Inf Retrieval (2006) 9: 71–93
DOI 10.1007/s10791-005-5722-4

A merging strategy proposal: The 2-step retrieval status value method

Fernando Martínez-Santiago · L. Alfonso Ureña-López · Maite Martín-Valdivia

Received: December 9, 2003 / Revised: December 10, 2004 / Accepted: January 12, 2005

Abstract A usual strategy to implement CLIR (Cross-Language Information Retrieval) systems is the so-called query translation approach. The user query is translated for each language present in the multilingual collection in order to compute an independent monolingual information retrieval process per language. Thus, this approach divides documents according to language. In this way, we obtain as many different collections as languages. After searching in these corpora and obtaining a result list per language, we must merge them in order to provide a single list of retrieved articles.

In this paper, we propose an approach to obtain a single list of relevant documents for CLIR systems driven by query translation. This approach, which we call 2-step RSV (RSV: Retrieval Status Value), is based on the re-indexing of the retrieval documents according to the query vocabulary, and it performs noticeably better than traditional methods.

The proposed method requires query vocabulary alignment: given a word for a given query, we must know the translation or translations to the other languages. Because this is not always possible, we have researched on a mixed model. This mixed model is applied in order to deal with queries with partial word-level alignment. The results prove that even in this scenario, 2-step RSV performs better than traditional merging methods.

Keywords CLIR · Merging strategies · Pseudo-relevance feedback · 2-step RSV · Mixed 2-step RSV

1. Introduction

The typical CLIR requirement is for the user to input a free form query, usually a brief description of a topic, into a search or retrieval engine which returns a list, in ranked order, of documents or web pages that are relevant to the topic. The search engine matches the terms in the query to indexed terms, usually keywords previously derived from the

F. Martínez-Santiago · L.A. Ureña-López · M. Martín-Valdivia
Department of Computer Science, University of Jaén, Jaén, Spain
e-mail:{dofer, laurena, maite}@ujaen.es

target documents. Unlike monolingual information retrieval, CLIR requires query terms in one language to be matched to indexed terms in another. Issues in CLIR are how to translate query terms into index terms, how to eliminate alternative translations (e.g. to decide that French "traitement" in a query means "treatment" and not "salary"), and how to rank or weight translation alternatives that are retained (e.g. how to order the French terms "aventure", "business", "affaire", and "liaison" as relevant translations of English "affair") (Grefenstette 1998).

A new issue arises as to whether queries are translated into each language present in the multilingual document collection. The user query is translated into each language present in the multilingual collection in order to compute a different monolingual information retrieval process per language. Thus, this approach divides document sources according to language. In this way, we obtain as many different collections as languages. Searching in these corpora and obtaining a result list per language is only the first step in proposing CLIR systems. The second step must merge monolingual result lists in order to provide users with a single list of retrieved documents. Obtaining an optimal multilingual list by means of monolingual list is not an easy problem, since the score assigned to each document (the so called Retrieval Status Value - RSV) is calculated not only according to the relevance of the document and the IR model, but also the rest of monolingual corpus is determinant (Dumais 1994).

The rest of the paper is organized as follows. Firstly, we present a brief revision of the most well-used methods for merging strategies. Section 3 and 4 describe our proposed method. In Section 5, we detail the experiments carried out and the results obtained. Finally, we present our conclusions and future lines of work.

2. Traditional merging strategies

There are various approaches in order to carry out the merging of monolingual collections, but even so a large decrease of precision is generated in the process (depending on the collection, between 20% and 40%) (Savoy 2002). Perhaps for this reason, CLIR systems based on document translation tend to obtain results noticeably better than system driven by query translation. The most popular approaches are briefly depicted below:

1. Round-Robin fashion. The documents are interleaved according to rank obtained for each document by means of monolingual information retrieval processing. Thus, given a multilingual collection and N languages, the first document for each monolingual retrieval list will constitute M first documents, the second document of each list will constitute the next M documents, and so on. In this case, the hypothesis is the homogeneous distribution of relevant documents across the collections. This merging process decreases precision about 40% because of the merging process (Callan et al. 1995, Voorhees et al. 1995a).
2. Raw-scoring. This method produces a final list sorted by document score computed independently for each monolingual collection. This method works well whether each collection is searched by the same or a very similar search engine and query terms are distributed homogeneously over all the monolingual collections. Heterogenous term distribution will mean that query weights may vary widely among collections (Dumais 1994), and therefore this phenomenon may invalidate the raw-score merging hypothesis.
3. Normalized scoring. An attempt to make document scores comparable is by normalizing in some way the document score reached for each document:

- Given a monolingual collection, by dividing each RSV by the maximum RSV reached in such a collection:

$$RSV'_i = \frac{RSV_i}{\max(RSV)}, \quad 1 <= i <= N \tag{1}$$

- A variant of the previous method is to divide each RSV by the difference between the maximum and minimum document score values (Powell et al. 2000) reached for each collection:

$$RSV'_i = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)}, \quad 1 <= i <= N \tag{2}$$

4. Perhaps the most original approach is by creating a single index of all the documents without taking into account the multilingual nature of the collection (Gey et al. 2000). In this way, a single and multilingual index is obtained with all the documents of every language. Given a user query, a single multilingual query is obtained in the same way as a single term index was obtained. That is, the query must be translated into each language present in the multilingual collection. A query for each translation is not generated but all the translations are joined thereby forming a composite query. Finally, this composite query is searched across the entire multilingual term index. In the same way as the approach based on document translation, this method will return a single list of documents for each query. But the background problem is not eliminated. Although a single index is generated, the vocabulary of each language is practically exclusive. Two different languages rarely share terms (a noticeable exception are proper nouns). For this reason, the weight obtained by each term will depend on the language. Therefore, the rank and scoring will be fully comparable between documents expressed in the same language. However, the obtained results with this method are disappointing (Nie and Jin 2002, McNamee and Mayfield 2002).
   Note that CLIR merging problem is very similar to *the collection fusion problem* (Voorhees et al. 1995a) of Distributed Information Retrieval (DIR), although DIR environments tend to be monolingual and uncooperative (information about resources such as size of the collection, documental frequency of the terms are not detailed). Thus, some collection fusion techniques have been applied to CLIR with several degrees of success:
5. Towell et al. (1995),Voorhees et al. (1995b) apply several learning algorithms in order to predict whether a document is relevant or not, depending on the question and the collection. One such learning algorithm is based on neural networks. A similar approach, has been applied to CLIR environments with good results (Martín et al. 2003). The main difference between both works is the neural network topology (LVQ neural networks instead of feed-forward neural networks). In addition, Voorhees et al. (1995b) create only one network with one output network per collection. Martín et al. (2003) create one network per collection with a single output according to the document relevance.
6. One of the most extended DIR models is CORI (Callan et al. 1995). Some of the CORI metodology was applied in Savoy (2002), Moulinier and Molina-Salgado (2003) with poor results. However, recent variations of the CORI model have obtained better results applied to CLIR (Savoy 2004).
7. Finally, Calvé and Savoy (2000), Savoy (2003a) propose a merging approach based on logistic regression. Logistic regression is a statistical methodology for predicting the probability of a binary outcome variable according to a set of independent explanatory

variables. The probability of relevance to the corresponding document D_i will be estimated according to both the original score and logarithm of the ranking. Based on these estimated probabilities of relevance, the monolingual list of documents will be interleaved forming a single list:

$$\text{Prob}[D_i \textit{ is rel} \mid \textit{rank}_i, \textit{rsv}_i] = \frac{e^{\alpha+\beta_1 \cdot \ln(\text{rank}_i)+\beta_2 \cdot \textit{rsv}_i}}{1 + e^{\alpha+\beta_1 \cdot \ln(\text{rank}_i)+\beta_2 \cdot \textit{rsv}_i}} \tag{3}$$

The coefficients α, β_1 and β_2 are unknown parameters of the model. The usual methods when fitting the model tend to be maximum likelihood or iteratively re-weighted least squares methods.
Because this approach requires fitting the underlying model, the training set (topics and their relevance assessments) must be available for each monolingual collection (in the same way as approaches based on neural networks). Relevance assessments are usually available only for academic collections such as TREC or CLEF[1] campaigns.

3. The 2-step retrieval status value method

The basic 2-step RSV idea is straightforward: given a query term and its translations into the other languages, their document frequencies are grouped together (Martínez-Santiago et al. 2003). 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. Because re-indexing all the documents in the 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 of translating and searching the query on each monolingual collection, D_i, as is 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. Thus, we obtain a T' vocabulary made up by "concepts". A concept consists of each term together with its corresponding translation. In other words, a concept represents a word expressed independently of the language. Concepts are a bit like a small interlingua vocabulary.
   - A single multilingual collection of preselected documents (D' collection) as result of the union of the first 1000 retrieved documents for each language.Thus, such a multilingual collection D' has about $N * 1000$ documents, where N is the number of languages.
2. The re-indexing phase consists of re-indexing the multilingual collection D', but considering solely the T' vocabulary. Only the concepts are re-indexed: given a concept, its document frequency is the result of grouping together the document frequencies of the terms which makes up the concept. Finally, a new query formed by concepts in T' is generated and this query is carried out against the new index. Thus for example, if

[1] Cross Language Evaluation Forum (CLEF) is an annual activity working with European languages, first held in 2000 and coordinated by DELOS Network of Excellence for Digital Libraries conferences, in collaboration with the NIST.

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", English and Spanish terms represent exactly the same concept. 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"[2]
Note that the new "2-step RSV" calculus does not need to download retrieved documents. The method only needs to know which terms matched in which documents, and their associated frequencies in those documents.

The hypothesis of this method is as follows. Given two documents, the score of both documents will be comparable when 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 compliance with the hypothesis.

3.1. A more formal definition of the approach

A large number of retrieval methods are based on this structure (Sheridan et al. 1997):

$$< T, \Phi, D; \quad ff, df >$$

where:

- D es is the document collection to be indexed.
- Φ is the vocabulary used in the indices generated from D.
- A token τ is a specific occurrence of a term in a document. Let T be the set of all tokens representing an occurrence of a term $\varphi_i \in \Phi$ in a document $d_j \in D$.
  Thus, the function

  $$\varphi : T \to \Phi, \quad \tau \to \varphi(\tau)$$

  maps the set of all tokens, T, to the indexing vocabulary Φ. The function φ can be a simple process such as removing accents or another more complex such as root extraction (stemming), lemmatization ... In addition, stopwords will be removed ($\varphi(\tau) = \emptyset$, if τ belongs to the stopwords set.)
- ff is the feature frequency and denotes the number of occurrences of φ_i in a document d_j:

  $$ff(\varphi_i, d_j) :=\mid \{\tau \in T \mid \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\} \mid$$

  where d is the function that makes each token τ correspond to its document:

  $$d : T \to D, \quad \tau \to d(\tau)$$

[2] There is a particular case. When a document contains both terms,"house" and "home", this document is not counted twice. Thus, the document frequency of the concept "casa" is the sum of the document frequency of "casa", "house" and "home" minus the number of English documents containing both translations, "home" and "house".

- df is the document frequency and denotes the number of documents containing the feature φ_i at least once:

$$df(\varphi_i) :=\mid \{d_j \in D \mid \exists \tau \in T : \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\} \mid$$

For each monolingual collection we begin with the already-known structure:

$$< T_i, \Phi_i, D_i, ff, df >, \quad 1 <= i <= N$$

Where N is the number of languages present in the multilingual collection to be indexed. Let $Q = \{Q_i, 1 <= i <= N\}$, be the set formed by the original query together with its translation to the other languages, in such a way that Q_i is the query expressed in the same language as the collection D_i. After each translation Q_i has been run against its corresponding structure $< T_i, \Phi_i, D_i, ff, df >$, it is possible to obtain a new and single structure:

$$< T', \Phi', D, D', ff', df' >$$

where:

- D is the complete multilingual document collection: $D = \{D_i, 1 <= i <= N\}$.
- D' is the set of retrieved multilingual documents as consequence of running the query Q.
- T' is the set of concepts τ_j, and denotes the vocabulary of the D' collection. Since each query Q_i is a translation of another, it is possible to align queries at term level[3].

  $$\tau_j := \{\tau_{ij} \in Q_i, 1 <= i <= N\}, \quad 1 <= j = M, M = |Q|$$

  where τ_{ij} represents all the retained translations (usually synonymous) of the term j of the query Q to the language i and M indicates the query length. Thus, τ_j denotes the concept j of the query Q independently of the language.
- Φ' is a new vocabulary to be indexed, such that each $\varphi_j \in \Phi'$ is generated as follows:

  $$\varphi_j := \{\varphi(\tau_{ij}), \quad 1 <= i <= N\}, \quad 1 <= j <= M$$

- The ff' function and df' function are interpreted as usual:
  - ff' is the number of occurrences of the concept j, expressed in the i language, in the k document.

    $$ff'(\varphi_j, d_k) := ff(\varphi_{ij}, d_k), \quad d_k \in D'$$

  - φ_{ij} is the index token j of the translation to the language i of the query Q.

[3] Two or more queries are aligned at term level if every query is translated to the rest, and for each term query, we know the term translation or translations to the rest of queries. If the alignment process is not possible for every term, the query is *partially* aligned (see Section 4).

– df' is the number of documents with the concept j in the D collection. That is, the sum of the documents with the term ij in the query, expressed in language i:

$$df'(\varphi_j) := \mid \{d_k \in D_i \mid \exists \tau \in T : \varphi(\tau) = \varphi_j \wedge d(\tau) = d_k\} \mid$$
$$:= \Sigma df(\varphi_{ij}), \forall \varphi_{ij} \in \varphi_j, d_k \in D, \quad 1 <= i <= N$$

where $df(\varphi_{ij})$ is all documents that contain the term ij in the monolingual collection D_i.

Given this structure, a new index is generated in search time, but only taking into account the documents that are found in D'. The df function operates on the whole collection D, not only on the retrieved documents in the first phase, D', because of we found empirically that the obtained results are slightly better when the whole collection is considered in order to calculate the new document frequency. This is not surprising since D' is made up by searching documents with the T' vocabulary and differences between document frequency of concepts will be artificially lower than by taking into account D instead of D'.

Note that 2-step RSV approach is quite different from the computation performed by Pirkola in his experiments based on synonym operator (Pirkola 1998). The model proposed by Pirkola treated the translations of a query term as if they were synonyms, by using InQuery synonym operator #syn, which means grouping target words derived from the same source word into the same facet. The main difference between this approach and 2-step RSV is that 2-step RSV treats both the original term and the translated terms as synonymous. On the other hand, Pirkola's approach only applies the synonym operator between translated terms. Another important difference is that 2-step RSV calculus only operates on a small subset of documents (D') rather than on the whole set of documents, which is much faster than the InQuery synonym. Finally, Pirkola and others (Sperer and Oard 2000, Airio et al. 2003) have applied the InQuery synonym operator on bilingual experiments instead of on full multilingual environments.

In some way, this method shares some ideas with the CLIR systems based on corpus translation, but instead of translating the complete corpus, it only translates non-empty words appearing in the query in the retrieved documents. These two simplifications allow the deployment of the system in search time since the necessary re-indexing process in the second phase is computationally possible due to small size of D' collection and to the scarce vocabulary T' (approximately, non-empty query terms multiplied by the average number of retained translations by term and by the number of languages present in D').

4. Mixed 2-step RSV and not aligned words

Perhaps the strongest constraint for this method is that every term in the query must be aligned with the rest of its translations. However, this information is not always available:

- Several translation techniques such as Machine Translation make word-level alignment of the queries difficult.
- The second step of the proposed method does not make use of automatic query expansion techniques such as relevance feedback (RF) or pseudo-relevance feedback (PRF) applied to monolingual queries. Since RF and PRF extend every monolingual query with collection-dependent words, the reindexing process (second step of 2-step RSV) will not take into account all of these words. Because such words are not the same for each monolingual

collection, and the translation to the other languages is unknown, 2-step RSV method ignores these new terms for the second step. However, the overall performance will also improve since PRF and RF improve on monolingual experiments and usually some extended terms coincide with terms of the original query, and such terms will be aligned.

- Sometimes, a word is translated as two or more words (i.e., a multi word expression). For example, the word "Pope is translated to Russian by "Rimskij papa (Cyrillic alphabet has been transliterated with ASCII characters, following the standard Library of Congress transliteration scheme). In this case, word and translation are not aligned. Nevertheless, if a good multiword expressions list is available, then the most frequent multiword expressions are mapped as a single token (Rimskij papa → Rimskij Papa). Thus, "Pope is successfully aligned with "Rimskij Papa. If this is not possible, the word and that translation remain unaligned. We have used multiword expressions list only for some European agglutinative languages: Dutch, Finnish and German (Martínez-Santiago et al. 2004).

As a way to deal with partially aligned queries, we propose four approaches by mixing evidence from aligned and not aligned terms:

- Raw mixed 2-step RSV method: A 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.
  In this way, the second step of the 2-step RSV method manages two vocabularies for each language: the concept dictionary T', and the new local term vocabulary T'_i. T'_i contains every unaligned query-term expressed in the i language. Thus, for a given τ_{ij}, term j into the monolingual collection i, the document frequency value will be:

  – $df'(\varphi_i)$, if $\varphi(\tau_{ij})$ belongs to the concept φ_i. In other words, φ_{ij} is aligned.
  – $df(\varphi_{ij})$, if τ_{ij} is not an aligned word. The translation of τ_{ij} into the other languages is unknown.

  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:

$$RSV'_i = \alpha \cdot RSV_i^{\text{align}} + (1 - \alpha) \cdot RSV_i^{\text{nonalign}} \tag{4}$$

  where RSV_i^{align} is the calculated score calculated by means of aligned terms, such as original 2-step RSV method depicts. On the other hand, RSV_i^{nonalign} is calculated locally. Finally, α is a constant (usually fixed to $\alpha = 0.75$ because we have found empirically that this value obtains the best results for many queries. Nevertheless, in spite of values behind 0.6 and 0.8 obtain very similar results, the α value should be fixed for each particular multilingual system whether relevance assessments are available).
- 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. In an attempt to make comparable RSV_{align} and RSV_{nonalign}, we are able to normalize those

values, as is shown in Formula 1:

$$RSV'_i = \alpha \cdot \frac{RSV_i^{\text{align}} - \min(RSV^{\text{align}})}{\max(RSV^{\text{align}}) - \min(RSV^{\text{align}})} + (1-\alpha) \cdot \frac{RSV_i^{\text{nonalign}} - \min(RSV^{\text{nonalign}})}{\max(RSV^{\text{nonalign}}) - \min(RSV^{\text{nonalign}})}, \quad 1 <= i <= N \tag{5}$$

- Learning-based algorithms (logistic regression and neural networks). In the same way that the score and $\ln(rank)$ evidence was integrated by using logistic regression (Formula 3), we are able to integrate $\ln(rank)$, RSV^{align} and RSV^{nonalign}:

$$\text{Prob}\left[D_i \; is \; rel \middle| \text{rank}_i, rsv_i^{\text{align}}, rsv_i^{\text{nonalign}}\right] = \frac{e^{\alpha+\beta_1 \cdot \ln(\text{rank}_i)+\beta_2 \cdot rsv_i^{\text{align}}+\beta_3 \cdot rsv_i^{\text{nonalign}}}}{1+e^{\alpha+\beta_1 \cdot \ln(\text{rank}_i)+\beta_2 \cdot rsv_i^{\text{align}}+\beta_3 \cdot rsv_i^{\text{nonalign}}}} \tag{6}$$

where RSV_i^{align} and RSV_i^{nonalign} are calculated as Formula 4, and RSV_i^{rank} is the local rank reached by D_i at the end of the first step.
Again, training data must be available in order to fit the model. This a serious drawback, but this approach allows integrating not only aligned and non-aligned scores but also the original rank of the document. In addition this approach can be applied with fully-aligned queries ($rsv_i^{\text{nonalign}} = 0$) in a way to improve the original 2-step RSV by using extra information extracted from the first step: the rank of the document obtained through the monolingual searching process.
In addiction, we also used neural networks to integrate aligned and non-aligned queries (Martín et al. 2003). The neural network architecture is based on the Learning Vector Quantization (LVQ) algorithm. LVQ is supervised competitive learning which needs a training data set to adjust the model.

4.1. 2-step RSV and machine translation

In this study, machine translation is perceived as a black box which receives English phrases and generates translations of theses phrases to the other languages. We have developed a straightforward and quite effective algorithm in order to align phrases and their translations. In order to explain how it works, suppose the phrase "Pesticides in baby food" is translated to Spanish as "Pesticidas en alimentos para niños". The question is what English word is translated by what Spanish word?. The algorithm works as follows:

1. Let the original phrase P_{en}, "Pesticides in baby food".
   (a) $Unigrams_{P_{en}}$ is the set of unigrams into P_{en} (stopwords are eliminated):
   - $Unigrams_{P_{en}} = \{\text{Pesticides, baby, food}\}$

   (b) $Bigrams_{P_{en}}$ is the set of bigrams into P_{en} (stopwords are eliminated):
   - $Bigrams_{P_{en}} = \{\text{Pesticides baby, baby food}\}$

2. Translate $P_{en} + Unigrams_{P_{en}} + Bigrams_{P_{en}}$ by using a machine translation resource such as Babelfish[4]. Thus, the translated expression is:

- EXP_{en}=\{Pesticides in baby food\}\{Pesticides,baby, food\}\{Pesticides baby,baby food \}

[4] Babelfish is a Machine Translation available at http://babelfish.altavista.com.

The translation of EXP_{en} to Spanish will be:

- $EXP_{sp} = \{$Pesticidas en alimento para niños$\}\{$Pesticidas, bebé, alimento$\}\{$Pesticidas bebés, alimento para niños$\}$

Thus, we obtain (Spanish stop-words are eliminated):

- $P_{sp} = \{$Pesticidas alimento niños$\}$
- $Unigrams_{P_{sp}} = \{$Pesticidas, bebé, alimento$\}$ ($Unigrams_{P_{sp}}$ is the translation of $Unigrams_{P_{en}}$)
- $Bigrams_{P_{sp}} = \{$Pesticidas bebés, alimento niños$\}$ ($Bigrams_{P_{sp}}$ is the translation of $Bigrams_{P_{en}}$)

At this point, P_{sp} represents the set of non aligned words. When a word is aligned, this word is removed from P_{sp} and both the original and translated word are added to the *ALIGNED* set. The alignment will be perfect when $P_{sp} = \emptyset$. $Unigrams_{P_{sp}}$ is aligned with $Unigrams_{P_{en}}$ at word level, and $Bigrams_{P_{sp}}$ is aligned with $Bigrams_{P_{en}}$ at bigram level:

$word_i^{sp}$ is translation of $word_i^{en}, \forall word_i^{sp} \in Unigrams_{P_{sp}}, word_i^{en} \in Unigrams_{P_{en}}$

$bigram_i^{sp}$ is translation of $bigram_i^{en}, \forall bigram_i^{sp} \in Bigrams_{P_{sp}}, bigram_i^{en} \in Bigrams_{P_{en}}$

3. For each $word_i^{sp} \in Unigrams_{P_{sp}}$ do

   (a) if $word_i^{sp} \in P_{sp}$, then remove $word_i^{sp}$ from P_{sp}, and add $(word_i^{sp}, word_i^{en})$ to the set of aligned words *ALIGNED*

   Thus, we obtain:

   - $P_{sp} = \{$niños$\}$
   - $ALIGNED = \{$(pesticidas,pesticides),(alimento,food)$\}$

4. For each bigram $bigram_i^{sp} \in Bigrams_{P_{sp}}$ ($bigram_i^{sp} = (word_1^{sp}, word_2^{sp})$, $bigram_i^{en} \in Bigrams_{P_{en}}$ ($bigram_i^{en} = (word_1^{en}, word_2^{en})$ do:

   (a) if $(word_1^{sp}, word_1^{en}) \in ALIGNED$ ($word_1^{sp}$ is aligned with $word_1^{en}$) and $word_2^{sp} \in P_{sp}$ then remove $word_2^{sp}$ from P_{sp} and add $(word_2^{sp}, word_2^{en})$ to *ALIGNED* set.
   (b) if $(word_1^{sp}, word_2^{en}) \in ALIGNED$ and $word_2^{sp} \in P_{sp}$ then remove $word_2^{sp}$ from P_{sp} and add $(word_2^{sp}, word_1^{en})$ to *ALIGNED* set.
   (c) if $(word_2^{sp}, word_1^{en} \in ALIGNED$ and $word_1^{sp} \in P_{sp}$ then remove $word_1^{sp}$ from P_{sp} and add $(word_1^{sp}, word_2^{en})$ to *ALIGNED* set.
   (d) if $(word_2^{sp}, word_2^{en} \in ALIGNED$ and $word_1^{sp} \in P_{sp}$, then remove $word_1^{sp}$ from P_{sp} and add $(word_1^{sp}, word_1^{en})$ to *ALIGNED* set.

   Since the bigram (alimento niños) is aligned with the bigram (baby food), and "alimento" is aligned with "food" and "niños" $\in P_{sp}$, then "niños" is aligned with "baby":

   - $P_{sp} = \emptyset$
   - $ALIGNED = \{$(pesticidas,pesticides),(alimento,food) (niños,baby)

Table 1 Percent of aligned non-empty words (CLEF2001 + CLEF2002 + CLEF2003 query set, Title + Description fields, Babelfish machine translation)

Spanish	German	French	Italian
91%	87%	86%	88%

This algorithm fails if there are bigrams without any aligned term. For example, bigram "baby food" could be translated to Spanish as "alimentos niños" instead of "alimento niños".[5] Thus, the alignment process fails since neither "alimentos" nor "niños" are previously aligned words. In order to improve the matching process, words are stemmed by removing at least genre and number.

Finally, agglutinative languages such as German usually translate (adjective, noun) bigrams by using a compound word. For example, "baby food" is translated by "Säuglingsnahrung" instead of "Säugling Nahrung" (Babelfish translation). We decompound compound words if possible by using the algorithm depicted in Martínez-Santiago et al. (2004).

We have tested the proposed algorithm with the CLEF query set (Title+Description) of the last three years. It aligns about 85–90% of non-empty words (Table 1).

5. Experiments and results

The experiments have been carried out for eight languages: English, Spanish, German, French, Italian, Swedish, Dutch and Finnish. CLEF 2001, 2002 and 2003 collection data and relevance assessments have been used in our experiments. The Cross-Language Evaluation Forum (CLEF) supports global digital library applications by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.[6] A brief description of test collections and structure of queries is depicted as follows:

- CLEF 2001 and CLEF 2002 editions have the same test collection for the multilingual task. This collection is made up by news published for 1994 in Agencia EFE (Spanish), Der Spiegel (German), Frankfurter Rundschau (German), La Stampa (Italian), Los Angeles Times (English), Le Monde (French) and SDA (French, German and Italian). CLEF 2003 edition has two different multilingual tasks: the main multilingual task CLEF 2003-8 is made up by news published for 1994 and 1995 in eight different European languages. CLEF 2003-8 test collection is a superset of CLEF 2001, 2002 test collections. Thus, CLEF 2003-8 is made up by the whole of CLEF 2001, 2002 sources (extended to 1995 year) and news from Algemeen Dagblad (Dutch), Aamulehti (Finnish), Glasgow Herald (English), Handelsblad (Dutch) and Tidningarnas Telegrambyrå (Swedish). Finally, the other CLEF 2003 multilingual task, CLEF 2003-4, is a subset of CLEF 2003-8 task, since CLEF 2003, 4 collection set is limited to four languages (English, French, German and Spanish) (see Table 3).

[5] The translation is "alimento para niños", but "para" is eliminated because "para" is a stop-word.

[6] Text cited from CLEF site: http://clef.iei.pi.cnr.it:2002 (available at June 2004).

Table 2 CLEF-2003 English query

```
<top>
  <num>141</num>
  <EN-title>Letter Bomb for Kiesbauer</EN-title>
  <EN-desc>
  Find information on the explosion of a letter bomb
  in the studio of the TV channel PRO7 presenter
  Arabella Kiesbauer.
  </EN-desc>
  <EN-narr>
  A letter bomb from right-wing radicals sent to the
  black TV personality Arabella Kiesbauer exploded in
  a studio of the TV channel PRO7 on June 9th, 1995.
  An assistant was injured. All reports on the explosion
  and police inquiries after the event are relevant.
  Other reports on letter bomb attacks are of no interest.
  </EN-narr>
</top>
```

Table 3 Brief description of test-collections

Collection	Languages	Source	# of documents
CLEF 2001-2002	EN, SP, DE, FR, IT	Newswires and national newspapers (1994)	Over 800,000
CLEF 2003	EN, SP, DE, FR, IT, SV, NL, FI	Newswires and national newspapers (1994 & 1995)	Over 1.5 million

Table 4 Brief description of query sets (only Title+Description fields)

Query-set	Languages	Collection	# of queries
CLEF 2001	EN, SP, DE, FR, IT	CLEF 2001–2002	50
CLEF 2002	EN, SP, DE, FR, IT	CLEF 2001–2002	50
CLEF 2003–4	EN, SP, DE, FR	CLEF 2003	60
CLEF 2003–8	EN, SP, DE, FR, IT, SV, NL, FI	CLEF 2003	60

- CLEF queries have three sections: title, description and narrative. The title is a very sort phrase (4–5 words). The description is a slightly longer phrase (15–20 words). Finally the narrative section is a more detailed paragraph about the topic of the query (see Table 2 and Table 4). Note that CLEF 2003-4 query set is the same as the CLEF 2003-8 one but only for four languages.

Every collection has been pre-processed as usual, using stopword lists and stemming algorithms available across the Web.[7] Stopword lists have been increased with terms such as "retrieval", "documents", "relevant" ... Due to the German, Dutch Swedish and Finnish morphological complexity, compound words have been reduced to simple words by using a simple probabilistic procedure (Martínez-Santiago et al. 2004). Once collections have been

[7] http://www.unine.ch/info/clef (available at June 2004).

pre-processed, they are indexed with the Zprise IR system, using the OKAPI probabilistic model (fixed at $b = 0.75$ and $k1 = 1.2$) Robertson et al. (2000). The OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV.

5.1. Translation strategies and bilingual results

We have used several translation approaches. For each query we have taken into account only *Title* and *Description* query fields.

- Machine Readable Dictionary (MDR, Babylon[8]) has been used to translate the query word for word. This 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 (under the heading "Babylon_1") or the first two terms (indicated under the label "Babylon_2"). Since Babylon_1 and Babylon_2 are word for word translations, such translations are fully-aligned. Thus, both translation approaches are used to test original 2-step RSV.
- Machine Translation (MT, Babelfish) translates phrases better than words, and then word level alignment is not possible at all. We propose a simple and quite effective approach in order to align phrase translations at word level (Section 4.1). Partially aligned phrases are used to test mixed 2-step RSV.
- Mixed MT and MDR. This third approach translates every phrase by taking together Babelfish and Babylon_1 translations.

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

Tables 5 and 6 show the bilingual precision obtained re-indexing by means of several translation approaches. Babelfish+Babylon_1 bilingual experiments are noted by the "BB" label.

Babylon_1+Babelfish approach performs slightly better than the rest. On the other hand, Babylon_1 slightly outperforms Babylon_2. Since Babylon_2 keeps two translations per word, precision could be worse because of bad translations kept by Babylon_2. Babelfish obtains results between Babylon_1+Babelfish and Babylon. These results are according to other reports over the same query set (Savoy 2002 2003b). Better improvements are usually reached by means of mixing several MT and MDR resources. The aim of this work is not to obtain the best translation possible but to experiment using the 2-step RSV technique with several translation approaches.

Expansion queries were carried out by means of pseudo-relevance feedback (blind expansion). In this study, we adopted Robertson-Croft's approach (Harman 1992) where the system expands the original query generally by 10–15 search keywords, extracted from the 10-best ranked documents. We have choosen this configuration because empirically we have obtained better results than with other configurations available at ZPrise system.

5.2. Multilingual results with fully aligned queries at term level

The obtained bilingual results list is 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 in which

[8] Babylon is a Machine Dictionary Readable available at http://www.babylon.com. This dictionary is not available for Finnish, thus for this language we used http://www.tracetech.net/sanat/ instead of Babylon.

Table 5 Bilingual experiments (without expansion query)

Approach	EN	SP	DE	FR	IT	NL	SV	FI
CLEF 2001								
Babylon_1	0.54	0.43	0.30	0.42	0.26	–	–	–
Babylon_2	0.54	0.36	0.30	0.43	0.24	–	–	–
Babelfish	0.54	0.43	0.31	0.46	0.26	–	–	–
BB	0.54	0.42	0.32	0.47	0.27	–	–	–
CLEF 2002								
Babylon_1	0.51	0.36	0.28	0.39	0.23	–	–	–
Babylon_2	0.51	0.33	0.29	0.39	0.24	–	–	–
Babelfish	0.51	0.37	0.31	0.38	0.28	–	–	–
BB	0.51	0.40	0.33	0.42	0.30	–	–	–
CLEF 2003 (4 & 8 languages)								
Babylon_1	0.46	0.31	0.29	0.37	0.24	0.25	0.21	0.29
Babylon_2	0.46	0.29	0.28	0.38	0.23	0.21	0.20	0.30
Babelfish	0.46	0.32	0.31	0.38	0.26	–	–	–
BB	0.46	0.34	0.32	0.42	0.29	–	–	–

Table 6 Bilingual experiments (with expansion query)

Approach	EN	SP	DE	FR	IT	NL	SV	FI
CLEF 2001								
Babylon_1	0.46	0.45	0.33	0.42	0.31	–	–	–
Babylon_2	0.46	0.41	0.31	0.44	0.30	–	–	–
Babelfish	0.46	0.47	0.34	0.45	0.34	–	–	–
BB	0.46	0.49	0.32	0.43	0.35	–	–	–
CLEF 2002								
Babylon_1	0.50	0.39	0.32	0.47	0.28	–	–	–
Babylon_2	0.50	0.37	0.33	0.43	0.29	–	–	–
Babelfish	0.50	0.42	0.33	0.47	0.35	–	–	–
BB	0.50	0.40	0.40	0.42	0.36	–	–	–
CLEF 2003 (4 & 8 languages)								
Babylon_1	0.45	0.35	0.32	0.40	0.29	0.26	0.31	0.25
Babylon_2	0.45	0.32	0.34	0.46	0.30	0.27	0.28	0.23
Babelfish	0.45	0.37	0.34	0.43	0.33	–	–	–
BB	0.45	0.36	0.38	0.44	0.35	–	–	–

query and translation are fully aligned. This scenario is not possible with machine translation approaches. In addition, given a query, if expansion query techniques are independently applied for each language collection, then those new terms added to the original query will be not aligned since such terms are language-dependent. Thus, in order to study original

Table 7 Multilingual experiments with fully aligned queries

	Avg. Prec. CLEF 2001		Avg. Prec. CLEF 2002	
Merging strategy	Babylon_1	Babylon_2	Babylon_1	Babylon_2
Round-Robin	0.248 (67.6%)	0.229 (68.1%)	0.221 (64.0%)	0.219 (65.0%)
Raw scoring	0.255 (69.4%)	0.248 (73.8 %)	0.238 (69.0%)	0.229 (68.8%)
Normalized score eq 1	0.259 (70.6 %)	0.241 (71.7%)	0.239 (69.3%)	0.239 (71.8%)
Normalized score eq 2	0.259 (70.6 %)	0.250 (74.4%)	0.240 (69.6%)	0.245 (73.6%)
Logistic regression	data training		0.261 (75.6%)	0.252 (75.7%)
LVQ NN	data training		0.265 (76.8%)	0.250 (75.0%)
2-step RSV	**0.312 (85.0%)**	**0.289 (86.0%)**	**0.296 (85.8%)**	**0.288 (86.5%)**
Optimal performance	*0.367*	*0.336*	*0.345*	*0.333*

2-step RSV, we have used Babylon_1 and Babylon_2 query set without any expansion query techniques. In this way, the queries are fully aligned at term level.

The merging approach has been made up by using several approaches: round-robin, raw scoring, normalized score (Formula 1), variation of normalized score (Formula 2), logistic regression (Formula 3), neural networks and 2-step RSV approach. In addition, theoretical optimal performance has been calculated by using the procedure proposed in Chen (2003) (label "Optimal performance" in Table 7). This 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. The procedure obtains the upper-bound performance for a set of ranked list of documents, and this information is useful in measuring 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 overcome, at least theoretically.

The 2-step RSV merging approach improves on all the other approaches, reaching about 85% of theoretical optimal performance (Table 7). On the other hand, traditional methods perform at about 70%.

Logistic regression and neural networks obtain the second best result, but both approaches require data training (we have used CLEF-2001 queries and relevance assessments).

We have implemented the logistic regression model with the *R* package. The model is adjusted by an iterative re-weighted least squares algorithm (it is part of the *R* package).

We have carried out the experiments with the LVQ algorithm by using the implementation described in LVQ_PAK documentation with default parameters.

Babylon_1 and Babylon_2 obtain a very similar precision. Since bilingual experiments by using Babylon_2 translation approach are a little worse than Babylon_1 (see Table 5), multilingual experiments based on Babylon_1 are a little better than the ones based on Babylon_2.

Finally, we have carried out several experiments with CLEF2003-4 and CLEF2003-8 tasks in order to evaluate several merging approaches with four and eight languages. The improvement of 2-step RSV with respect to other approaches holds in the same way as previous results, or is even increased (Table 8).

Table 8 CLEF2003-4 and CLEF2003-8 experiments with fully aligned queries

	Avg. Prec. CLEF 2003-4		Avg. Prec. CLEF 2003-8	
Merging strategy	Babylon_1	Babylon_2	Babylon_1	Babylon_2
Round-Robin	0.216 (65.3%)	0.210 (64.0%)	0.160 (56.1%)	0.154 (55%)
Raw scoring	0.269 (81.2%)	0.263 (80.1%)	0.203 (71.2%)	0.203 (72.5%)
Normalized scoring	0.232 (70.1%)	0.231 (70.4%)	0.182 (63.9%)	0.180 (64.3%)
2-step RSV	**0.291 (87.8%)**	**0.287 (87.5%)**	**0.242 (85.0%)**	**0.239 (85.4%)**
Optimal performance	*0.331*	*0.328*	*0.285*	*0.280*

Table 9 Three queries with worst precision by using 2-step RSV strategy (Babylon_1 query set)

Query	Query Title	Round-Robin	Formula 2	2-step RSV	*Optimal performance*
44	Indurain Wins Tour	0.190	0.195	0.167	*0.303*
49	Fall in Japanese Car Exports	0.120	0.131	0.079	*0.211*
50	Revolt in Chiapas	0.430	0.436	0.290	*0.722*

Table 10 Details of query 50 by using 2-step RSV merging strategy

Language	Relevant docs.	Retrieved relevant docs.	Retrieved docs.
English	107	67	72
German	112	31	48
French	49	26	53
Spanish	228	203	811
Italian	95	14	16
Total	591	341	1000

5.3. Analysis of failures

Sometimes, 2-step RSV technique works worse than traditional merging strategies. For example, Table 9 shows three such queries, translated by using Babylon_1 query set.

Maybe that the most representative case of error is the query 50. The lost of precision is over 50%. Thus, whether we examine this query more carefully we obtain the data summarized in Table 10.

Table 10 shows the great impact of Spanish collection for this query. There are 591 relevant documents and 228 are written in Spanish (38.5%). This percentage is too high for the importance assigned by the IR system to Spanish documents. By far, Spanish is the language with more documents retrieved (81.1% of retrieved documents are written in Spanish) and more relevant documents retrieved (203 of 341, 59.5%). In other words, there are too many documents from the Spanish collection for this query. The question is why this happens and how it affects the 2-step RSV approach.

Table 11 shows document frequency for each concept of the query. To ensure clarity we represent each concept by the corresponding English term. The 2-step RSV approach uses the global document frequency (grouping together the document frequencies of aligned

Table 11 Local and global document frequency for concepts of query 50

Concept	Eng. df	Ger. df	Fr. df	Sp. df	It. df	Global df
Chiapas	268	374	231	1940	272	3085
Indians	2210	3763	1219	2364	1124	10680
Uprising	373	982	1104	979	175	3613
Mexico	4098	1573	812	14063	53	20590
Revolt	302	301	718	723	3904	5948

query terms) re-indexing with the global document frequency. The concepts "Chiapas" and "uprising" are the most meaningful concepts since they have the lowest document frequency. The concept "uprising" is smoothly distributed among the German, French and Spanish collections. On the other hand, 62.9% of documents containing the concept "Chiapas" are written in Spanish. This high percentage could explain, at least partially, the very high number of retrieved documents coming from the Spanish data collection. In other words, 2-step RSV performance is damaged when a meaningful term is excessively present in a collection, and the proportion of relevant documents for that collection is inferior to the proportion of meaningful terms that the collection contains. Other approaches such as round-robin or normalized score could present less sensibility to this situation because the score of each document is discarded (round-robin sort by ranking only) or is calculated and normalized locally.

5.4. Multilingual results with partially aligned queries at term level

In this section we study the mixed 2-step RSV merging strategy by means of queries partially aligned at term level. Given a language and a query, the query usually contains non-aligned words when the translation approach works at phrase level better than word level, or expansion query techniques such as blind-feedback are applied. The pseudo-relevance feedback technique expands the original query by adding search keywords extracted from the first N documents ranked. Since some of these keywords are new terms (not appearing in the original query), these terms are not aligned.

5.4.1. Experiments based on MDR translation approach and pseudo-relevance feedback

Tables 12 and 13 show results obtained with MDR translations and CLEF 2001–2002 and 2003 corpora. Pseudo-relevance feedback is applied for each query and language. Again, the merging approach has been formed by using several approaches: round-robin, raw scoring, normalized score (Formula 1), variation of normalized score (Formula 5), logistic regression (Formula 3) 2-step RSV, and mixed 2-step RSV approach (raw, normalized and neural network and logistic regression).

The proposed 2-step RSV merging approach improves on all the other approaches. Raw mixed 2-step RSV and normalized mixed 2-step RSV have been calculated by means of Formula 4 and Formula 5, with $\alpha = 0.75$. These values have been fixed because empirically we have found good results. Mixed 2-step by means of logistic regression is implemented as shown in Formula 6. Thus, the unknown parameters α, β_1, β_2 and β_3 must be estimated in the same way as shown in Equation 6 by using iteratively re-weighted least squares method. The

Table 12 CLEF 2001–2002 multilingual experiments with partially aligned queries by means of pseudo-relevance feedback

	Avg. Prec. CLEF 2001		Avg. Prec. CLEF 2002	
Merging strategy	Babylon_1	Babylon_2	Babylon_1	Babylon_2
Round-Robin	0.245 (65.0%)	0.254 (64.3%)	0.251 (68.4%)	0.246 (66.5%)
Raw scoring	0.291 (69.2%)	0.285 (72.1%)	0.281 (76.6 %)	0.269 (72.7%)
Normalized score eq 1	0.271 (64.5%)	0.269 (68.1%)	0.235 (64.0%)	0.252 (68.1%)
Normalized score eq 2	0.297 (70.7%)	0.278 (70.4%)	0.272 (74.1%)	0.272 (73.5 %)
Logistic regression	data training		0.289 (78.7%)	0.289 (78.1%)
LVQ NN. mixed 2-step RSV	data training		**0.293 (79.8%)**	**0.291 (78.6%)**
Original 2-step RSV	**0.327 (77.8%)**	**0.297 (75.2%)**	**0.308 (83.9 %)**	**0.304 (82.2%)**
Raw mixed 2-step RSV	**0.348 (82.8%)**	**0.316 (80.0%)**	**0.320 (87.2%)**	**0.322 (87.0%)**
Norm. mixed 2-step RSV	**0.322 (76.7%)**	**0.294 (74.4%)**	**0.300 (81.7 %)**	**0.301 (81.3%)**
Log.reg. mixed 2-step RSV	data training		**0.323 (88.0%)**	**0.324 (87.6%)**
LVQ NN. mixed 2-step RSV	data training		**0.333 (90.7%)**	**0.320 (86.5%)**
Optimal performance	*0.420*	*0.395*	*0.367*	*0.370*

Table 13 CLEF 2003 multilingual experiments with partially aligned queries by means of pseudo-relevance feedback

	Avg. Prec. CLEF 2003-4		Avg. Prec. CLEF 2003-8	
Merging strategy	Babylon_1	Babylon_2	Babylon_1	Babylon_2
Round-Robin	0.245 (66.0%)	0.244 (67.2%)	0.181 (51.7%)	0.180 (52.3%)
Raw scoring	0.294 (79.2%)	0.285 (78.5%)	0.239 (68.3%)	0.229 (76.6%)
Normalized score eq 2	0.283 (76.3%)	0.268 (73.8%)	0.222 (63.4%)	0.215 (62.5%)
Raw mixed 2-step RSV	**0.335 (90.3%)**	**0.326 (89.8%)**	**0.296 (84.6%)**	**0.266 (84.3%)**
Norm. mixed 2-step RSV	**0.315 (84.9%)**	**0.294 (81.0%)**	**0.266 (76.0%)**	**0.261 (75.9%)**
Optimal performance	*0.371*	*0.363*	*0.350*	*0.344*

best result is obtained with LVQ neural network mixed 2-step RSV approach and Babylon-1 translation. Unfortunately, learning-based algorithms are not applied to CLEF2003-4 and CLEF2003-8 tasks because training data is not currently available.

Perhaps the most surprising result is the good performance of raw-mixed 2-step RSV, even overcoming the normalized version of the approach, and obtaining a result very near to the result reached by means of logistic regression and neural networks. This result is counterintuitive since the method adds two values which are not directly comparable: the score obtained by both aligned and non-aligned terms. Some of the reasons for this good result are:

- α parameter of Formula 4 limits the weight of the unaligned factor.
- Not all the 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. Thus, these terms are

Table 14 Performance by considering all extended terms (Babylon_2 query set and CLEF 2002 data) and aligned terms only

Merging strategy	All terms	Only aligned terms	Loss of precision
Round-Robin	0.2456	0.2354	4.2 %
Raw-scoring	0.2693	0.2568	4.7%
Normalized score f. 1	0.2525	0.2322	8.1%
Normalized score f. 2	0.2717	0.2463	9.4%
Raw mixed 2-Step RSV	0.3220	0.3039	5.7%

aligned terms. In the same way this explains the good performance of 2-step RSV original method with expanded queries.

- CLEF uses comparable document collections (news stories from the same period). The results might be different if collections have vastly different sizes and/or properties.

In order to study the real impact of non aligned words in the final performance of the system, we have carried out an experiment by using original expanded queries and expanded queries without non aligned terms (non aligned terms have been removed from the original expanded query) . The results are summarized in Table 14.This table shows that the improvement is moderate, but this improvement holds when using mixed 2-step RSV approach.

Another interesting result is that the performance obtained by raw mixed 2-step is about 85% of the theoretical optimal performance. This percentage is very similar to the percentage obtained with the original 2-step RSV method (Table 7).

In short, 2-step RSV approach seems to work well with non aligned terms, when the proportion of such terms is not too large.

5.4.2. Experiments based on MT translation approach

In order to evaluate the proposed approach with other translation approaches, we have carried out several experiments with the CLEF 2001–2002 test collection and CLEF2001+CLEF 2002+CLEF2003 query set (160 queries, five languages, EN, SP, DE, FR, IT) by using MT (Babelfish label) and MT+MDR (Babelfish+Babylon_1 label) with and without pseudo-relevance feedback (Table 16). Since the CLEF2003 collection is a superset of previous CLEF collections, and we are using the CLEF2001–2002 collection test, we have removed from the relevance assessments the documents belonging to the CLEF 2003 document collection.

Since MT does not obtain fully aligned queries (see Table 1) 2-step RSV method is not directly applicable, so we have used a raw-scoring 2-step RSV variant with $\alpha = 0.65$. The most interesting result is that MT and MT+MDR approaches are near 90% of optimal performance.

Again, results by merging with raw-scoring are noteworthy. CLEF corpora are comparable and the indices have been created by using the same IR approach (OKAPI). Thus, this contributes to make the document score comparable to a certain extent.

Babelfish+Babylon_1 obtains noticeably better results when no expanded queries are used. On the other hand, the best translation approach is not clear when PRF is taken into account.

Table 15 CLEF 2002 Extended queries statistic (Babylon_2 query set)

	# of terms	Average	Maximum	Minimum	Standard deviation
			Aligned terms		
English	879	17.58	27	6	5.07
German	1348	26.96	48	9	8.83
French	955	20.76	35	9	6.09
Spanish	1086	21.72	35	7	6.34
Italian	1029	20.58	32	8	6.14
Total	5297	21.53	48	6	7.2
			non aligned terms		
English	409	8.18	13	5	1.60
German	165	3.3	9	0	2.78
French	722	15.69	19	11	1.89
Spanish	751	15.02	19	9	2,29
Italian	778	15.56	19	10	2.08
Total	2825	11.48	19	0	5.45

Table 16 CLEF 2001+CLEF2002+CLEF2003 experiments with Machine Translation

Translation strategy	Round-Robin	Raw-scoring	Raw mixing 2-step RSV	*Optimal performance*
		Without pseudo-relevance feedback		
Babylon_1	0.219 (69.5%)	0.241 (76.5%)	0.275 (87.3%)	*0.315*
Babylon_2	0.226 (71.3%)	0.245 (77.3%)	0.280 (88.3%)	*0.317*
Babelfish	0.241 (75.3%)	0.255 (79.7%)	0.281 (87.7%)	*0.320*
Babelfish + Babylon_1	0.253 (74.2%)	0.270 (79.2%)	0.291 (85.3%)	*0.341*
		With pseudo-relevance feedback		
Babylon_1	0.246 (67.6%)	0.263 (72.3%)	0.326 (89.6%)	*0.364*
Babylon_2	0.260 (69.0%)	0.269 (71.4%)	0.342 (90.7%)	*0.377*
Babelfish	0.272 (70.8%)	0.301 (78.4%)	0.343 (89.3%)	*0.384*
Babelfish + Babylon_1	0.279 (71.7%)	0.310 (79.7%)	0.341 (87.7%)	*0.389*

6. Conclusion and future work

In this paper we propose a new approach, 2-step RSV as a way of solving the problem of merging relevant documents in CLIR systems. This approach has performed noticeably better than other traditional approaches in a wide range of scenarios, irrespective of query set, collections, languages or translation resources. The proposed method reaches about 85–90% of the theoretical optimal performance (traditional merging strategies obtain about 65–70%). In order to achieve this performance, queries must be aligned at term level. In addition, we

suspect that the best results are obtained when meaningful terms are distributed throughout document collections approximately in the same proportion as relevant documents.

On the other hand, a drawback for the proposed method is that, given a query, every word must be aligned with the other words, for every language. Thus, we study four approaches for the integration of aligned and non-aligned terms in the same query. The best results are obtained re-indexing by means of logistic regression and neural networks, but this approach depends on data training (mainly assessments of relevance) for each collection, usually scarce. Good results are obtained with both raw and normalized mixed 2-step RSV approaches.

Our future efforts are directed towards the following aspects:

- Dealing with translation probabilities. The original term and translations are treated in exactly the same way in the proposed model. When translation probabilities are available, the calculation of the document frequency and term frequency for a given concept should be reconsidered by means of the translation probability. This can be modelled as follows:

$$ff'(\varphi_j, d_k) := \Sigma ff(\varphi_{ij}, d_k) * w(\tau_{ij}), \forall \varphi_{ij} \in \varphi_j,$$
$$\varphi(\tau_{ij}) = \varphi_{ij}, 1 <= i <= N$$
$$df'(\varphi_j) := \Sigma df(\varphi_{ij}) * w(\tau_{ij}), \forall \varphi_{ij} \in \varphi_j, d_k \in D, 1 <= i <= N$$

  where $w(\tau_{ij})$ represents the translation probability of each translation of term j in the query to language i.
- Testing the method with other translation strategies such as the Multilingual Similarity Thesaurus.
- Index terms used in the 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 representations of concepts, n-grams cannot be assimilated directly as concepts since an n-gram is usually contained within several unrelated terms. In addition, we have carried out preliminary experiments, and the obtained results suggest that an n-gram is not a direct representation of a concept.
- Finally, we will continue studying strategies in order to deal with aligned and non-aligned term queries: the integration of both sorts of terms by means of bayesian networks (although this structure requires data training) and the development of global rather than local pseudo-relevance feedback constitute interesting areas to explore.

Acknowledgements This work has been supported by Spanish Government with grant FIT-150500-2003-412 (MCYT) and TIC2003-07158-C04-04 (CICYT).

References

Airio E, Keskustalo H, Hedlund1 T and Pirkola A (2003) UTACLIR @ CLEF 2002—Bilingual and Multilingual Runs with a Unified Process. In C Peters, M Braschler, J Gonzalo, and M Kluck, (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers, vol. 2785 of Lecture Notes in Computer Science, pp. 91–100. Springer Verlag.

Callan JP, Lu Z and Croft WB (1995) Searching distributed collections with inference networks. In Proceedings of the 18th International Conference of the ACM SIGIR'95, pp. 21–28, New York. The ACM Press.

Calvé A and Savoy J (2000) Database merging strategy based on logistic regression, Information Processing & Management, 36:341–359.

Chen A (2003) Cross-language retrieval experiments at CLEF-2002, In C Peters, M Braschler, J Gonzalo, and M Kluck, (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19–20, 2002. Revised Papers, vol. 2785 of Lecture Notes in Computer Science, pp. 26–48. Springer Verlag.

Dumais S (1994) Latent semantic indexing (LSI) and TREC-2, In Proceedings of TREC'2, volume 500-215, pp. 105–115, Gaithersburg. NIST, D. K. Harman.

Gey F, Jiang H, Chen A and Larson R (2000) Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In EM Voorhees and DK Harman (Eds.), Proceedings of the Seventh Text REtrieval Conference (TREC-7), vol. 500-242, pp. 527–540. NIST.

Grefenstette G, ed. (1998) Cross-language information retrieval, Kluwer academic publishers, Boston, USA.

Harman DK (1992) Relevance feedback revisited. In NJ Belkin, P Ingwersen, and AM Pejtersen (Eds.), Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92), pp. 1–10. ACM.

Martín M, Martínez-Santiago F and Ureña L (2003) Aprendizaje neuronal aplicado a la fusión de colecciones multilingües en CLIR, Procesamiento del Lenguaje Natural, 1(31):227–234.

Martínez-Santiago F, Martín M and Ureña L (2003) SINAI at CLEF 2002: Experiments with merging strategies. In C Peters, M Braschler, J Gonzalo, and M Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers, vol. 2785 of Lecture Notes in Computer Science, pp. 103–110.

Martínez-Santiago F, Montejo-Ráez A, Ureña L and Diaz M (2004) SINAI at CLEF 2003: Merging and decompounding. Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer Verlag, pp. 192–200.

McNamee P and Mayfield J (2002) JHU/APL Experiments at CLEF: Translation resources and score normalization. In C Peters, M Braschler, J Gonzalo, and M Kluck, (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers, volume 2406 of Lecture Notes in Computer Science, pp. 193–208. Springer Verlag.

Moulinier I and Molina-Salgado H (2003) Thomson Legal and Regulatory experiments for CLEF 2002. In C Peters, M Braschler, J Gonzalo, and M Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers, volume 2785 of Lecture Notes in Computer Science, pp. 155–163. Springer Verlag.

Nie J and Jin F (2002) Merging different languages in a single document collection. In C Peters, M Braschler, J Gonzalo, and M Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers, volume 2406 of Lecture Notes in Computer Science, pp. 59–62. Springer Verlag.

Pirkola A (1998) The efects of query structure and dictionary setups in dictionarybased cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.

Powell AL, French JC, Callan J, Connell M and Viles CL (2000) The impact of database selection on distributed searching. In Press TA (ed.), Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, pp. 232–239, New York.

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

Savoy J (2002) Report on CLEF-2001 experiments In C Peters, M Braschler, J Gonzalo and M Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3–4, 2001, Revised Papers, vol. 2406 of Lecture Notes in Computer Science, pp. 27–43. Springer Verlag.

Savoy J (2003a) Cross-language information retrieval: Experiments based on CLEF 2000 corpora, Information Processing & Management, 39:75–115.

Savoy J (2003b) Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence, In C Peters, M Braschler, J Gonzalo, and M Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19–20, 2002. Revised Papers, vol. 2785 of Lecture Notes in Computer Science, pp. 31–46. Springer Verlag.

Savoy J (2004) Combining multiple strategies for effective cross-language retrieval, Information Retrieval, 7(1/2):121–148.

Sheridan P, Braschler P and Schäuble P (1997) Cross-Language information retrieval in a multilingual legal domain, In Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 253–268.

Sperer R and Oard DW (2000) Structured translation for cross-language information retrieval. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120–127. ACM Press.

Towell G, Voorhees EM, Gupta NK and Johnson-Laird B (1995) Learning collection fusion strategies for information retrieval. In Proceedings of the Twelfth Annual Machine Learning Conference, Lake Tahoe.

Voorhees E, Gupta NK and Johnson-Laird B (1995a) The collection fusion problem. In Harman, D. K., (Ed.), Proceedings of the 3th Text Retrieval Conference TREC-3, vol. 500–225, pp. 95–104, Gaithersburg. National Institute of Standards and Technology, Special Publication.

Voorhees E, Gupta NK and Johnson-Laird B (1995b) Learning collection fusion strategies. In ACM, editor, Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 172–179, Seattle.