MIRACLE Progress in Monolingual Information Retrieval at Ad-Hoc CLEF 2007

José-Carlos González-Cristóbal[1,3], José Miguel Goñi-Menoyo[1], Julio Villena-Román[2,3], and Sara Lana-Serrano[1,3]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.
josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es
jvillena@daedalus.es, slana@diatel.upm.es

Abstract. This paper presents the 2007 MIRACLE's team approach to the Ad-Hoc Information Retrieval track. The main work carried out for this campaign has been around monolingual experiments, in the standard and in the robust tracks. The most important contributions have been the general introduction of automatic named-entities extraction and the use of wikipedia resources. For the 2007 campaign, runs were submitted for the following languages and tracks: a) Monolingual: Bulgarian, Hungarian, and Czech. b) Robust monolingual: French, English and Portuguese.

1 Introduction

The MIRACLE[1] Information Retrieval toolbox is made of basic components in a classical pipeline architecture: stemming, transformation (transliteration, elimination of diacritics and conversion to lowercase), filtering (elimination of stop and frequent words), proper nouns detection and extracting, and paragraph extracting, among others. Some of these basic components can be used in different combinations and order of application for document indexing and for query processing. Standard stemmers were used from Porter [8] for English, and from Neuchatel [11] for Hungarian, Bulgarian and Czech. In the 2007 experiments, only OR combinations of the search terms were used. The retrieval model used is the well-known Robertson's Okapi BM-25 [9] formula for the probabilistic retrieval model, without relevance feedback. Through our participation in previous campaigns, the integration procedure of the different modules is stable and, to some point, optimized. MIRACLE toolbox has already been described in previous campaigns papers [2], [3], [7].

MIRACLE makes use of its own indexing and retrieval engine, which is based on the trie data structure [1]. Tries have been successfully used by the MIRACLE team for years, as an efficient storage and retrieval of huge lexical resources, combined

[1] 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 fifth participation in CLEF.

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 156–159, 2008.

with a continuation-based approach to morphological treatment [6]. For the 2007 campaign, runs were submitted for the following languages and tracks:

- Monolingual: Bulgarian, Hungarian, and Czech.
- Robust monolingual: French, English and Portuguese.

The most relevant work carried out in this campaign was the incorporation of modules for the recognition of named entities in the tokenizing process, besides the compiling of extended resources adequate for this task.

2 Results for the Monolingual and Robust Tasks

The following table and figure summarize the performance of our official experiments in the monolingual tasks (using the topic fields title/description).

The most relevant work carried out for the 2007 campaign was the integration of components for multilingual Named Entities Recognition. In particular, entities from Wikipedia were extracted for all languages of interest in the framework of CLEF. In

Table 1. Average precision for monolingual experiments

lang	Average Precision	Prec. at 0	Prec. At 100
BG	0.2717	0.5946	0.0531
CZ	0.3203	0.6697	0.0701
HU	0.3499	0.7672	0.987

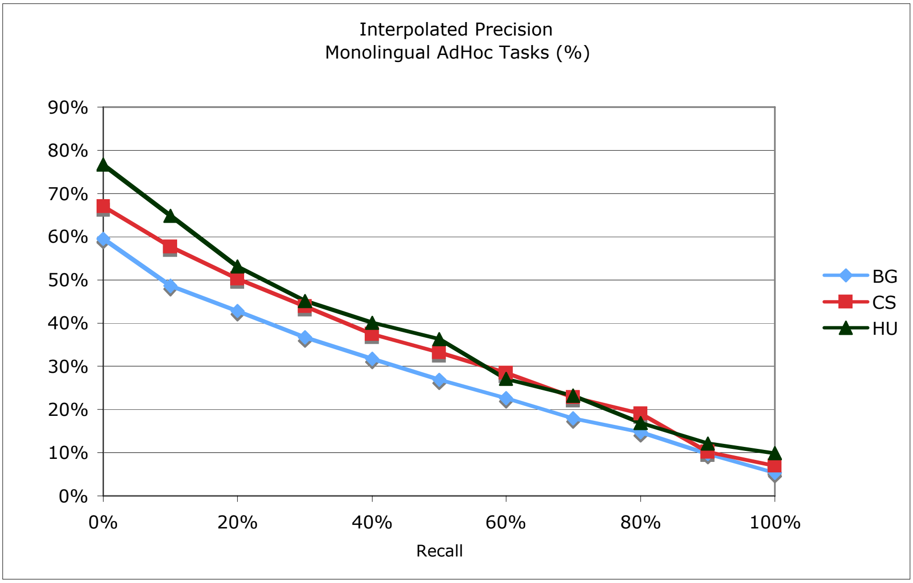

Fig. 1. Interpolated precision for monolingual experiments

the case of English, the number of entities used was above 500,000. An additional improvement was made through normalization of the recognized entities. Under this approach, the terms *United Nations, UN, U.N.* and *U. N.* were automatically substituted by an identifier associated with this international organization. For evaluation purposes, one baseline system was implemented that applied a simple Porter stemmer with lowercase reduction. The results of these experiments were fully available after the end of the campaign, and are shown in Table 2.

The results are discouraging, showing no improvement associated with the usage of these techniques, although a detailed analysis shows that the number of correctly retrieved texts is slightly higher.

Table 2. Precision figures for robust monolingual experiments in English

Run	Average Precision	Prec. at 0	Prec. at 1
Simple stemming	0.3966	0.6457	0.1688
Wikipedia-based named-entity recognition	0.3892	0.6398	0.1622
Named-entity recognition + Normalization	0.3920	0.6428	0.1658

3 Conclusions and Future Work

For the 2007 campaign, the processing scheme was maintained from previous ones, starting some improvements regarding proper nouns and entities detection and indexing. The results presented here indicate that Named Entity Recognition techniques have no impact on the TREC-based precision measures used for CLEF experiments. Although new experiments have to be conducted, it seems obvious that stemming provides a simple, fast and robust way for information retrieval of English texts. Further work includes extending this approach for languages other than English, integrated with other sets of external resources apart from Wikipedia, or through automatic learning of entities from the collections.

We still think that a high-quality entity recognition (proper nouns or acronyms for people, companies, countries, locations, and so on) can improve the precision and recall figures in some information retrieval tasks, as well as a correct recognition and normalization of dates, times, numbers, etc. In particular, such techniques can reveal a higher impact on cross-lingual tasks.

Regarding Wikipedia-based resources, three main usages are foreseen. The first one is the identification of relevant multiword expressions. The second is the expansion of acronyms. The last one is the translation of expressions between languages, to be used in future bilingual tasks. For these specific tasks, the use of multilingual thesaurus (e.g. Eurovoc) will be also generalized, as the IR platform is ready to incorporate such resources.

Acknowledgements. This work has been partially supported by the Spanish R&D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by the Madrid's R&D Regional Plan, by means of the MAVIR project (Enhancing the

Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

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, J.M., González, J.C., Villena, J.: MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. In: Peters, C., et al. (eds.) Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria. LNCS, vol. 4022, pp. 44–53. Springer, Heidelberg (2006)
3. Goñi, J.M., González, J.C., Villena, J.: Miracle's 2005 Approach to Monolingual Information Retrieval. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria (2005)
4. Goñi, J.M., González, J.C., Martínez, J.L., Villena, J.: MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. In: Peters, C., Clough, P., Gonzalo, J., et al. (eds.) Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004. LNCS, vol. 3491, pp. 188–199. Springer, Heidelberg (2005)
5. Goñi, J.M., González, J.C., Martínez, J.L., Villena, J., García, A., Martínez, P., de Pablo, C., Alonso, J.: MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. In: Peters, C., Borri, F. (eds.) Working Notes for the CLEF 2004 Workshop, Bath, United Kingdom, pp. 141–150 (2004)
6. Goñi, J.M., González, J.C., Fombella, J.: An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid (2004)
7. González, J.C., Goñi, J.M., Villena, J.: Miracle's 2005 Approach to Cross-lingual Information Retrieval. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria (2005)
8. Porter, M.: Snowball stemmers and resources page, http://www.snowball.tartarus.org [Visited 18/07/2006]
9. Robertson, S.E., et al.: Okapi at TREC-3. In: Harman, D.K. (ed.) Overview of the Third Text REtrieval Conference (TREC-3), April 1995. NIST, Gaithersburg (1995)
10. Savoy, J.: Report on CLEF-2003 Multilingual Tracks. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 64–73. Springer, Heidelberg (2004)
11. University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers, etc.), http://www.unine.ch/info/clef [Visited 18/07/2006]