MIRACLE Retrieval Experiments with East Asian Languages

Julio Villena-Román[1,2] José Miguel Goñi-Menoyo[3]
José C. González-Cristóbal[1,3] José Luis Martínez-Fernández[1, 2]
[1]DAEDALUS, [2]Universidad Carlos III de Madrid, [3]Universidad Politécnica de Madrid
jvillena@daedalus.es, josemiguel.goni@upm.es
jgonzalez@dit.upm.es, jmartinez@daedalus.es

Abstract

This paper describes the participation of MIRACLE in NTCIR 2005 CLIR task. Although our group has a strong background and long expertise in Computational Linguistics and Information Retrieval applied to European languages and using Latin and Cyrillic alphabets, this was our first attempt on East Asian languages. Our main goal was to study the particularities and distinctive characteristics of Japanese, Chinese and Korean, specially focusing on the similarities and differences with European languages, and carry out research on CLIR tasks which include those languages. The basic idea behind our participation in NTCIR is to test if the same familiar linguistic-based techniques may also applicable to East Asian languages, and study the necessary adaptations.

Keywords: *MIRACLE, Asian languages, NTCIR, Cross-language Information Retrieval Task, linguistic approach, segmenter, precision.*

1 Introduction

MIRACLE (Multilingual Information Retrieval for the CLEF campaign) is a Spanish research group, made up of three public university research groups (UPM, UC3M and UAM) and DAEDALUS, a private company founded in 1998 as a spin-off of two of these groups which is now a leading company in linguistic technologies in Spain. MIRACLE was born specifically to participate in CLEF (Cross Language Evaluation Forum), the European homologue of NTCIR, in which we have taken part since 2003 and submitted experiments for all tasks, including bilingual, monolingual and cross lingual retrieval tasks [5][6] [11][16], image [10], web and geographic information retrieval, question answering [13] and interactive task.

East Asian languages have some factors and differential characteristics with respect to European languages which make them very appealing: a complex writing system made up of a mixture of scripts, the morphological structure which poses a hard challenge to perform accurate segmentation and conflation, lack of a standard orthography and/or the presence of numerous orthographic variants which force to use cross-orthographic searching, and other miscellaneous technical requirements such as transcoding between multiple character sets and encodings and support for Unicode and input method editors [8].

The main idea behind our participation in NTCIR, as a first and naive attempt to approach East Asian languages, is to compare the effectiveness of similar linguistic-based techniques to the ones which are applied to European languages with good results [16]. Those techniques include segmentation, stop and frequent word filtering, compound detection and named entities recognition (such as dates), proper noun extraction, word conflation or lemmatization, paragraph extraction, semantic expansion with synonyms, etc. Our secondary goal was to improve our basic processing and indexing tools, adapting them to new languages with different encoding schemes, and, specifically, Unicode support.

We submitted runs in the CLIR task for monolingual Japanese and Chinese, bilingual English, Chinese and Korean to Japanese and bilingual-pivot for Chinese and Korean to Japanese. In the following sections, we describe our approach and the system that was developed to carry out the experiments, comment about the evaluation results and, finally, present some conclusions and our future lines of work.

2 System Description

During 2004, our group had been working hard to improve an indexing system based on the trie data structure [1]. Although tries had been successfully used by MIRACLE team for years, as an efficient storage and retrieval of huge lexical resources, combined with a continuation-based approach to morphological treatment [7], the adaptation of these structures to efficiently manage

document indexing and retrieval for IR applications had been a hard task, mainly in the issues concerning the performance of the construction of the index and the execution of queries. Apart from performance, other key points of our new engine are that both the vector space model and the BM25 probabilistic model are implemented, and also has native Unicode support, which was essential in this case for the languages which are involved. The Xapian [17] retrieval engine which had been successfully used in previous CLEF editions was no more needed.

On the other hand, all the auxiliary modules to carry out the basic text-handling operations were available from previous participation in CLIR experiments. Those modules cover text extraction and XML handling, character transformation (transliteration, character encoding conversion, elimination of diacritics, conversion to lowercase – when applicable), filtering (elimination of stop and frequent words, detection of introductory sentences such as "give me documents about..."), paragraph extraction, etc. Apart from adding the necessary linguistic resources, only very simple modifications had to be made to the modules to adapt them to the new languages

For core linguistic-processing modules, mainly segmentation, lemmatization and translation, we resorted to publicly available tools. In the case of Japanese, Mecab [12] (in the first experiments) and finally Chasen [2] were used for segmentation and lemmatization and Kakasi [9] for script transformation. For Chinese, the segmenter [3] by Chinese Computing and MMSEG [15] were used for segmentation, and no conflation was done due to the lack of an appropriate tool. In the end, as we were not able to find a segmentation tool for Korean, we couldn't participate in monolingual Korean runs. For bilingual experiments, Excite [4] was used as the translation engine in all cases (CKE→J, CK→E→J).

For stop and frequent word filtering, two lists were built for Japanese and Chinese with the 1,000 most frequent words appearing in a random set of 5,000 documents in each collection, obtained after processing the text with the previously mentioned tools. Character encoding conversion was only applied when was required to run the tools (GB in Chinese segmenter).

Our system is designed with a modular architecture which allows to reuse and combine the different components to perform different experiments, with a similar idea to the Unix pipeline. Thus, each experiment is defined by the modules which are included and the order in which they are used for document indexing and for query processing.

3 Experiment Design

Japanese and Chinese document collections were preprocessed before indexing, with the following sequence of steps:

1. Text extraction: ad-hoc scripts select the contents of the desired XML elements of each news article, in this case, HEADLINE and TEXT. The resulting text is the concatenation of both elements, without further distinction to feed subsequent processing steps.
2. Segmentation and lemmatization
3. Keyword selection: lemmas for Japanese, words for Chinese
4. Frequent word filtering

When all the documents processed through a combination of the former steps are ready for indexing, the resulting keyword lists are fed into our trie-based indexing engine to build the document collection index.

For European languages, in which we have a broader range of tools, we usually define experiments which address different preprocessing strategies (various combinations of elementary processes, each one oriented to a particular characteristic of the language). In this case, we were limited by the available resources and know-how.

Topics are processed in the same way as documents, but, in this case, several experiments may be defined depending on the topic field that is selected for the query. Each topic has four fields: 'T' (title), 'D' (description), 'N' (narrative) and 'C' (keywords), which may be combined to design different experiments.

According to our experience in CLEF, it would have been very convenient to apply an additional filter in the extraction step for topics in the case of experiments using the narrative field: some patterns should be eliminated since they are recurrent and misleading in the retrieval process. For example, for English, we may mention patterns as "…are not relevant", "…are to be excluded" or "...will be regarded as irrelevant if...". Our experience shows [7] that retrieval precision improves when sentences that contain such patterns are filtered out. Unfortunately, no resources for developing such patterns for Japanese, Chinese and Korean were available at the moment of performing these experiments.

Finally, a wide set of 30 experiments was finally submitted to NTCIR official competition.

We designed 5 runs for each monolingual task (Japanese and Chinese, 10 runs in all), named after the selected topic fields:

- T-run and D-run (mandatory)
- DN-run (recommended)
- TD-run
- TDC-run

The same set of 5 runs was submitted for pure (not pivot) bilingual tasks, except for Korean were DN-run was not submitted (the reason is that the resulting text was so long that the Excite translation engine refused to translate it properly) –a total of 14 runs. For the bilingual-pivot task, only 2 runs for each language pair (6 in all) were submitted: T-run and D-run (the mandatory ones).

The resulting keyword list corresponding to each topic is fed to an ad-hoc front-end of the trie-based retrieval engine to search the formerly built document collection index. In all experiments, only OR combinations of the search keywords were used and the retrieval model which was finally selected was the well-known Robertson's Okapi BM-25 [14] formula for the probabilistic retrieval model, without relevance feedback.

4 Evaluation Results

Evaluation in the NTCIR CLIR task is based on a TREC-like procedure using results of relevance judgements of each pool of retrieved documents for topics. The trec_eval program is then used to score search results submitted by participants.

In addition, to allow another analysis dimension, two kinds of relevance judgements are provided by the task organizers: rigid judgements (which include documents judged as "highly relevant" and "relevant") and relaxed judgements (which also include "partially relevant" documents).

Next sections show the results of our experiments, with the evaluation measures provided by the task organizers. For each of the tasks, a table shows the run identifier along with the interpolated precision-values at 0 and 1 points of recall and the average precision. The results are sorted in average precision ascending order, but an asterisk marks all the best precision values for each column.

4.1 Monolingual runs

The results for the monolingual task in Japanese are shown in Table 1. The best results are obtained when using description and narrative fields from topics (DN-run). Also, according to this table, precision decreases as the length of the queries is lower (in general, the longest field is narrative and then description).

Table 1. Japanese monolingual runs

Relaxed relevance			
Run	**At0**	**At1**	**AvgP**
MIRAa-J-J-DN	**0.8835*	**0.0607*	**0.4573*
MIRAa-J-J-TDC	*0.8213*	*0.0672*	*0.4060*
MIRAa-J-J-TD	*0.7882*	*0.0486*	*0.3479*
MIRAa-J-J-T	*0.7524*	*0.0348*	*0.3062*
MIRAa-J-J-D	*0.7061*	*0.0347*	*0.2864*
Strict relevance			
Run	**At0**	**At1**	**AvgP**
MIRAa-J-J-DN	**0.7477*	**0.0703*	**0.3758*
MIRAa-J-J-TDC	*0.6705*	*0.0711*	*0.3179*
MIRAa-J-J-TD	*0.6279*	*0.0592*	*0.2649*
MIRAa-J-J-T	*0.5631*	*0.0455*	*0.2239*
MIRAa-J-J-D	*0.5691*	*0.0379*	*0.2135*

Table 2 compares these results with the results of the rest of participants in NTCIR. It is very encouraging for us to observe that our best run scored high above the average, in an intermediate position between the average and the best experiment, with a reduction in precision of only 15% (relaxed relevance).

Table 2. Overall results (J-J)

Run	**Relaxed AvgP**	**Rigid AvgP**
MIRACLE best	*0.4573*	*0.3758*
Average J-J	*0.3856*	*0.2991*
Min J-J	*0.1591*	*0.1164*
Max J-J	*0.5427*	*0.4480*

Table 3 shows the results for the monolingual task in Chinese. The same that happened with Japanese, the best results are obtained when using long queries (TCD and DN run).

Table 3. Chinese monolingual runs

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-C-TDC	*0.6917*	**0.0106*	**0.2694*
MIRAa-C-C-DN	**0.7587*	*0.0009*	*0.2626*
MIRAa-C-C-TD	*0.6113*	*0.0100*	*0.2175*
MIRAa-C-C-T	*0.5079*	*0.0039*	*0.1772*
MIRAa-C-C-D	*0.5484*	*0.0084*	*0.1582*

Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-C-DN	**0.6973*	*0.0113*	**0.2378*
MIRAa-C-C-TDC	*0.6115*	**0.0284*	*0.2274*
MIRAa-C-C-TD	*0.5302*	*0.0078*	*0.1876*
MIRAa-C-C-T	*0.4730*	*0.0031*	*0.1610*
MIRAa-C-C-D	*0.4494*	*0.0069*	*0.1307*

These results are compared in Table 4 with the results of the rest of participants in NTCIR. Our figures are clearly below the average, but, due to lack of time, we haven't still been able to deeply analyze the reasons for these low precision values, but we tend to think that there may a problem with the character encoding conversion or with the output of the segmenter.

Table 4. Overall results (C-C)

Run	Relaxed AvgP	Rigid AvgP
MIRACLE best	*0.2694*	*0.2378*
Average C-C	*0.3613*	*0.3090*
Min C-C	*0.0112*	*0.0060*
Max C-C	*0.5441*	*0.5047*

Monolingual experiments with Japanese and Chinese can be compared with similar monolingual experiments that were carried out by MIRACLE with the four European languages which were covered by CLEF 2005 tasks.

Table 5 includes the average precision values for several languages and shows that similar values for average precision (except for Chinese) are obtained in all cases.

Table 5. Comparison with other languages

Language	AvgP
Japanese	*0.3758*
Chinese	*0.2378*
Bulgarian	*0.2819*
Hungarian	*0.3536*
Portuguese	*0.3698*
French	*0.3921*

4.2 Bilingual runs

Table 6 shows the results of the Chinese to Japanese bilingual runs. Again, the best results were obtained in the experiments which made use of longest queries (TDC and DN runs).

Table 6. Chinese to Japanese bilingual runs

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-J-TDC	*0.7009*	**0.0259*	**0.2634*
MIRAa-C-J-DN	**0.7049*	*0.0205*	*0.2601*
MIRAa-C-J-TD	*0.5710*	*0.0162*	*0.2171*
MIRAa-C-J-D	*0.4976*	*0.0113*	*0.1883*
MIRAa-C-J-T	*0.5627*	*0.0155*	*0.1764*
Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-J-DN	**0.5694*	*0.0273*	**0.2068*
MIRAa-C-J-TDC	*0.5510*	**0.0325*	*0.2057*
MIRAa-C-J-TD	*0.4809*	*0.0216*	*0.1677*
MIRAa-C-J-D	*0.4117*	*0.0165*	*0.1388*
MIRAa-C-J-T	*0.4427*	*0.0143*	*0.1324*

Table 7 compares the experiments of the rest of participants. The results of our experiments are on the average, which, given that our monolingual results were far above the average, indicates that our election of the translation engine was not appropriate and the translations provided by that engine are not very good for this bilingual task, at least with the given topics.

Table 7. Overall results (C-J)

Run	Relaxed AvgP	Rigid AvgP
MIRACLE best	*0.2634*	*0.2068*
Average C-J	*0.2674*	*0.1995*
Min C-J	*0.1136*	*0.0816*
Max C-J	*0.3607*	*0.2747*

A comparison between the monolingual and bilingual domains shows that the decrease in precision in the bilingual scenario with respect to the monolingual one is about 40% (38% for the best groups and 42% in our case).

This turns to be a very important figure when compared with the case of European languages, in which the decrease between the monolingual and bilingual scenario is only about 15%. This may show that there is still a large space for improvement in automatic translation between Japanese and Chinese, or, in general, among East Asian languages, but this has to be studied with further detail.

The same analysis as before is presented for the English to Japanese bilingual runs, in the following tables (Table 8 and 9).

Table 8. English to Japanese bilingual runs

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-E-J-DN	**0.8067*	*0.0280*	**0.3601*
MIRAa-E-J-TDC	*0.7615*	**0.0381*	*0.3198*
MIRAa-E-J-TD	*0.7415*	*0.0302*	*0.2879*
MIRAa-E-J-T	*0.6583*	*0.0163*	*0.2497*
MIRAa-E-J-D	*0.6395*	*0.0244*	*0.2353*
Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-E-J-DN	**0.6845*	**0.0544*	**0.2973*
MIRAa-E-J-TDC	*0.6329*	*0.0503*	*0.2472*
MIRAa-E-J-TD	*0.5771*	*0.0436*	*0.2121*
MIRAa-E-J-T	*0.4964*	*0.0305*	*0.1783*
MIRAa-E-J-D	*0.5003*	*0.0324*	*0.1728*

Table 9. Overall results (E-J)

Run	Relaxed AvgP	Rigid AvgP
MIRACLE best	*0.3601*	*0.2973*
Average E-J	*0.3234*	*0.2394*
Min E-J	*0.1023*	*0.0784*
Max E-J	*0.4076*	*0.3139*

Our results are in an intermediate position 10% above the average and 10% away from the best runs, which is again very satisfactory for us. In this case, it seems that the selection of the translation engine was right, or, at least, not wrong, comparing with the other participants in this task.

Although the loss in precision with respect to the monolingual domain (30%) is lower than in the case of Chinese, the translation from English to Japanese still suffers from the same problem.

Last, but not least, the analysis for the bilingual Korean to Japanese runs is presented in the following tables (Table 10 and 11). In this case, as explained before, no TD-run was submitted. As it happened for Japanese and Chinese, the run which makes use of the longest queries offer the best results.

Table 10. Korean to Japanese bilingual runs

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-K-J-TDC	**0.7832*	**0.0434*	**0.3439*
MIRAa-K-J-TD	*0.6984*	*0.0317*	*0.3016*
MIRAa-K-J-T	*0.5665*	*0.0250*	*0.2456*
MIRAa-K-J-D	*0.5876*	*0.0236*	*0.2174*
Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-K-J-TDC	**0.6232*	**0.0462*	**0.2735*
MIRAa-K-J-TD	*0.5510*	*0.0332*	*0.2304*
MIRAa-K-J-T	*0.4421*	*0.0356*	*0.1995*
MIRAa-K-J-D	*0.4748*	*0.0189*	*0.1724*

Table 11. Overall results (K-J)

Run	Relaxed AvgP	Rigid AvgP
MIRACLE best	*0.3439*	*0.2735*
Average K-J	*0.2139*	*0.1635*
Min K-J	*0.0368*	*0.0338*
Max K-J	*0.4643*	*0.3795*

Again, our results are above the average and intermediate between the average and the best runs.

4.3 Bilingual-pivot runs

Finally, Table 12 and 13 show the results of the bilingual experiments from Chinese and Korean to Japanese, using English as the pivot language. Precision values, as expected, are much lower than the same values for the pure bilingual runs (-54% for Chinese and -67% for Korean).

Table 12. Bilingual-pivot runs (C-J)

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-J-Tb	*0.3569*	**0.0069*	**0.1202*
MIRAa-C-J-Db	**0.4013*	*0.0024*	*0.1136*
Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-C-J-Tb	*0.3100*	**0.0114*	**0.0868*
MIRAa-C-J-Db	**0.3285*	*0.0024*	*0.0816*

Table 12. Bilingual-pivot runs (K-J)

Relaxed relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-K-J-Tb	**0.3374*	**0.0076*	**0.0896*
MIRAa-K-J-Db	*0.2905*	*0.0022*	*0.0737*
Strict relevance			
RunId	**At0**	**At1**	**AvgP**
MIRAa-K-J-Tb	**0.2739*	**0.0110*	**0.0695*
MIRAa-K-J-Db	*0.2006*	*0.0041*	*0.0491*

No comparison with the other participants is possible as there is no data available.

5 Conclusions and Future Work

As this was our first participation in NTCIR, our main effort was mainly dedicated to study and learn the basics and distinctive characteristics of the languages involved and to build all the necessary linguistic infrastructure to be able to submit our experiments in time. Anyway, the obtained results are very satisfactory for us because they support our hard work and, at the same time, show a large space for improvement. Of course, starting from scratch is always difficult, so there are many aspects that we could not address in our experiments this year, due to evident limitations of computing resources, time and expertise.

However, our main conclusion is that East Asian languages are not as difficult as we thought before NTCIR, and they are not very far from our more familiar European languages in some aspects of Computational Linguistics. For instance, the rich morphology may be comparable to the one of Latin-derived languages such as Spanish or French, the orthographic variations may be similar to those in Italian, or even the character encoding difficulty is also present in Bulgarian.

Future work of the MIRACLE team in NTCIR tasks will be directed to address specifically several lines of research: (a) Tuning our trie-based indexing and retrieval engine in order to get even better performance in the indexing and retrieval phases, (b) improving the segmentation step, as we are aware that this is the key point when dealing with East Asian languages, (c) study other cross-lingual strategies, and (d) perform more complex experiments involving pseudo-relevance feedback and combined experiments.

A good entity recognition and normalization is still missing in our processing scheme. We need better performance of the retrieval system to drive runs that are efficient when the query has some hundred terms. This occurs when using pseudo-relevance feedback, in which, after a first retrieval step, the first N retrieved documents are used to get their M top-ranked indexing terms which then, with a given combination method, are fed back to a second retrieval step, which gives the final results.

After our experiments in CLEF 2005, we are also interested in more complex experiments, in which the results from some basic experiments are combined in different ways to improve precision. The underlying hypothesis is that, to some extent, the documents with a good score in many all experiments are more likely to be relevant than other documents that have a good score in a few experiments but a bad one in others. This kind of combination has offered promising results for Hungarian and Bulgarian but their impact in Japanese or Chinese still has to be studied.

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.

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

References

[1] Aoe, Jun-Ichi; Morimoto, Katsushi; Sato, Takashi. *An Efficient Implementation of Trie Structures.* Software Practice and Experience 22(9): 695-721, 1992.

[2] Chasen. On line http://chasen.naist.jp/hiki/ChaSen/ [Visited 14/8/2005]

[3] Erik Peterson. *Chinese segmenter.* On line http://www.mandarintools.com/segmenter [Visited 14/8/2005]

[4] Excite.co.jp. http://www.excite.co.jp/world/english [Visited 14/8/2005]

[5] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. *MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval.* CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005.

[6] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. *MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval.* Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[7] Goñi-Menoyo, José Miguel; González-Cristóbal, José Carlos and Fombella-Mourelle, Jorge. *An optimised trie index for natural language processing lexicons.* MIRACLE Technical Report. Universidad Politécnica de Madrid, 2004.

[8] Halpern, Jack. *The Challenges of Intelligent Japanese Searching.* The CJK Dictionary Institute, Inc. On line http://www.cjk.org/cjk/joa/joapaper.htm [Visited 15/09/2005].

[9] Kakasi. *Kanji Kana simple inverter.* On line http://kakasi.namazu.org/ [Visited 14/8/2005]

[10] Martínez-Fernández, José L.; García-Serrano, Ana; Villena, J. and Méndez-Sáez, V. *MIRACLE approach to ImageCLEF 2004: merging textual and*

content-based Image Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005.

[11] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. *MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research.* Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[12] Mecab. *Yet Another Part-of-Speech and Morphological Analyzer.* On line http://chasen.org/~taku/software/mecab/ [Visited 14/8/2005]

[13] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P.; Villena, J.; García-Serrano, A. M.; Goñi, J. M.; and González, J. C. *miraQA: Initial experiments in Question Answering.* CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005.

[14] Robertson, S.E. et al. *Okapi at TREC-3.* In Overview of the Third Text REtrieval Conference (TREC-3). D.K. Harman (Ed.). Gaithersburg, MD: NIST, April 1995.

[15] Tsai, Chih-Hao. *MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm.* On line http://www.geocities.com/hao510/mmseg/ [Visited 14/8/2005]

[16] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. *Image Retrieval: The MIRACLE Approach.* Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[17] *Xapian: an Open Source Probabilistic Information Retrieval library.* On line http://www.xapian.org [Visited 13/07/2005].