SINAI at CLEF 2003: decompounding and merging

Fernando Martínez-Santiago
Department of Computer Science. University of Jaén, Jaén, Spain
dofer@ujaen.es

Arturo Montejo-Ráez
Scientific Information Service, European Organization for Nuclear Research, Geneva, Switzerland
Arturo.Montejo@cern.ch

L. Alfonso Ureña-López, Manuel Carlos Díaz-Galiano
Department of Computer Science. University of Jaén, Jaén, Spain
{*laurena,mcdiaz*}@*ujaen.es*

Abstract

This paper describes the application of the *two-step RSV* and *mixed two-step RSV* merging methods over 8 and 4 multilingual tasks in CLEF 2003. We study their performance compared to previous studies and approaches. Furthermore, a new strategy for dealing with compound words is presented and evaluated within our methods, allowing automatic decomposition by using predefined vocabularies.

1 Introduction

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

For query translation two architectures are known: centralized and distributed architectures [1]. Centralized architecture handles document collection in different languages as a single document collection, replacing the original query by the sum of translations in all possible languages found in collection. In the distributed architecture, documents in different languages are indexed and retrieved separately. Later on, all ranked lists are 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 of the ranked lists at every monolingual list. The new RSV, called two-step RSV, is calculated reindexing 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 revision of merging strategies and the 2-step RSV approach, a description of the proposed decompounding algorithm and a description of ours experiments. Finally, section 5 outlines some conclusions, and also future research lines.

2 Merging strategies and 2-step RSV approach

IR distributed architectures require result merging: 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 statics such as inverse document frequencies, and may be different representations and/or retrieval algorithms that usually cannot be compared directly.

2.1 Traditional merging strategies

There are various approaches in order to carry out the merging of monolingual collections, anyhow a large decrease of 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 using query translation are round-robin algorithms and computing normalized scores.
Other approach is depicted in [6]: a single and multilingual index is obtained with the whole of documents of 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 will be searched across the entire multilingual term index. The idea is coherent, but current researches with this method are disappointing[7, 8].
Finally, learning-based algorithms are very interesting, but they requires learning data (relevance judgments) and it is not always available. Thus, Le Calvé and Savoy [9, 10] propose a merging approach based on logistic regression and Martínez-Santiago et al.[11] improve slightly regression logistic results by using LVQ neural networks.

2.2 2-step RSV and mixed 2-step RSV

Last year we obtain good results by using a new approach called 2-step RSV [2]. The hypothesis of this method is as follows: given two documents, the score of both documents will be comparable whenever the document frequency is the same for each meaningful term query and their translations. By grouping together the document frequency for each term and its own translations, we ensure the hypothesis compliance.

The basic 2-step RSV idea is straightforward: given a query term and their translations to the rest of languages, their document frequencies are grouping together [2]. In this way, the method requires recalculating the document score by changing the document frequency of 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 their translations. Since reindexing the whole multilingual collection could be computationally expensive, given a query only the retrieved documents for each monolingual collection are re-indexed. These two steps are:

1. The document pre-selection phase consists in translating and searching the query on each monolingual collection as usual in CLIR systems based on query translation. This phase produces two results:
    - The translation to the rest of languages for each term from the original query as 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 1000 first retrieved documents for each language
2. The re-indexing phase consists of re-indexing the multilingual retrieved collection, but considering solely the query vocabulary, by grouping together their document frequencies.
   Finally, the query is 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].

[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.

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 neither by machine translation (which produces translations at phrase level) nor by automatic query expansion techniques such as pseudo-relevance feedback.

As a way to deal with partially aligned queries (i.e. queries with some terms not aligned), we propose three approaches by mixing evidence from aligned and not aligned terms [12, 13]:

- 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, just 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 original 2-step RSV method depicts. In the other hand, $RSV_i^{nonalign}$ 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 idea for the raw mixed 2-step RSV seems counterintuitive. As an attempt to make RSV_{align} and $RSV_{nonalign}$ comparable, we normalize those values:

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

- mixed 2-Step RSV method and learning-based algorithms such as logistic regression or neural networks [14]. Training data must be available in order to fit the model. This a serious drawback, but this approach allows integrating not only aligned and not aligned scores but also the original rank of the document.

3 Decompounding algorithm

In some languages like Dutch, Finnish, German and Swedish there are words formed up by 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 the recall, compound words must be decompounded. Unfortunally there is no straighforward method to do so, due to high number of possible decompositions exhibited by many compound words.

Chen [1] 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, decomposition is performed with a minimal set of rules and a dictionary which must contain no compound words. Chen has applied this algorithm only on German corpora, so no data about its effectiveness on other languages is available. Also we find that applying decomposition on every compound word may not be desirable, since some of these words have a meaning which, when decomposed, is lost.

Hollink et al. [15] provide a review on compound words for Dutch, German and Swedish, giving the connectives used for compositioning by each of these languages. They apply an existing recursive algorithm for finding all possible decompositions using a dictionary generated from the collection of documents. This work is very illustrative for decomposition of words, but lacks of a proposal for selection.

Our adopted solution is based mainly on Chen approach, but preserving compound word in some cases and extending the algorithm to Dutch and Swedish. We stablish three main rules as core of the algorithm. First, the word is decomposed to all possible compositions as done by Hollink et al. Then, given a compound word cw formed from composites $w_1, w_2 ... w_n$ we select a decomposition by applying following rules:

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

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

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

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

   being $tfc(w_i)$ the number of ocurrences 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, spell dictionary, Babylon	387735
Finnish	CLEF data, spell dictionary	359117
German	CLEF data, spell dictionary, Babylon, MORPHIX	657452
Swedish	CLEF data, spell dictionary, Babylon	294151

4 Experiments and results

We have participated on 4-Multi and 8-Multi tasks. Every collection has been pre-processed as usual, using stopword lists and stemming algorithms available across the Web[2]. Stopword lists have been increased with terms such as “retrieval”, “documents”, “relevant”.... Once the collections have been pre-processed, they are indexed with the Zprise IR system, using the OKAPI probabilistic model[16]. This OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV.

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

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

4.1 Translation strategy and Bilingual Results

The translation approach is very simple. We have used Babylon[3] to translate English query terms. Since English to Finnish dictionary is not available in Babylon site, we use *FinnPlace* online dictionary [4]. Both bilingual dictionary may suggest not only one, but several terms for the translation of each word. In our experiments, we decide to pick the first translation available.
In addition, we have retrieved documents by using non-expanded and expanded queries (pseudo-relevance feedback, PRF). Non-expanded queries are fully aligned queries. Queries expanded by pseudo-relevance feedback are expanded with monolingüal collection-depended words. Usually, such words will be not aligned. The first group of queries is used by testing original 2-Step RSV. Mixed 2-Step RSV is tested by considering second group of queries.

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

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

The expansion queries were carried out by means of pseudo-relevance feedback (blind expansion). In this study, we adopted Robertson-Croft's approach[17] where the system expands the original query generally no more than 15 search keywords, extracted from the 10-best ranked documents.

4.2 Multilingual results

The obtained bilingual results list are the starting point, the first step in order to provide users with a single list of retrieved documents. In this section, we study the second step. Suddenly, an implementation error has damaged dramatically over own official runs based en 2-Step RSV approach [5].We have decided to include both official and fixed runs.
The merging approach has been made up by using several approaches: round-robin, raw scoring, normalized score and 2-step RSV approach. In addition, theoretical optimal performance has been calculated by using the procedure proposed in [1] (label "Optimal performance" in table 4) . Such procedure computes the optimal performance that could possibly be achieved by a CLIR System by merging bilingual and monolingual results, under the constraint that the relative ranking of the documents in the individual ranked list is preserved. The relevances of documents must be known previously. Thus it is not useful to predict ranks of documents in the multilingual list of documents. Anyhow, the procedure obtains the upper-bound performance for a set of ranked list of document, and this information is useful to measure the performance of several merging strategies. Note that 2-step RSV calculus does not ensure the preservation of the relative ranking of documents, the upper-bound performance calculated by such procedure could be overcame, at least theoretically. The detailed description of the algorithm is available in[1].

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

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

[5] The error was as follows: we use two indices per collection: Okapi index and term frequency index. Okapi index is used by monolingual runs. TF index is used by the second step of 2-step RSV method: since re-weighting query terms is required, such re-weighting process get term-frequency statistics from TF-index files. In some languages such as English, we make a mistake by taking into account OKAPI-index files instead of TF-index files.

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	0.223	0.249
2-step RSV(official)	0.1423	-
raw mixed 2-step RSV(official)	-	0.168
2-step RSV	**0.242**	-
raw mixed 2-step RSV	-	**0.287**
norm. mixed 2-step RSV	-	0.266
optimal performance	*0.285*	*0.350*

The proposal 2-step RSV merging approach improves the whole of the rest of approaches. Raw mixed 2-step RSV and normalized mixed 2-step RSV have been calculated by means of eq. 1 and eq. 2, with $\alpha = 0.75$. Mixed 2-step by means of logistic regression and neural networks are not available in this work because training data(relevance judgments) for the new collections of this year is not available.

The good performance of raw-mixed 2-step RSV is counterintuitive. Nevertheless , not the whole of terms to be added to the original query are new terms since some terms obtained by means of pseudo-relevance feedback are in the initial query. In the other hand, as table 4 shows, raw-scoring works relatively fine 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 whether collections are very irregular 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 overcome slightly raw-scoring as a way to integrate both available values when mixed 2-step is used[14]. Anyway, the mixing of both local and global score obtained for each document by means of mixed 2-step RSV is an open problem about the integration of several sources of information, and it remembers to the same collection fusion problem.

Maybe the most interesting issue obtained for us this year is depicted in figures 1 and 2. As we suspected last year, round-robin and raw-scoring performs worse when the number of languages is increased. In the other hand, 2-step RSV holds about 85 % of optimal performance.

5 Conclusion and future work

This year, merging approaches and decompounding algorithms have been treated . We have tested 2-step RSV and mixed 2-step RSV with 4-Multi a 8-Multi tasks. Results show that the proposed method scales well with four, five and eight languages, overcoming traditional approaches.

Our next efforts are directed towards three aspects:

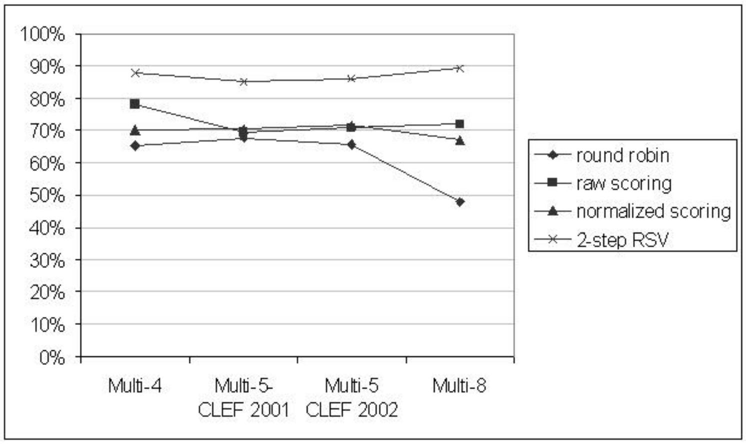

Figure 1: Performance of traditional merging strategies respect of several set of languages (fully aligned queries). Case base (100%) is the *optimal performance*.

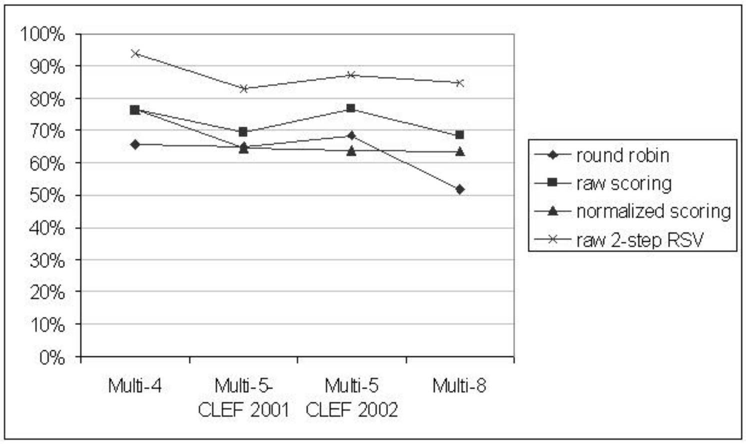

Figure 2: Performance of traditional merging strategies respect of several set of languages (partially aligned queries by means of PRF). Case base (100%) is the *optimal performance*.

- Since decompounding algorithm is highly depend of the wordlists used, we aim to obtain a better wordlist.

- Testing the method with other translation strategies such as Machine Translation or Multilingual Similarity Thesaurus.

- Index terms used in reported experiments are basically obtained by means of stemming. We are very interested in the application of the proposed approach to n-grams indexing. While stemming terms are directly assimilable as feasible representative of concepts, n-grams are not able to be assimilated directly as concepts since given a n-gram usually is contained by several unrelated terms. In addition, we have carried out preliminary experiments, and the obtained results suggest that a n-gram is not a representant of a concept directly.

- Finally, we will keep on studying strategies in order to deal with aligned and not-aligned queries term. The integration of both sort of terms by means of neural networks (although these structures require training data) and development of global pseudo-relevance feedback, and not locally for each monolingual collection, constitutes interesting ways to explore.

6 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant FIT-150500-2002-416.

References

[1] A. Chen. Cross-language Retrieval Experiments at CLEF-2002. In Carol Peters, editor, *Proceedings of the CLEF 2002 Cross-Language Text Retrieval System Evaluation Campaign.*, pages 5–20, 2002.

[2] F. Martínez-Santiago, M. Martín, and L.A. Ureña. SINAI at CLEF 2002: Experiments with merging strategies. In Carol Peters, editor, *Proceedings of the CLEF 2002 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 103–110, 2002.

[3] J. Savoy. Report on CLEF-2001 Experiments. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 27–43. Springer Verlag, 2002.

[4] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th International Conference of the ACM SIGIR'95*, pages 21–28, New York, 1995. The ACM Press.

[5] E. Voorhees. The collection fusion problem. In NIST, editor, *Proceedings of the 3th Text Retrieval Conference TREC-3*, volume 500, pages 95–104, Gaithersburg, 1995.

[6] F. Gey, H. Jiang, A. Chen, and R. Larson. Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 527–540, 2000.

[7] J.Y. Nie and F. Jin. Merging different languages in a single document collection. In *Proceedings of the CLEF 2001*, pages 59–62, 2002.

[8] P. McNamee and J. Mayfield. JHU/APL Experiments at CLEF:Translation Resources and Score Normalization. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 193–208. Springer-Verlag, 2001.

[9] Le Calvé and J. A., Savoy. Database merging strategy based on logistic regression. *Information Processing & Management*, 36:341–359, 2000.

[10] J. Savoy. Cross-language information retrieval: experiments based on clef 2000 corpora. *Information Processing & Management*, 39:75–115, 2003.

[11] M. Martín, F. Martínez-Santiago, and L.A. Ureña. Merging strategy for cross-lingual information retrieval based on learning vector quantization. Technical report, University of Jaén, 2003.

[12] F. Martínez-Santiago and L.A. Ureña. SINAI experience at CLEF. *Revista Iberoamericana de Inteligencia Artificial*, page In press., 2003.

[13] F. Martínez-Santiago and L.A. Ureña. A merging strategy proposal: the 2-step retrieval status value method. Technical report, University of Jaén, 2003.

[14] M. Martín, F. Martínez-Santiago, and L.A. Ureña. Aprendizaje neuronal aplicado a la fusión de colecciones multilingües en CLIR. *Procesamiento del Lenguaje Natural*, page In press., 2003.

[15] C. Monz V. Hollink, J. Kamps and M. de Rijke. Monolingual rertieval for european languages, 2003.

[16] S. E Robertson, S. Walker., and M. Beaulieu. Experimentation as a way of life:Okapi at TREC. *Information Processing and Management*, 1(36):95–108, 2000.

[17] Donna Harman. Relevance feedback revisited. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92)*, pages 1–10, 1992.