GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview

Fredric Gey[1], Ray Larson[1], Mark Sanderson[2], Hideo Joho[2], Paul Clough[2] and Vivien Petras[1]

[1]University of California, Berkeley, CA, USA
gey@berkeley.edu , ray@sims.berkeley.edu , vivienp@sims.berkeley.edu
[2]Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk, h.joho@sheffield.ac.uk, p.d.clough@sheffield.ac.uk

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval -- Query Formulation; H.3.4 Systems and Software -- Performance evaluation (efficiency and effectiveness); H.3.7 Digital Libraries

General Terms
Measurement, Performance, Experimentation

Keywords
Geographic Information Retrieval, Cross-language Information Retrieval

Introduction
GeoCLEF is a new track for CLEF 2005. GeoCLEF was run as a pilot track to evaluate retrieval of multilingual documents with an emphasis on geographic search. Existing evaluation campaigns such as TREC and CLEF do not explicitly evaluate geographical relevance. The aim of GeoCLEF is to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants were offered a TREC style ad hoc retrieval task based on existing CLEF collections. GeoCLEF was a collaborative effort by research groups at the University of California, Berkeley and the University of Sheffield. Twelve research groups from a variety of backgrounds and nationalities submitted 117 runs to GeoCLEF.

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Given that many documents contain some kind of spatial reference, there are examples where geographical references (geo-references) may be important for IR. For example, to retrieve, re-rank and visualize search results based on a spatial dimension (e.g. “find me news stories about riots near Dublin City”). In addition to this, many documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. This would require an additional translation step to enable successful retrieval.

For this pilot track 2 languages, German and English, were chosen to be the document languages, while topics were developed in English with topic translations provided for German, Portuguese and Spanish. There were two Geographic Information Retrieval tasks: monolingual (English to English or German to German) and bilingual (language X to English or language X to German, where X was one of English, German, Portuguese or Spanish).

Document collections used in GeoCLEF
The document collections for this year's GeoCLEF experiments are all newswire stories from the years 1994 and 1995 used in previous CLEF competitions. Both the English and German collections contain stories covering international and national news events, therefore representing a wide variety of geographical regions and places. The English document collection consists of 169,477 documents and was composed of stories from the British newspaper The Glasgow Herald (1995) and the American newspaper The Los Angeles Times (1994). The German document collection consists of 294,809 documents from the German news magazine Der Spiegel (1994/95), the German newspaper Frankfurter Rundschau (1994) and the Swiss news agency SDA (1994/95). Although there are more documents in the German collection, the average document length (in terms of words in the actual text) is much larger for the English collection. In both collections, the documents have a common structure: newspaper-specific information like date, page, issue, special filing numbers and usually one or more titles, a byline and the actual text. The document collections were not geographically tagged or contained any other location-specific information.

Generating Search Topics
A total of 25 topics were generated for this year’s GeoCLEF. Ten of them were extended from the past CLEF topics and 15 of them were newly created. This section will discuss the processes taken to create the spatially-aware topics for the track.

Format of topic description

We used the format to describe the search topics, which we proposed in the introductory presentation of Geo Track in CLEF 2004. The format was designed to highlight the geographic aspect of the topics so that the participants can exploit the information in the retrieval process without extracting the geographic references from the description. A sample topic was shown in Figure 1.

```
<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>
```

Figure 1 Topic GC001: Shark Attacks off Australia and California

As can be seen, after the standard data such as the title, description, and narrative, the information about the main concept, locations, and spatial relation which were manually extracted from the title were added to the topics. The above example has the original topic ID of CLEF since it was created based on the past topic. The process of selecting the past CLEF topics for this year's GeoCLEF will be described below.

Analysis of past CLEF topics

Creating a subset of topics from the past CLEF topics had several advantages for us. First of all, it would reduce the amount of effort required to create new topics. Similarly, it would save the resource required to carry out the relevance assessment of the topics. The idea was to revisit the past relevant documents with a greater weight on the geographical aspect. Finally, it was anticipated that the distribution of relevant documents across the collections would be ensured to some extent.

The process of selecting the past CLEF topics for our track was as follows. Firstly, two of the authors went through the topics of the past Ad-Hoc tracks (except Topic 1-40 due to the limited coverage of document collections) and identified those which either contained one or more geographical references in the topic description or asked a geographical question (i.e., Which countries are …?). A total of 72 topics were found from this analysis.

The next stage involved examining the distribution of relevant documents across the collections chosen for this year's track. A cross tabulation was run on the qrel files for the document collections to identify the topics that covered our collections. A total of 10 topics were then chosen based on the above analysis as well as the additional manual examination of the suitability for the track.

One of the characteristics we found from the chosen past CLEF topics was a relatively low granularity of geographical references used in the descriptions. Many referred to countries. This is not surprising given that a requirement of CLEF topics is that they are likely to retrieve relevant documents from as many of the CLEF collections as possible (which are predominately newspaper articles from different countries). Consequently, the geographic references in topics were likely to be to well-known locations, i.e. countries.

However, we felt that the topics with a finer granularity should also be devised to make the track geographically more interesting. Therefore, we decided to create the rest of topics by focusing on each of the chosen collections. 7 topics were created based on the articles of LA Times, and 8 topics were created based on Glasgow Herald. The new topics were then translated into other languages by one of the organisers and the volunteers from the participants.

Geospatial processing of document collections

Geographical references found in the document collections were automatically tagged. This was done for two reasons: firstly, it was thought that highlighting the geographic references in the documents would facilitate the topic generation process; secondly, it would help assessors identify relevant documents more quickly if such references were highlighted. In the end though only some assessments were conducted using such information.

Tagging was conducted using a geo-parsing system developed in the Spatially-Aware Information Retrieval on the Internet (SPIRIT) project (http://www.geospirit.org/). The implementation of the system was built using the information extraction component from the General Architecture for Text Engineering (GATE) system (Cunningham, 2002) with the additional contextual rules especially designed for the geographical entities. The system used several gazetteers such as the SABE (Seamless Administrative Boundaries of Europe) dataset, the Ordnance Survey 1:50,000 Scale Gazetteer for the UK, and the Getty Thesaurus of Geographic Names (TGN). The detail of the geo-parsing system can be found in Clough (2005).

Relevance assessment

Assessment was shared by Berkeley and Sheffield Universities. Sheffield was assigned topics 18-25 for the English collections (LA Times, Glasgow Herald); Berkeley assessed topics 1-17 for English and topics 1-25 for the German collections. Assessment resources were restricted for both groups, which influenced the manner in which assessments were conducted.

Berkeley used the conventional approach of judging documents taken from the pool formed by the top-*n* documents from participants' submissions. In TREC the tradition is to set *n* to 100. However, due to a limited number of assessors, Berkeley set *n* to 60, consistent with the ad-hoc CLEF cutoff. English judgments were conducted by Berkeley authors of this paper, and half of the German judgments were conducted by an external assessor paid €1000 (from CLEF funds). Although restricting the number of documents assessed by so much appears to be a somewhat drastic measure, it was observed at last year's TRECVID that reducing pool depth to as little as 10 had little effect on the relative ordering of runs submitted to that evaluation exercise (Kraaji, Smeaton, Over and Arlandis, 2004). More recently Sanderson and Zobel (2005) conducted a large study of the levels of error in effectiveness measures based on shallow pools and again showed that error levels were little different from those based on much deeper pools.

Sheffield was able to secure some funding to pay students to conduct relevance assessments, but the money had to be spent before geoCLEF participants were due to submit their results. Assessments had to be conducted before the submission date; therefore, Sheffield used the Interactive Searching and Judging (ISJ) method described by Cormack, Palmer and Clarke (1998) and broadly tested by Sanderson and Joho (2004). With this approach to building a set of relevance judgments, assessors for a topic become searchers, who were encouraged to search the topic in as broad and diverse a way as possible, marking any relevant documents found. To this end, an ISJ system was previously built for the SPIRIT project was modified for GeoCLEF (see Figure 4).

Sheffield employed 17 searchers (mostly University students), paying each of them (£40) for a half-day session; one searcher worked for three sessions. In each session, two topics were covered. Before starting, searchers were given a short introduction to the system. The authors of the paper also contributed to the assessing process. As so many searchers were found, Sheffield moved beyond the eight topics assigned to it and contributed judgments to the rest of the English topics, overlapping with Berkeley's judgments. For the judgments used in the GeoCLEF exercise, if two documents were found to judged by both Sheffield and Berkeley, Berkeley's judgment was used. The reason for producing such an overlap is the plan to compare judgment quality between the ISJ process and the more conventional pooling approach, which will be forthcoming.

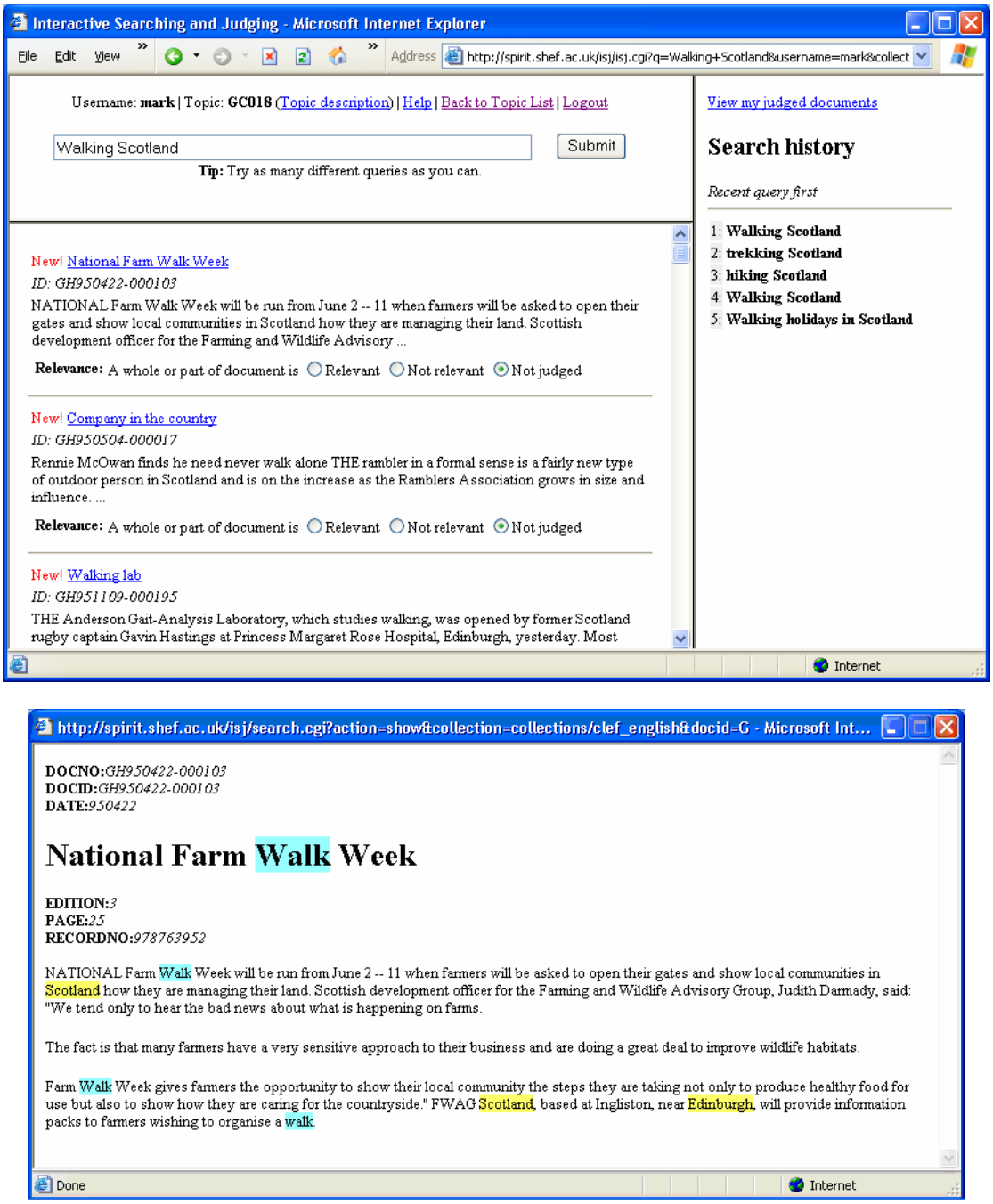

Figure 2: Interactive Search and Judging system. Queries are entered in the top panel with search results displayed below, where documents can be marked as relevant/not relevant. A history of the searcher's queries is listed in the right-hand panel. The user can display the full text of a document (loaded in a new window) with query words and place names highlighted.

The participants used a wide variety of approaches to the GeoCLEF tasks, ranging from basic IR approaches (with no attempts at spatial or geographic reasoning or indexing) to deep NLP processing to extract place and topological clues from the texts and queries. As Table 1 shows, all of the participating groups submitted runs for the Monolingual English task. (Note that Linguateca did not submit runs, but worked with the organizers to translate the GeoCLEF queries to Portuguese, which were then used by other groups). The bilingual X->EN task actually represents 3 separate tasks, depending on whether the German, Spanish, or Portuguese query sets were used (and likewise for X->DE from English, Spanish or Portuguese). The University of Alicante is the only group that submitted runs for all possible Monolingual and Bilingual tasks including Spanish and Portuguese to both English and German. The least participation was for the Bilingual X->DE task.

Participants

Twelve groups participated in the GeoCLEF task this year, the following table shows the group names and the sub-tasks in which they submitted runs:

Group Name	Mono EN	Mono DE	Bi ->EN	Bi ->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)	0	0	0	0	0
Linguit GmbH. (Germany)	16	0	0	0	16
MetaCarta Inc.	2	0	0	0	2
MIRACLE (Universidad Politécnica de Madrid)	5	5	0	0	10
NICTA, University of Melbourne	4	0	0	0	4
TALP Research Center (Universitat Politècnica de Catalunya)	4	0	0	0	4
Universidad Politécnica 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 (FernUniversität 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

GeoCLEF Performance

Monolingual performance:

Since the largest number of runs (57) were submitted for monolingual English, it is not surprising that that evalution is represented by the largest number of groups (11). Monolingual German was carried out by 6 groups submitting 25 runs. The following is a ranked list of performance and results by overall mean average precision using the TREC_Eval software, displaying best English against best German. We choose only the single best run from each participating group (independent of method used to produce the best run):

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		

One immediately apparent observation is that German performance is substantially below that of English performance. This derives from two sources: Many of the topics were “English” news story-oriented and had few, if any, relevant documents in the German language. Four topics (1, 20, 22, and 25) had no relevant German documents. Topics 18 and 23 had 1 and 2 relevant documents, respectively. By contrast, no English version of the topic had less than 3 relevant documents. The German task seems to have been inherently more difficult, with fewer geographic resources available in the German language to work with.

Performance Comparison on Mandatory Tasks:

A fairer comparison (one usually used in CLEF, TREC and NTCIR) is to compare system performance on identical tasks. The two runs expected from each participating group were a Title-Description run which used only these fields and a Title-Description-Geotags run which utilized the geographic tag triples (Concept-Location-Operator-Location). The precision scores for best Title-Description runs for monolingual English are as follows.

Recall	CSUSM	Berkeley2	Alicante	Berkeley	NICTA
0.0	0.7634	0.7899	0.7889	0.6976	0.6680
0.1	0.6514	0.6545	0.6341	0.5222	0.5628
0.2	0.5348	0.5185	0.4972	0.4321	0.4209
0.3	0.4883	0.4584	0.4315	0.3884	0.3456
0.4	0.4549	0.3884	0.3776	0.3435	0.2747
0.5	0.3669	0.3562	0.3258	0.2783	0.2217
0.6	0.3039	0.2967	0.2728	0.2221	0.1715
0.7	0.2439	0.2563	0.2072	0.1877	0.1338
0.8	0.1834	0.1963	0.1591	0.1168	0.0908
0.9	0.1040	0.1169	0.0701	0.0525	0.0624
1.0	0.0484	0.0603	0.0314	0.0194	0.0272
MAP	0.3613	0.3528	0.3255*	0.2794*	0.2514*

*CSUSM run is a statistically significant improvement over this run using a paired t-test at 5% probability level

The next mandatory run was to also include (in addition to Title and Description) the contents of the Geographic tags in the topic description. The next table provides performance comparison for the best 5 runs with TD+GeoTags:

Recall	Berkeley2	Alicante	CSUSM	Berkeley	Miracle
0.0	0.8049	0.7856	0.7017	0.6981	0.5792
0.1	0.7144	0.6594	0.5822	0.5627	0.4932
0.2	0.5971	0.5318	0.4612	0.4804	0.4266
0.3	0.5256	0.4675	0.4204	0.4149	0.3516
0.4	0.4534	0.4138	0.3803	0.3460	0.3184
0.5	0.3868	0.3580	0.2937	0.2960	0.2815
0.6	0.3464	0.2924	0.2293	0.2257	0.2231
0.7	0.2913	0.2342	0.1974	0.1869	0.1889
0.8	0.2301	0.1779	0.1451	0.1198	0.1450
0.9	0.1318	0.0823	0.1084	0.0534	0.0928
1.0	0.0647	0.0317	0.0281	0.0243	0.0344
MAP	0.3937	0.3471	0.3032*	0.2924*	0.2653*

*Berkeley2 run is a statistically significant improvement over this run using a paired t-test 1% probability level

Bilingual performance

Fewer groups accepted the challenge of bilingual retrieval. There were a total of 22 bilingual X to English runs submitted by 5 groups and 17 bilingual X to German runs submitted by 3 groups. The table below shows the performance of bilingual best runs by each group for both English and German, independent of method used to produce the run.

Best bilingual-X→English-run	MAP	Best bilingual-X→German-run	MAP
berkeley-2_BKGeoDE2	0.3715	berkeley-2_BKGeoED2	0.1788
csu-sanmarcos_csusm3	0.3560	alicante_irua-ende-syn	0.1752
alicante_irua-deen-ner	0.3178	berkeley_BERK1BLENDENOL01	0.0777
berkeley_BERK1BLDEENLOC01	0.2753		

Conclusions and Future Work

While the results of the GeoCLEF 2005 pilot track were encouraging, both in terms of number of groups/runs participating, but also in terms of interest, there is some question as to whether we have truly identified what constitutes the proper evaluation of geographic information retrieval. One participant has remarked that "The geographic scope of most queries had the granularity of Continents or groups of countries. It should include queries with domain of interest restricted to much smaller areas, at least to the level of cities with 50000 people." In addition, the best performance was achieved by groups using standard keyword search techniques. If we believe that *GIR ≠ Keyword Search*, then we must find a path which distinguishes between the two. GeoCLEF will probably continue in 2006 and expand the number of document languages (likely Portuguese and perhaps Spanish) as well as the scope of the task (i.e. consider more difficult topics such as "find stories about places within 125 kilometers of [Vienna, Viena, Wien]").

Possible directions which we might foresee for 2006 are:

1. *Additional languages: which and how many*? Since Portuguese was suggested this year, it seems a natural for next year. Spanish was also considered this year and would be fairly easy to integrate? The inclusion of another language assumes that some group will be willing to do the relevance judgments.
2. *Multilinguality?* Currently the tasks are monolingual and bilingual. Should we have a multilingual task where the documents are ranked independent of language?
3. *Task difficulty:* Should we increase the challenge of GeoCLEF 2006? One possible direction to increase task difficulty is to include geospatial distance or locale in the topic, i.e. "find stories about places within 125 kilometers of Vienna" or "Find stories about wine-making along the Mosel River" or "what rivers pass through Koblenz Germany?". Should the task become more of a named entity extraction task (see the next point on evaluation)?
4. *Evaluation:* Do we stick with the relative looseness of ranking documents according to subject and geographic reference? Or should we make the task more of an entity extraction task, like the shared task of the Conference on Computational Natural Language Learning 2002/2003 (CoNLL) found at http://www.cnts.ua.ac.be/conll2003/ner/ . This task had a definite geographic component. See also the background lecture by Marti Hearst at http://www.sims.berkeley.edu/courses/is290-2/f04/lectures/lecture15.ppt. In this instance we might have the evaluation be to extract a list of unique geographic names and the recall/precision measures are on the completeness of the list (how many relevant found) and (I guess) how many are found at rank x (precision) as well as the F measure. I'm not sure if this measure is also used for the list task in TREC question answering. Paul Clough and Mark Sanderson have proposed a MUC style evaluation for GIR (Clough and Sanderson, 2004).

Acknowledgments:
All effort done by the GeoCLEF organizers both in Sheffield and Berkeley was volunteer labo(u)r – none of us has funding for GeoCLEF work. The English assessment was evenly divided with Ray Larson and myself at University of California taking half and the Sheffield group taking the other half. Vivien Petras did half the German assessments. At the last minute Carol Peters found some funding for part of the German assessment, which might not have been completed otherwise. Similarly groups volunteered the topic translations into Portuguese (thanks to Diana Santos and Paulo Rocha of Linguateca) and Spanish (thanks Andres Montoyo of U. Alicante). In addition a tremendous amount of work was done above and beyond the call of duty by the Padua group (thanks Giorgio Di Nunzio and Nicola Ferro) – we owe them a great debt. Funding to help pay for assessor effort and travel came from the EU projects, SPIRIT and BRICKS. The future direction and scope of GeoCLEF will be heavily influenced by funding and the amount of volunteer effort available.

References

Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K. and Liberman, M. (2002). Corpora for Topic Detection and Tracking. In Allan, J. (ed.), Topic Detection and Tracking: Event-based Information Organization, 33-66, Kluwer.

Clough, P.D., Sanderson, M. (2004). A Proposal for Comparative Evaluation of Automatic Annotation for Geo-referenced Documents. In Proceedings of Workshop on Geographic Information Retrieval, SIGIR, 2004.

Clough, P.D. (2005). Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In Proceedings of GIR'05 Workshop at CIKM2005, Nov 4, Bremen, Germany, on-line.

Cormack, G.V., Palmer, C.R. and Clarke, C.L.A. (1998). Efficient Construction of Large Test Collections. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 282-289.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of ACL'02. Philadelphia.

Kraaij, W., Smeaton, A.F., Over, P., Arlandis, J. (2004). TRECVID 2004 - An Overview. In TREC Video Retrieval Evaluation Online Proceedings, http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

Kuriyama, K., Kando, N., Nozue, T. and Eguchi, K. (2002). Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop. Information Retrieval, 5 (1), 41-59.

Sanderson, M. and Joho, H. (2004). Forming Test Collections with No System Pooling. In Järvelin, K., Allan, J., Bruza, P., and Sanderson, M. (eds), Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 33-40, Sheffield, UK.

Sanderson, M and Zobel, J. (2005). Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In Proceedings of the 28th ACM SIGIR conference.