Monolingual and Bilingual Experiments in GeoCLEF2006

Rocio Guillén
California State University San Marcos
rguillen@csusm.edu

Abstract

This paper presents the results of our initial experiments in the monolingual English, Spanish and Portuguese tasks and the Bilingual Spanish → English, Spanish → Portuguese, English → Spanish and Portuguese → Spanish tasks. Twenty runs were submitted as official runs, thirteen for the monolingual task and seven for the bilingual task. We used the Terrier Information Retrieval Platform to run experiments for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization 2. Experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out using gazetteers (Alexandria Digital Library, European Parliament and GEOnet Names Server), some of them containing translations in languages other than English, others containing the latitude, longitude and area which allow for semi-automated spatial analysis (proximity analysis). For the bilingual task we developed a component based on the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Then topics in the source language were translated to the target language. A major problem we detected after submitting our results was that we did not include the Spanish newspaper collection for the year 95 (EFE 95) for indexing and retrieval purposes. Therefore, the results of our experiments with Spanish for the monolingual and bilingual tasks were affected in terms of recall and precision. We are currently re-running experiments with the full Spanish collection for the monolingual and bilingual task to obtain a more accurate evaluation of the retrieval performance.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Geographical information retrieval, term translation

1 Introduction

Geographic Information Retrieval (GIR) is aimed at the retrieval of geographic data based not only on conceptual keywords, but also on spatial information. Building GIR systems with such capabilities requires research on diverse areas such as information extraction of geographic terms from structured and unstructured data; word sense disambiguation, which is geographically relevant; ontology creation; combination of geographical and contextual relevance; and geographic term translation, among others.

Research efforts on GIR are addressing issues such as access to multilingual documents, techniques for information mining (i.e., extraction, exploration and visualization of geo-referenced information), investigation of spatial representations and ranking methods for different representations, application of machine learning techniques for place name recognition, development of datasets containing annotated geographic entities, among others. [4]. Other researchers are exploring the usage of the World Wide Web as the largest collection of geospatial data.

The purpose of GeoCLEF 2006 is to experiment with and evaluate the performance of GIR systems when topics includes geographic locations such as rivers, regions, seas, continents. Two tasks were considered, a monolingual and a bilingual. We participated in the monolingual task in English, Portuguese and Spanish; for the bilingual task we worked with topics in Spanish and documents in English and Portuguese, and with topics in English and Portuguese and documents in Spanish.

In this paper we describe our initial experiments in the monolingual task and the bilingual task. We used the Terrier Information Retrieval (IR) platform to run our experiments, and built an independent module for the translation of the topics based on the transfer approach in machine translation. We used Terrier because it has performed successfully in monolingual information retrieval tasks in CLEF and TREC. Our goal is to have a baseline for further experiments with our component for translating georeferences and improved spatial analysis.

The paper is organized as follows. In Section 2 we present our work in the monolingual task including an overview of Terrier. Section 3 describes our setting and experiments in the bilingual task. We present conclusions and current work in Section 4.

2 Monolingual Task

In this section we first give an overview of Terrier (TERabyte RetRIEveR) an information retrieval (IR) platform used in all the experiments. Then we describe the monolingual experiments for English, Portuguese and Spanish.

Terrier is a platform for the rapid development of large-scale Information Retrieval (IR) systems. It offers a variety of IR models based on the Divergence from Randomness (DFR) framework ([3],[7],[8]). The framework includes more than 50 DFR models for term weighting. These models are derived by measuring the divergence of the actual term distribution from that obtained under a random process ([2]). Terrier provides automatic query expansion with 3 documents and 10 terms as default values; additionally the system allows to choose a specific query expansion model.

Both indexing and querying of the documents was done with Terrier. The document collections indexed were the LA Times (American) 1994 and the Glasgow Herald (British) 1995 for English, efe94 for Spanish, publico94, publico95, folha94 and folha95 for Portuguese. There were 25 topics for each of the languages tested. Documents and topics in English were processed using the English stopwords list (571 words) built by Salton and Buckley for the experimental SMART IR system [1], and the Porter stemmer. Stopwords lists for Spanish and Portuguese were also used. No stemmers were applied to the Portuguese and Spanish topics and collections,

We worked with the InL2 term weighting model, which is the Inverse Document Frequency model with Laplace after-effect and normalization 2. Our interpretation of GeoCLEF's tasks was that they were not exactly classic ad-hoc tasks, hence we decided to use a model for early precision. We experimented with other models and found out that this model generated the best results when analyzing the list of documents retrieved.

The risk of accepting a term is inversely related to its term frequency in the document with respect to the elite set, a set in which the term occurs to a relatively greater extent than in the rest of the documents. The more the term occurs in the elite set, the less the term frequency is due to randomness. Hence the probability of the risk of a term not being informative is smaller. The Laplace model is utilized to compute the information gain with a term within a document. Term frequencies are calculated with respect to the standard document length using a formula referred to as normalization 2 shown below.

$$tfn = tf.log(1 + c\frac{sl}{dl})$$

tf is the term frequency, *sl* is the standard document length, and *dl* is the document length, c is a parameter. We used $c = 1.5$ for short queries, which is the default value, $c = 3.0$ for short queries with automatic query expansion and $c = 5.0$ for long queries. Short queries in our context are those which use only the topic title and topic description; long queries are those which use the topic title, topic description and topic narrative. We used these values based on the results generated by the experiments on tuning for BM25 and DFR models done by He and Ounis [6]. They carried out experiments for TREC (Text REtrieval Conference) with three types of queries depending on the different fields included in the topics given. Queries were defined as follows: 1) short queries are those where the title and the description fields are used; and 2) long queries are those where title, description and narrative are used.

2.1 Experimental Results

We submitted 4 runs for English, 4 runs for Portuguese and 5 runs for Spanish. For some of the runs we used the automatic query expansion capability of terrier with the default values of 3 documents and 10 terms. We are running experiments with new values to determine which number of documents and terms perform better. Results for the monolingual task in English, Portuguese and Spanish are shown in Table 1, Table 2 and Table 3, respectively.

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoEN1	title, description	automatic	yes	26.37	28.57
SMGeoEN3	title, description, narrative	automatic	yes	28.57	33.66
SMGeoEN4	title, description	automatic	no	26.37	28.57
SMGeoEN5	title, description, narrative	automatic	no	23.77	25.81

Table 1: English Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoPT1	title, description, narrative	automatic	yes	10.98	13.91
SMGeoPT2	title, description	automatic	yes	13.44	15.02
SMGeoPT3	title, description, narrative	automatic	no	10.98	13.91
SMGeoPT4	title, description	automatic	no	10.63	13.57

Table 2: Portuguese Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoES1	title, description, narrative	automatic	yes	14.71	20.44
SMGeoES2	title, description	automatic	yes	14.71	20.44
SMGeoES3	title, description	manual	yes	14.71	20.44
SMGeoES4	title, description	automatic	no	13.78	18.63
SMGeoES5	title, description, narrative	automatic	no	14.71	20.44

Table 3: Spanish Monolingual Retrieval Performance

3 Bilingual Task

For the bilingual task we worked with Spanish topics and English and Portuguese documents, and English and Portuguese topics and Spanish documents. We built a component, independent of Terrier, based on the transfer approach in machine translation to translate topics from the source language to the target language using mapping rules. All the information in the topics within the title, description and narrative was translated. Topics in English, Spanish, and Portuguese were preprocessed by removing diacritic marks and using stopwords lists. Diacritic marks were also removed from the stopwords lists and duplicates were eliminated. Plural stemming was then applied.

Automatic and manual query construction was carried out with the aid of the German Alexandria Digital Library gazetteer [9], the Spanish Toponymy from the European Parliament [10], and the Names files of countries and territories from the GEOnet Names Server (GNS) [11]. The German gazetteer was particularly helpful because it included information such as latitude, longitude and area. Thus, English Topic 027 with narrative "Relevant documents discuss cities within 100 kilometers of Frankfurt am Main Germany, latitude 50.11222, longitude 8.68194..." lend itself to spatial analysis using a distance measure to find out the cities within 100 kilometers of Frankfurt.

3.1 Experimental Results

Seven runs were submitted as official runs for the GeoCLEF2006 bilingual task. In Table 4 we report the results for X-Spanish (X={English, Portuguese}) and in Table 5 the results for Spanish-X (X={English,Portuguese}).

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoENES1	title, description	automatic	no	12.82	16.89
SMGeoPTES2	title, description	automatic	no	10.89	14.67
SMGeoPTES3	title, description, narrative	automatic	no	11.50	15.27

Table 4: X-Spanish Bilingual Retrieval Performance (X = {English,Portuguese})

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoESEN1	title, description	automatic	no	12.82	16.89
SMGeoESEN2	title, description, narrative	automatic	no	12.82	16.89
SMGeoESPT1	title, description	automatic	no	10.89	14.67
SMGeoESPT2	title, description, narrative	automatic	no	11.50	15.27

Table 5: Spanish-X Bilingual Retrieval Performance (X = {English,Portuguese})

4 Conclusions

In this paper we presented work on monolingual and bilingual geographical information retrieval. We used Terrier to run our experiments, and an independent translation component built to map source language (English, Portuguese or Spanish) topics into target language (English, Portuguese or Spanish) topics. Results were affected because we did not include one of the Spanish collections for indexing and retrieval purposes. We are currently re-running experiments with the entire collection of Spanish documents and testing the automatic query expansion capabilities of terrier with new values and new weighting models.

References

[1] http://ftp.cs.cornell.edu/pub/smart/.

[2] Lioma, C., He, B., Plachouras, V., Ounis, I.: The University of Glasgow at CLEF2004; French monolingual information retrieval with Terrier. In Working notes of the CLEF 2004 Workshop, Bath, UK, 2004.

[3] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In Proceedings of the 27th European Conference on Information Retrieval (ECIR 05), 2005. http://ir.dcs.ga.ac.uk/terrier/

[4] Purves, R., Jones, C. editors : SIGIR2004: Workshop on Geographic Information Retrieval, Sheffield, UK, 2004.

[5] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: the CLEF2005 Cross-Language Geographic Information Retrieval Track Overview Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer Science, 2006 (in this volume).

[6] He, B., Ounis, I. : A study of parameter tuning for the frequency normalization. Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, 2003.

[7] Amati, G., van Rijsbergen, C.J. : Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems.* Vol. 20(4), pp:357-389.

[8] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings *ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006).*

[9] Alexandria Digital Library Gazetteer. 1999- . Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents. http://www.alexandria.ucsb.edu/gazetteer (ADL Gazetteer Development page with links to various clients and protocols that access the ADL Gazetteer).

[10] European Parliament. Tools for the External Translator. http://www.europarl.europa.eu/transl_es/plataform/pagina/toponim/toponimo.htm

[11] http://earth-info.nga.mil/gns/html/index.html