GEOUJA System. University of Jaén at GEOCLEF 2007

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, Arturo Montejo-Ráez
SINAI Group. Department of Computer Science. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,magc,mgarcia,amontejo}@ujaen.es

Abstract

This paper describes the second participation of the SINAI group of the University of Jaén in GeoCLEF 2007. We have developed a system different from the one presented in GeoCLEF 2006. Our architecture is made up of five main modules. The first one is the Information Retrieval Subsystem, that works with collections and queries in English and returns relevant documents for a query. The queries that are not in English are translated by the Translation Subsystem. All the queries are filtered by the Geo-Relation Finder Subsystem, that finds any spatial relation in the topic, and NER (Named Entities Recognition) Subsystem, that looks for any location in the topic. The most important module is the Geo-Relation Validator Subsystem, it applies some heuristics to filter documents recovered by the IR Subsystem. We have made several runs, combining these modules to resolve the monolingual and the bilingual tasks. The results obtained show that the heuristics applied are quite restrictive and therefore it must be generated new heuristics and to improve the definition of new rules to filter recovered documents.

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 objective of GeoCLEF is to evaluate Geographical Information Retrieval (GIR) systems in tasks that involve both spatial and multilingual aspects. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections in English, but with topics in English, Spanish, German or Portuguese [3]. This is our second participation in GeoCLEF, after the previous year[2].

In the last edition we studied the behavior of query expansion. The results obtained showed us that filtering improves precision and recall. For this reason, our system consists of five subsystems: Translation, Geographical Relations Finder, NER, Validator and Information Retrieval.

Experiment	Mean Average Precision	R-Precision
Sinai_GEEN_Exp1_fb_okapi	0.0686	0.0704
Sinai_PTEN_Exp1_fb_okapi	0.1568	0.1519
Sinai_SPEN_Exp1_fb_okapi	0.2362	0.2238
Sinai_GEEN_Exp1_fb_tfidf	0.0572	0.0606
Sinai_PTEN_Exp1_fb_tfidf	0.1080	0.1133
Sinai_SPEN_Exp1_fb_tfidf	0.1511	0.1533
Sinai_GEEN_Exp1_simple_okapi	0.0484	0.0569
Sinai_PTEN_Exp1_simple_okapi	0.1544	0.1525
Sinai_SPEN_Exp1_simple_okapi	0.2310	0.2476
Sinai_GEEN_Exp1_simple_tfidf	0.0435	0.0420
Sinai_PTEN_Exp1_simple_tfidf	0.1053	0.1117
Sinai_SPEN_Exp1_simple_tfidf	0.1447	0.1513
Sinai_PTEN_Exp2_fb_tfidf	0.0695	0.1074

Table 2: Summary of results for the bilingual task

that have been recovered are valid but the GR Validator Subsystem has filtered some ones that must not have eliminated.

For the future, we will try to add more heuristics to the GR Validator Subsystem making use of Geonames Gazetteer. Also we will define more precise rules so that the system is less restrictive for the selection of recovered documents. Finally, we will also explore a larger number of retrieved documents by the IR Subsystem, in the aim of providing a larger variety of documents to be checked by the GR Validator Subsystem.

6 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIN2006-15265-C06-03.

References

[1] Miguel A. García-Cumbreras, L. Alfonso Ureña-López, Fernando Martínez Santiago, and José M. Perea-Ortega. Bruja system. the university of jaén at the spanish task of qa@clef 2006. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[2] Manuel García-Vega, Miguel A. García-Cumbreras, L.A. Ureña-López, and José M. Perea-Ortega. Geouja system. the first participation of the university of jaén at geoclef 2006. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[3] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, and Paulo Rocha. Geoclef 2006: the clef 2006 cross-language geographic information retrieval track overview. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[4] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.

[5] S.E. Robertson and S.Walker. Okapi-Keenbow at TREC-8. In *Proceedings of the 8th Text Retrieval Conference TREC-8, NIST Special Publication 500-246*, pages 151–162, 1999.

[6] G. Salton and G. Buckley. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 21:288–297, 1990.