R2D2 at GeoCLEF 2006: a Combined Approach

Manuel García-Vega¹, Miguel A. García-Cumbreras¹
L. Alfonso Ureña-López¹, José M. Perea-Ortega¹, F. Javier Ariza-López¹
Oscar Ferrández², Antonio Toral², Zornitsa Kozareva²
Elisa Noguera², Andrés Montoyo², Rafael Muñoz²
Davide Buscaldi³, Paolo Rosso³

¹ University of Jaén

{mgarcia,magc,laurena,jmperea,fjariza}@ujaen.es

² University of Alicante

{ofe,atoral,zkozareva,elisa,montoyo,rafael}@dlsi.ua.es

³ Polytechnical University of Valencia

{dbuscaldi,proso}@dsic.upv.es

Abstract. This paper describes the participation of a combined approach in GeoCLEF-2006. We have participated in Monolingual English Task and we present joint work of the three groups or teams belonging to the project R2D2⁴ with a new system, combining the three individual systems of these teams. We consider that research in the area of GIR is still in its very early stages, therefore, although a voting system could improve the individual results of each system, we have to further investigate different ways to achieve a better combination of these systems.

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 relevant 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 GIR system that helps these systems to obtain documents with relevant geographic information.

In this paper we present joint work of the three groups belonging to the project R2D2: UJA⁵, UA⁶ and UPV⁷.

⁴ In Spanish “Recuperación de Respuestas en Documentos Digitalizados” —Answering Retrieval in Digitized Documents—. Financed project by Spanish Government with grant TIC2003-07158-C04.

⁵ University of Jaén with RIM subproject and reference TIC2003-07158-C04-04

⁶ University of Alicante with BRUJULA subproject and reference TIC2003-07158-C04-04-01

⁷ Polytechnical University of Valencia with SIRICM subproject and reference TIC2003-07158-C04-03

Our approach consists of a combined system based on the three systems' scores of the aforementioned teams. These scores are treating individually, generating a new one by means of voting strategy. 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 UJA system, secondly the UA system and finally the UPV system. Section 5 describes the voting system and section 6 shows the results. Finally, section 7 illustrates some conclusions and future work proposals.

2 Description of the System of the University of Jaén

The SINAI team at University of Jaén propose a Geographical Information Retrieval System that is composed 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, a dedicated module has been used, called SINTRAM (SINai TRAnslation Module), that works with several online machine translation services, 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 been alternatively translated by another translation service (using the other translators).
- **Named Entity Recognition-Gazetteer Subsystem:** is the query geo-expansion module. The main goal of the 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 the NER-Gazetteer Subsystem. The NER-Gazetteer Subsystem generates labels the original topics by adding the locations found.
- **Geographical Information Subsystem:** is the module that stores the geographical data. This information has been obtained from the Geonames⁸ 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 are 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”,

⁸ <http://www.geonames.org/>

“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 a thesaurus developed at the University of Jaén. A collection of thesauri was generated from the GeoCLEF training corpus. This subsystem was looking for words with a very often cooccurring in documents. 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 cosine similarity of 0.9 turned out to be the best parameter setting based on the recall and precision of the results.
- **IR Subsystem:** is the Information Retrieval module. The English collection dataset has been indexed using the LEMUR IR system⁹. It is a toolkit¹⁰ that supports indexing of large-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.

3 Description of the System of the University of Alicante

The aim of the University of Alicante approach is to evaluate the impact of the application 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. Here we show a brief description of them, for further details see [5].

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 Geonames¹¹ was used, which is freely available and may be used through web services or downloaded as a database 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 a self develop taxonomy, country code, population and so on. This module is used to enrich the initial information provided by topics with related geographic items.

⁹ <http://www.lemurproject.org/>

¹⁰ 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.

¹¹ <http://www.geonames.org>

Fig. 1. UJA system architecture

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 and the run enriched with this knowledge was used for the combined system, we

Fig. 2. UA system architecture

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 Description of the System of the Polytechnical University of Valencia

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)¹² contains 3094 entities of type “city”.

The indexing process is performed by the Lucene¹³ 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

¹² http://www.getty.edu/research/conducting_research/vocabularies/tgn/

¹³ <http://lucene.apache.org>

entities only. 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* project¹⁴.

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 Hemisphere*} 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 Combined System

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

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

1. The systems have their own scoring method and we need a normalized version of them for a correct fusion. 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 combined one. We have used the addition of the normalized values as the final score. If a document appears in more than one system, it has a score result of the sum of each score.

¹⁴ <http://opennlp.sourceforge.net>

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

6 Experimental results

Our combined approach has only participated in monolingual task and the official results are shown in Table 1 and Table 2. The results shown that the combined approach must be improved because the average precision obtained is poor. This could be done 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.

Table 1. Average precision in monolingual task

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%

7 Conclusions and Future work

We have presented a combined approach using other three GeoCLEF-2006 systems. Although the results have not been very high we consider them quite promising. The combined approach takes advantages of them, although it acquires their defects too. The linearity of the voting system biases documents common in all three individual list of documents whether relevant or not.

As future work we intend to improve the combined approach using three individual systems considering other voting strategies. Using some techniques from artificial intelligence for detecting the quality of the individual systems could improve the final results. A neural network can tune the precision of the individual system for improve the behavior of the combined one.

Table 2. R-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%

Acknowledgments

This work has been supported by Spanish Government with grant TIC2003-07158-C04-04 (University of Jan), 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*, volume 4022, pages 939–946, 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 Recuperacin de Informacin Basado en Pasajes. *Procesamiento del Lenguaje Natural*, 30:127–128, 2003.
4. G. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.
5. Antonio Toral, Oscar Ferrández, Elisa Noguera, Zornitsa Kozareva, Andrés Montoyo, and Rafael Muñoz. Geographic IR Helped by Structured Geospatial Knowledge Resources. In Alessandro Nardi, Carol Peters, and Jose Luis Vicedo, editors, *Working Notes at CLEF 2006*, 2006.