SINAI experience at CLEF

Fernando Martínez, L. Alfonso Ureña, M. Teresa Martín
Dpto. Informática. Universidad de Jaén
Av. Madrid, 35
Jaén, E-23071 Spain
fdofer,laurena,maite @ujaen.es

This paper reports our work on CLEF and CrossLanguage Information Retrieval using CLEF resources. We aim to construct a highly language-independent CLIR model. To accomplish this objective, several problems must be overcome: text translation or pseudo-translation and merging the obtained results for each language for a given query. Three issues of text-translation are investigated: the impact of translation probabilities, automatic multi-word recognition, and the generation of similarity thesauri from a Web corpus. Because the proposed model is query-translation driven, it is necessary to merge several monolingual results in a unique multilingual list of documents. To accomplish this task, we propose a new approach, which we call 2-step RSV, and we show that it performs better than more traditional approaches.

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol. VIII, No.22 (2004),pp. 111-121
ISSN: 1137-3601. © AEPIA (http://www.aepia.org/revista).

SINAI experience at CLEF

Fernando Martínez, L. Alfonso Ureña, M. Teresa Martín

Dpto. Informática. Universidad de Jaén
Av. Madrid, 35
Jaén, E-23071 Spain
{dofer,laurena,maite}@ujaen.es

Abstract

This paper reports our work on CLEF and Cross-Language Information Retrieval using CLEF resources. We aim to construct a highly language-independent CLIR model. To accomplish this objective, several problems must be overcome: text translation or pseudo-translation and merging the obtained results for each language for a given query. Three issues of text-translation are investigated: the impact of translation probabilities, automatic multi-word recognition, and the generation of similarity thesauri from a Web corpus. Because the proposed model is query-translation driven, it is necessary to merge several monolingual results in a unique multilingual list of documents. To accomplish this task, we propose a new approach, which we call 2-step RSV, and we show that it performs better than more traditional approaches.

Keywords: Cross-Lingual Information Retrieval, Retrieved Status Value, Merging strategies, Multi-word, Similarity thesaurus.

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 target documents. Unlike monolingual information retrieval, CLIR requires query terms in one language to be matched to indexed terms in another. Matching can be done by bilingual dictionary lookup, full machine translation, or by applying statistical methods. A query's success is measured in terms of recall (how many potentially relevant target documents are found) and precision (what proportion of documents found are relevant). Issues in CLIR are ([Grefenstette, 1998]) 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').

Three issues of text-translation are investigated: the impact of translation probabilities,automatic multi-word recognition, and the generation of similarity thesauri from a Web corpus. Because the proposed model is query-translation driven, it is necessary to merge several monolingual results in a unique multilingual list of documents. To accomplish this task, we propose a new approach, which we call 2-step RSV, and we show that it performs better than more traditional approaches.

The rest of the paper is organized as follows: section 2 shows our first participation at CLEF: the calculation of translation probabilities by means of the integration of EuroWorNet and SemCor. This work is addressed to the investigation of the pruning of bad translations in a term-by-

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No 20 (2003), pp. 34-52.
ISSN: 1137-3601.

term translation fashion. Our second participation was aimed at the multilingual task: section 3 shows a merging document strategy called 2-step RSV. In addition, we also propose a variation called mixed 2-step RSV. Section 4 shows briefly two works tested using CLEF collections: multi-word detection by using neural networks and the development of multilingual similarity thesauri elaborated with corpora extracted from the Web. In this way, some results have been reported though both works are still at a preliminary stage. Finally, section 5 outlines some conclusions, and also future research lines.

2 CLEF 2001: Calculating translation probabilities

This section depicts an approach for a bilingual Spanish-English information retrieval based on EuroWordNet [Vossen, 1997] but also using another linguistic resource known as SemCor [Fellbaum, 1998]. It was our first participation at CLEF [Martínez-Santiago et al., 2002a].

2.1 EuroWordNet and SemCor

EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American WordNet for English [Miller, 1995] in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each WordNet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton WordNet 1.5 [Miller, 1995]. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets.

SemCor is a subset of the Brown Corpus [Francis, 1982]. The Brown Corpus is formed by documents about politics, sports, music, films, philosophy, etc. In SemCor, all the nouns, verbs, adjectives and adverbs defined in WordNet are sense tagged.

2.2 Translation approach

The proposed method is an Spanish-English bilingual information retrieval system: the query language is Spanish and the document language is English, although the query language may be any EuroWordNet language.

The translation approach consists of the word for word translation into English of the query, using synonymy relationship for the translation of the words. There are works which also make use of "similar meaning" in the translation [Gollins and Sanderson, 2000]. We take a more restrictive approach since we use only the synonymy relationship. This straightforward approach has several disadvantages. EuroWordNet, and WordNet, make a very fine-grained distinction of meanings available for each word. For instance, the word "capacidad" (capacity) has up to twelve possible translations into English, shared among the five meanings of the source word.

One way of solving this problem may be by classifying, i.e. identifying where the difference is irrelevant to the needs of Information Retrieval [Gonzalo et al., 1998]. The difficulty of this approach lies in knowing when two or more meanings must be joined into just one. Our approach differs considerably from the idea of grouping according to meaning, although the two methods are not incompatible. The method suggested here attempts to filter the query obtained through a word for word translation using EuroWordNet, disposing of the words we consider to be very rare translations of the Spanish word. It is important to point out that no disambiguation of the original word in Spanish is being made as all the possible meanings of the word are taken into account. What we are trying to achieve is to discard all the words in English which are highly unlikely as translations of the original word in Spanish. In short, we are trying to establish for a given word T in Spanish, and its corresponding translation into English $\{S_1, ...S_n\}$, how probable it is that S_i is a translation of T.

The method is as follows: SemCor labels every word with its sense, so it is possible to calculate how many times a term is used with a given sense. Thus the probability of every sense for a given term is known automatically: the sense proba-

bility is the number of occurrences of the term in SemCor with the given sense divided by the total occurrences of the term in SemCor. Thus, it is easy to build up a frequency table of senses which shows how often a particular sense is assigned to each term. Table 1 shows an example of this process corresponding to the word "absolute". This term belongs to three synsets, so it receives three senses in SemCor, of which the third is the most unusual.

Meaning	Freq	Sense Prob
1	10	0.6665
2	4	0.2667
3	1	0.0667

Table 1: **Weights for the 3 meanings of the word *absolute***

For the translation, we make use of the frequency table of senses as follows: EuroWordNet gives the translation of every term into different languages by means of the synonymy relationship. For instance, the Spanish word "sanatorio" may be translated into English as "sanatorium" or "home" because both pairs (sanatorio, sanatorium) and (sanatorio, home) share a particular meaning (English term and Spanish term belong to the same synset). However the problem is: should we translate the term "sanatorio" with both words?

Word	Meaning	synset	Sense Prob
sanatorium	1	03023310n	0.85
home	6	02628978n	0.05

Table 2: Translations of the Spanish word "sanatorio".

The solution to this problem depends on the strength of the synonymy relationship. The term "sanatorio" has two senses in EuroWordNet so "sanatorio" belongs to the synsets 03023310n and 02628978n. Since there is no disambiguation, we take into account both senses, but not all the translations for each sense: for each sense, translations with a low sense probability will be discarded. Thus, "home" and "sanatorio" belong to the synset 02628978n. But, "home" is a polysemous word (it belongs to many synsets). Is the 02628978n sense frequent for this word?. We conclude that the probability of translating "sanatorio" by "home" is the probability of the sense 02628978n by the word "home". This frequency information is that stored in our table of senses. Moreover, according to this table, we know that the sense 0268978n is very unusual for the word "home". So we considerer "home" with the meaning of "sanatorio" as irrelevant. Thus, the translation of "sanatorio" by "home" is discarded. In addition, "sanatorium" most frequent sense is the shared synset with "sanatorio" (synset 03023310n). We conclude that "sanatorio" must be translated to "sanatorium".

2.3 Results and conclusions

In our experiment we used the ZPrise[1] Information Retrieval System, and *tf.idf* standard weighting schema. This choice was determined by its availability and because this system has been recommended by CLEF organization in the evaluation of linguistic resources in CLIR tasks such as the one presented here [Gonzalo, 2001]. The test corpus was "Los Angeles Times, 1994", made available by CLEF. This collection has 113.005 documents from the 1994 editions of the "Los Angeles Times". The title, heading and article text were extracted. The official experiments carried out were as follows:

1. *sinai_org* run: original set of queries in English. This is taken as the best case and it is used as reference for the rest of the runs.
2. *sinai-ewn* run: using the query obtained through the word by word EuroWordNet translation.
3. *sinai-ewn_semcor* run: a filter was applied based on the probabilities of translation obtained with SemCor: to the set of translated queries. The aim was to eliminate all the target term candidates below a threshold of 0.25 in their probability of translation. It is important to point out that those words that do not appear in SemCor are retained in the original query, as we have no information for them.

The obtained precision we obtained for each of the following experiments is shown in the table 3.

If we take *sinai-org*, as the reference experiment we notice that the loss of precision in

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

Official run	Avg Prec
sinai-org	0.4208
sinai-ewn	0.1701
sinai-ewn_semcor	0.1941

Table 3: Avg precision obtained

the *sinai-ewn* experiment is 59.5% compared with a 53.8% loss in the *sinai-ewn_semcor* (EuroWordNet+SemCor) run. Therefore the use of probabilities of translation calculated on SemCor reduces the lack of precision by 6.3% compared to that obtained using EuroWordNet without filtering (*sinai-ewn* experiment). It is likely this percentage would improve if we had a corpus containing all the meanings from EuroWordNet with a number of words far superior to that of SemCor.

	cons_exp		cons_exp+MW	
Appear	PT $\geq$ 0.25	344	PT $\geq$ 0.25	42
SC	PT $\leq$ 0.25	295	PT $\leq$ 0.25	12
	Sum:	639	Sum:	54
Not in SC	196		137	
Sum	735		191	

Table 4: Breakdown of words in the queries translated from EuroWordNet

Table 4 shows how many times SemCor provided information that helped to eliminate noise. Thus, we note that of a total of 735 words, which is the sum of words in the *cons-exp* queries, on 196 occasions we do not obtain any information from SemCor. This means that 27% of times we cannot decide whether the word is a good translation or not. This situation becomes considerably worse when we consider the Multi-Words (MW). The percentage of indecision in this case rises to 72%. However, for those multi-words that we do not find in SemCor, we note that 77.8% are assigned a Probability of Translation (PT) superior to 0.25 compared to 53.8% of simple words. This could be read as meaning that multi-words tend to be a more precise translation of the original word, as in general a multi-word tends to be monosemous or have very few meanings.

Finally, SemCor has two serious drawbacks. The first is its relatively small size (SemCor 1.6 has approximately 31.600 —word, meaning— pairs) and the second is that it is only available for English.

3 CLEF 2002: A new approach to merging problem. Two-Step Retrieval Status Value

The experience on CLEF 2001 encourages us to try the multilingual task. The objective was the construction of a multilingual framework and the experimentation with merging strategies. This section depicts the highlights of our participation on CLEF 2002.

3.1 The problem of merging the retrieved documents

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 the monolingual corpus to which the document belongs is determinant [Dumais, 1994].

There are various approaches to standardize the RSV, but even so a large decrease in precision is generated in the process (depending on the collection, between 20% and 40%) [Voorhees, 1995, Savoy, 2001]. Perhaps for this reason, CLIR systems based on document translation tend to obtain results which are noticeably better than those which only translate the query.

The rest of this section 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.

3.2 A brief revision of 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 documents. 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) [Kwok et al., 1995, Moffat and Zobel, 1995]. 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 indexing techniques take into account not only the term frequency in the document (tf),but also consider how frequent such a term is in the rest of the documents, that is, the inverse document frequency (idf) [Salton and McGill, 1983]. Thus, the idf depends on each particular monolingual collection. A first attempt to make these values comparable is to standardize in some way the RSV reached by each document:

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

$$RSV_i' = \frac{RSV_i}{\max(RSV)}$$

- 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 [Powell et al., 2000]:

$$RSV_i' = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)}$$

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.

However, the problem is only partially solved, 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 high.

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 which that document belongs.

3.3 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 [Sheridan et al., 1997]:

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

where:

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

3.4 Two-Step Retrieval Status Value

The proposed method [Martínez-Santiago et al., 2002c], [Martínez-Santiago and Ureña, 2002] 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. 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 run 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".

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 languages present in D').

3.5 Mixed Two-Step RSV and not aligned words

Perhaps the strongest constraint for this method is that, given a query, every word must be aligned with the rest of the words, for every language. But this information is not always possible:

- 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 add some collection-dependent words for each monolingual query, the reindexing process (second step of 2-step RSV) will not take into account these words. Because such words are not the same for each monolingual collection, and the translation to the rest of languages is unknown, 2-step RSV method

ignores these new terms for the second step. However, because PRF and RF improve the monolingual experiments, the overall performance will also improve a little.

An straightforward and effective way to partially solve this problem is by taking non-aligned words into account locally, just as terms of a given monolingual collection. In this way, the second step of the two-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} to the rest of the languages is unknown.

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

3.6 Experiments and Results

The experiment has been carried out for the five languages of the multilingual 2002 CLEF task: English, Spanish, German, French and Italian. Each collection has been pre-processed as usual, by 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. 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 [Finkler and Lutzky, 1994]. Once the collections have been pre-processed, they are indexed with the Zprise IR system, using the OKAPI probabilistic model [Robertson et al., 2000]. This OKAPI model has also been used for the on-line re-indexing process required by the calculation of two-step RSV.

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 [Hull and Grefenstette, 1996b]. For each term, we have considered the first two translations available in Babylon. Words not found in the dictionary have not been translated. This approach allows us to carry out query alignment at term level easily.

Strategy	AvgP	R-Prec	Overall /Recall
Raw Scoring	0.2038	0.2787	4246/8068
Round Robin	0.2038	0.2787	4246/8068
N. Scoring	0.2068	0.2647	4297/8068
2-Step RSV	0.2774	0.3280	4551/8068

Table 5: Performance using different merging strategies

The obtained results show that the calculation of the two-step RSV improves more than seven points (36% more) the precision reached compared to other approaches (Table 5).

3.6.1 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.

Language	AvgP	AvgP with PRF
english → spanish	0.2991	0.3243
english → german	0.2747	0.3402
english → french	0.3467	0.4021
english → italian	0.2438	0.3308

Table 6: Bilingual experiments (Title+Description)

Note that the multilingual document list has been calculated starting from the document lists obtained in the bilingual experiments. The accuracy obtained by using the 2-step RSV is similar to that obtained in the bilingual experiments(Table

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

6), surpassing even the accuracy for German and Italian, and only two points short of that reached in Spanish.

3.6.2 Experiments with mixed 2-step RSV and pseudo-relevance feedback

We have tested the mixed 2-step RSV by means of pseudo-relevance feedback (blind expansion). We have used the Rocchio's approach [Buckley et al., 1996] where the expanded terms are extracted from the 10-best documents for every query for each language. The average precision improvement reached is about 20% (Table 6). We have then merged these multilingual results as shown in the previous section.

Strategy	AvgP	
	No PRF	PRF
Round Robin	0.2038	0.2763
N. Scoring	0.2068	0.2428
2-Step RSV	0.2774	0.2905
mixed 2-Step RSV	0.2774	0.3315

Table 7: Performance with and without pseudo-relevance feedback

The obtained multilingual results are shown in Table 7. An interesting result is that 2-step RSV only slightly improves the result obtained without PRF. The reason is mentioned in section 3.5: the second step of the method doesn't use the whole of the available terms, it only uses aligned terms, concepts. On the other hand, mixed 2-step RSV is 19.5 % better than 2-step RSV without PRF. This improvement is very similar to the improvement obtained in bilingual queries. Thus, mixed 2-step RSV is a valid strategy for the integration of aligned terms and non-aligned terms in the same query.

4 Other Experiments with CLEF collections

In this section we briefly show other experiments carried out with Spanish and English CLEF 2000 and 2001 collections. Multi-word recognition and pseudo-translation of queries using Multilingual Similarity are the topics.

4.1 Using Neural Networks for Multiword Recognition in IR

A multiword is a succession of words whose sense taken as a whole differs from the sum of the senses of its single words. Thus, a multiword can be considered in fact as a new concept.

Multiword recognition has been explored by many researchers as a way to improve traditional Text Retrieval, in general with a moderate degree of success. However, David Hull and Gregory Grefenstette [Hull and Grefenstette, 1996a] show that multiword detection and correct translation largely improve the precision in a CLIR system.

A supervised neural network (the LVQ algorithm) has been used to classify pairs of terms as being multiwords or non-multiwords [Martínez-Santiago et al., 2002b]. Classification is based on the values yielded by different estimators, currently available in literature, used as inputs for the neural network. Lists of multiwords and non-multiwords have been built to train the net. Afterward, many other pairs of terms have been classified using the trained net.

Original CLEF 2000 query set	CLEF 2000 query set with multi-word
0.375	0.410

Table 8: Performance with and without multi-word recognition

Results obtained in this classification have been used to perform information retrieval tasks by using CLEF 2000 English collection to test the results. Experiments show that detecting multiwords results in better performance of the IR methods. Nevertheless, the method used must obtain higher accuracy, because incorrect detection of multiwords damages the precision of the IR system.

4.2 Generating a corpus from the Web

We describe in [García et al., 2002] the construction of a multilingual similarity thesaurus [Sheridan et al., 1997] by using a comparable corpus extracted from the Web. Selected multilingual Web sites are downloaded with a web crawler called WebReader, which generates structured, homogeneous and low noise documents from the

semi-structured, heterogeneous and noisy Web. Also, we describe a method to align the obtained multilingual documents by using clustering techniques. The aligned documents are used to create a multilingual similarity thesaurus, with English and Spanish documents from several online newspapers. Finally, we test the quality of the thesaurus by means of a bilingual IR experiment: we use the multilingual similarity thesaurus to accomplish the pseudo-translation of CLEF-2001 English query set to Spanish.

Site	Pages
ABC	28,173
CNN	36,691
El Mundo	29,828
El Pais	31,863
The Observer	29,153
W. Post	32,683

Table 9: Sites and downloaded pages

Translation tool	AvgP
SYSTRAN	0.26
EuroWordNet	0.19
Multilingual Similarity Thesaurus	0.16

Table 10: Average precision

In order to compare our translation method, the query set is translated with both Systran and EuroWordNet. The query set translated by Systran obtains the best results whereas the query set obtained by means of the multilingual similarity thesaurus obtains the worst results. Instead of the relatively poor results achieved by the thesaurus, we think that a multilingual similarity thesaurus generated from a comparable corpus extracted from the Web is a valid approach where there is no more sophisticated resources available, such as parallel corpora or machine translation. In addition, the small size of the generated corpus is a serious drawback. We expect that a larger comparable corpus will improve the precision of the method.

5 Conclusion and future work: CLEF 2003

This paper shows our work at CLEF 2001 and CLEF 2002. Our first participation was addressed to the initiation of our group at CLIR tasks: three bilingual experiments were submitted and the reached result was satisfactory for us. We proposed a method to calculate translation probabilities by means of integration of EuroWordNet and SemCor. The proposed method has two serious drawbacks: the small coverage of SemCor, and the fact that SemCor is only available for English.

Our second participation was a little more ambitious: the development of a multilingual system. In spite of the very simple dictionary translation approach and the use of a poorly-tuned IR model without query expansion or other refinements, we were pleased to reach fourth position in the competition. The proposed innovation was a new approach to solving the problem of merging relevant documents in CLIR systems. We call the method 2-step RSV. A constraint for this method is that, given a query, every word must be aligned with the rest of the words, for every language. In the paper we propose a variant of the original 2-step RSV called mixed 2-step RSV, which supports aligned and non-aligned words. Thus, both are used to calculate the RSV for a given document. In this way, our next efforts are addressed to standardising in some way the weight obtained by non-aligned words. Finally, we are anxious to use our method with eight languages, because we suspect that it will improve our results compared to traditional merging approaches.

Translation probabilities calculation, multi-word recognition, pseudo-translation of queries by using Multilingual Similarity Thesauri and new merging documents approaches are several aspects of our work at CLIR task. The integration of all of them to reach a multilingual IR model is our main objective, and CLEF is the best way to improve the result of our efforts.

6 Acknowledgments

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

References

[Buckley et al., 1996] Buckley, C., Singhal, A., Mitra, M., and Salton, G. (1996). New Retrieval Approaches using SMART. In Harman, D. K., editor, *Overview of the 4th Text*

REtrieval Conference TREC-4, pages 25–48, Gaithersburg. NIST.

[Dumais, 1994] Dumais, S. (1994). Latent Semantic Indexing (LSI) and TREC-2. In NIST, editor, *Proceedings of TREC'2*, volume 500, pages 105–115, Gaithersburg.

[Fellbaum, 1998] Fellbaum, C. (1998). Wordnet:an electronic lexical database. *The MIT Press*.

[Finkler and Lutzky, 1994] Finkler, W. and Lutzky, O. (1994). MORPHIX. In Hausser, R., editor, *Linguistische Verifikation. Dokumentation zur ersten Morpholympics*, pages 67–88, Tbingen:Niemeyer.

[Francis, 1982] Francis, W. (1982). *Problems of Assembling and Computerizing Large Corpora*. Norwegian Computing Centre for the Humanities, Bergen.

[García et al., 2002] García, M., Martínez-Santiago, F., Martín, M., and Ureña, L. (2002). Generación de un tesauro de similitud multilingüe a partir de un corpus comparable aplicado a CLIR. Pendiente de publicar.

[Gollins and Sanderson, 2000] Gollins, T. and Sanderson, M. (2000). Sheffield University CLEF 2000 Submission - Bilingual Track: German to English. In Peters, C., editor, *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*. Springer-Verlag.

[Gonzalo, 2001] Gonzalo, J. (2001). Language resources in cross-language information retrieval: a CLEF perspective. In *Cross-Language Information Retrieval and Evaluation: Proceedings of the First Cross-Language Evaluation Forum*.

[Gonzalo et al., 1998] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarrán, J. (1998). Indexing with WordNet synsets can improve Text Retrieval. In *Procedings of COLING/ACL Workshop on usage of WordNet in NLP Systems*.

[Grefenstette, 1998] Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Kluwer academic publishers, Boston, USA.

[Hull and Grefenstette, 1996a] Hull, D. and Grefenstette, G. (1996a). Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[Hull and Grefenstette, 1996b] Hull, D. and Grefenstette, G. (1996b). 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.

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

[Martínez-Santiago et al., 2002a] Martínez-Santiago, F., Díaz, M., García, M., Martín, M., and Ureña, L. (2002a). SINAI on CLEF 2001: Calculating translation probabilities with SemCor. In Peters, C., editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 185–192. Springer-Verlag.

[Martínez-Santiago et al., 2002b] Martínez-Santiago, F., Díaz, M., Martín, M., Rivas, V., and Ureña, L. (2002b). Using neural networks for multiword recognition in IR. In *Proceedings of 2002 Conference of International Society of Knowledge Organization (ISKO)*.

[Martínez-Santiago et al., 2002c] Martínez-Santiago, F., Martín, M., and Ureña, L. (2002c). SINAI on CLEF 2002: Experiments with merging strategies. In Peters, C., editor, *Proceedings of the CLEF 2002 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*. Springer-Verlag. Pendiente de Publicaci'on.

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

[Miller, 1995] Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11).

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

[Powell et al., 2000] Powell, A. L., French, J. C., Callan, J., Connell, M., and Viles, C. L. (2000). The impact of database selection on distributed

searching. In Press., T. A., editor, *Proceedings of the 23rd International Conference of the ACM-SIGIR'2000*, pages 232–239, New York.

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

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

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

[Sheridan et al., 1997] 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*, pages 253–268.

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

[Vossen, 1997] Vossen, P. (1997). EuroWordNet: A Multilingual Database for Information Retrieval. In *THIRD DELOS WORKSHOP Cross-Language Information Retrieval*, pages 85–94. European Research Consortium For Informatics and Mathematics.