Carol Peters
Julio Gonzalo
Martin Braschler
Michael Kluck (Eds.)

LNCS 3237

Comparative Evaluation of Multilingual Information Access Systems

4th Workshop of the
Cross-Language Evaluation Forum, CLEF 2003
Trondheim, Norway, August 2003, Revised Papers

CLEF

Springer

Volume Editors

Carol Peters
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche (ISTI-CNR)
Via G. Moruzzi 1, 56124 Pisa, Italy
E-mail: carol.peters@isti.cnr.it

Julio Gonzalo
Universidad Nacional de Educación a Distancia (UNED)
Departamento de Lenguajes y Sistemas Informáticos
c/ Juan del Rosal, 16, 28040 Madrid, Spain
E-mail: julio@lsi.uned.es

Martin Braschler
Eurospider Information Technology AG
Schaffhauser Str. 18, 8006 Zürich, Switzerland
E-mail: martin.braschler@eurospider.com

Michael Kluck
Informationszentrum Sozialwissenschaften der
Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (IZ)
Lennéstr. 30, 53113 Bonn, Germany
E-mail: kluck@bonn.iz-soz.de

Library of Congress Control Number: 2004115726

CR Subject Classification (1998): H.3, I.2, H.4

ISSN 0302-9743
ISBN 3-540-24017-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11342052 06/3142 5 4 3 2 1 0

Table of Contents

SINAI at CLEF 2003: Decompounding and Merging

Fernando Martínez-Santiago[1], Arturo Montejo-Ráez[2], Luis Alfonso Ureña-López[1], and M. Carlos Díaz-Galiano[1]

[1] Dpto. Computer Science. University of Jaén. Avda. Madrid 35. 23071 Jaén, Spain
{dofer,laurena,mcdiaz}@ujaen.es
[2] Scientific Information Service, European Organization for Nuclear Research, Geneva, Switzerland
{Arturo.Montejo}@cern.ch

Abstract. This paper describes the application of the *two-step RSV* and *mixed two-step RSV* merging methods in the multilingual-4 and multilingual-8 tasks at CLEF 2003. We study the performance of these methods compared to previous studies and approaches. A new strategy for dealing with compound words which uses predefined vocabularies for automatic decomposition is also presented and evaluated.

1 Introduction

The aim for CLIR (Cross-Language Information Retrieval) systems is to retrieve a set of documents written in different languages in answer to a query in a given language. Several approaches exist for this task, such as translating the whole document collection into an intermediate language or translating the quesry into every language found in the collection.

Two architectures are known for query translation: centralized and distributed architectures [1]. A centralized architecture handles the document collections in different languages as a single collection, replacing the original query by the sum of translations in all possible languages found in the collection. In a distributed architecture, documents in different languages are indexed and retrieved separately. All ranked lists are then merged into a single multilingual ranked list.

We use a distributed architecture, focusing on a solution for the merging problem. Our merging strategy consists in calculating a new RSV (Retrieval Status Value) for each document in the ranked lists for each monolingual collection. The new RSV, called the two-step RSV, is calculated by re-indexing the retrieved documents according to a vocabulary generated from query translations, where words are aligned by meaning, i.e. each word is aligned with its translations [2].

The rest of the paper has been organized into three main sections: a brief review of merging strategies and the 2-step RSV approach, a description of the proposed decompounding algorithm and a description of our experiments. Finally, Section 5 provides some conclusions, and also outlines future research lines.

C. Peters et al. (Eds.): CLEF 2003, LNCS 3237, pp. 192–201, 2004.

2 Merging Strategies and the 2-Step RSV Approach

Distributed IR architectures require result merging in order to integrate the ranked lists returned by each database/language into a single, coherent ranked list. This task can be difficult because document rankings and scores produced by each language are based on different corpus statistics such as inverse document frequencies, and possibly also different representations and/or retrieval algorithms that usually cannot be compared directly.

2.1 Traditional Merging Strategies

There are various approaches to the merging of monolingual collections. In all cases, a large decrease in precision is generated in the process (depending on the collection, between 20% and 40%) [3]. Perhaps for this reason, CLIR systems based on document translation tend to obtain results noticeably better than system driven by query translation. Most popular approaches to merging using query translation are round-robin algorithms and computing normalized scores. Other approach is depicted in [4]: a single and multilingual index is obtained for the whole set of documents in every language, without any translation. Then, the user query is translated for each language present in the multilingual collection. A query for each translation is not generated but all the translations are concatenated making up a composite query. Finally, this composite query is used to search across the entire multilingual term index. The idea is coherent, but current results with this method are disappointing [5, 6].

Finally, learning-based algorithms are very interesting, but they require training data (relevance judgments) and this is not always available. Thus, Le Calvé and Savoy [7, 8] propose a merging approach based on logistic regression and Martínez-Santiago et al. [9] improve slightly regression logistic results by using LVQ neural networks.

2.2 2-Step RSV and Mixed 2-Step RSV

Last year we obtained good results at CLEF 2002 by using a new approach called 2-step RSV [2]. This method is based on the hypothesis that: given two documents, the score of both documents will be comparable whenever the document frequency is the same for each meaningful query term and its translations. By grouping together the document frequency for each term and its translations, we ensure the compliancy of the hypothesis.

The basic idea underlying 2-step RSV is straightforward: given a query term and its translations to the other languages in the document collection, the document frequencies are grouped together [2]. In this way, the method requires recalculating the document score by changing the document frequency for each query term. Given a query term, the new document frequency will be calculated by means of the sum of the monolingual document frequency of the term and its translations. Since re-indexing the whole multilingual collection could be computationally expensive, given a query only the documents retrieved for each monolingual collection are re-indexed. These two steps are as follows:

1. The document pre-selection phase consists in translating and searching the query on each monolingual collection, in the usual way for CLIR systems based on query translation. This phase produces two results:
   - The translation of each term from the original query to the other languages as a result of the translation process. In this way, we have *queries aligned at term level*.
   - A single multilingual collection of preselected documents as result of the union of typically the first 1000 retrieved documents for each language.
2. The re-indexing phase consists of re-indexing the retrieved multilingual collection, but considering solely the query vocabulary, by grouping together their document frequencies. The query is then executed against the new index. Thus for example, if we have two languages, Spanish and English, and the term "casa" is part of the original query and it is translated to "house" and "home", both terms represent exactly the same index token. Given a document, the term frequency will be calculated as usual, but the document frequency will be the sum of the document frequency of "casa", "house" and "home"[1].

Perhaps the strongest constraint for this method is that every query term must be aligned with its translations. But this information is not always available whether using machine translation (which produces translations at phrase level) or automatic query expansion techniques such as pseudo-relevance feedback.

As a way of dealing with partially aligned queries (i.e. queries with some terms not aligned), we propose three approaches which mix evidence from aligned and not aligned terms [10, 11]:

- Raw mixed 2-step RSV method: An straightforward and effective way to partially solve this problem is by taking non-aligned words into account locally, only as terms of a given monolingual collection. Thus, given a document, the weight of a non-aligned term is the initial weight calculated in the first step of the method.

  Thus, the score for a given document d_i will be calculated in a mixed way by means of the weight of local terms and global concepts present in the query:

  $$RSV_i' = \alpha \cdot RSV_i^{align} + (1-\alpha) \cdot RSV_i^{nonalign} \tag{1}$$

  where RSV_i^{align} is the score calculated by means of aligned terms, such as the original 2-step RSV method proposes, while $RSV_i^{nonalign}$ is calculated locally. Finally, α is a constant (usually fixed to $\alpha = 0.75$).
- Normalized mixed 2-step RSV method: Since the weights of the aligned and non-aligned words are not comparable, the proposal of a raw mixed 2-step RSV seems counterintuitive. AIn an attempt to make RSV_{align} and $RSV_{nonalign}$ comparable, we normalize those values:

[1] Actually, we subtract the number of documents where both "house" and "home" terms appear. Thus, given a document which contains both terms, we avoid counting the same document twice.

$$RSV'_i = \alpha \cdot \frac{RSV_i^{align} - \min(RSV^{align})}{\max(RSV^{align}) - \min(RSV^{align})} + (1-\alpha) \cdot \frac{RSV_i^{nonalign} - \min(RSV^{nonalign})}{\max(RSV^{nonalign}) - \min(RSV^{nonalign})} \quad (2)$$

- Mixed 2-Step RSV method and learning-based algorithms such as logistic regression or neural networks [9]. Training data must be available in order to fit the model. This a serious drawback, but this approach allows the integration of not only aligned and not aligned scores but also the original rank of the document.

3 Decompounding Algorithm

In some languages, such as Dutch, Finnish, German and Swedish, words are formed by the concatenation of others. These are the so-called *compound words* which, if untreated, may bias the performance of our multilingual system. In order to increase recall, compound words must be decompounded. Unfortunately there is no straightforward method for this due to the high number of possible decompositions exhibited by many compound words.

Chen [12] proposes an approach towards a maximal decomposition applied on German documents: decompositions with a minimal number of components and, in case of multiple options, the one with highest probability, are chosen. In this way, decompounding is performed with a minimal set of rules and a dictionary which must contain no compound words. Chen has applied this algorithm only to German corpora, so no data about its effectiveness on other languages is available. However, we find that applying decomposition to every compound word may not be desirable, since some of these words have a meaning which, when decomposed, is lost.

Hollink et al. [13] provide a review of compound words for Dutch, German and Swedish, giving the connectives used for compounding by each of these languages. They apply an existing recursive algorithm to find all possible decompositions, using a dictionary generated from the document collection. This study is very illustrative with respect to the decomposition of words, but lacks a proposal for selection.

The solution we have adopted is based mainly on the Chen approach, but preserves compound words in some cases and extends the algorithm to Dutch and Swedish. We establish three main rules as the core of our algorithm. First, the word is decompounded in all possible compositions as in [13]. Then, given a compound word cw formed by composites $w_1, w_2 ... w_n$, we select a decomposition by applying following rules:

1. **Rule 1.** We do not decompound if the probability of the compound word is higher than any of its composites.

   $P(cw) \leq P(w_1) \wedge P(cw) \leq P(w_2) \wedge ... \wedge P(cw) \leq P(w_n) \longrightarrow cw$ is returned

2. **Rule 2.** Shortest decomposition (that one with the lowest number of composites) is selected. For example, if we find that cw can be decomposed into two forms $w_1 + w_2$ or $w_3 + w_4 + w_5$ the first decomposition would be selected.
3. **Rule 3.** In case several decompositions have the same number of composites, that one with highest probability will be chosen. The probability of a composition is the same as proposed by Chen: the product of the probabilities of its composites:

$$P(w_1 + w_2 + ... + w_n) = P(w_1) \cdot P(w_2) \cdot ... \cdot P(w_n)$$

where the probability for a word w_i in a collection is

$$P(w_i) = \frac{tfc(w_i)}{\sum_{j=1}^{N} tfc(w_j)}$$

where $tfc(w_i)$ is the number of occurrences of word w_i in a collection whose dictionary contains N different words.

Table 1. Length of wordlist used by the decompounding algorithm

Language	Main word sources	Size
Dutch	CLEF data, spelling dictionary, Babylon	387735
Finnish	CLEF data, spelling dictionary	359117
German	CLEF data, spell.dictionary, Babylon, MORPHIX	657452
Swedish	CLEF data, spelling dictionary, Babylon	294151

4 Experiments and Results

We participated in the Multi-4 and Multi-8 tasks. Each collection was preprocessed as usual, using stopword lists and stemming algorithms available on the Web[2]. Stopword lists were increased with terms such as "retrieval", "documents", "relevant"... . Once the collections had been pre-processed, they were indexed with the Zprise IR system, using the OKAPI probabilistic model [14]. This OKAPI model was also used for the on-line re-indexing process required by the calculation of 2-step RSV.

The rest of this section describes our bilingual and multilingual experiments driven by query-translation with fully and partially aligned queries.

4.1 Translation Strategy and Bilingual Results

Our translation approach is very simple. We used Babylon[3] to translate English query terms. Since an English to Finnish dictionary is not available on the Babylon site, we used the *FinnPlace* online dictionary [4]. Both bilingual dictionaries

[2] http://www.unine.ch/info/clef

[3] Babylon is a Machine Readable Dictionary available at http://www.babylon.com

[4] available at http://www.tracetech.net/db.htm

may suggest more than one translation for the translation of each query term. In our experiments, we decided to take the first translation listed.

We retrieved documents using non-expanded and expanded queries (pseudo-relevance feedback, PRF). Non-expanded queries are fully aligned queries. By this we mean that a translation is obtained for each term in the query. Queries expanded by pseudo-relevance feedback are expanded with monolingual collection-depended words. Such words will usually not be aligned. The first type of queries was used when testing original 2-Step RSV. Mixed 2-Step RSV was tested by using the second type of queries.

Table 2 shows the bilingual precision obtained by means of both translation approaches. We have taken only *Title* and *Description* query fields into account.

Table 2. English and Bilingual experiments

	Avg. Prec. without PRF	Avg. Prec. with PRF
English → Dutch	0.251	0.310
English	0.464	0.453
English → Finnish	0.286	0.253
English → French	0.371	0.400
English → German	0.288	0.321
English → Italian	0.237	0.292
English → Spanish	0.310	0.348
English → Swedish	0.212	0.259

In this study, we adopted Robertson-Croft's approach to pseudo-relevance feedback (blind expansion) [15], where the system expands the original query with generally no more than 15 search keywords, extracted from the 10-best ranked documents.

4.2 Multilingual Results

The bilingual results list obtained were the starting point - the first step towards providing users with a single list of retrieved documents. In this section, we study the second step. Unfortunately, an implementation error damaged dramatically our own official runs based on the 2-Step RSV approach[5] In the following, we present the results of both official and corrected runs.

Our approach to merging combined several approaches: round-robin, raw scoring, normalized score and 2-step RSV approach. In addition, a theoretical

[5] The error was as follows: we use two indices per collection: Okapi index and term frequency (TF) index. The Okapi index was used by monolingual runs. The TF index was used by the second step of the 2-step RSV method: in order to re-weight the query terms, term-frequency statistics were obtained from the TF-index files. In some languages such as English, we made a mistake by taking the OKAPI-index files instead of the TF-index files.

optimal performance was calculated by using the procedure proposed in [12] (label "Optimal performance" in Table 3). This procedure computes the optimal performance that could be achieved by a CLIR System merging bilingual and monolingual results, under the constraint that the relative ranking of the documents in the individual ranked list is preserved. In this procedure, the relevance of documents must be known a-priori. Thus it is not useful to predict ranks of documents in the multilingual list of documents, but it gives the upper-bound performance for a set of ranked lists of documents, and this information is useful to measure the performance of different merging strategies. Note that 2-step RSV calculus does not guarantee the preservation of the relative ranking of documents, theoretically the upper-bound performance calculated by this procedure could be surpassed. A detailed description of the algorithm is available in [12].

Table 3. Multi-4 experiments with fully and partially aligned queries

	Avg. Prec. without PRF	Avg. Prec. with PRF
round-Robin	0.216	0.245
raw scoring	0.269	0.294
normalized scoring	0.232	0.283
2-step RSV (official)	0.1724	-
raw mixed 2-step RSV (official)	-	0.211
2-step RSV (fixed)	**0.291**	-
raw mixed 2-step RSV (fixed)	-	**0.335**
norm. mixed 2-step RSV (fixed)	-	0.315
optimal performance	*0.331*	*0.371*

Table 4. Multi-8 experiments with fully and partially aligned queries

	Avg. Prec. without PRF	Avg. Prec. with PRF
round-Robin	0.160	0.1815
raw scoring	0.213	0.239
2-step RSV (official)	0.1423	-
raw mixed 2-step RSV (official)	-	0.168
2-step RSV (fixed)	**0.242**	-
raw mixed 2-step RSV (fixed)	-	**0.296**
norm. mixed 2-step RSV (fixed)	-	0.266
optimal performance	*0.285*	*0.350*

The proposed 2-step RSV merging approach achieves a better performance than any of the other approaches. Raw mixed 2-step RSV and normalized mixed 2-step RSV were calculated by means of eq. 1 and eq. 2, with $\alpha = 0.75$. Mixed 2-step results using logistic regression and neural networks are not given in this paper because training data (relevance judgments) for this years new collections are not available.

The good performance of raw-mixed 2-step RSV is counterintuitive. However, not all of the terms to be added to the original query are new terms since some terms obtained by means of pseudo-relevance feedback are already in the initial query. On the other hand, as Table 3 shows, raw-scoring works relatively well for this experiment. Thus, the percent (0.25) of local RSV added to each document score is partially comparable. However, normalized mixed 2-step RSV should improve raw mixed 2-step RSV results when collections are very different in size or very different weighting schemas are used for each collection. Finally, experiments carried out with CLEF 2001 (training) and CLEF 2002 (evaluation) relevance judgments show that learning-based algorithms perform slightly better than raw-scoring as a way to integrate both available values when mixed 2-step is used [11]. In any case, the mixing of both the local and global scores obtained for each document by means of mixed 2-step RSV is an open problem with respect to the integration of several sources of information, and again refers to the collection fusion problem.

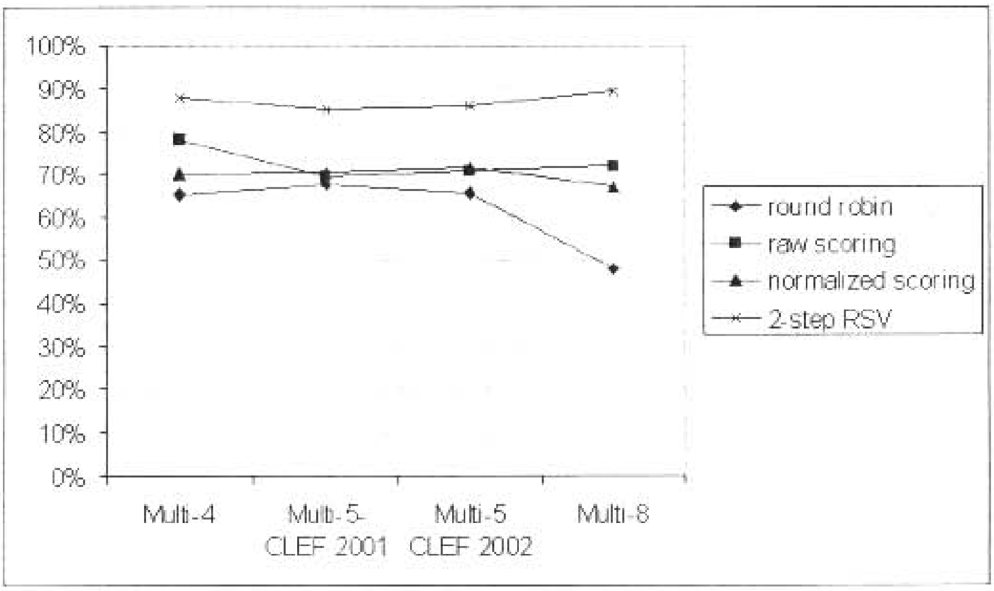

Fig. 1. Performance of traditional merging strategies with respect to several sets of languages (fully aligned queries). The 100% case represents *optimal performance*.

Perhaps our most interesting result this uear is shown in Figures 1 and 2. As we suspected last year, the perfomance of round-robin and raw-scoring decreases as the number of languages increases. On the other hand, 2-step RSV maintains about 85% of optimal performance.

5 Conclusion and Future Work

At CLEF 2003 we focused on merging approaches and decompounding algorithms. We have tested 2-step RSV and mixed 2-step RSV in the Multi-4 and

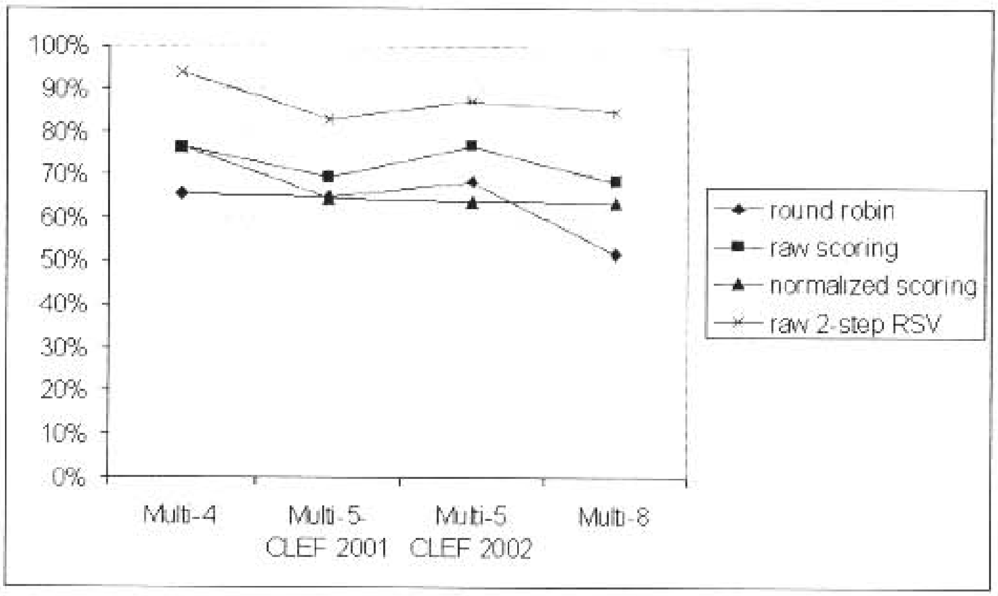

Fig. 2. Performance of traditional merging strategies with respect to several sets of languages (partially aligned queries using PRF). The 100% case represents *optimal performance*

Multi-8 tasks. Results show that the proposed method scales well with four, five and eight languages, overcoming traditional approaches.

Our next efforts will be aimed in a number of directions:

- Since our decompounding algorithm is highly dependent on the wordlists used, we intend to obtain a better wordlist.
  We mean to test the method described here using other translation strategies such as Machine Translation or Multilingual Similarity Thesaurus.
- The index terms used in the experiments reported here are basically obtained by means of stemming. We are very interested in the application of an n-gram indexing approach. However, while stemming terms are directly assimilable as feasible representations of concepts, n-grams cannot be assimilated directly as concepts since a given n-gram is usually contained by several unrelated terms. We have carried out some preliminary experiments, and the results obtained so far confirm that an n-gram cannot function as a direct representation of a concept.
- Finally, we will keep on studying strategies in order to deal with aligned and not-aligned query terms. The integration of both types of terms by means of neural networks (although these structures require training data) and the development of global pseudo-relevance feedback procedures, and not locally for each monolingual collection, should be interesting areas for investigation.

Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant FIT-150500-2003-412.

References

1. Chen, A.: Multilingual Information Retrieval using English and Chinese Queries. In Peters, C., ed.: Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2406, Springer Verlag (2002) 44–58
2. Martínez-Santiago, F., Martín, M., Ureña, L.: SINAI at CLEF 2002: Experiments with merging strategies. In Peters, C., ed.: Proceedings of the CLEF 2002 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2785, Springer Verlag (2003) 187–198
3. Savoy, J.: Report on CLEF-2001 Experiments. In Peters, C., ed.: Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2406, Springer Verlag (2003) 27–43
4. Gey, F., Jiang, H., Chen, A., Larson, R.: Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In Voorhees, E.M., Harman, D.K., eds.: Proceedings of the Seventh Text REtrieval Conference (TREC-7). (2000) 527–540
5. Nie, J., Jin, F.: Merging Different Languages in a Single Document Collection. In Peters, C., ed.: Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2406., Springer Verlag (2003) 59–62
6. McNamee, P., Mayfield, J.: JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In Peters, C., ed.: Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2406., Springer-Verlag (2002) 193–208
7. Calvé, L., A., Savoy, J.: Database merging strategy based on logistic regression. Information Processing & Management. 36 (2000) 341–359
8. Savoy, J.: Cross-language information retrieval: experiments based on CLEF 2000 corpora. Information Processing & Management **39** (2003) 75–115
9. Martín, M., Martínez-Santiago, F., Ureña, L.: Aprendizaje neuronal aplicado a la fusión de colecciones multilingües en CLIR. Procesamiento del Lenguaje Natural (2003) In press.
10. Martínez-Santiago, F., Ureña, L.: SINAI experience at CLEF. Revista Iberoamericana de Inteligencia Artificial (2003) In press.
11. Martínez-Santiago, F., Ureña, L.: A merging strategy proposal: the 2-step retrieval status value method. Technical report, University of Jaén (2003)
12. Chen, A.: Cross-language Retrieval Experiments at CLEF-2002. In Peters, C., ed.: Proceedings of the CLEF 2002 Cross-Language Text Retrieval System Evaluation Campaign. LNCS 2785. (2003) 5–20
13. V. Hollink, J. Kamps, C.M.M.d.R.: Monolingual retrieval for European languages. Information Processing Retrieval **7(1-2)** (2004)
14. Robertson, S.E., Walker., S., Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. Information Processing and Management 36(1) **1** (2000) 95–108
15. Harman, D.: Relevance feedback revisited. In: Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92). (1992) 1–10