

Enhanced Gazetteer Development for Multilingual Geographic Information Retrieval of Natural Language Text

Ray Larson and Fredric Gey
University of California, Berkeley
ray@sim.s.berkeley.edu, gey@berkeley.edu

Geographic information retrieval (GIR) from text is the subject of active research in the information retrieval research community. GIR focuses upon search and retrieval with a geographic component, e.g. *Find stories about cities near the Danube and Rhine rivers in Europe*. There have been GIR research workshops in 2004 (SIGIR, Sheffield UK), 2005 (CIKM Hanover, Germany) and 2006 (SIGIR Seattle USA). GeoCLEF (<http://ir.shef.ac.uk/geoclef/>) is a component track in the European Cross-Language Evaluation Forum (CLEF) which evaluates performance of Geographic Information Retrieval on multilingual text by creating test topics in multiple languages which are run against document collections in those languages. For the GeoCLEF 2006 evaluation just concluded and results presented in Alicante Spain, the languages were English, German, Portuguese and Spanish (additionally, topics were translated into Japanese for cross-language search from that language). The document collections consisted of news stories from USA, Swiss and Germany, Portugal and Brazil, and Spain. The total number of documents being searched exceeds 1 million documents. GeoCLEF has emerged as the standard by which GIR for text research advances can be objectively evaluated.

Among the components of GIR search topics which differ from ordinary information retrieval are:

- Geographic challenge:
<EN-title>**Cities within 100km of Frankfurt**</EN-title>
<DE-title>**Städte im Umkreis von 100 km um Frankfurt**</DE-title>
<PT-title>**Cidades a menos de 100 quilómetros de Francoforte**</PT-title>
<ES-title>**Ciudades a menos de 100 kilómetros de Fráncfort**</ES-title>
<JP-title>□□□□□□□□□□ **100km** □□□□□□□□□□</JP-title>
- Geographic location disambiguation for vaguely defined entities:
<EN-title>**Scientific research in New England Universities**</EN-title>
- Geotemporal disambiguation for vague references
<EN-title>**Credits to the former Eastern Bloc aka the Warsaw Pact**</EN-title>
- Approximate regional restriction:
<EN-title>**Forest fires in Northern Portugal**</EN-title>

Research issues which have been explored by GeoCLEF participants and GIR researchers include named entity extraction in multiple languages, place name disambiguation, geographic hierarchy and expansion, as well as examining issues about the granularity of gazetteer information (e.g., when expanding queries using placenames derived from gazetteer lookup, should only major populated areas be used, or should all toponyms in

the referenced area). Researchers have also explored how to combine text ranking and geographic ranking schemes in retrieval.

Digital gazetteers form a critical infrastructure for GIR and for related Digital library applications. Obviously, for cross-language retrieval digital gazetteers must include place names in multiple languages, with appropriate point coordinates or footprints for the places. At UC Berkeley we have been involved in a variety of projects ranging from research and development of GIR ranking methods to the development of time and space-based retrieval systems for library catalogs and internet resources. Much of this work has been conducted in collaboration with the Electronic Cultural Atlas Initiative (ECAI). One of the goals of ECAI is to have an interactive Map based search and discovery front end so that the initial search for distributed metadata does not have to be a text search. This requires that there be geo-temporal metadata in the primary metadata for searching which will allow identification of the area of coverage of an object or collection and information about where to get the additional metadata and data required to provide the transactional geo-temporal browsing functionality. For ECAI the geo-temporal metadata serves as the union catalog for accessing a wide range of distributed transactional objects. Digital Gazetteers, in conjunction with the Time Period Directory developed as a prototype over the past 2 years, provide key elements for linking events, places and the people and subjects related to them. (See <http://www.ecai.org/ims2004/ims4w/> The time period directory is a digital structure analogous to a digital gazetteer, but which associates named periods and events with date ranges, as well as referencing gazetteer information for the places associated with those time periods and events.) The combination of Time Period Directories and Gazetteers provides a metadata infrastructure for ECAI projects and systems.

Our interests in attending the meeting are twofold. First we will represent the needs of the cross-language IR research community (Ray Larson and Fred Gey are co-chairs of the GeoCLEF evaluation track), and second we will represent the interests of ECAI in gazetteer standards development and cultural and humanities applications of geo-spatial information resources. It is our contention that ordinary gazetteers are a necessary but insufficient component for resolving the problems of geographic search of natural language text and supporting the required access mechanisms for systems like those being developed for ECAI. Among the needed enhancements to Digital Gazetteers are:

- To enhance gazetteers with thesaurus-like co-references including the ordinary language expressions (Latin America, Middle East, etc.) which can be tied to exact gazetteer geographic components.
- Additionally historical names and footprints for places should be added whenever possible to provide a temporal dimension for gazetteers
- To link geographic information to its historical context, either by directly embedding it in the gazetteer, or by reference to time period directories.

Having attended (Larson) the previous gazetteer workshop in Washington DC, we look forward to an interesting and productive meeting.