SINAI-GIR System. University of Jaén at GeoCLEF 2008

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, L. Alfonso Ureña-López
SINAI Research Group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,magc,mgarcia,laurena}@ujaen.es

Abstract

This paper describes the third participation of the SINAI research group from University of Jaén in GeoCLEF track. We have tried to improve the system proposed last year in GeoCLEF 2007. The main developments are related to the use of query reformulation, keywords recognition, hyponyms extraction and query geo-expansion. On the other hand, new rules have been applied in the *Validator* subsystem in order to filter the documents recovered by the IR subsystem. We have run several experiments, combining these developments in order to resolve the monolingual and bilingual tasks. The results obtained shown that filtering does not reach yet to improve the baseline case. However, the use of *keywords* and *hyponyms* in the re-ranking process seems to improve the filtering results. On the other hand, the use of query reformulation and geo-expansion does not improve the baseline case either.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic Information Retrieval (GIR), Named Entity Recognition (NER), Spatial Relations, Geo-referencing, Filtering Documents, GeoCLEF

1 Introduction

GeoCLEF is a cross-language Geographic Information Retrieval (GIR) task at the Cross-Language Evaluation Forum (CLEF) campaign since 2005. The aim of GeoCLEF is to evaluate GIR systems. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections using queries into several languages [1, 2]. Queries are textual descriptions with three fields (*title*, *description* and *narrative*), including spatial relations and geographic locations such us continents, seas, rivers or regions.

This paper describes the approaches taken by the SINAI[1] research group from the University of Jaén for the main GeoCLEF 2008 subtasks: mono and bilingual retrieval. In 2006 [3], we studied

[1] http://sinai.ujaen.es

5 Conclusions

In this paper we have presented the experiments carried out in our third participation in the GeoCLEF track, following the basic architecture used in the 2007 experiments. We have tried to improve the filtering and re-ranking process introduced the previous year, adding new developments related to several techniques such as query reformulation, *keywords* and *hyponyms* extraction and even query geo-expansion. Moreover, we have established predefined weights for each manual rule in *Validator* in order to improve the final score of the valid documents. However, we still get the best result without applying any of these techniques. This is because we have not used an optimal method to raise valid documents in the final ranking, depending on the *geo-information* that recovered documents and topics have in common.

About the new developments employed in the experiments, only the use of *keywords* in the re-ranking process seems to improve the filtering results in some cases. Instead, the use of *hyponyms* does not improve the results. Therefore, the proper use of *keywords* for the re-ranking process could be interesting in the future.

With respect to the experiments in which we have used the *fusion list*, the results obtained indicates that the query reformulation does not seem to work well in this field, although in some topics the Q_2 and Q_3 query types add valid documents to the final list which have not been found by the IR subsystem using the default query (Q_1).

6 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.

References

[1] 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.

[2] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. Geoclef 2007: the clef 2007 cross-language geographic information retrieval track overview. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, 2007.

[3] 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 *Lecture Notes in Computer Science*, volume 4730 of LNCS Series, pages 913–917. Springer-Verlag, 2007.

[4] José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and Arturo Montejo-Ráez. GEOUJA System. University of Jaén at GEOCLEF 2007. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, page 52, 2007.

[5] Zhisheng Li, Chong Wanga, Xing Xie, and Wei-Ying Ma. Query Parsing Task for GeoCLEF 2007 Report. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, 2007.

[6] 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 *Lecture Notes in Computer Science*, volume 4730 of LNCS Series, pages 328–338. Springer-Verlag, 2007.

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