SINAI*on CLEF 2002: Experiments with merging strategies

Fernando Martínez-Santiago, Maite Martín, Alfonso Ureña
Department of Computer Science, University of Jaén, Jaén, Spain
{*dofer,maite,laurena*}*@ujaen.es*

Abstract

For our first participation in CLEF multilingual task, We present a new approach to obtain a single list of relevant documents for CLIR systems based on query translation. This new approach, which we call two-step RSV, is based on the re-indexing of the retrieval documents according to the query vocabulary, and it performs noticeably better than traditional methods.

1 Introduction

A usual approach in CLIR is to translate the query to each language present in the corpus, and then run a monolingual query in each language. It is then necessary to obtain a single ranking of documents merging the individual lists from the separate retrieved documents. However, a problem is how to carry out such a merge?. This is known as merging strategies problem and it is not an unimportant problem, since the weight assigned to each document (Retrieval Status Value - RSV) is calculated not only according to the relevance of the document and the IR model used, but also the rest of monolingual corpus to which the document belongs is determinant [2].

There are various approaches to standardise the RSV, but even so a large decrease of precision is generated in the process (depending on the collection, between 20% and 40%) [15, 13]. Perhaps for this reason, CLIR systems based on document translation tend to obtain results which are noticeably better than which only translate the query.

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

2 A brief revision of the merging strategies

For each N language, we have N different lists of relevant documents each obtained independently from the others. The problem is that it is necessary to obtain a single list by merging all the relevant languages. If we suppose that each retrieved document of each list has the same probability to be relevant and the similarity values are therefore directly comparable, then an immediate approach would be simply to order the documents according to their RSV (this method is known as raw scoring) [5, 8]. However, this method is not adequate, since the document scores computed by each language are not comparable. For example, a document in Spanish that includes the term "información", can calculate a radically different RSV from another document in English with the same term, "information". In general, this is due to the fact that the different indexation techniques take into account not only the term frequency in the document (tf),but also consider

*Sistemas Inteligentes de Acceso a la Información, Intelligent Information Access Systems

how frequent such a term is in the rest of the documents, that is, the inverse document frequency (*idf*) [12]. Thus, the *idf* depends on each particular monolingual collection. A first attempt to make these values comparable is to standardise in some way the RSV reached by each document:

- By dividing each RSV by the maximum RSV reached in each collection:

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

- A variant of the previous method is to divide each RSV by the difference between the maximum and minimum document score values reached in each collection [10]:

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

in which RSV_i is the original retrieval status value, and $\max(RSV)$ and $\min(RSV)$ are the maximum and minimum document score values achieved by the first and last documents respectively. N is the number of documents in the collection.

However, the problem only is solved partially, since the normalization of the document score is accomplished independently of the rest of the collections, and therefore, the differences in the RSV are still great.

Another approach is to apply a round-robin algorithm. In this case, the RSV obtained for each retrieved document is not taken into account, but rather the relative position reached by each document in their collection. A single list of documents is obtained and the document score m is in the position m in the list. Thus for example, if we have five languages and we retrieve five lists of documents, the first five documents of the single result list will coincide with the first document of each list; the next five, with the second document of each list; and so on. This approach is not completely satisfactory because the position reached by each document is calculated exclusively considering the documents of the monolingual collection to the one which belongs.

Finally, another approach, perhaps the most original, it is to generate a single index with all the documents without taking into account the multilingual nature of the collection [1, 3, 7]. In this way, a single index is obtained in which the terms from each language are intermixed. In the same way in that all the documents in a single index are merged, we obtain a single query where the terms in several languages also are intermixed. That is, the query must be translated to each of the languages present in the multilingual collection. However, we do not generate a query for each translation, but we merge all the translations forming a single query. This query will then be the one which we contrast with the document collection. As with the approach based on document translation, in this approach the system will always return a single list of documents for each query. In spite of this, the problem is not eliminated: the ranking of each document is dependent on the language in which it is written. Although a single index is generated, the vocabulary of each language is practically exclusive. Two different languages rarely share terms. For this reason, the weight obtained by each term will refer to the language to which it belongs, and therefore, the similarity between documents will be correct with respect to the documents expressed in the same language. Finally, it is necessary to mention that a notable exception are proper names, which are frequently invariable in different languages. In such a case, this approach proves very effective.

3 A useful structure to describe IR models

In this section we present a notation that will be used to describe the proposed model. A large number of retrieval methods are based on this structure [14]:

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

where:

- D es is the document collection to be indexed.
- Φ is the vocabulary used in the indices generated from D.
- T is the set of all tokens τ present in the collection D, commonly the words or terms. Thus, the function

  $$\varphi : T \rightarrow \Phi, \tau \rightarrow \varphi(\tau)$$

  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...
- 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:

  $$d : T \rightarrow D, \tau \rightarrow d(\tau)$$

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

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

4 Two-Step Retrieval Status Value

The proposed method [6] is a system based on query translation and it calculates RSV in two phases, a pre-selection phase and a re-indexing phase. Although the method is independent of the translation technique, it is necessary to know how each term translates.

1. 1. The document pre-selection phase consists of translating and running the query on each monolingual collection, D_i, as is usual in CLIR systems based on query translation. This phase produces two results:
   - we obtain a single multilingual collection of preselected documents (D' collection) as a result of joining all retrieved documents for each language.
   - we obtain the translation to the other languages for each term from the original query as a result of the translation process. That is, we obtain a T' vocabulary, where each element τ is called "concept" and consists of each term together with its corresponding translation. Thus, a concept is a set of terms expressed independently of the language.
2. The re-indexing phase consists of re-indexing the multilingual collection D', but considering solely the T' vocabulary. That is, only the concepts are re-indexed. Finally, a new query formed by the concepts in T' is generated and this query is executed against the new index. Thus for example, if we have two languages, Spanish and English, and the term "casa" is in the original query and it is translated by "house", both terms represent exactly the same concept. If "casa" occurs a total of 100 times in the Spanish collection, and "house" occurs a total of 150 times in the English collection, then the term frequency would be 250. From a practical point of view, in this second phase each occurrence of "casa" is treated exactly just as each occurrence of "house".

Formally, the method can be described as follows:
For each monolingual collection we begin with the already-known structure:

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

Where N is the number of present languages 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 the queries at term level.

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

  where τ_{ij} represents all the translations of the term j of the query Q to the language i. 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:

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

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

    $$\begin{aligned} ff'(\varphi_j, d_k) &:=\mid \{\tau \in T \mid \varphi(\tau) = \varphi_j \wedge d(\tau) = d_k\} \mid \\ &:= \Sigma ff(\varphi_{ij}, d_k), \forall \varphi_{ij} \in \varphi_j, d_k \in D', 1 <= i <= N \end{aligned}$$

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

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

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

Given this structure, a new index is generated in run 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'. This is so because in practice, we have found that the obtained results have been slightly better when the whole collection has been considered to calculate the idf factor. Once the indices have been generated in this way, the query Q formed by concepts, not by terms, is re-run on the D' collection.

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

Some relevant aspects of two-step RSV are:

- It is easily scalable to several languages.
- The system requires the term-level alignment of the original query and the translation of its terms. Depending on the approach followed for the translation, this process can prove more or less complex.
- A term together with its translation are treated in exactly the same way in the proposed model. This is not too realistic since it is usual for the original term and its translations not to be equally weighted. For example, it is possible that for a given language i, we maintain more than one translation for a given concept of the original query. Consequently, the concept frequency will be increased artificially in the documents expressed in the i language. In this case, if we know the translation probability of each term, we can weight each term according to its translation probability with respect to the original term. This can be modelled as follows:

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

where $w(\tau_{ij})$ represents the translation probability of each translation of term j in the query to language i, default it will be 1.

5 Description of Experiments and Results

5.1 Multilingual Experiments

The experiment has been carried out for the five languages of the multilingual task. Each collection has been pre-processed as usual, using the stopword lists and stemming algorithms available for the participants, except for Spanish, in which we have used a stemming algorithm provided by the ZPrise system [1]. We have added to the stopword lists terms such as "retrieval", "documents", "relevant"... Due to the German morphological wealth, compound words have been reduced to simple words with the MORPHIX package [9]. Once the collections have been pre-processed, they are indexed with the Zprise IR system, using the OKAPI probabilistic model [11]. This OKAPI model has also been used for the on-line re-indexing process required by the calculation of two-step RSV.

Table 1: Description of official Experiments

Experiment	Task	Form	Query	Merging Strategy
UJAMLTDRR	Multilingual	automatic	Title+Description	Round-Robin
UJAMLTDNORM	Multilingual	automatic	Title+Description	Normalized score
UJAMLTDRSV2	Multilingual	automatic	Title+Description	2-Step RSV
UJAMLTDRSV2RR	Multilingual	automatic	Title+Description	2-Step RSV+Round-Robin
UJABITD{SP,DE,FR,IT}	Bilingual	automatic	Title+Description	

For each query, we have used the Title and Description sections. The method of query translation is very simple: we have used the Babylon[2] electronic dictionary to translate query terms [4]. For each term, we have considered the first two translations available by Babylon. Words not found in the dictionary have not been translated. This approach allows us to carry out query alignment at term level easily.

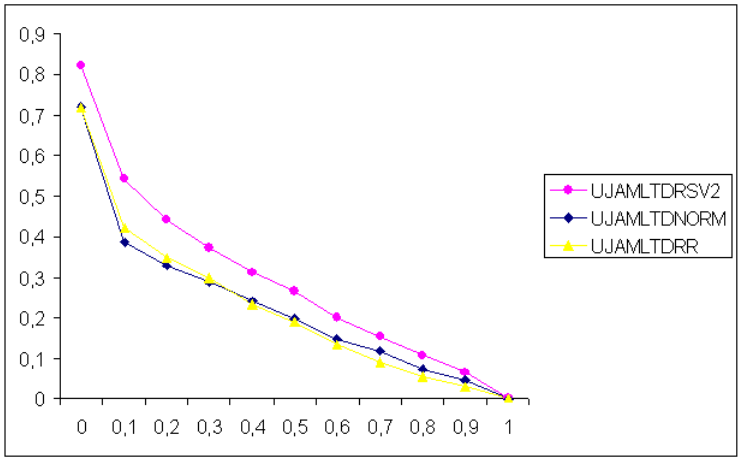
[1] ZPrise, developed by Darrin Dimmick (NIST). Available on demand at http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html

[2] Babylon is available at http://www.babylon.com

Table 2: Performance using different merging strategies (official runs)

Experiment	Avg. prec.	R-Precision	Overall Recall
UJAMLTDRR	0.2038	0.2787	4246/8068
UJAMLTDNORM	0.2068	0.2647	4297/8068
UJAMLTDRSV2	0.2774	0.3280	4551/8068

Figure 1: 11-Pt precision



The obtained results show that the calculation of the two-step RSV improves more than seven points (36% more) the precision reached with respect to other approaches (table 2). This improvement is approximately constant with short, medium and large queries (table 3).

Table 3: Average precision using different merging strategies and query lengths

Merging strategy	Tit.	Tit.+Desc.	Tit.+Desc.+Narr.
round-robin	0.1593	0.2038	0.2425
normalized score	0.1592	0.2068	0.2554
2-step RSV	0.2159	0.2774	0.3209

5.2 Bilingual Experiments

The differences in accuracy between the bilingual experiments may be due to the stemming algorithms used, the quality of which varies according to language. Thus, the simplest stemming algorithm is used for Italian: it removes only inflectional suffixes such as singular and plural word forms or feminine and masculine forms, and it is in this language where the lowest level of accuracy is achieved.

Note that the multilingual document list has been calculated starting from the document lists obtained in the bilingual experiments. The accuracy obtained in the UJAMLTDRSV2 experiment is similar to that obtained in the bilingual experiments(table 4), surpassing even the accuracy for German and Italian, and only two points short of that reached in Spanish.

Table 4: Bilingual experiments (Title+Description)

Experiment	Language	Avg. prec.	R-Precision
UJABITDSP	english → spanish	0.2991	0.3141
UJABITDDE	english → german	0.2747	0.3077
UJABITDFR	english → french	0.3467	0.3365
UJABITDIT	english → italian	0.2438	0.2620

5.3 Merging Several Approaches

Finally, we have carried out an experiment merging several approaches through a simple linear function. specifically, we have calculated document relevance with the function:

$$Pos'_i = 0.6 * Pos_i^{rsv2} + 0.4 * Pos_i^{merge-approach}$$

Where Pos'_i is the new document position i. Pos_i^{rsv2} is the document position reached using two-step RSV, and $Pos_i^{merge-approach}$ is the document position using the Round-Robin or normalized score approach. As shown in table 5, not only is there no improvement, but the accuracy even decreases slightly.

Table 5: Merge of two-step RSV and round-robin/normalized score (Title+Description)

Experiment	Merging strategies	Avg. prec.	R-Precision
UJAMLTDRSV2	RSV2	0.2774	0.3280
UJAMLTDRSV2RR	RSV2 and round-robin	0.2758	0.3265
ujamltdrsv2norm	RSV2 and normalized score	0.2631	0.3162

6 Future work

We have presented a new approach to solve the problem of merging relevant documents in CLIR systems. This approach has performed noticeably better than other traditional approaches. To achieve this performance, it is necessary to align the query with its respective translations at term level. Our next efforts are directed towards three aspects:

- We suspect that with the inclusion of more languages, the proposed method will perform better than other approaches. Our objective is therefore to confirm this suspicion.
- To test the method with other translation strategies. We have a special interest in the Multilingual Similarity Thesaurus, since this provides a measure of the semantic proximity of two terms. That semantic proximity can be used by our method as the translation probability of a term.
- Finally, we could study the effect of the pseudo-relevance feedback in the first and second phase of the proposed method.

References

[1] A. Chen. Multilingual Information Retrieval Using English and Chinese Queries. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 44–58. Springer Verlag, 2002.

[2] S.T. Dumais. Latent Semantic Indexing (LSI) and TREC-2. In NIST, editor, *Proceedings of TREC'2*, volume 500, pages 105–115, Gaithersburg, 1994.

[3] 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.

[4] D.A. Hull and G. Grefenstette. Querying across languages. a dictionary-based approach to multilingual information retrieval. In *Procedings of 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.

[5] K. L. Kwok, L. Grunfeld, and D. D. Lewis. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In NIST, editor, *Proceedings of TREC'3*, volume 500, pages 247–255, Gaithersburg, 1995.

[6] F. Martínez-Santiago and L.A. Ureña. Proposal for a Language-Independent CLIR System. In *JOTRI'2002*, pages 141–148, 2002.

[7] 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.

[8] A. Moffat and J. Zobel. Information retrieval systems for large document collections. In NIST, editor, *Proceedings of TREC'3*, volume 500, pages 85–93, Gaithersburg, 1995.

[9] G. Neumann. Morphix software package. http://www.dfki.de/~neumann/morphix/morphix.html.

[10] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selection on distributed searching. In The ACM Press., editor, *Proceedings of the 23rd International Conference of the ACM-SIGIR'2000*, pages 232–239, New York, 2000.

[11] 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.

[12] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, London, U.K., 1983.

[13] 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, 2001.

[14] P. Sheridan, P. Braschler, and P. Schäuble. Cross-language information retrieval in a multilingual legal domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, 1997.

[15] 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.