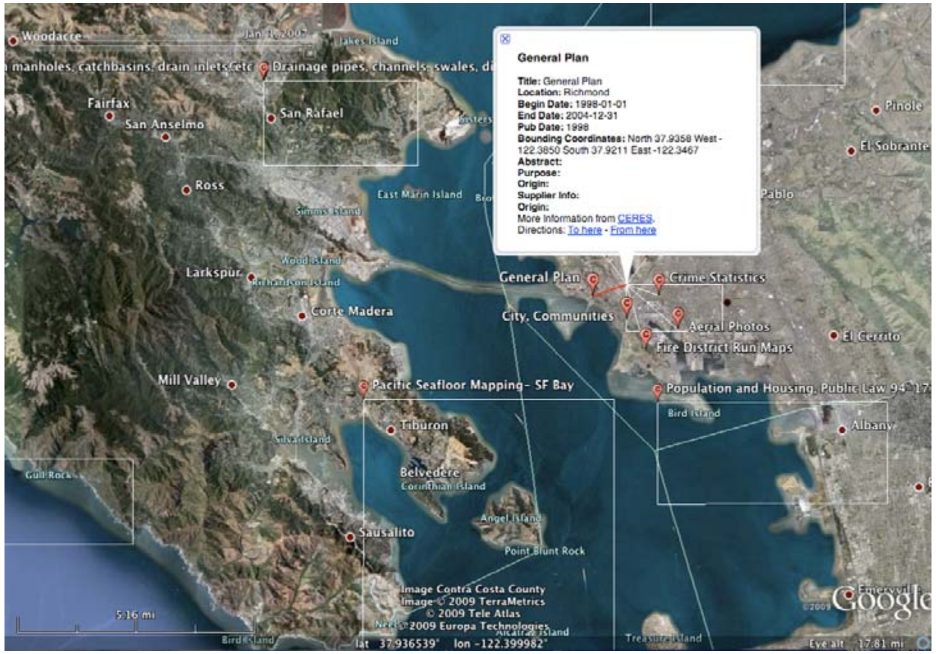
GIR Examples

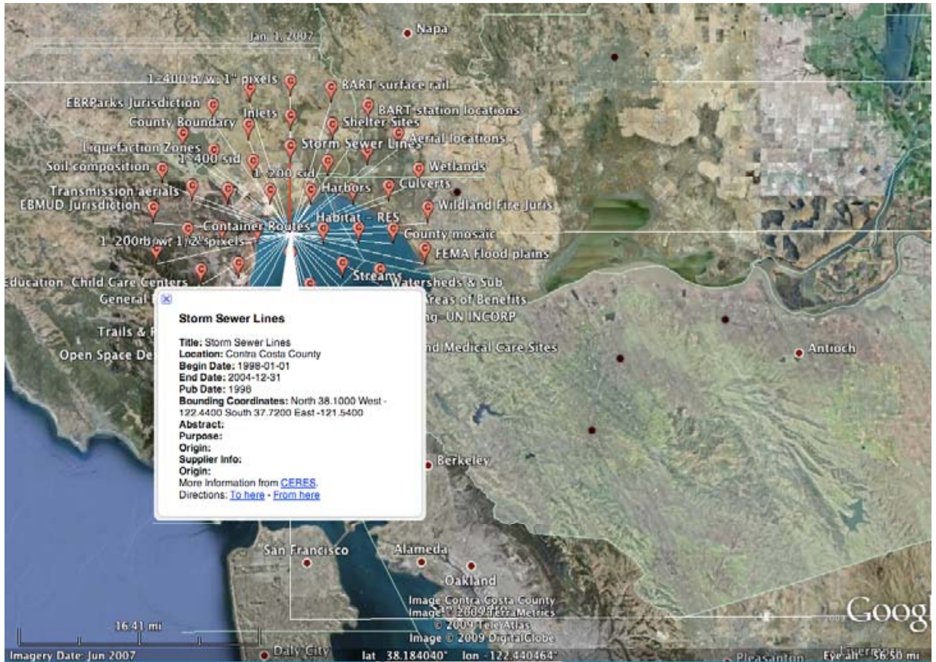
- The following screen captures are from a GIR application using the algorithms (2 variable logistic regression model) and data (the CIEC database data)
- Uses a Google Earth network link to provide a GIR search interface

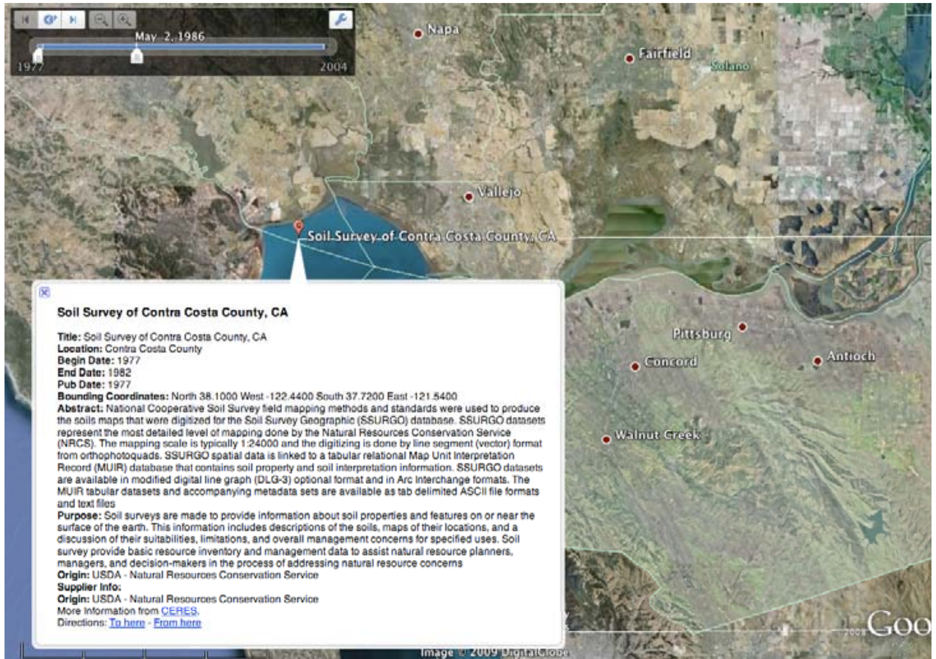












GIR Evaluations

- The GeoCLEF track of CLEF conducted evaluations of GIR systems using text-based queries
  - One finding was that good text retrieval methods may work as well, or better, than more complex geographic modeling and query expansion approaches
- The GikiCLEF track of CLEF
- New NTCIR-GEOTIME track focuses GeoTemporal Information starting -- see http://metadata.berkeley.edu/NTCIR-GeoTime/

GeoCLEF Overview

- Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Given that many documents (and queries) contain some kind of spatial reference, there are examples where geographical references (geo-references) may be important for IR.
- 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. This would require an additional translation step to enable successful retrieval.
- Existing evaluation campaigns such as TREC and CLEF do not explicitly evaluate geographical IR relevance.
- The aim of GeoCLEF was to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects.

Organizers of GeoCLEF

- Fred Gey and Ray Larson, University of California, Berkeley, USA (gey@berkeley.edu, ray@sims.berkeley.edu)
- Mark Sanderson, Department of Information Studies, University of Sheffield, UK (m.sanderson@sheffield.ac.uk)
- Hideo Joho, University of Glasgow, UK (hideo@dcs.gla.ac.uk)
- Thomas Mandl and Christa Womser-Hacker of U. Hildesheim Germany (German language coordinators)
- Diana Santos and Paulo Rocha of Linguateca (Portuguese coordinators)
- Andrés Montoyo of U. Alicante (Spanish coordinator)

GeoCLEF

- Proposed 2004, first evaluation 2005
- The last GeoCLEF was held in 2008, the new GikiCLEF task is taking its place
- This overview will focus on the topics, participants and performance for GeoCLEF 2005 and 2006, with some looks at 2007 and 2008

Topic for GeoCLEF 2005

Topics translated for both English and German

```
<top>
<num> GC001 </num>
<orignum> C084 </orignum>
<EN-title>Shark Attacks off Australia and
California</EN-title>
<EN-desc> Documents will report any information
 relating to shark attacks on humans. </EN-desc>
<EN-narr> Identify instances where a human was
attacked by a shark, including where the attack
took place and the circumstances surrounding the
attack. Only documents concerning specific attacks
are relevant; unconfirmed shark attacks or
suspected bites are not relevant. </EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept> Shark attacks </EN-concept>
<EN-spatialrelation>near</EN-spatialrelation>
<EN-location> Australia </EN-location>
<EN-location> California </EN-location>
</top>
```

GeoCLEF 2005 Collections

- The document collections for GeoCLEF 2005 are all newswire stories from the years 1994 and 1995 used in previous CLEF competitions.
- The English document collection consists of 169,477 documents from the Glasgow Herald (1995) and the Los Angeles Times (1994).
- The German document collection consists of 294,809 documents from Der Spiegel (1994/95), the Frankfurter Rundschau (1994) and the Swiss news agency SDA (1994/95)
- The same collections were used for all GeoCLEF evaluations 2005-2008

GeoCLEF 2005 Documents

- In both collections, the documents have a common structure:
- newspaper-specific information like:
  - date
  - page
  - issue
  - special filing numbers
  - one or more titles
  - a byline
  - the actual text.
- The document collections were not explicitly geographically tagged or contained any other location-specific information.

GeoCLEF 2005 Runs

Group Name	Mono EN	Mono DE	Bilin X→E	Bilin X→DE	Total Runs
California State University, San Marcos	2	0	2	0	4
Grupo XLDB (Universidade de Lisboa)	6	4	4	0	14
Linguateca (Portugal and Norway)	-	-	-	-	-
Linguit GmbH. (Germany)	16	0	0	0	16
MetaCarta Inc.	2	0	0	0	2
MIRACLE (Universidad Polit cnica de Madrid)	5	5	0	0	10
NICTA, University of Melbourne	4	0	0	0	4
TALP (Universitat Polit cnica de Catalunya)	4	0	0	0	4
Universidad Polit cnica de Valencia	2	0	0	0	2
University of Alicante	5	4	12	13	34
University of California, Berkeley (Berkeley 1)	3	3	2	2	10
University of California, Berkeley (Berkeley 2)	4	4	2	2	12
University of Hagen (FernUniversitŠt in Hagen)	0	5	0	0	5
Total Submitted Runs	53	25	22	17	117
Number of Groups Participating in Task	11	6	5	3	12

† Linguateca helped with evaluation, but did not submit runs

GeoCLEF 2006 Topics

Topics in English, German, Spanish and Portuguese

<top>
<num>GC026</num>
<EN-title>Wine regions around rivers in Europe</EN-title>
<EN-desc>Documents about wine regions along the banks of European rivers</EN-desc>
<EN-narr>Relevant documents describe a wine region along a major river in European countries. To be relevant the document must name the region and the river.</EN-narr>
</top>
<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>

GeoCLEF 2006 Topics

<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>

<top>
<num>GC042</num>
<EN-title>Regional elections in Northern Germany</EN-title>
<EN-desc>Documents about regional elections in Northern Germany</EN-desc>
<EN-narr>Relevant documents are those reporting the campaign or results for the state parliaments of any of the regions of Northern Germany. The states of north ern Germany are commonly Bremen, Hamburg, Lower Saxony, Mecklenburg-Western Pomerania and Schleswig-Holstein. Only regional elections are relevant; municipal, national and European elections are not.</EN-narr></top>

GeoCLEF 2006 Collections

- Same English and German documents as 2005
- Added Spanish and Portuguese collections
  - Spanish: EFE 1994-1995
  - Portuguese: Público 1994-1995, Folha de São Paulo 1994-1995
- For 2007 and 2008 the Spanish collection was dropped

GeoCLEF 2006 Runs

NAME	DE	EN	ES	PT	X2DE	X2EN	X2ES	X2PT	Total
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				13
imp-coll		2							2
jaen		5							5
ms-china		5							5
nicta		5							5
rfia-upv		4							4
sanmarcos		5	5	4			3	2	19
talp		5							5
u.buffalo		4							4
u.groningen		5							5
u.twente		5							5
unsw		5							5
xldb		5		5					10
TOTALS (17)	**16**	**73**	**15**	**13**	**11**	**0**	**5**	**4**	**137**

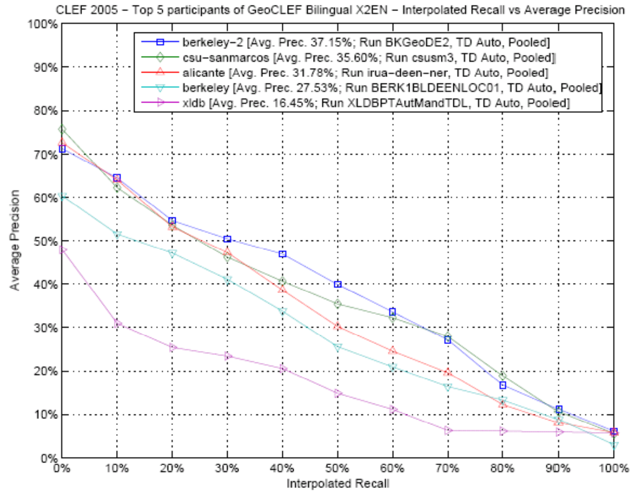
Techniques used by various groups in 2005 and 2006

- Ad-hoc text retrieval techniques (blind feedback, German word de-compounding, etc.)
- Question-answering modules
- Gazetteer construction (GNIS, World Gazetteer)
- Toponym Named Entity Extraction
- Term expansion using Wordnet, geographic thesauri
- Toponym resolution
- NLP – Geofiltering predicates
- Latitude-longitude assignment
- Gazetteer-based query expansion

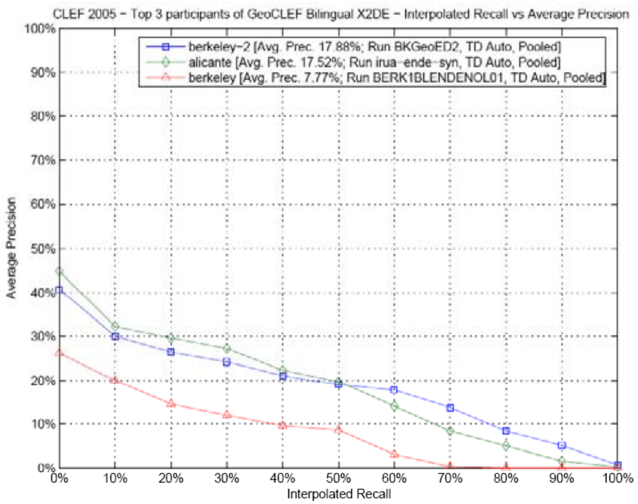
Best-Performing Monolingual Runs: GeoCLEF 2005

Best monolingual-English-run	MAP	Best monolingual-German-run	MAP
berkeley-2_BKGeoE1	0.3936	berkeley-2_BKGeoD3	0.2042
csu-sanmarcos_csusm1	0.3613	alicante_irua-de-titledescgeotags	0.1227
alicante_irua-en-ner	0.3495	miracle_GCdeNOR	0.1163
berkeley_BERK1MLENLOC03	0.2924	xldb_XLDBDEManTDGKBm3	0.1123
miracle_GCenNOR	0.2653	hagen_FUHo14td	0.1053
nicta_i2d2Run1	0.2514	berkeley_BERK1MLDELOC02	0.0535
linguit_LTITLE	0.2362		
xldb_XLDBENManTDL	0.2253		
talp_geotalpIR4	0.2231		
metacarta_run0	0.1496		
u.valencia_dsic_gc052	0.1464		

Bilingual English Performance



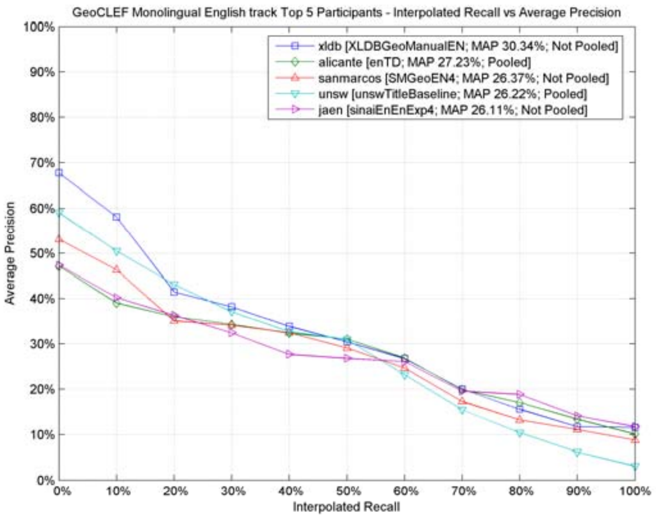
Bilingual German Performance



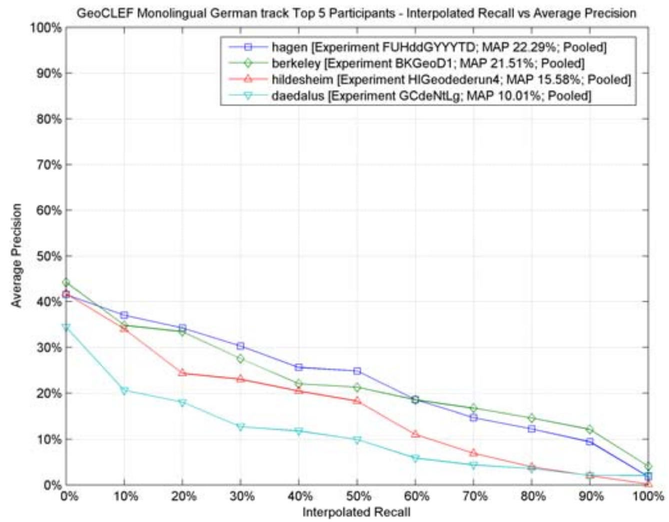
GeoCLEF 2006 Top Mono. Runs

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	unswTitleB aseline pooled	sinaiEnEnEx p4 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	HIGeodeder un4 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%

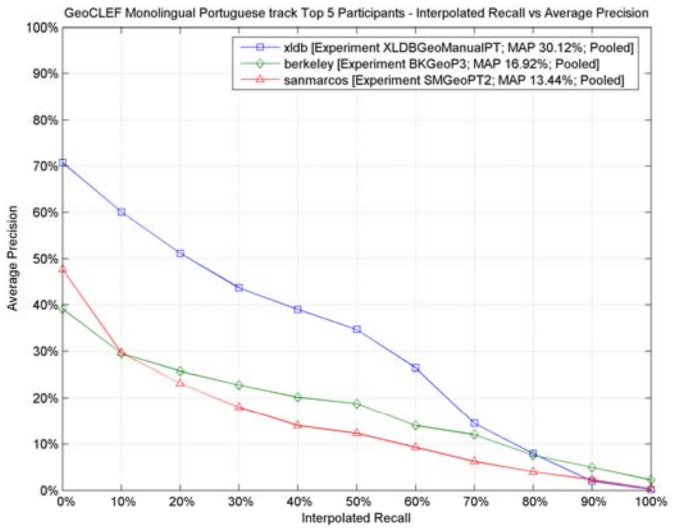
Monolingual English 2006



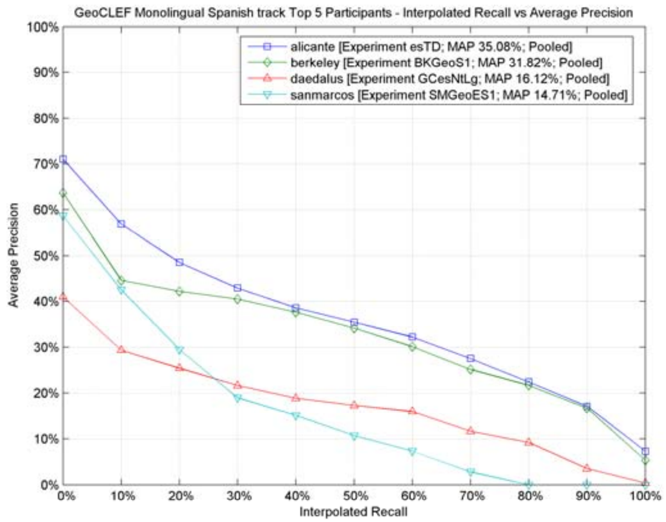
Monolingual German 2006



Monolingual Portuguese 2006



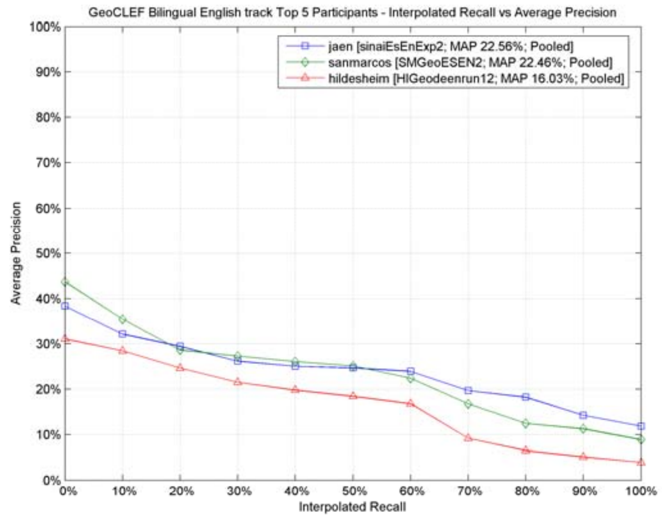
Monolingual Spanish 2006



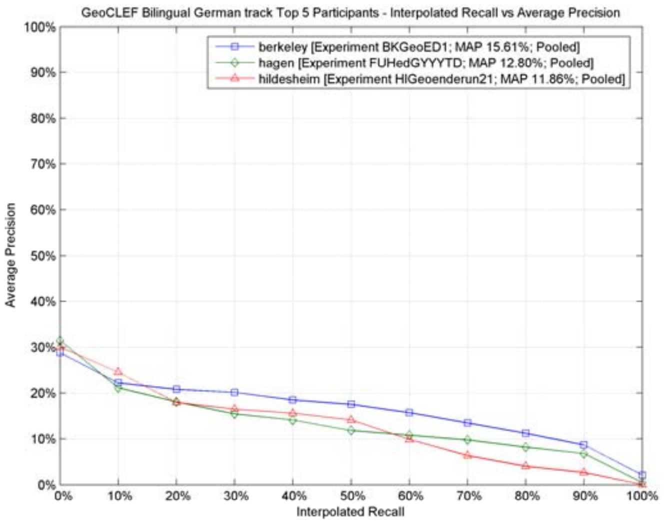
GeoCLEF 2006 Top Biling. Runs

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Bilingual English	Part.	jaen*	sanmarcos	hildesheim*			
	Run	sinaiESENEXP2 pooled	SMGeoESEN2 pooled	HIGeodeenrun12 pooled			
	Avg. Prec.	22.56%	22.46%	16.03%			40.74%
Bilingual German	Part.	berkeley	hagen	hildesheim*			
	Run	BKGeoED1 pooled	FUHedGYYYTD pooled	HIGeoenderun21 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%

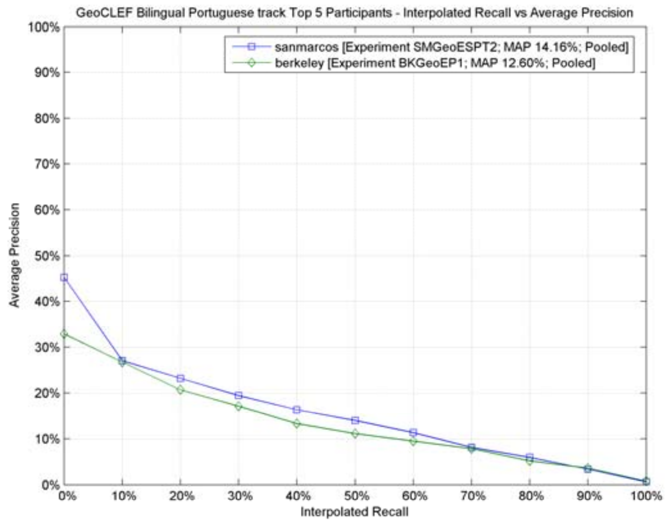
Bilingual English 2006



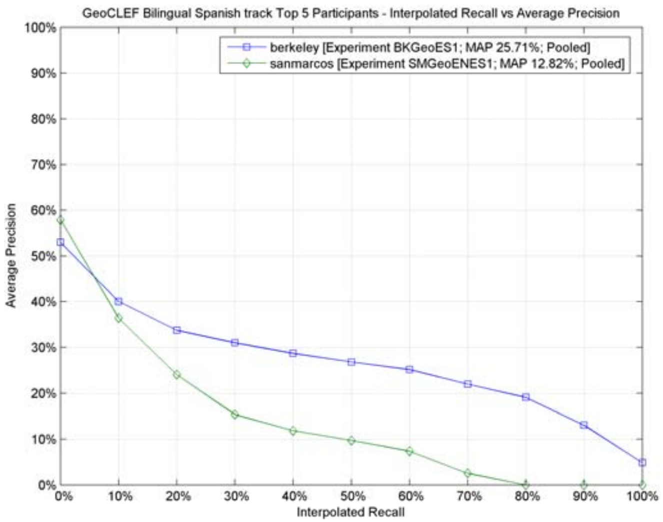
Bilingual German 2006



Bilingual Portuguese 2006



Bilingual Spanish 2006



GeoCLEF Collections 2007

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

Example Topics 2007

<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

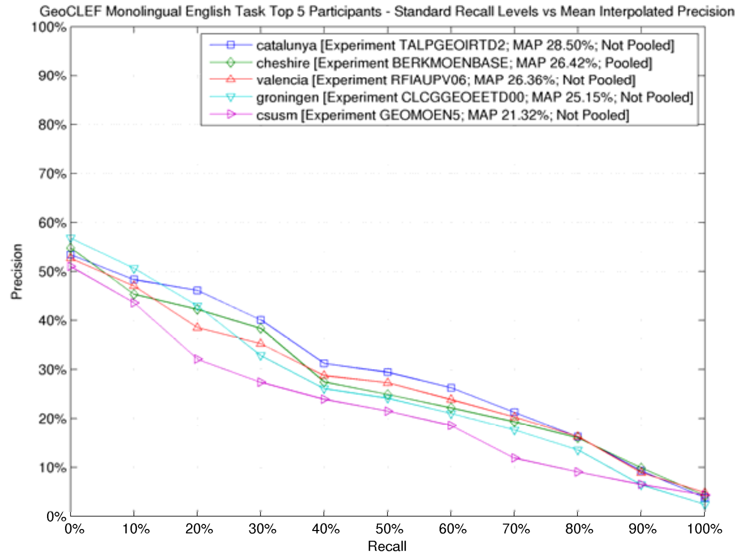
Participant Approaches 2007

- 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

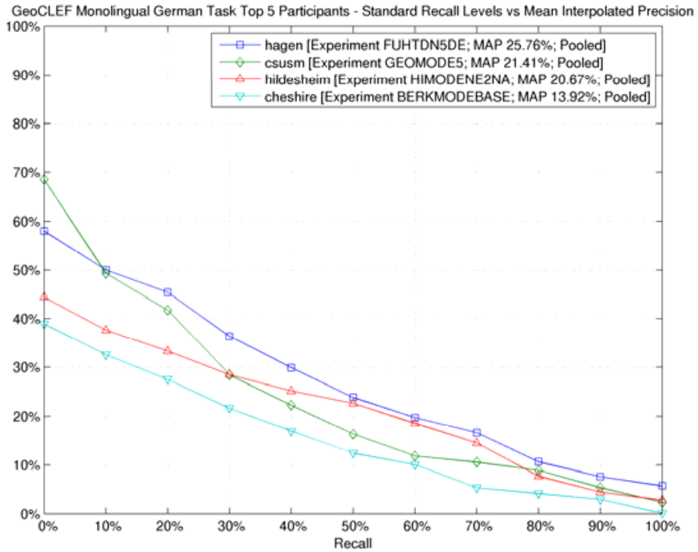
Monolingual Results 2007

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%
	–			**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%
	–			**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%
	–			**442 %**

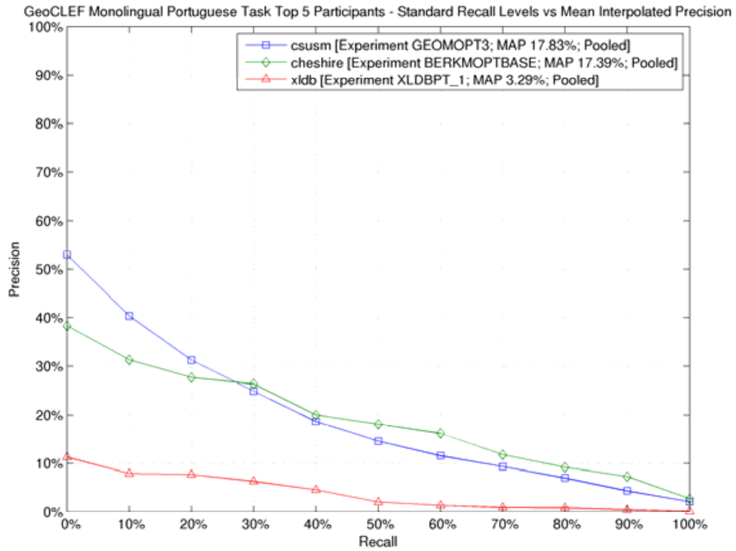
Monolingual English 2007



Monolingual German 2007



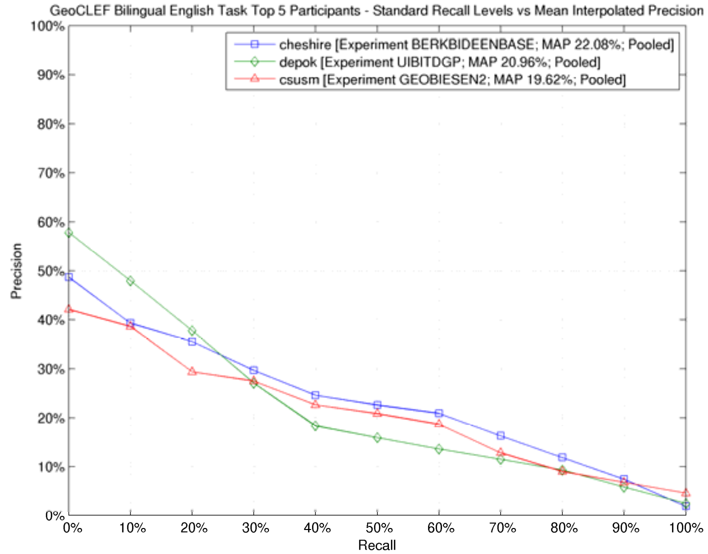
Monolingual Portuguese 2007



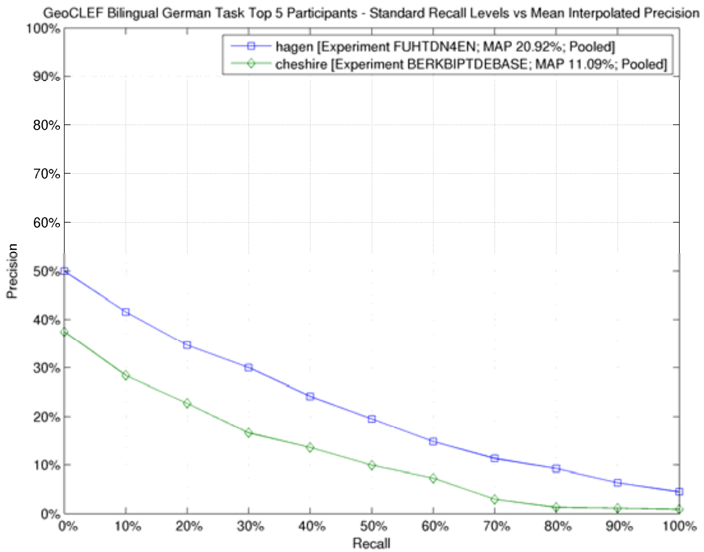
Bilingual results 2007

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%

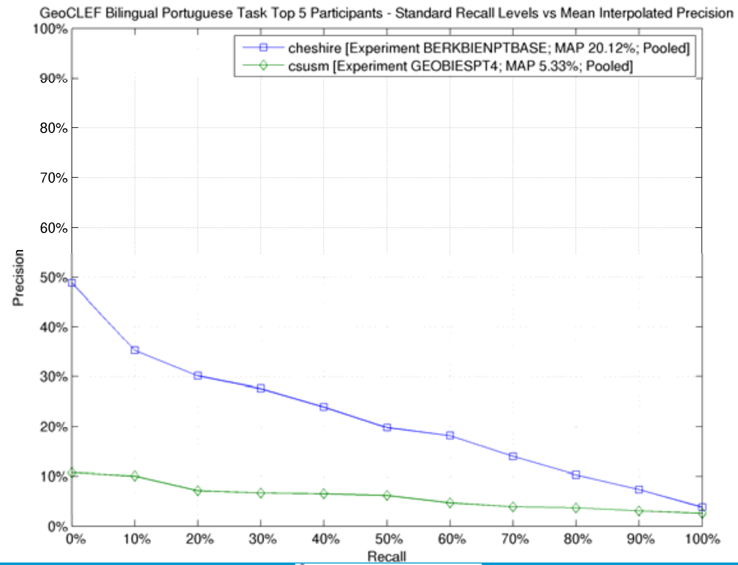
Bilingual English 2007



Bilingual German 2007



Bilingual Portuguese 2007



GeoCLEF 2008

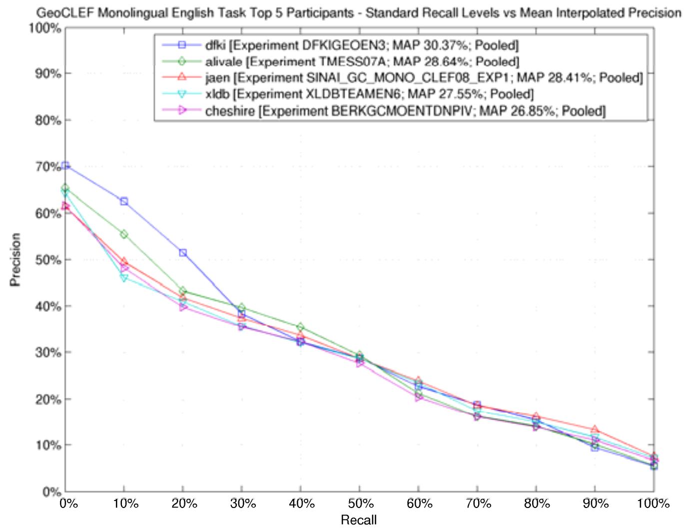
- The 2008 evaluation continued the same basic approach to topics and results with the same test collections
- In 2008 more of the topics were originally formulated in Portuguese, and then translated to English and German

Example Topics 2008

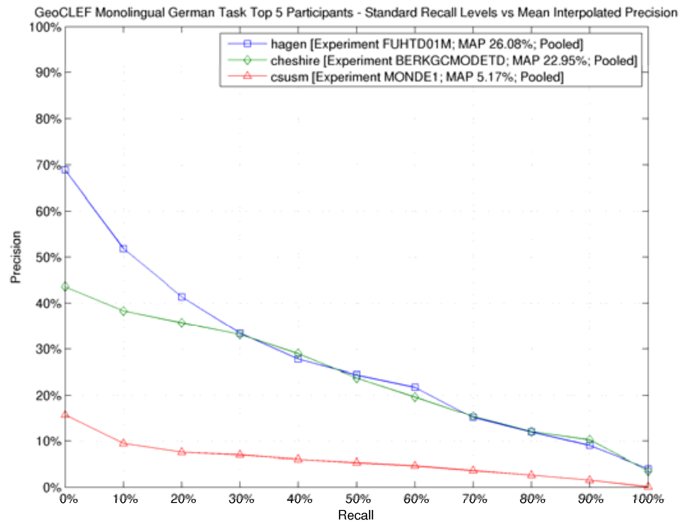
Tab. 3: Topics GC08958 and GC08475

<num>10.2452/89-GC</num> Ê<title>Trade fairs in Lower Saxony </title> Ê<desc>Documents reporting about industrial or cultural fairs in Lower Saxony. </desc> Ê<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, OsnabrŸck, Oldenburg and Gšttingen. </narr> Ê</top>	<num>10.2452/84-GC</num> Ê<title>Atentados ˆ bo mba na Irlanda do Norte </title> Ê<desc>Os documentos relevantes mencionem atentados bombistas em localidades da Irlanda do Norte </desc> Ê<narr>Documentos relevantes devem mencionar atentados ˆ bomba na Irlanda do Norte, indicando a localiza□‹o do atentado. </narr> Ê</top>

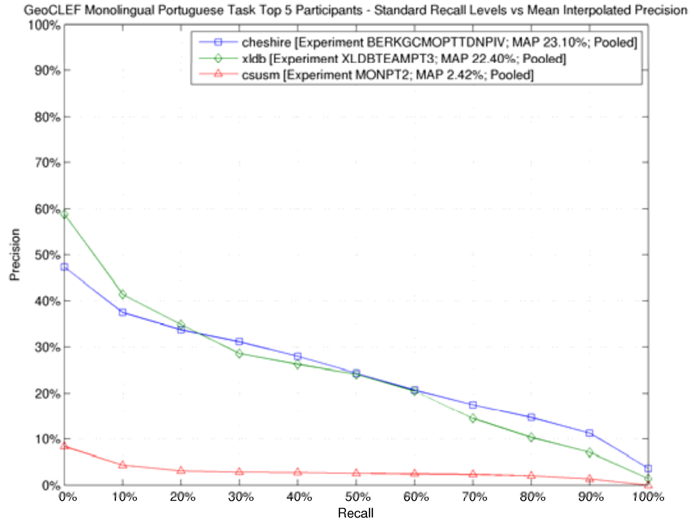
Monolingual English 2008



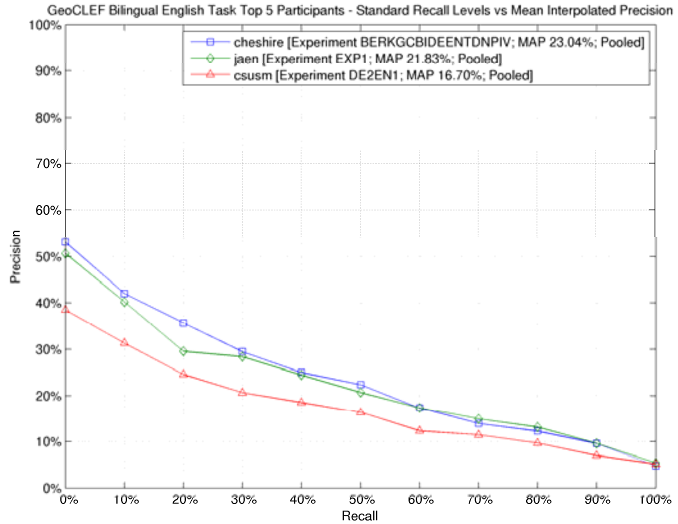
Monolingual German 2008



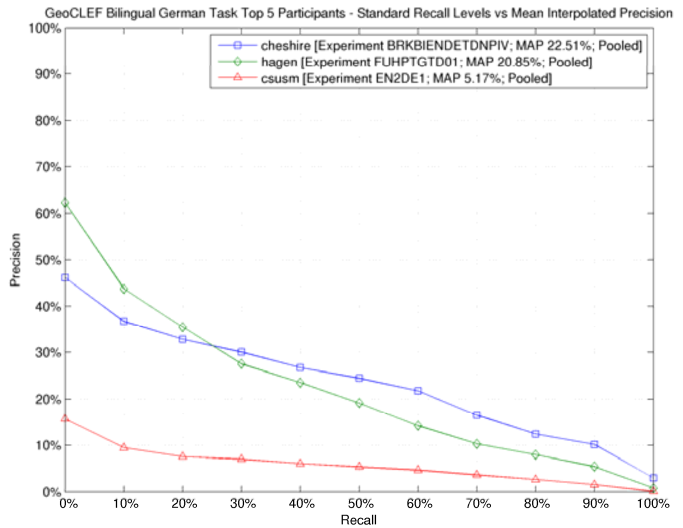
Monolingual Portuguese 2008



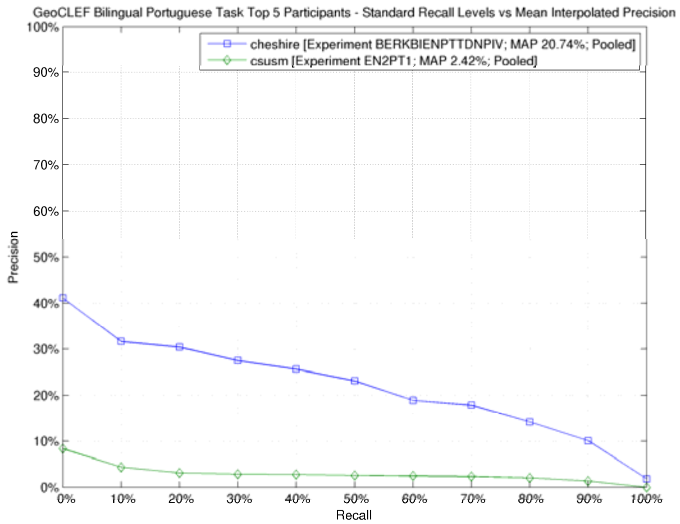
Bilingual English 2008



Bilingual German 2008



Bilingual Portuguese 2008



Cheshire Results 2007-2008

- The good results obtained in 2007 and 2008 by our system were not due to explicit geographic processing (such as explicit geographic query expansion or geometric approaches)
- We used only text retrieval methods as used in other text retrieval tasks
  - Logistic regression text retrieval with psuedo relevance feedback
- For GeoCLEF type queries, place names searched as text appears to perform as well or better than more complex geographic processing (but good machine translation software is essential)

Comparison of Cheshire Runs

Cheshire Runs 2006-2008

TASK	MAP 2006	MAP 2007	MAP 2008	Diff. '06-'07	Diff. '07-'08	Diff. '06-'08
Monolingual English	0.250	0.264	0.268	5.303	1.493	6.716
Monolingual German	0.215	0.139	0.230	-54.676	39.565	6.522
Monolingual Portuguese	0.162	0.174	0.231	6.897	24.675	29.870
Bilingual English⇒German	0.156	0.090	0.225	-73.333	60.00	30.667
Bilingual English⇒Portuguese	0.126	0.201	0.207	37.313	2.899	39.130

GikiCLEF 2009

- GikiCLEF has replaced GeoCLEF for GIR-related retrieval in the 2009 CLEF Evaluation
- GikiCLEF uses the Wikipedia database in 10 different languages
  - Bulgarian, Dutch, English, German, Italian, Norwegian (Bokmål and Nynorsk), Portuguese, Romanian and Spanish

GikiCLEF 2009

- For GikiCLEF, systems need to answer or address geographically challenging topics, on the Wikipedia collections, *returning Wikipedia document titles as list of answers*
- The user model for which GikiCLEF systems intend to cater for is anyone who is interested in knowing something that might be already included in Wikipedia, but has not enough time or imagination to browse it manually

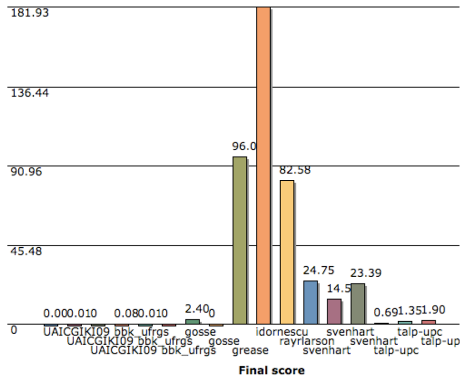
GikiCLEF 2009 Example Topics

- <topic id="GC-2009-01">List the Italian places where Ernest Hemingway visited during his life.</topic>
- <topic id="GC-2009-07"> What capitals of Dutch provinces received their town privileges before the fourteenth century? </topic>
- <topic id="GC-2009-21"> List the left side tributaries of the Po river. </topic>

GikiCLEF Results (just released)

Final score

#	Participant	RunScore	#answers	#Corrects	Precision	Score	
1	idornescu	1	813	385	0.4736	181.9329	Details
2	grease	1	1161	332	0.2860	96.0070	Details
3	rayrlarson	1	564	214	0.3794	82.5861	Details
4	svenhart	1	38	31	0.8158	24.7583	Details
5	svenhart	3	985	142	0.1442	23.3919	Details
6	svenhart	2	994	107	0.1076	14.5190	Details
7	gosse	1	638	36	0.0564	2.4053	Details
8	talp-upc	3	356	26	0.0730	1.9018	Details
9	talp-upc	2	295	20	0.0678	1.3559	Details
10	talp-upc	1	526	18	0.0342	0.6964	Details
11	bbk_ufrgs	1	726	8	0.0110	0.0882	Details
12	UAICGIKI09	2	6420	8	0.0012	0.0156	Details
13	bbk_ufrgs	2	734	3	0.0041	0.0123	Details
14	UAICGIKI09	1	1133	2	0.0018	0.0062	Details
15	gosse	2	272	0	0.0000	0.0000	Details
16	bbk_ufrgs	3	686	0	0.0000	0.0000	Details
17	UAICGIKI09	3	4910	0	0.0000	0.0000	Details



NTCIR GeoTime 2010

- The introductory NTCIR GeoTime track will explore GIR with the added complexity of temporal (time-based) elements
- Will use both English and Japanese collections
- Still open for participation

NTCIR GeoTime Example Topics

<TOPIC ID="ACLIA1-JA-T119">
- <QUESTION LANG="EN">
- <![CDATA[What is the controversy surrounding the use of the Stealth Fighter in Yugoslavia?]]>
</QUESTION>
+ <QUESTION LANG="JA">
ユーゴスラビアに関わるステルス戦闘機の話題にはどんなものがありますか?
- <NARRATIVE LANG="EN">
- <![CDATA[I would like to know about the dates and times of events and places in which there was a controversy surrounding the use of the Stealth Fighter in Yugoslavia.]]>
</NARRATIVE>
- <NARRATIVE LANG="JA">
- <![CDATA[ユーゴスラビアに関わるステルス戦闘機の話題について日時、場所なども含め知りたい。]]>
</NARRATIVE></TOPIC>

This means that identification of dates and geography are an essential pre-requisite to successfully answering this question.

GeoTime Web Site: http://metadata.berkeley.edu/NTCIR-Ge

Thank you.

ありがとう。

Questions?