CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks

Rocio Guillén
Computer Science Department
rguillen@csusm.edu

Abstract

This paper presents the results of our initial experiments in the monolingual English task and the Bilingual Spanish → English 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. For the bilingual task we developed a component to first translate the topics from Spanish into English. No spatial analysis was carried out for any of the tasks. One of our goals is to have a baseline to compare further experiments with term translation of georeferences and spatial analysis. Our initial results show that geographic information does not improve significantly 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

Experimentation, Measurement, Performance

Keywords

geographical information retrieval, term translation

1 Introduction

Geographic Information Retrieval is aimed at the retrieval of geographic data based not only on conceptual keywords, but also on spatial information. Building GIRs with such capabilities requires of research on diverse areas such as information extraction of geographic terms from structured and unstructured data; word sense disambiguation 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]

GeoCLEF 2005, which is part of the Cross Language Evaluation Forum (CLEF) aims to provide the necessary framework for evaluating GIRs in two aspects spatial and multilingual. The purpose of GeoCLEF is to experiment with and evaluate the performance of GIRS when geographic

operators and geographic locations are added. Four tasks were considered, two monolingual and two bilingual. The monolingual tasks consist in searching and retrieving relevant documents in English given a set of descriptions of a spatial user need (topic) in English; searching and retrieving relevant documents in German given a set of descriptions of a spatial user need (topic) in German. The bilingual tasks consist in searching and retrieving relevant documents in English given a set of topics in one of the following languages: German, Spanish, Portuguese; searching and retrieving relevant documents in German given a set of topics in one of the following languages: English, Spanish, Portuguese.

In this paper we describe our initial experiments in the English monolingual task and the bilingual task with topics in Spanish and documents in English. We used the Terrier Information Retrieval (IR) platform to run our experiments, and built an independent module for the translation of the topics from Spanish into English. We used Terrier because it has performed successfully in monolingual information retrieval tasks (CLEF2004 and TREC2004). Our goal is to have a baseline for further experiments with our component for translating georeferences from Spanish into English, and addition of spatial analysis.
The paper is organized as follows. In Section 2 we present our work in the English monolingual task including an overview of Terrier. Section 3 describes our setting and experiments in the bilingual task. Experimental results are discussed in Section 4, and we present our conclusions and future work in Section 5.

2 Monolingual Task

In this section we first give an overview of the IR platform used in our experiments. 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]). 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]).

Both indexing and querying of the documents were done with Terrier. The document collections to be indexed were the LA Times (American) 1994 and the Glasgow Herald (British) 1995. There were 25 topics. Documents and topics were processed using the English stopword list (571 words) built by Salton and Buckley for the experimental SMART IR system [1], and the Porter stemmer. We worked with the InL2 term weighting model, which is the Inverse Document Frequency model with Laplace after-effect and normalization 2.

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 $c = 1.0$ for short queries, which is the the default value, and $c = 7.0$ 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, topic concept, topic spatial-relation and topic location.

3 Bilingual Task

For the bilingual task we worked with Spanish topics and English documents. We built a component, independent of Terrier, to translate the topics from Spanish into English. All the information within the tags was translated except for the narrative, because we did not consider the narrative for any of the two runs that we submitted. Topics in Spanish were preprocessed by removing diacritic marks, and stopwords using a list of 351 Spanish words from SMART. Diacritic marks were also removed from the stopwords list, duplicates were eliminated. Plural stemming was then applied. The last step was to perform word-by-word translation without considering word ordering and syntactic differences between Spanish and English. For instance, Topic 8 *Consumo de leche en Europa* was mapped into "consumption milk europe". However, we took into account abbreviations and different spellings. For instance, in Topic 3 the title *AI en Latinoamerica* was mapped to "amnesty international latinamerica latin-america latin america". The new set of topics thus created was then used to run two monolingual tasks in English (see section above).

4 Experimental Results

Four runs were submitted as official runs, two for the GeoCLEF2005 monolingual task, and two for the GeoCLEF2005 bilingual task. In Table 1 we report the results for the English monolingual task.

Run Id	Topic Fields	Avg Prec.	Recall Prec.
csusm1	title, description	36.13	37.61
csusm2	title, description, geographic tags	30.32	33.66

Table 1: English Monolingual Retrieval Performance

Our retrieval results for the first run performed well above average for most of the topics, except for Topics 7 and 8. The performance of the run adding geographic tags did not improve as it was originally expected. It decreased in general. The average precision of topics 5, 6, 7, 9 and 20 was below the MAP. We have run experiments with different values for the c parameter, but the relevant judgments are not available yet to evaluate the results.

The results for the two official runs submitted for the bilingual task are shown in Table 2. Our results performed well above average for 13 out of the 25 topics, and below average for 7 of the topics with only the title and the description. For the run using additional geographic tags the results performed above the median precision average for 15 topics. Word by word translation did not perform as well in general. One of the factors may be word ordering, another factor could be the addition of abbreviations and full names. The latter because the information gain starts decreasing at point. Further experiments and study of the search algorithm will provide us with more accurate data. One of our goals in the future is to extend the translation component to do more than word-to-word translation.

Run Id	Topic Fields	Avg Prec.	Recall Prec.
csusm3	title, description	35.60	37.17
csusm4	title, description, geographic tags	31.23	33.50

Table 2: Bilingual (Spanish $\rightarrow$ English) Retrieval Performance

Comparison of retrieval performance between the monolingual task and the bilingual task, which was ran as a monolingual task shows that the average precision using the title and description performed better in the former task. The average precision using the title, description and geographic tags was better in the bilingual task than in the monolingual task.

5 Conclusions and Future Work

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 Spanish topics into English topics. The results were evaluated in the GeoCLEF2005 track of CLEF2005.

Experimental results show the following. The use of geographical information did not improve retrieval performance significantly in our work. Translated geographic information performed better than using the original English tags. Further experimentation with other DFR models of Terrier and other IR systems is necessary to determine to what extent this is true. We are starting to investigate the translation of spatial relations and contextual relevance to GIR.

References

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

[2] Christian Lioma, Ben He, Vassilis Plachouras, and Iadh Ounis. The university of glasgow at clef2004; french monolingual information retrieval with terrier. In *Working notes of the CLEF 2004 Workshop, Bath, UK*, 2004.

[3] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *27th European Conference on Information Retrieval (ECIR 05)*, 2005.

[4] Ross Purves and Chris Jones, editors. *SIGIR2004: Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004.