MIRACLE's Ad-Hoc and Geographical IR Approaches for CLEF 2006

José M. Goñi-Menoyo[1], José C. González-Cristóbal[2], Sara Lana-Serrano[1], and Ángel Martínez-González[2]

[1] Universidad Politécnica de Madrid
[2] DAEDALUS – Data, Decisions, and Language S.A.
josemiguel.goni@upm.es, jgonzalez@dit.upm.es, slana@diatel.upm.es, amartinez@daedalus.es

Abstract. This paper presents the 2006 Miracle team's approaches to the Ad-Hoc and Geographical Information Retrieval tasks. A first set of runs was obtained using a set of basic components. Then, by putting together special combinations of these runs, an extended set was obtained. With respect to previous campaigns some improvements have been introduced in our system: an entity recognition prototype is integrated in our tokenization scheme, and the performance of our indexing and retrieval engine has been improved. For GeoCLEF, we tested retrieving using geo-entity and textual references separately, and then combining them with different approaches.

1 Introduction

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is our fourth participation in CLEF since 2003. As well as bilingual, monolingual and robust multilingual tasks, the team has participated in the ImageCLEF, Q&A, and GeoCLEF tracks.

For this campaign, runs were submitted for the following languages and tracks:

- Ad-hoc monolingual: Bulgarian, French, Hungarian, and Portuguese.
- Ad-hoc bilingual: English to Bulgarian, French, Hungarian, and Portuguese; Spanish to French and Portuguese; and French to Portuguese.
- Ad-hoc robust monolingual: German, English, Spanish, French, Italian, and Dutch.
- Ad-hoc robust bilingual: English to German, Italian to Spanish, and French to Dutch.
- Ad-hoc robust multilingual: English to robust monolingual languages.
- Geo monolingual: English, German, and Spanish.

2 MIRACLE Toolbox

All document collections and topic files are processed before feeding the indexing and retrieval engine. This processing is carried out using different combinations of

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 877 – 880, 2007.

elementary components. The details can be consulted in the papers from 2006 workshop [2], [3].

These components include (i) extraction, which incorporates a special process to filter out some sentences from the topic narrative that matches a number of recurrent and misleading patterns ; (ii) tokenization, which extracts basic text components, such as single words, years, or entities; (iii) entity detection, integrated in the tokenization module, and having a central role in IR processes: for now, it detects previously collected entities and integrates them into a special resource; (iv) filtering, which eliminates stopwords and words without semantic content in the CLEF context; (v) transformation, which normalizes case and diacritics; (vi) stemming [4], [6]; (vii) indexing, using our own trie-based [1] tool; and (viii) retrieval engine, which implements the well-known Robertson's Okapi [5] BM-25 formula for probabilistic retrieval model, without relevance feedback.

After retrieval, a number of other special combination processes were used to define additional experiments. The results from several basic experiments are combined using two different strategies: average and asymmetric WDX combination (see [2]). The underlying hypothesis for these combinations is that highly scored documents in several experiments are more likely to be relevant than other documents that have good scores in some experiments but bad ones in others.

We used the traditional approach to multilingual information retrieval that translates topic queries into the target language of the document collections. The probabilistic BM25 approach used for monolingual retrieval gives relevance measures that depends heavily on parameters that are too dependent on the monolingual collection, so it is not very good for this type of multilingual merging, since relevance measures are not comparable between collections. In spite of this, we carried out merging experiments using the relevance figures obtained from each monolingual retrieval process, considering three cases (see [2]): (i) using original relevance measures for each document as obtained from the monolingual retrieval process; (ii) normalizing relevance measures with respect to the maximum relevance measure obtained for each topic query *i* (*standard normalization)*; and (iii) normalizing relevance measures with respect to the maximum and minimum relevance measure obtained for each topic query *i (alternate normalization)*. In the three cases, documents with higher resulting relevance are selected from all monolingual results lists. Round-Robin merging for results of each monolingual collection has not been used.

In addition to all this, we tried a different approach to multilingual merging: Considering that the more relevant documents for each of the topics are usually the first ones in the results list, we select a variable number of documents, proportional to the average relevance number of the first N documents from each monolingual results file. Thus, if we need 1,000 documents for a given topic query, we get more documents from languages where the average relevance of the first N relevant documents is greater. We implemented this process in two ways: The appropriate number of documents to be aggregated is computed from both non normalized and normalized runs (using the two normalization formulae).

For Geographical IR, a Gazetteer that drives a named Geo-entity identifier was built as well as a tagger [3]. The gazetteer is a list of geographical resources,

compiled from two existing gazetteers, GNIS and NGA, which required the development of a geographical ontology.

3 Ad-Hoc Experiments

Both in the monolingual and the bilingual cases, the results obtained for "related" languages, such as French and Portuguese, are better than those obtained for Bulgarian and Hungarian. In the bilingual case, French experiments have best average precision.

In all cases, the best results are obtained from the experiments that take the topic narrative into account. Unfortunately, the official published reports only consider experiments that use topic title and description exclusively.

In the robust monolingual case, results for Spanish are much better than those obtained for the other the languages. In all cases the use of baseline runs has obtained better results than the use of combined ones. Curiously, the Dutch target language runs have better results than those in other languages. Note that in all cases, the experiments taking the topic narrative into account show the best results, as happened in the non-robust case.

4 Geographical Retrieval Experiments

The main objective or us in this campaign was to test the effects of geographical IR in documents containing geographical tags. We designed experiments to try to isolate geographical retrieval from textual retrieval. We have replaced all geo-entity textual references with associated tags in each topic, and then we searched all documents for these tags. This is done sequentially by combining a Geo-query Identifier process, a Spatial Relation identifier, and an Expander. These results are combined with those obtained using the usual ad-hoc text retrieval process. For combinations, several techniques were used: union (OR), intersection (AND), difference (AND NOT), and external join (LEFT JOIN). These techniques re-rank the output results by computing new relevance measure values for each document.

5 Conclusions

We still need to work harder to improve a number of aspects of our processing scheme, entity recognition and normalization being the most important ones. It is clear that the quality of the tokenization step is of paramount importance for precise document processing. A high-quality entity recognition (proper nouns or acronyms for people, companies, countries, locations, and so on) could improve the precision and recall figures of the overall retrieval, as well as a correct recognition and normalization of dates, times, numbers, etc. Although we have introduced some improvements to our processing scheme, a good multilingual entity recognition and normalization tool is still missing both for Ad-hoc and Geo IR.

We are also improving the architecture of our indexing and retrieval *trie*-based engine in order to get an even better performance in the indexing and retrieval phases,

tuning some data structures and algorithms. We are now implementing pseudo-relevance feedback and document filtering modules.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, through the RIMMEL project (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by Madrid's R+D Regional Plan, through the MAVIR project (Enhancing the Access and the Visibility of Networked Multilingual Information for Madrid Community), S-0505/TIC/000267.

Special mention of our colleagues from the MIRACLE team should be made (in alphabetical order): Ana María García-Serrano, Ana González-Ledesma, José Mª Guirao-Miras, José Luis Martínez-Fernández, Paloma Martínez-Fernández, Antonio Moreno-Sandoval and César de Pablo-Sánchez.

References

1. Aoe, J.-I., Morimoto, K., Sato, T.: An Efficient Implementation of Trie Structures. Software Practice and Experience 22(9), 695–721 (1992)
2. Goñi-Menoyo, J.M., González, J.C., Villena-Román, J.: Report of MIRACLE Team for the Ad-Hoc Track in CLEF. In: Working Notes for the CLEF 2006 Workshop. Alicante, Spain (2006) http://www.clef-camp aign.org/
3. Lana-Serrano, S., Goñi-Menoyo, J.M., González, J.C.: Report of MIRACLE Team for Geographical IR in CLEF, Working Notes for the CLEF 2006 Workshop. Alicante, Spain, (2006), http://www.clef-campaign.org/
4. Porter, M.: Snowball stemmers and resources page. [Visited 30/09/2006] http://www.snowball. tartarus.org
5. Robertson, S.E., et al.: Okapi at TREC-3. In: Harman, D.K. (ed.) Overview of the Third Text REtrieval Conference (TREC-3), Gaithersburg, MD: NIST (1995)
6. University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers..). [Visited 30/09/2006] On line http://www.unine.ch/info/clef