Using Query Reformulation and Keywords in the Geographic Information Retrieval Task

José Manuel Perea-Ortega, L. Alfonso Ureña-López, Manuel García-Vega, and Miguel Angel García-Cumbreras

SINAI Research Group*, Computer Science Department, University of Jaén, Spain
{jmperea,laurena,mgarcia,magc}@ujaen.es

Abstract. This paper describes the use of query reformulation to improve the Geographic Information Retrieval (GIR) task. This technique also includes the geographic expansion of the topics. Moreover, several experiments related to the use of keywords and hyponyms in the filtering process are performed. We also use a new approach in the re-ranking process based on the original position of each document in the ranking. The results obtained show that our query reformulation sometimes retrieves valid documents that the default query is not able to find, but on average it does not improve the baseline case. The best result is obtained considering the geographic entities in the traditional retrieval process.

1 Introduction

In this paper we describe our system to resolve the Geographic Information Retrieval task. GeoCLEF is a cross-language GIR task, whose aim is to evaluate GIR systems. It belongs to the Cross-Language Evaluation Forum[1] (CLEF) campaign since 2005. GIR is concerned with improving the quality of geographically specific information retrieval with a focus on access to unstructured documents [4].

For GeoCLEF 2006 [3], we studied the behavior of query expansion using a gazetteer and a thesaurus. Those experiments showed us that the method we used to make the query expansion was not very good. For GeoCLEF 2007 [6], we changed the approach and we applied a filtering process to the documents retrieved by the IR subsystem without using any query expansion. The results for this approach were better than those achieved in 2006 using query expansion.

In the new system, we have added some Natural Language Processing (NLP) techniques such as query reformulation, keywords and hyponyms extraction and even query geo-expansion. In the next section we describe the general architecture. In Section 3 we present the experiments and results and finally we expound the conclusions and future work.

2 SINAI-GIR System Overview

As we can see in Figure 1, our GIR system is made up of five main subsystems: *Translator*, *Collection Preprocessing*, *Query Analyzer*, *Information Retrieval* and

* http://sinai.ujaen.es
[1] http://www.clef-campaign.org/

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 855–862, 2009.

used an optimal method to raise valid documents in the final ranking. However, if we do not consider geo-entities in the topics for the text retrieval, the use of the filtering and re-ranking process improves the results.

In relation to the use of *keywords* in the re-ranking process, it seems to improve slightly the filtering results for some experiments. Instead, the use of *hyponyms* does not improve the results. This is because we have tried to find each keyword (or hyponym) from the topics, into each document retrieved by the IR subsystem through a simple matching. It is difficult that appears the keyword or hyponyms exactly in the document. Nevertheless, the proper use of *keywords* in the re-ranking process could be interesting for the future.

On the other hand, the type of query reformulations we have used in the experiments does not seem to work well, although in some topics the Q_2 and Q_3 query types add valid documents to the final list which have not been found using the default query (Q_1). The main reason to explain the low results obtained with query geo-expansion is that we expand the topics with all geo-entities related to the "*where*" component, for the same query reformulation, so it is introducing a lot of noise in the retrieval process. For the future, we will generate a query geo-expansion for each geo-term related to the "*where*" component detected in the topic and we will retrieve them separately. In addition, we have tried to improve the query reformulation and the geographic expansion analyzing when they should be expanded and how.

Acknowledgments

This work has been supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by the University of Jaén (Spain).

References

1. Brill, E.: A simple rule-based part-of-speech tagger. In: Proceedings of the third Conference on Applied Natural Language Processing (ANLP 1992), Trento, Italy, pp. 152–155 (1992)
2. García-Cumbreras, M.A., Ureña-López, L.A., Martínez-Santiago, F., Perea-Ortega, J.M.: BRUJA System. In: The University of Jaén at the Spanish task of QA@CLEF 2006. LNCS, vol. 4730, pp. 328–338. Springer, Heidelberg (2007)
3. García-Vega, M., García-Cumbreras, M.A., Ureña-López, L.A., Perea-Ortega, J.M.: GEOUJA System. In: The first participation of the University of Jaén at GEOCLEF 2006. LNCS, vol. 4730, pp. 913–917. Springer, Heidelberg (2007)
4. Jones, C.B., Purves, R.S.: Geographical Information Retrieval. International Journal of Geographical Information Science 22, 1365–8816, 219–228 (2008)
5. Mandl, T., Carvalho, P., Gey, F., Larson, R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In: Proceedings of the Cross Language Evaluation Forum, CLEF 2008 (2008)