SINAI at CLEF 2004: Using Machine Translation resources with mixed 2-step RSV merging algorithm

Fernando Martínez-Santiago, Miguel A. García-Cumbreras
Manuel C. Díaz-Galiano, L. Alfonso Ureña
Department of Computer Science. University of Jaén, Jaén, Spain
{*dofer,magc,mcdiaz,laurena*}*@ujaen.es*

Abstract

This year, we have participated in multilingual CLEF task. Our main interest has been testing Machine Translation (MT) with mixed 2-step RSV merging algorithm. Since 2-step RSV requires grouping together the document frequency for each term and its own translations, and MT translates the whole of the phrase better than word for word, MT is not directly feasible with 2-step RSV merging algorithm (given a word of the original query, its translation to the rest of languages must be known). Thus, we propose a straightforward and effective algorithm in order to align the original query and its translation at term level.

1 Introduction

The aim of 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. There are several approaches for this task, such as translating the whole document collection into an intermediate language or translating the question into every language found in the collection. Moreover, for query translation two architectures are known: centralized and distributed architectures [2]. We use a distributed architecture, where documents in different languages are indexed and retrieved separately. Later on, all ranked lists are merged into a single multilingual ranked list. We focus on a solution for the merging problem. Our merging strategy consists of 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 by 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 [5]. The query is translated using an approach based on Machine Translation (MT), when available. Note that since MT translates the whole of the phrase better than word for word, the 2-step RSV merging algorithm is not directly feasible with MT. 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 word-level alignment algorithm based on MT and a description of our experiments. Section 4 proposes a new way to apply blind relevance feedback (BRF). The last section outlines some conclusions, and also future research lines.

2 Mixed 2-step RSV merging algorithm and Machine Translation

The basic 2-step RSV idea is straightforward: given a query term and the translation of such term into the other languages, the document frequencies are grouped together[5]. Therefore, 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 retrieved document frequency of the term and their translations. In the first step the query is translated and searched in each monolingual collection. This phase produces a T_0 vocabulary made up by ”concepts. A concept consists of each term together with its corresponding translation. Moreover, we obtain a single multilingual collection D_0 of preselected documents as a result of the union of the first 1000 retrieved documents for each language. The second step consists of re-indexing the multilingual collection D_0, but considering solely the T_0 vocabulary. Finally, a new query formed by concepts in T_0 is generated and this query is carried out against the new index.

2.1 An algorithm in order to align at term level a phrase and its translation by using Machine Translation

Since 2-step RSV requires grouping together the document frequency for each term and its own translations, and MT translates the whole of the phrase better than word for word, the 2-step RSV merging algorithm is not directly feasible with MT (given a word of the original query, its translation to the rest of languages must be known). Thus, we propose a straightforward and effective algorithm in order to align the original query and its translation at term level. In this paper, machine translation is perceived as a black box which receives English phrases and generates translations of theses phrases to the other languages. Briefly, for each translation the algorithm works as follows (a more detailed description is available in [6]):

1. Let the original phrase be in English. The phrases is translated to the target language with an MT resource.

2. To extract unigrams and bigrams from the English phrase. Both of them are translated with the same MT resource used in 1.

3. To remove stopwords. Non stopwords are stemmed.

4. To test the alignment of terms by matching terms into the translated phrase with the translation based on unigrams (note that the translation based on unigrams is fully aligned. Thus, if a word of the translated phrase is translated in the same way with a word for word translation method, then we know the translation of the word in the translated phrase. Thus, this word is aligned).

5. After the alignment based on the translation of unigrams is finished, if any term in the translated phrase is not aligned, use the bigrams with exactly one term aligned in order to align the other term of the bigram.

This algorithm fails if there are bigrams without any aligned term after the step 3. In addition, 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 (adjetive, 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 with the algorithm depicted in [7].

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

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 year, we have used MT resources in order to translate the original English query into French and Russian language. However, we have not found quality free Finnish MT, so we have used a Machine Dictionary Readable (MDR) approach (see section 3.1 for more details about translation strategies). The percentage of aligned words is shown in table 2.

Table 2: Percentage of aligned non-empty words (CLEF2004 query set, Title+Description fields, MT for French and Russian. MDR for Finnish)

Finnish	French	Russian
100%	85%	80%

2.2 Mixed 2-step RSV

Although the proposed algorithm to align phrases and translations at term level works well, it does not obtain fully aligned queries. In order to improve the system performance when some terms of the query are not aligned, we make two subqueries. The first one is made up by the aligned terms only and the other one is formed with the non-aligned terms. Thus, for each query every retrieved document obtains two scores. The first score is obtained by using the 2-step RSV merging algorithm over the first subquery. In contrast, the second subquery is used in a traditional monolingual system with the respective monolingual list of documents. Therefore, we have two scores for each query, one is global for all languages and the other is local for each language. Thus we have to integrate both values. As a way to deal with partially aligned queries (i.e. queries with some terms not aligned), last year we proposed several approaches by mixing evidence from aligned and non-aligned terms [7]. This year we have used raw mixed 2-step RSV and logistic regression:

- Raw mixed 2-step RSV method:

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

  where RSV_i^{align} is the score calculated by means of aligned terms, as original 2-step RSV method shows.On the other hand, $RSV_i^{nonalign}$ is calculated locally. Finally, α is a constant (usually fixed to $\alpha = 0.75$).

- Logistic regression: [1, 10] 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:

  $$Prob[D_i \ is \ rel | rank_i, rsv_i] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}} \tag{2}$$

  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 that the score and $\ln(rank)$ evidence was integrated by using logistic regression (Formula 2), we are able to integrate RSV^{align} and $RSV^{nonalign}$ values:

  $$Prob[D_i \ is \ rel | rank_i, rsv_i^{align}, rsv_i^{nonalign}] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}}{1+e^{\alpha+\beta_1 \cdot rsv_i^{align}+\beta_2 \cdot rsv_i^{nonalign}}} \tag{3}$$

where RSV_i^{align} and $RSV_i^{nonalign}$ are calculated as Formula 1. Again, training data must be available in order to fit the model. This is a serious drawback, but this approach allows integrating not only aligned and non-aligned scores but also the original rank of the document:

$$Prob[D_i \ is \ rel|rank_i, rsv_i^{align}, rsv_i^{nonalign}] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}} \tag{4}$$

where RSV_i^{rank} is the local rank reached by D_i at the end of the first step.

3 Experiments and Results

Our Multilingual Information Retrieval System uses English as the selected topic language, and the goal is to retrieve relevant documents for all languages in the collection, listing the results in a single, ranked list. In this list there are a set of documents written in different languages retrieved as an answer to a query in a given language, English in our case. There are several approaches for this task, such as translating the whole document collection to an intermediate language or translating the question to every language found in the collection. Our approach is the latter: we translate the query for each language present in the multilingual collection. Thus, every monolingual collection must be preprocessed and indexed separately. The preprocessing and indexing tasks are shown below.

3.1 Language-dependent features

In CLEF 2004 the multilingual task is made up by four languages: English, Finnish, French and Russian. These languages are very heterogeneous: agglutinative languages such as Finnish, Cyrillic alphabet of the Russian and finally the morphologic complexity of French make difficult the application of a homogeneous strategy for preprocessing and translation tasks:

- English has been preprocessed as usual in other years. Stop-words have been eliminated and we have used the Porter algorithm[8] as it is implemented in the ZPrise system.
- Finnish is an agglutinative language. Thus, we have used the same decompounding algorithm as last year [7]. Stopword list and stemmer algorithm have been obtained in the snowball site [1]. Since we have not found any good free machine translation for Finnish, we use *FinnPlace* online dictionary [2].
- The resources for French have been updated by using the stop-word list and French stemmer from http://www.unine.ch/info/clef. The translation from English has been carried out by using Reverso[3] software.
- For Russian, stop-word list and stemmer algorithm have been obtained in the snowball site. Cyrillic alphabet has been transliterated with ASCII characters, following the standard Library of Congress transliteration scheme. We have used the Prompt MT [4] in order to translate the queries from English into Russian

[1] Snowball is a small string-handling language in which stemming algorithms can be easily represented. Its name was chosen as a tribute to SNOBOL. Available at http://www.snowball.tartarus.org

[2] FinnPlace is available on-line at http://www.tracetech.net/db.htm

[3] Reverso is available on-line at translation2.paralink.com

[4] Prompt is available on-line at http://www.online-translator.com/text.asp?lang=en

Table 3: Language preprocessing and translation approach

	English	Finnish	French	Russian
Preprocessing	stop words removed and stemming			
Additional preprocessing		decompounding		Cyrillic → ASCII
Translation approach		FinnPlace MDR	Reverso MT	Prompt MT

3.2 Language-independent features

Once collections have been pre-processed, they are indexed with the ZPrise IR system[5], using the OKAPI probabilistic model (fixed at $b = 0.75$ and $k1 = 1.2$) [9]. OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV. This year, we have not used blind feedback because the improvement is very poor for these collections, the precision is even worse for some languages (English and Russian).

3.3 Results

Table 4 shows the obtained result by means of several merging approaches. Experiments UJAMLRSV2, UJAMLRL2P and UJAMLRL3P are based on mixed 2-step RSV which requires the combination of two scores per retrieved query (see section 2.2 for details). Perhaps the most surprising result is the poor performance achieved by logistic regression. The reason for this result could be that this merging approach requires relevance assessments for each collection in order to fit the underlying model. Nevertheless, we have no relevance assessment for 1995 *Le Monde* document collection (this collection is available for the first time this year). Thus, we have trained the model with the rest of the French collections. For this reason, we think that the model has been trained poorly. In this way, this explains that the best result is obtained by using the most straightforward mixed 2-step RSV approach (UJAMLRSV2), since the rest of approaches are based on the combination of logistic regression with 2-step RSV.

Table 4: Results using several merging approaches.

Merging strategy	Experiment	AvgPrec
Round robin	unofficial	0.220
Raw scoring	unofficial	0.280
Formula 2 (logistic regression)	UJAMLRL	0.277
Formula 1 (raw mixed 2-step RSV)	**UJAMLRSV2**	**0.334**
Formula 3 (logistic regression and 2-step RSV)	UJAMLRL2P	0.333
Formula 4 (logistic regression and 2-step RSV)	UJAMLRL3P	0.301

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

4 Global relevance blind feedback

This year, we have not used blind feedback because the obtained improvement is poor. We have tested a new way to apply blind feedback *globally* better than *locally. Local relevance blind feedback* is the expansion of the query applied by every monolingual IR system. *Global relevance blind feedback* is the expansion of the query applied by the multilingual IR system. In this way, we analyze the top-N documents ranked into the multilingual list of documents. This idea is applied to 2-step RSV merging algorithm as follows:

1. Merge the document rankings using 2-step RSV.
2. Apply blind relevance feedback to the top-N documents ranked into the multilingual list of documents.
3. Add the top-N more meaningful terms to the query. Since there are documents written in very different languages, the list of selected terms will be multilingual.
4. Expand the concept query[6] with the selected terms.
5. Apply again 2-step RSV over the ranked lists of documents, but by using the expanded query instead of the original query.

Note that blind relevance feedback (we have used Okapi BM25 in this experiment) usually selects terms that are in the initial query. Thus, such terms will probably be aligned. The rest of the selected terms are integrated by using mixed 2-step RSV.

Table 5: Results using global blind relevance feedback (top 10 documents, best 10 terms, Okapi BM25).

Merging strategy	AvgPrec	
	without global BRF	with global BRF
Formula 1 (raw mixed 2-step RSV)	0.334	0.331
Formula 3 (logistic regression and 2-step RSV)	0.333	0.332
Formula 4 (logistic regression and 2-step RSV)+global BRF	0.301	0.309

Table 5 shows that there is no improvement with the application of global relevance blind feedback. We think that there are several possible reasons for this result:

1. Usually, blind relevance feedback is poorly suited to CLEF document collections.
2. We use the expanded query to apply 2-step RSV re-weighting the documents retrieved for each language, but the list of retrieved documents does not change (it only changes the score of such documents). We can also test the improvement of the results by sending the expanded query for each monolingual collection. Thus, the monolingual lists of documents will be modified. Then, we could apply 2-step RSV with the expanded query by recalculating the score of these modified monolingual lists of documents instead of the lists retrieved by means of the non-expanded query. In this way, new documents will be retrieved and evaluated.

5 Conclusions and future work

In past years, we have used a merging approach called 2-step RSV with translations based on MDR. This year we have used the proposed method with several Machine Translation resources. In addition, the multilingual task requires working with very different languages (very different

[6] The concept query is the query used by 2-step RSV with aligned terms. A concept represents a term independently of the language

alphabets and morphological structures). Other years we have tested the performance of 2-step RSV with MDR, blind feedback and other languages and collections. In every experiment, the proposed merging algorithm works well. It outperforms traditional merging approaches about 20-40%. Thus, 2-step RSV is a very stable and scalable merging strategy. Another aim for this year is the integration of learning based algorithms such as logistic regression with 2-step RSV. The obtained results have been not so good. We think that the idea is good but the model could be trained poorly because we have no relevance assessments for one document collection (*Le Monde* 1995). A study in progress is evaluating this approach but filtering 2004 CLEF relevance assessment by eliminating relevant documents of *Le Monde* 1995. Thus, the whole of the multilingual collection would be covered by the relevance assessments used for training.

In spite of the bad results we think that the idea of global blind relevance feedback should improve the performance of the our CLIR model, so we will continue working on this point.

Finally, we are interested in the application of other learning algorithms instead of logistic regression, such as Support Vector Machines (SVM)[11, 3] and Perceptron Learning Algorithm with Uneven Margins (PLAUM)[4].

6 Acknowledgments

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

References

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

[2] A. Chen. Cross-language retrieval experiments at CLEF-2002. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *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*, pages 26–48. Springer Verlag, 2003.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[4] Y. Li, H. Zaragoza, R. Herbrich, J. Shawe-Taylor, and J. Kandola. The perceptron algorithm with uneven margins. In *Proceedings of the International Conference of Machine Learning.(ICML'2002)*, 2002.

[5] F. Martínez-Santiago, M. Martín, and L.A. Ureña. SINAI at CLEF 2002: Experiments with merging strategies. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *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*, pages 103–110, 2003.

[6] F. Martínez-Santiago, M. Martín, and L.A. Ureña. A merging strategy proposal: the 2-step retrieval status value method. *Technical Report. Department of Computer Science of University of Jaén*, 2004.

[7] F. Martínez-Santiago, A. Montejo-Ráez, L.A. Ureña, and M.C. Diaz. SINAI at CLEF 2003: Merging and decompounding. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Proceedings of the CLEF 2003 Cross-Language Text Retrieval System Evaluation Campaign*, pages 99–109, 2003.

[8] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.

[9] 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.

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

[11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.