R2D2 at GeoCLEF 2006: a mixed approach

Manuel García-Vega, Miguel A. García-Cumbreras
L. Alfonso Ureña-López, José M. Perea-Ortega, F. Javier Ariza-López
University of Jaén
{mgarcia,magc,laurena,jmperea,fjariza}@ujaen.es
Oscar Ferrández, Antonio Toral, Zornitsa Kozareva
Elisa Noguera, Andrés Montoyo, Rafael Muñoz
University of Alicante
{ofe,atoral,zkozareva,elisa,montoyo,rafael}@dlsi.ua.es
Davide Buscaldi, Paolo Rosso
Polytechnical University of Valencia
{dbuscaldi,prosso}@dsic.upv.es

Abstract

This paper describes the participation of a mixed approach in GeoCLEF-2006. We have participated in Monolingual English Task and we present a joint work of three groups or teams belonging to project R2D2 [1] with a new system, mixing the 3 individual systems of the teams.

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; H.3.7 Digital Libraries

General Terms

Algorithms, Languages, Performance, Experimentation

Keywords

Information Retrieval, Geographic Information Retrieval, Named Entity Recognition, GeoCLEF

1 Introduction

The aim of GeoClef 2006 monolingual and bilingual tasks is to retrieve relevant documents from a monolingual collection. These documents are retrieved by using geographic tags like geographic places, geographic events and so on. Nowadays, the fast development of Geographic Information Systems (GIS) involves the need of Geographic Information Retrieval system (GIR) that helps these systems to obtain documents with relevant geographic information.

In this paper we present a joint work of three groups or teams belonging to project R2D2: UJA[2], UA[3] and UPV[4].

[1] In Spanish "Recuperación de Respuestas en Documentos Digitalizados" —Answering Retrieval in Digitized Documents—. Financed project by Spanish Government with grant TIC2003-07158-C04.

[2] University of Jaén with RIM subproject and reference TIC2003-07158-C04-04

[3] University of Alicante with BRUJULA subproject and reference TIC2003-07158-C04-04-01

[4] Polytechnical University of Valencia with SIRICM subproject and reference TIC2003-07158-C04-03

Our approach has been a mixed system based on individual scores, generating a new one with the voted final results. Since this is the first year of the voting system, we have used a simple method.

The rest of the paper is organized as follows: Sections 2, 3 and 4 describes the individual systems, firstly the UA system, following, UJA system and finally the UPV system. Section 5 describes the voting system and section 6 shows the results. Finally, section 7 describes the conclusions and future work.

2 UJA System

The SINAI team at University of Jaén propose a Geographical Information Retrieval System that is made up of five subsystems:

- **Translation Subsystem**: is the query translation module. This subsystem translates the queries to the other languages and it is used for the following bilingual tasks: Spanish-English, Portuguese-English and German-English. For the translation an own module has been used, called SINTRAM (SINai TRAnslation Module), that works with several online Machine Translators, and implements several heuristics. For these experiments we have used an heuristic that joins the translation of a default translator (the one that we indicate depends of the pair of languages), with the words that have another translation (using the other translators).
- **Named Entity Recognition-Gazetteer Subsystem**: is the query geo-expansion module. The main goal of NER-Gazetteer Subsystem is to detect and recognize the entities in the queries, in order to expand the topics with geographical information. We are only interested in geographical information, so we have used only the locations detected by the NER module. The *location* term includes everything that is *town, city, capital, country* and even *continent*. the information about locations is loaded previously in the Geographical Information Subsystem, that is related directly to NER-Gazetteer Subsystem. The NER-Gazetteer Subsystem generates some labelled topics, based on the original one, adding the found locations.
- **Geographical Information Subsystem**: is the module that stores the geographical data. This information has been obtained from Geonames[5] gazetteer. The objective of this module is to expand the locations of the topics, using geographical information. The expansion that we do is the *automatic query expansion*[2]. The Geonames database contains over six million entries for geographical names, whereof 2.2 million cities and villages. It integrates geographical data such as names, altitude, population and others from various sources. When a location is recognized by the NER subsystem we look for in the Geographical Information Subsystem. In addition, it is necessary to consider the spatial relations found in the query (“near to", “within X miles of", “north of", “south of", etc.). Depending on the spatial relations, the search in the Geographical Information subsystem is more or less restrictive.
- **Thesaurus Expansion Subsystem**: is the query expansion module using an own Thesaurus. A collection of thesauri was generated from the GeoCLEF training corpus. This subsystem was looking for words with a very high rate of document co-location. These words were treated like synonyms and added to the topics. An inverse file with the entire collection was generated for comparing words. Training with GeoCLEF-2005 files, a 0.9 cosine similarity was the best rate that obtain the best precision/recall results.
- **IR Subsystem**: is the Information Retrieval module. The English collection dataset has been indexed using LEMUR IR system[6]. It is a toolkit[7] that supports indexing of large-

[5] http://www.geonames.org/

[6] http://www.lemurproject.org/

[7] The toolkit is being developed as part of the Lemur Project, a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

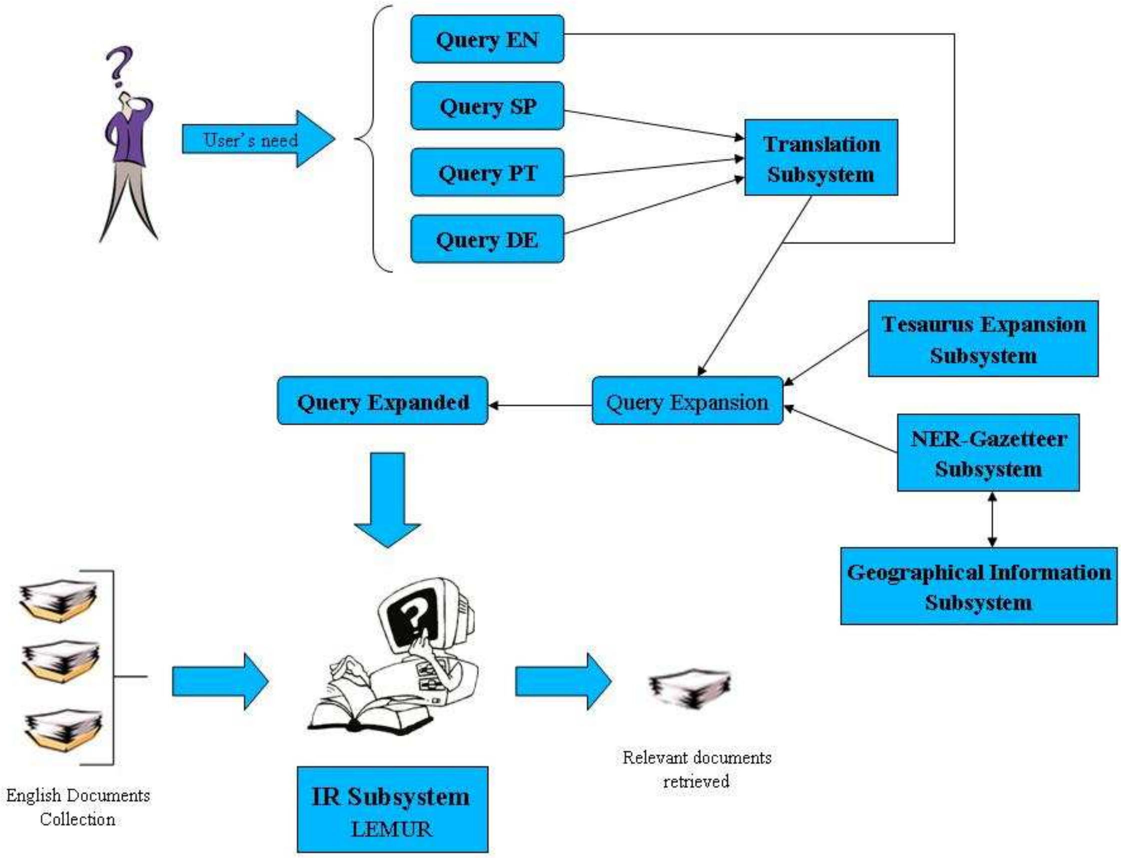

Figure 1: UJA system architecture

3 UA System

The aim of University of Alicante approach is to evaluate the impact of the appliance of geographic knowledge extracted from a structured resource over a classic Information Retrieval (IR) system. In the Figure 2, an overview of the system is depicted.

The GIR system developed by the University of Alicante for the second edition of GeoCLEF is made up of two main modules:

IR A Passage Retrieval (PR) module called IR-n [3] has been used for several years in the GeoCLEF campaign. It allows using different similarity measures. The similarity measure used has been dfr as it has been the one that obtained the best results when trained over CLEF corpora.

Geographic knowledge Also, a geographic database resource called Geonames [8] was used, which is freely available and may be used through web services our downloaded as a database

[8] http://www.geonames.org

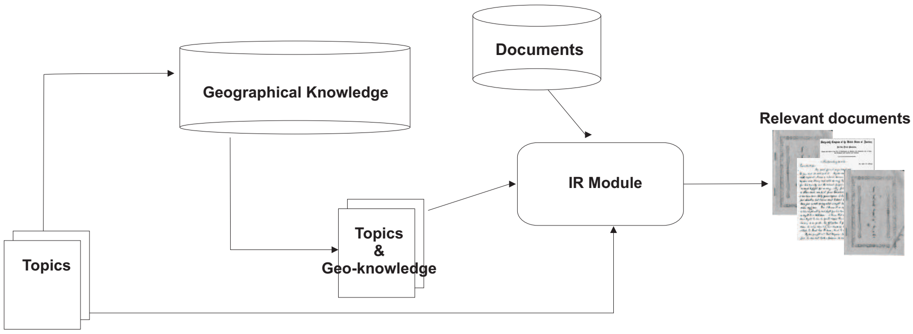

Figure 2: UA system architecture

dump. It is a structured resource built from different sources such as NGA, GNIS and Wikipedia. For each entry it contains several fields which provide name, additional translations, latitude, longitude, class according to an own taxonomy, country code, population and so on. This module is used to enrich the initial information provided by topics with related geographic items.

The system works in the following way:

1. Topic processing: the topic is processed in order to obtain the relevant geographic items and the geographical relations among them. Besides, all the nouns in the topic which are not of a geographic nature are marked as required words, belonging to topics widely used words (e.g. document) nor stop words.

2. Geographic query: once the geographic information from the topic is obtained, a query to the Geonames database is methodically built. This has the aim of obtaining related geographic items from this resource. In order to build this query information, longitude, latitude or country names are considered.

3. Topic enrichment: a new topic is composed of the information contained in the provided topic and the geographic items obtained in the previous step.

4. Information Retrieval: finally, the relevant documents from the collection according to the topic and its related geographic items are obtained. For a document to be considered in the retrieval, it should contain at least one of the required words. The objective of this is to lessen the noise that could be introduced by adding big lists of geographic items to the topic.

Even though we have incorporated geographic knowledge in our system, we can conclude that the research in GIR is at the very first steps. This claim is supported by the fact that the systems that obtained the best results in the first edition of GeoCLEF were the ones using classic IR without any geographic knowledge. Therefore, we plan to continue this line of research, whose principal aim is to make out how to incorporate information of this nature so that the systems can benefit from it.

4 UPV System

The GeoCLEF system of the UPV is based on a WordNet-based expansion of the geographical terms in the documents, that exploits the synonymy and holonymy relationships. This can be seen as an "inverse" approach with respect to the UPV's 2005 system [1] which exploited the meronymy and synonymy in order to perform a query expansion. Query expansion was abandoned due to the poor results obtained in the previous edition of the GeoCLEF.

WordNet [4] is a general-domain ontology, but includes some amount of geographical information that can be used for the Geographical Information Retrieval task. However, it is quite difficult to calculate the number of geographical entities stored in WordNet, due to the lack of an explicit annotation of the synsets. We retrieved some figures by means of the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. As a comparison, a specialized resource like the Getty Thesaurus of Geographic Names (TGN)[9] contains 3094 entities of type “city".

The indexing process is performed by means of the Lucene[10] search engine, generating two index for each text: a *geo* index, containing all the geographical terms included in the text together with those obtained by means of WordNet, and a *text* index, containing the stems of text words that are not related to geographical entities. Thanks to the separation of the indices, a document containing “John Houston” will not be retrieved if the query contains “Houston”, the city in Texas. The adopted weighting scheme is the usual *tf-idf*. The geographical names were detected by means of the Maximum Entropy-based tool available from the *openNLP* project[11].

Since the tool does not perform a classification of the named entities, the following heuristics is used in order to identify the geographical ones: when a Named Entity is detected, we look in WordNet if one of the word senses has the *location* synset among its hypernyms. If this is true, then the entity is considered a geographical one.

For every geographical location l, the synonyms of l and all its holonyms (even the inherited ones) are added to the *geo* index. For instance, if *Paris* is found in the text, its synonyms *City of Light*, *French capital*, *capital of France* are added to the *geo* index, together with the holonyms: {*France, French republic*}, {*Europe*}, {*Eurasia*}, {*Northern Emisphere*} and {*Eastern Hemisphere*}. The obtained holonyms tree is:

```
Paris, City of Light, French capital
=>France, French republic
 =>Europe
  =>Eurasia
  =>Northern Hemisphere
  =>Eastern Hemisphere
```

The advantage of this method is that knowledge about the enclosing, broader, geographical entities is stored together with the index term. Therefore, any search addressing, for instance, *France*, will match with documents where the names *Paris*, *Lyon*, *Marseille*, etc. appear, even if *France* is not explicitly mentioned in the documents.

5 The Mixed System

These three systems have returned their own final scores. Our approach has been a mixed system based on the individual scores generating a new one with the voted final results.

Since this is the first year of the voting system, we have used a simple method:

1. The systems have its own scoring method and we need a normalized version of them for a correct composition. After this procedure, all scores will have values between 0 and 1.

2. If a document is very well considered in the three systems, we want it to have a good value in the mixed one. We have used the addition of the normalized values like final score. If a document appears in more than one system, it has an score result of the sum of each score.

3. A list of new scores was generated from final results of all systems. Finally, we obtained our system final score by sorting this list and cutting in the 1,000th position.

[9] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[10] http://lucene.apache.org

[11] http://opennlp.sourceforge.net

Interpolated Recall (%)	Precision Averages (%)
0%	43,89%
10%	36,13%
20%	34,85%
30%	33,53%
40%	33,09%
50%	32,25%
60%	24,06%
70%	14,30%
80%	11,60%
90%	7,30%
100%	6,08%
Average precision	24,03%

Table 1: Average precision in monolingual task

Docs Cutoff Levels	Precision at DCL (%)
5 docs	21,60%
10 docs	17,60%
15 docs	16,80%
20 docs	16,60%
30 docs	13,73%
100 docs	7,40%
200 docs	4,42%
500 docs	2,09%
1.000 docs	1,25%
R-Precision	23,19%

Table 2: R-precision in monolingual task

6 Results

Our mixed approach has only participated in monolingual task and the official results are shown in Table 1 and Table 2. The results shown that the mixed approach must improve more because average precision obtained is poor. This could become analyzing the advantages of each system in each case or topic and to obtain what system works better for each spatial relation or type of region (location, country, etc.). The future system of voting would have to consider the previous analysis and to weigh with greater or smaller score the results of each system depending on the type of question or spatial relation.

7 Conclusions and Future work

We have presented a mixed approach using other 3 GeoCLEF-2006 systems and the results are promising. The mixed approach takes advantages of them, although it acquires their defects too. The linearity of the voting system has benefit to the common documents in the 3 individual list of document, relevant or not.

The future work must be about the system of voting. Using some technique of artificial intelligence for detecting the quality of the individual systems will improve the results. A neural network can tune the precision of the individual system for improve the behavior of the mixed one.

8 Acknowledgments

This work has been supported by Spanish Government with grant TIC2003-07158-C04-04 (University of Jaén), TIC2003-07158-C04-04-01 (University of Alicante) and TIC2003-07158-C04-03 (Polytechnical University of Valencia).

References

[1] D. Buscaldi, P. Rosso, and E. Sanchis. Using the wordnet ontology in the geoclef geographical information retrieval task. In *Proceedings of the CLEF 2005*, 2005.

[2] D. Buscaldi, P. Rosso, and E. Sanchis-Arnal. A wordnet-based query expansion method for geographical information retrieval. *Working Notes for the CLEF 2005 Workshop*, 2005.

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

[4] G. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.