GeoCLEF2007 Experiments in Query Parsing and Cross-language GIR

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 Spanish → English, Spanish → Portuguese tasks. We also present initial results on the recognition, extraction and categorization of web-based queries for the Query Parsing task. Twenty-three runs were submitted as official runs, 16 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 and the Ponte-Croft language model. Experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out for the bilingual task using the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Results show that automatic relevance feedback with 5 terms and 20 documents performs better, in general. The initial approach used in the Query Parsing task is a pattern-based approach. Due to the ungrammaticality, multilinguality and ambiguity of the language in the 800,000 web-based queries in the collection, we started by building a list of all the different words in the queries, similar to creating an index. Next, a lookup of the words was done in a list of countries to identify potential locations. Because many locations were missed, we further analyzed the queries looking for spatial prepositions and syntactic cues. Queries were processed by combining search in gazetteers with a set of patterns. Categorization was also based on patterns. Results were low in terms of recall and precision.

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; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Geographical Information Retrieval (GIR), Query Processing, Information Extraction

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. [2]. Other researchers are exploring the usage of the World Wide Web as the largest collection of geospatial data.

The tasks in GeoCLEF 2007 were Cross-language GIR and Query Parsing. The focus of the first task was on experimenting with and evaluating the performance of GIR systems when topics include geographic locations such as rivers, regions, seas, continents. 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 Spanish and documents in English and Portuguese.

The query parsing task consisted of parsing queries to recognize and extract georeferences. The output was structured as a frame, which included geographical such as "where", "geospatial relation" (e.g., in, west, ...), type of geographical query (information, map, yellow page) "latitude-longitude".

In this paper we describe experiments in the cross-language monolingual and bilingual task. We used the Terrier Information Retrieval (IR) platform to run our experiments. This platform has performed successfully in monolingual information retrieval tasks in CLEF and TREC. In addition, we ran initial experiments in the query parsing task. We initially applied pattern-based parsing that did not generate accurate results. We are currently working on inferring a grammar to improve recognition and extraction of geographical references in web-based queries.

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. Pattern-based parsing applied to web-based queries is discussed in Section 4. Finally, we present conclusions and current work in Section 5.

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 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 ([5],[6]) and supports classic retrieval models like the Ponte-Croft language model ([4]). 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. 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 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, GeoCLEF2006 and GeoCLEF2005[9], successfully.
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 [3]. 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.
Additionally, we queried the documents in all the collections using the Ponte-Croft language model ([4]). A language model is inferred for each document and the probability of generating the query according to these models is estimated. The documents are then ranked according to these probabilities. In this approach, term frequency, document length and document frequency are integral part of the language model and are not used as in many other approaches.
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.

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 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 German and Portuguese were also used. No stemming was applied to the German and Portuguese topics and collections,

2.2 Experimental Results Monolingual Task

We submitted 6 runs for English, 6 runs for German, and 4 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. The third column shows the model used in each experiment, InL2 or LM (Language Model). The fourth column indicates whether the experiment was run with relevance feedback. For relevance feedback we choose 15 terms and 20 documents to expand the query. This choice was arbitrary and more experiments are needed to find the combination that yields the best performance.

Run Id	Topic Fields	Model	Rel. Fb.	MAP	Recall Prec.	Mean Rel. Ret.
GEOMOEN1	title, desc.	InL2	no	0.14	0.16	18.4
GEOMOEN2	title, desc., narr.	LM	no	0.13	0.15	16.48
GEOMOEN3	title, desc.	LM	no	0.14	0.14	16.48
GEOMOEN4	title, desc.	InL2	yes	0.18	0.19	21.2
GEOMOEN5	title, desc., narr.	InL2	yes	0.21	0.20	21.36
GEOMOEN6	title, desc., narr.	InL2	no	0.19	0.21	19.12

Table 1: English Monolingual Retrieval Performance InL2

Run Id	Topic Fields	Model	Rel. Fb.	MAP	Recall Prec.	Mean Rel. Ret.
GEOMODE1	title, desc.	InL2	no	0.20	0.22	25.12
GEOMODE2	title, desc., narr.	LM	no	0.11	0.14	15.12
GEOMODE3	title, desc.	LM	no	0.11	0.14	15.12
GEOMODE4	title, desc.	InL2	yes	0.21	0.19	26.84
GEOMODE5	title, desc., narr.	InL2	yes	0.21	0.22	26.84
GEOMODE6	title, desc., narr.	InL2	no	0.20	0.22	25.12

Table 2: German Monolingual Retrieval Performance

Results for experiments querying the collection with the language model option are not accurate because we did not index the collection using the language model. Therefore we cannot compare the results between the two models as originally planned.
Comparison of the results using the InL2 model shows, for the three languages, that relevance feedback with 15 terms and 20 documents improves performance retrieval.

Run Id	Topic Fields	Model	Rel. Fb.	MAP	Recall Prec.	Mean Rel. Ret.
GEOMOPT1	title, desc.	InL2	no	0.17	0.18	20.36
GEOMOPT2	title, desc.	InL2	yes	0.17	0.18	20.56
GEOMOPT3	title, desc.	InL2	no	0.17	0.18	20.36
GEOMOPT4	title, desc., narr.	InL2	yes	0.17	0.18	20.56

Table 3: Portuguese Monolingual Retrieval Performance

3 Bilingual Task

For the bilingual task we worked with Spanish topics and English and Portuguese documents. We translated the topics applying the transfer approach in machine translation using rules to map from the source language to the target language. 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 Spanish Toponymy from the European Parliament [7], and the Names files of countries and territories from the GEOnet Names Server (GNS) [8].

3.1 Experimental Results

Eight runs were submitted as official runs for the GeoCLEF2007 bilingual task. In Table 4 we report the results on runs with topics in Spanish and documents in English and in Table 5 the results on runs with Spanish topics and documents in Portuguese.

Run Id	Topic Fields	Model	Rel. Fb.	MAP	Recall Prec.	Mean Rel. Ret.
GEOBIESEN1	title, desc., narr	LM	no	0.15	0.16	17.44
GEOBIESEN2	title, desc.	InL2	yes	0.19	0.20	20.92
GEOBIESEN3	title, desc.	InL2	no	0.18	0.21	19
GEOBIESEN4	title, desc., narr	LM	no	0.15	0.16	17.44

Table 4: Spanish→English Retrieval Performance

Runs 1 and 4 are the same. We had problems uploading the correct run and deleting the duplicate experiment. Similar to the monolingual task, comparison of the results using the InL2 model shows that relevance feedback with 15 terms and 20 documents improves performance retrieval.
Documents were indexed with InL2 only. Therefore, the results for experiments querying the collection with the language model option are not accurate and a proper comparison with the parametric-based InL2 model could not be made.

Run Id	Topic Fields	Model	Rel. Fb.	MAP	Recall Prec.	Mean Rel. Ret.
GEOBIESPT1	title, desc.	InL2	no	0.05	0.06	7.64
GEOBIESPT2	title, desc.	InL2	yes	0.05	0.06	7.44
GEOBIESPT3	title, desc., narr.	InL2	no	0.05	0.06	7.44
GEOBIESPT4	title, desc., narr	InL2	yes	0.05	0.06	7.64

Table 5: Spanish→Portuguese 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 Query Parsing

Information Extraction (IE) has traditionally involved manual processing in the form of rules or tagging training examples where the user is requied to specify the potential relations of interest ([10]). The main focus of IE has been on extracting information from homogeneous corpora such as newswire stories. Hence, traditional IE systems rely on linguistic techniques applied to the domain of interest, such as syntactic parsers and named-entity recognizers. The problem of extracting information from Web-based corpora presents different challenges. The use of name-entity recognizers and syntactic parsers encounters problems when applied to heterogeneous text found on the Web, and web-based queries are no exception.

Current work on query processing for retrieving geographic information on the Web has been done by Chen *et. al* ([11]). Their approach requires a combination of text and spatial data techniques for usage in geographic web search engines. A query to such an engine consists of keywords and the geographic area the user is interested in (i.e., query footprint).

In our case we are working with a collection of heterogeneous queries and no documents. The task as defined by the organizers comprises three subtasks: 1) recognize geographic web-based queries; 2) extract the geographical location, the latitude and longitude, geographical relations; and 3) categorize the queries into three types, namely "map", "information" and "yellow page".

The initial approach used in the Query Parsing task combines information extraction and patterns.

Due to the ungrammaticality, multilinguality and ambiguity of the language in the 800,000 web-based queries in the collection, we started by building a list of all the different words, similar to creating an index, excluding stopwords. Next, a lookup of the words was done in a list of countries, main cities and states to identify potential locations. The list was created from the GEOnet Names Server database ([8]). One problem were multiword georeferences. Because many locations were missed, we selected those queries where spatial prepositions such as "in", "near" and syntactic cues, such as "lake", "cayo", "street", "piazza", "hotel", "accommodation", were present. We have considered these as good heuristics for recognizing multiword expressions as georeferences and create pattern-based rules to further process potential candidates.

Extraction of geographical information such as latitude and longitude was done as follows. We created a new list of words identified as potential geographic references. Latitude and longitude information was looked up in the GNS database. A problem that we found is related to ambiguity since a geographic reference may refer to a city, state, park, and the same geographic entity may be in different continents, countries, states, and cities.

Finally, categorization was done using patterns. If the only information available was the name of a place, the query was categorized as of type "Map". If words such as "college", "airport", "studio" were present, the query was categorized as of type "Yellow Page". If the query included words such as "flight", "survey", "company", the query was categorized as of type "Information".

Results were low in terms of recall and precision. We are currently working on inferring a grammar and eventually a language model that would improve the performance our initial system.

5 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 (Spanish) topics into target language (English or Portuguese) topics. In general, performance retrieval was improved with automatic relevance feedback using the InL2 model. Further experiments indexing the collection with the language model option and querying with this option will allow us to compare parameter-based vs. language-based models. Parsing of web-based queries is a difficult task because of the nature of the data. Further investigation and application of classical and statistical language processing techniques is needed to improve the performance of the approach presented.

References

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

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

[3] 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.

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

[5] 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.

[6] 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)*.

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

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

[9] Guillén, R.: CSUSM Experiments at GeoCLEF2005: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, R. Guillén. (2006). 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.

[10] Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open Information Extraction from the Web. In Proceedings *Twentieth International Joint Conference on Artificial Intelligence 2007*, pp. 2670-2676.

[11] Chen, Y., Suel, T., Markowetz, A.: Efficient Query Processing in Geographic Web Search Engines. In Proceedings *SIGMOD 2006*, June 2006, pp. 277-288.