Multi-lingual Geographical Information Retrieval

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

Abstract

This paper reports on the results of our experiments in the Monolingual English, German and Portuguese tasks and the Bilingual German topics on English collections, English topics on German collections and English topics on Portuguese collections tasks. Seven runs were submitted as official runs, four for the monolingual task and three for the bilingual task. We used the Terrier (TERabyte RetrIEveR) Information Retrieval Platform version 2.1 to index and query the collections. Experiments were performed for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization 2. Topics were processed automatically and the only fields considered were the title and the description. We included the title field only for an experiment with the Portuguese collection. The stopword list provided by Terrier was used to index all the collections. Results for both the monolingual and bilingual tasks were low in terms of precision and recall mainly due to the following reasons: 1) no manual processing was done; 2) no query expansion based on automated relevance feedback was added; 3) no experiments including the narrative field were run; 4) no terms were translated for the bilingual task; 5) no German and Portuguese stopword lists were used instead of the default stopword list; and 6) no pre-processing or removal of diacritic marks was performed. We are running new experiments to address some of the issues aforementioned and determine the impact they have on retrieval performance.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; Linguistic Processing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Geographical Information Retrieval (GIR)

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. [1]. Other researchers are exploring the usage of the World Wide Web as the largest collection of geospatial data.
The focus of one of the tasks was on experimenting with and evaluating the performance of GIR systems when topics include geographic references. Collections of documents and topics in different languages were available to carry out monolingual and bilingual experiments. We ran monolingual experiments in English, German, and Portuguese; for bilingual retrieval, we worked with topics in German and English and collections in English, German and Portuguese.
In this paper we describe experiments in the cross-language monolingual and bilingual task. We used the Terrier Information Retrieval (IR) platform version 2.1 to run our experiments. This platform has performed successfully in monolingual information retrieval tasks in CLEF and TREC.
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. Finally, we present conclusions and current work in Section 4.

2 Cross-lingual Geographical IR Task

In this section we present Terrier (TERabyte RetRIEveR) an information retrieval (IR) platform used in all the experiments. Then we describe experiments and results for monolingual GIR in English, German, and Portuguese. The final subsection includes the experiments and results for bilingual GIR with topics in German and English.
Terrier is a high performance and scalable search engine platform for the rapid development of large-scale retrieval applications. It offers a variety of IR models based on the Divergence from Randomness (DFR) framework ([4],[5]) and supports classic retrieval models like the Ponte-Croft language model ([3]).
The components of the DFR models are the following: 1) a randomness model; 2) an information gain model; and 3) a term frequency normalization model. The latter component adjusts the frequency of a term in a document based on the length of a document and the average document length in the entire collection. For example, the Normalization 2 term frequency normalization model assumes a decreasing density function of the normalized term frequency concerning the document length.
The normalized term frequency tfn is calculated as follows:

$$tfn = tf.log_2(1 + c\frac{avg\_len}{len})$$

tf is the term frequency, *avg_len* is the average document length in the collection, and *len* is the document length, and c is a hyper-parameter. We used $c = 1.5$ for short queries, which is the default value, $c = 3.0$ for short queries. Short queries in our context are those which use only the topic title field and the topic description field. We used these values based on the results generated by the experiments on tuning for BM25 and DFR models done by He and Ounis [2]. 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.
Each query term in a document is assigned a weight depending how important the term is to the document. Term weights are then used to match documents to a query. Documents are ranked according to their estimated relevance to the query. The formula to estimate the probability of producing the query for a given document is the sum of the probability of producing the terms in

the query plus the probability of not producing other terms.
Both indexing and querying of the documents in English, German, and Portuguese was done with Terrier using the InL2 term weighting model. This model is the Inverse Document Frequency model with Laplace after-effect and normalization 2. The InL2 model has been used in experiments in the past, GeoCLEF2005, GeoCLEF2006 and GeoCLEF2007[6, 7, 8], successfully.

2.1 Data

The document collections indexed were the LA Times (American) 1994 and the Glasgow Herald (British) 1995 for English, publico94, publico95, folha94 and folha95 for Portuguese, and der_spiegel, frankfurter and fr_rundschau for German. There were 25 topics for each of the languages tested. Documents and topics were processed using the English stopwords list and the Porter stemmer provided by Terrier. No stopwords lists for German and Portuguese were used.

2.2 Experimental Results for Monolingual Task

We submitted 1 run for English, 1 run for German, and 2 runs for Portuguese. Queries were automatically constructed for all the runs. Results for the monolingual task in English, German and Portuguese are shown in Table 1, Table 2 and Table 3, respectively.

Run Id	Topic Fields	MAP	Recall Prec.	Mean Rel. Ret.
monen1	title, desc.		0.16	18.4

Table 1: English Monolingual Retrieval Performance

Run Id	Topic Fields	MAP	Recall Prec.	Mean Rel. Ret.
monde1	title, desc.		0.22	25.12

Table 2: German Monolingual Retrieval Performance

Run Id	Topic Fields	MAP	Recall Prec.	Mean Rel. Ret.
monpt1	title, desc.	0.17	0.18	20.36
monpt2	title	0.17	0.18	20.56

Table 3: Portuguese Monolingual Retrieval Performance

3 Bilingual Task

For the bilingual task we worked with English and German topics and English, German and Portuguese documents. We did not translate or remove diacritic marks.

3.1 Experimental Results

Three runs were submitted as official runs for the GeoCLEF2008 bilingual task. In Table 4 we report the results on runs with topics in German and documents in English (de2en) and the results on runs with English topics and documents in German (en2de) and Portuguese (en2pt).

Run Id	Topic Fields	MAP	Recall Prec.	Mean Rel. Ret.
de2en	title, desc.	0.15	0.16	17.44
en2de	title, desc.	0.19	0.20	20.92
en2pt	title, desc.	0.18	0.21	19

Table 4: Bilingual Retrieval Performance

Unlike the monolingual runs and the Spanish →English run, relevance feedback did not improve performance retrieval. No querying was done with the language model option.

4 Conclusions

In this paper we presented work on monolingual and bilingual geographical information retrieval. We used Terrier to run our experiments using the InL2 parameter-based model. Comparing results with those obtained in the past three years (see [6, 7, 8] show that precision and recall are likely affected by the following factors: 1) not carrying out manual processing ; 2) excluding query expansion; 3) not including the narrative field content to generate the query; 4) leaving out the translation module for the bilingual task; and 5) not removing diacritic marks in the collection and the topics. We are running more experiments to determine the impact each of the above factors has on retrieval performance.

References

[1] Jones, C.B., Purves, R.S.: : Geographical information retrieval.(2008). In *International Journal of Geographical Information Science* 22(3), pp.219-228.

[2] 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.

[3] Ponte, J.M., Croft, W.B. : A Language Modeling Approach to Information Retrieval. SIGIR'98, Melbourne, Australia, 1998. p: 275-281.

[4] 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.

[5] Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research Directions in Terrier: A Search Engine for Advanced Retrieval on the Web. In UPGRADE *The European Journal for the Informatics Professional at http://www.upgrade-cepis.org* Vol. VII(1), February 2007, pp:49-56.

[6] Guillén, R.: CSUSM Experiments at GeoCLEF2005: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Peters, C.; Gey, F.; Gonzalo, J.; Mueller, H.; Jones, G.; Kluck, M.; Magnini, B.; de Rijke, M. (Eds.), Vienna, Austria, Revised Selected Papers. "Lecture Notes in Computer Science", vol. 4022. Springer-Verlag.

[7] Guillén, R.: Monolingual and Bilingual Experiments in GeoCLEF2006: Evaluation of Multilingual and Multi-modal Information Retrieval Cross-Language Information Forum, CLEF 2006, Revised Selected Papers. Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (Eds.). "Lecture Notes in Computer Science", vol. 4730. Springer-Verlag.

[8] Guillén, R.: GeoCLEF2007 Experiments in Query Parsing and Cross-language GIR: CLEF 2007 Working Notes. Alessandro Nardi and Carol Peters (Eds.) ISSN per Working Notes and CD: 1818-8044. ISBN Abstracts: 2-912335-31-0.