Geographic IR Helped by Structured Geospatial Knowledge Resources

A. Toral, O. Ferrández, E. Noguera, Z. Kozareva, A. Montoyo and R. Muñoz
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{atoral,ofe,elisa,zkozareva,montoyo,rafael}@dlsi.ua.es

Abstract

For the participation of the University of Alicante in the second edition of GeoCLEF, we have researched the incorporation of geographic knowledge into Geographic Information Retrieval (GIR). Our system is made up of an IR module used for several years in the CLEF competitions (IR-n) and a Geographic Knowledge module (Geonames). The latter is used to carry out an expansion of the initial topic by adding geographic items. The geographic items and relations are extracted from the topics and queries using the Geonames database are built from them. The returned information by this geographic resource is incorporated into the topics which at the end are processed by IR-n. We have submitted several runs, in order to compare the performance of the usage of a classic IR with the usage of geographic knowledge. The results show that the addition of geographic knowledge has negative impact on the obtained precision. However, the fact that for some topics the obtained results are better, makes us conclude that the addition of this knowledge could be useful but a lot of research effort is needed in order to determine how this knowledge should be correctly applied.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Geographic database, Experimentation, Measurement, Performance

Keywords

Information Retrieval, Geographic Information Retrieval, Geographic Database

1 Introduction

GeoCLEF is a track of the Cross-Language Evaluation Forum (CLEF) whose aim is to provide the necessary framework in which to evaluate Geographic Information Retrieval (GIR) Systems for search tasks involving both spatial and multilingual aspects.

The underlying and basic technology of GIR, Information Retrieval (IR), deals with the selection of the most relevant documents from a document collection given a query. Thus, GIR is a specialization of IR which introduces geospatial restrictions to the retrieval task.

Several approaches were followed in order to perform GIR within the first edition of GeoCLEF. Several systems used Named Entity Recognition [10] [4] [5] [8] [3] specialised to the geographic domain. Some used geographic knowledge resources [10] [5] [8] [3]. Some systems used Natural Language Processing tools such as Part-of-Speech tagging [5] or Text Mining [3]. Another approach was to perform a query expansion [10] [2]. Finally, there were approaches based on the classic IR without any treatment of geography [7] [6] [11].

Three out of the top-4 systems for the English monolingual run were based only on IR (the remaining one [4] used also geographic NER). This may be due to the fact that the systems which tried to use some kind of geographic reasoning did not do that in the correct way. Thus, the best results from GeoCLEF 2005, based on IR may be used as a baseline in order to test the performance of the geographic reasoning. This supports the claim that the research in GIR is still at the very first steps and so, there is a long way to go.

In our participation in GeoCLEF 2005 [4], we mentioned as an important aspect the lack of adequate ready-to-use structured knowledge resources of geographic items for our specific purpose. This is why for our participation in the second edition of this forum, we have centered our efforts on studying geographic resources and trying to determine how to use them within GIR. In a nutshell, we have researched the appliance of available resources of geospatial nature to GIR.

The GIR system developed with this purpose consists of exploiting geographic resources in order to make a query expansion with geographic knowledge. Obviously, we also want to evaluate the impact of the addition of these geographic items into our IR module.

The rest of this paper is organized as follows. The next section presents a detailed description of our system and the modules it is made of. Section 3 illustrates the carried out experiments and the obtained results by means of an example that shows the functioning of our system. Finally, Section 4 outlines our conclusions and future work proposals.

2 System Description

Our approach is based on IR with the appliance of geographic knowledge which is extracted from a structured knowledge resource. Figure 1 depicts an overview of our system as well as how the different modules interact among each other.

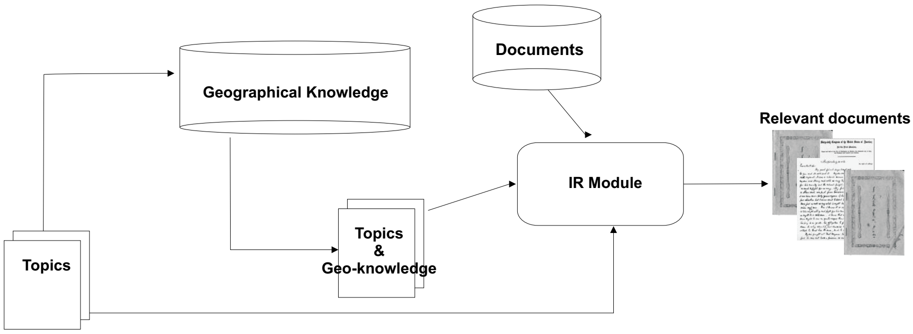

Figure 1: System architecture

The topics are processed and enriched with related geographic information which is obtained by exploiting the Geonames[1] resource. A SQL query is generated from the geographic information provided by the topic, and then the query is processed in the Geonames database in order to obtain the related geographic items.

Once all the geographic information is collected, we apply the IR module with this knowledge in order to retrieve all the relevant documents concerning this specific geospatial information.

The next subsections describe in details the two main modules of our system.

[1] www.geonames.org

2.1 IR Module

The IR module we used is called IR-n [12]. IR-n is a Passage Retrieval system (PR). These systems [9] study the appearance of query terms in contiguous fragments of the documents (also called passages). One of the main advantages of these systems is that they allow us to determine not only whether a document is relevant or not, but also to detect the relevant part of the document.

The passages are usually composed of a fixed number of sentences. This number depends on a measure obtained from the used document collection. To determine this value, the system has been trained on the GeoCLEF 2005 data collections. The number of sentences that obtains the best results is 8 both for English and Spanish. Furthermore, IR-n uses overlapping passages in order to avoid documents that could be considered as non relevant if there appear words of the question in adjacent passages.

IR-n allows the use of distinct similarity measures. With the aim to evaluate the most appropriate one, we have trained the system on the English and Spanish collections. For the both collections, the similarity measure which obtains the best results is dfr [1].

We have specifically adapted the IR-n system to incorporate geographic knowledge. In order to do this, we need to take into account two kinds of restrictions: required words and geographical items.

Required words These words are marked with '#'. Passages which do not contain at least one of these words are not included in the rank list.

Geographical places In addition, a query expansion has been done using the Geonames database (this is studied in depth in the section 2.2). They are added into a new label called <EN-geonames>.

As required words, we consider all the nouns of the topic (title, description and narrative) but geographic ones, stop words or other common words appearing in topic definitions (e.g. document, relevant). This is, we consider the words that define the main concept of the topic. The reason for doing this is to lessen the noise of the incorporation of big lists of geographic items to the IR query could introduce.

2.2 Geographic Knowledge Module

Geonames is a geographic database which contains more than 6 million entries for geographical names whereof 2.2 million are cities and villages. Geonames is built from different sources being the most important nga[2], gnis[3] and wikipedia[4] among others. Its data is freely available and may be used through web services or from database dumps which are periodically provided.

The information that Geonames provides for each entry is structured in several information fields from which we have used the following ones:

- Name: name of the geographical entry
- Alternate names: alternative names (different names for a geographical point that may include translations)
- Latitude: latitude in decimal degrees (wgs84)
- Longitude: longitude in decimal degrees (wgs84)
- Feature class: type of the entry according to Geonames taxonomy[5]
- Country code: ISO-3166 2-letter country code (two characters)

[2] http://gnswww.nga.mil/geonames/GNS/index.jsp

[3] http://geonames.usgs.gov/index.html

[4] http://www.wikipedia.org

[5] http://www.geonames.org/export/codes.html

- Population: number of inhabitants (only if the entry belongs to a populated place type)

Our approach regarding Geonames consisted of building a query to the Geonames database for each topic in a methodical way. We extract for each topic the geographic entities and relations, and we enrich the topic with the information that the query with this geographic info returns. We add an appendix in which the geographic queries for all topics are shown.

Due to the big size of Geonames, there is possible incorporation of noise into the topics. Therefore, we put some restrictions to the extracted data. From the returned entries to the query, we only consider those for which the population is bigger than 10,000 inhabitants and those that belong to a first-order administrative division (ADM1).

It should be noted that for some topics (26, 40 and 41), our method to build a query could not be applied, because the topics did not have any geographical restriction considered by Geonames. For instance topic 40 does not have any geographic restriction.

3 Experiments and Results

The organizers of GeoCLEF provide 25 topics in four languages (English, German, Portuguese and Spanish) for all participants, as well as different data collections for each target language (e.g. EFE 94 and EFE 95 for Spanish). In our participation in this edition of GeoCLEF we have evaluated our system for the English and Spanish monolingual tasks.

For each task, we have carried out three experiments. The first two ones just apply classic IR to the provided queries. The motivation is to provide an experiment which allow us to evaluate our approach which is practically carried out in our third experiment. The unique difference between these two experiments is that the first (called *uaTD*) uses only the topic title and description in order to retrieve the documents. However, the second experiment (called *uaTDN*)uses also the geographic information provided by the topic narrative section.

The third experiment (called *uaTDNGeo*), consists of IR module, but the queries which are passed to the system are previously enriched with geographic information. This information is obtained from the Geonames database. In the followings paragraphs, we show the whole process that is carried out with our system in this experiment.

The example with which we illustrate the process of our system is for topic 31.

1. Extract from the topic the required words and the geographic entities and relations:
   required words: combat, embargo, effect, fact
   geographic entities: Iraq
   geographic relations: north-of

2. Build the Geonames query

```
select name, alternames from geonames
 WHERE
  latitude>33 AND #average latitude of Iraq is 33 N
  country_code='IQ' AND
  ((feature_class='P' AND population > 10000) OR feature_code='ADM1');
```

3. Assemble a new IR query incorporating the extracted geographic knowledge

```
<num>GC031</num>
 <EN-title>Combats# and embargo# in the northern part of Iraq</EN-title>
 <EN-desc>Documents telling about combats# or embargo# in the northern
 part of Iraq</EN-desc>
 <EN-narr>Relevant documents are about combats# and effects# of the 90s
 embargo# in the northern part of Iraq.
 Documents about these #facts happening in other parts of Iraq are
 not relevant</EN-narr>
 <EN-geonames>Zakho Tozkhurmato Khurmati Touz Hourmato [...]</EN-geonames>
</num>
```

4. Retrieve the relevant documents using the IR-n system

Language	Run	AvgP
English	CLEF Average	0.1975
	uaTD	0.2723
	uaTDN	0.2985
	uaTDNGeo	0.1201
Spanish	CLEF Average	0.19096
	uaTD	0.3508
	uaTDN	0.3237
	uaTDNGeo	0.1525

Table 1: *Overall GeoClef 2006 officials results for the Monolingual tasks*

Topic	English AvgP			Spanish AvgP		
	uaTD	uaTDN	uaTDNGeo	uaTD	uaTDN	uaTDNGeo
026	49.07	50.09	48.34	16.67	15.18	15.18
027	0.22	0.66	0.04	2.56	3.89	10.35
028	16.85	3.93	0.59	30.56	37.65	0.47
029	9.07	12.84	4.91	57.58	68.63	0.07
030	91.67	95.83	0.00	43.05	37.18	0.00
031	43.10	35.57	2.03	61.18	69.22	41.28
032	88.41	90.05	64.59	90.00	95.91	90.16
033	0.30	0.50	2.59	6.00	2.31	54.64
034	37.68	51.52	1.44	13.51	20.67	0.00
035	5.07	2.40	0.00	21.05	8.17	0.15
036	0.00	0.00	0.00	60.91	21.43	0.00
037	9.16	0.06	1.30	20.69	0.00	0.00
038	1.03	0.57	0.00	20.00	11.78	0.00
039	3.94	36.53	6.28	38.81	30.75	4.34
040	36.93	32.11	26.93	70.50	77.24	77.55
041	0.17	0.18	0.67	40.00	38.12	22.45
042	45.00	100.00	6.55	32.08	45.30	4.58
043	1.13	0.95	0.05	12.50	8.54	0.11
044	11.34	8.38	0.00	46.60	5.69	0.01
045	10.04	29.89	34.89	8.33	1.66	0.11
046	71.43	69.05	4.08	60.71	77.26	6.47
047	5.48	1.69	0.00	3.39	0.00	0.00
048	80.82	82.66	81.69	66.04	81.18	52.18
049	36.11	35.00	0.00	51.15	38.69	0.06
050	26.79	15.85	13.23	22.00	12.85	1.04

Table 2: *Results topic by topic for the GeoClef 2006 Monolingual tasks*

The addition of geographical information has drastically decrement the precision. For English, the best run (uaTDN) obtains 29.85 while the geographic run (uaTDNGeo) achieves 12.01 (see Table 1). In the case of Spanish, the best run (uaTD) reaches 35.09 and the geographic one (uaTDNGeo) 15.25 (see Table 1). Although we implement the model of the required words in order to lessen the noise introduced by the large lists of geographic items to IR queries, this seems to be insufficient.

However, for both English and Spanish, the run with geographic information obtains the best results for the three topics (see Table 2): 33, 41 and 45 (EN) and 27, 33 and 40 (ES). Therefore, a more in-depth analysis should be carried out in order to achieve a better understanding on the behaviour of the geographic information incorporation and how it should be done.

It should be noted that the results for Spanish are slightly better than those for English. This is so for every run we have submitted (uaTD, uaTDN and uaTDNGeo). This happens because the IR module was initially designed for Spanish and, moreover, it has been used for this language for several years.

4 Conclusions and Future Work

For our participation in GeoCLEF 2006 we have proposed the expansion of IR queries with geographic information related to the topics. For this purpose we have studied knowledge geographic resources and we have used Geonames.

The proposal has obtained poor results compared to our simpler model in which we only use an Information Retrieval system. This is a paradigmatic example of the state of the art of the GIR field; it is just the beginning and more efforts are needed in order figure out how to introduce the geographic knowledge in a way that the basic IR systems could benefit from it.

Therefore, as future work we consider to research into different ways of providing the geographic knowledge to basic IR and evaluating the impact of each approach. Thus, our aim is to improve GIR results by applying existing geographic knowledge from structured resources.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV06-161.

References

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic Models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[2] Davide Buscaldi, Paolo Rosso, and Emilio Sanchis Arnal. A WordNet-based Query Expansion method for Geographical Information Retrieval. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[3] Nuno Cardoso, Bruno Martins, Marcirio Silveria Chaves, Leonardo Andrade, and Mario J. Silva. The XLDB Group at GeoCLEF 2005. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[4] Oscar Ferrández, Zornitsa Kozareva, Antonio Toral, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, and Fernando Llopis. The University of Alicante at GeoCLEF 2005. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[5] Daniel Ferrés, Alicia Ageno, and Horacio Rodríguez. The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis and a Geographical Thesaurus. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[6] Fredric Gey and Vivien Petras. Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[7] Rocio Guillé. CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[8] Baden Hughes. NICTA i2d2 at GeoCLEF 2005. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[9] M. Kaskziel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185, 1997.

[10] Sara Lana-Serrano and Jose M. Goñi-Menoyo. MIRACLE's 2005 Approachj to Geographical Information Retrieval. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[11] Ray R. Larson. Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR. *Working Notes in Cross-Language Evaluation Forum (CLEF) 2005*, 2005.

[12] Fernando Llopis. IR-n un Sistema de Recuperacin de Informacin Basado en Pasajes. Ph.D. tesis. *Procesamiento del Lenguaje Natural*, 30:127–128, Universidad de Alicante 2003.

A SQL queries

```
26. no geographic SQL-query was implemented;

27. (longitude>7.98 AND longitude<9.38 AND latitude>49.21 AND
latitude<51.01);

28. (country_code='CA' OR country_code='US' OR country_code='MX');

29. (country_code='AO' OR country_code='ZA');

30. (longitude>-6.32 AND longitude<-1.04 AND latitude>38.40 AND
latitude<42.40);

31. latitude>33.20 AND country_code='IQ';

32. country_code='CA' and admin1_code=10;

33. country_code='DE' and admin1_code=7;

34. (latitude>-23.51 AND latitude<23.51) AND ((feature_class='P' AND
population > 50000) OR feature_code='ADM1');

35. (country_code='BG' OR country_code='HU' OR country_code='CZ' OR
country_code='SK' OR country_code='PL' OR country_code='RO');

36. (country_code='JP' OR country_code='KP' OR country_code='KR' OR
country_code='RU');

37. (country_code='IR' OR country_code='IQ' OR country_code='TK' OR
country_code='EG' OR country_code='LB' OR country_code='SA' OR
country_code='JO' OR country_code='YE' OR country_code='QA' OR
country_code='KW' OR country_code='BH' OR country_code='IL' OR
country_code='OM' OR country_code='SY' OR country_code='AE' OR
country_code='CY' OR country_code='PS');

38. (country_code='BN' OR country_code='KH' OR country_code='TL' OR
country_code='ID' OR country_code='LA' OR country_code='MY' OR
country_code=’MM’ OR country_code=’PH’ OR country_code=’SG’ OR
country_code=’TH’ OR country_code=’VN’);

39. (country_code=’AZ’ OR country_code=’AM’ OR country_code=’GE’);

40. no geographic SQL-query was implemented;

41. no geographic SQL-query was implemented;

42. country_code=’DE’ and (admin1_code=’03’ or admin1_code=’04’ or
admin1_code=’06’ or admin1_code=’12’ or admin1_code=’10’);

43. country_code=’US’ and (admin1_code=’CT’ or admin1_code=’RI’ or
admin1_code=’MA’ or admin1_code=’VT’ or admin1_code=’NH’ or
admin1_code=’ME’);

44. (country_code=’SI’ OR country_code=’MK’ OR country_code=’HR’ OR
country_code=’YI’ OR country_code=’BK’);

45. country_code=’BR’ and (admin1_code=’02’ or admin1_code=’05’ or
admin1_code=’06’ or admin1_code=’13’ or admin1_code=’17’ or
admin1_code=’19’ or admin1_code=’20’ or admin1_code=’22’ or
admin1_code=’28’);

46. country_code=’PT’ and (admin1_code=’21’ or admin1_code=’17’ or
admin1_code=’04’ or admin1_code=’05’);

47. (country_code=’FR’ or country_code=’SP’ or country_code=’MC’ or
country_code=’IT’ or country_code=’MT’ or country_code=’SI’ or
country_code=’HR’ or country_code=’BA’ or country_code=’CS’ or
country_code=’AL’ or country_code=’GR’ or country_code=’TR’ or
country_code=’CY’);

48. (country_code=’GL’);

49. (country_code=’FR’);

50. (country_code=’DE’ or country_code=’AT’ or country_code=’SK’ or
country_code=’HU’ or country_code=’HR’ or country_code=’CS’ or
country_code=’BG’ or country_code=’RO’ or country_code=’UA’ or
country_code=’LI’ or country_code=’FR’ or country_code=’NL’ or
country_code=’CH’);
```