SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: query expansion using the Google search engine

Fernando Martínez-Santiago, Arturo Montejo-Ráez
Miguel Á. García-Cumbreras and L. Alfonso Ureña-López

SINAI Research Group. Computer Science Department. University of Jaén. Spain
{dofer,amontejo,magc,laurena}@ujaen.es

Abstract. This year, we have participated in the Ad-Hoc Robust Multilingual track with the aim of evaluating two important issues in Cross-Lingual Information Retrieval (CLIR) systems. This paper first describes the method applied for query expansion in a multilingual environment by using web search results provided by the Google engine in order to increase retrieval robustness. Unfortunately, the results obtained are disappointing. The second issue reported alludes to the robustness of several usual merging algorithms. We have found that 2-step RSV merging algorithms perform better than others algorithms when evaluating using geometric precision[1].

1 Introduction

Robust retrieval has been a task in the TREC evaluation forum [1]. One of the best systems proposed involves query expansion through web assistance [4, 3, 2]. We have followed the approach of Kwok and his colleagues and applied it to robust multilingual retrieval.

Pseudo-relevance feedback has been used traditionally to generate new queries from the results obtained from a given source query. Thus, the search is launched twice: one for obtaining first relevant documents from which new query terms are extracted, and a second turn to obtain final retrieval results. This method has been found useful to solve queries producing small result sets, and is a way to expand queries with new terms that can widen the scope of the search. But pseudo-relevance feedback is not that useful when queries are so difficult that very few or no documents are obtained at a first stage (the so-called weak queries). In that case, there is a straighforward solution: use a different and richer collection to expand the query. Here, Internet plays a central role: it is a huge amount of web pages where virtually any query, no matter how difficult it is, may be related to some subset of those pages. This approach has obtained remarkable

[1] This work has been supported by the Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

results in monolingual IR systems evaluated in TREC conferences. Unexpectedly, the results obtained in a multilingual are very poor and we think that our implementation of the approach must be tuned for CLEF queries, regardless of our conviction that an intensive tuning work is unrealistic for real-world systems. In addition, as we expected, the quality of the expanded terms depends on the selected language.

On the other hand, we have evaluated several merging algorithms from the perspective of robustness: round-Robin, raw scoring, normalized raw scoring, logistic regression, raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression and mixed 2-step RSV based on bayesian logistic regression. We have found that round-Robin, raw scoring and methods based on logistic regression perform worse than 2-step RSV merging algorithms.

The rest of the paper has been organized as three main sections: first, we describe the experimentation framework, then we report our bilingual experiments with web-based expansion queries, and finally we describe the multilingual experiments and how geometric precision affects several merging algorithms.

2 Experimentation framework

Our Multilingual Information Retrieval System uses English as the selected topic language. 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 is a set of documents written in different languages retrieved as an answer to a query in a given language (here, English). There are several approaches to 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, as is described below.

2.1 Preprocessing and translation resources

In CLEF 2006 the multilingual task is made up of six languages: Dutch, English, French, German, Italian and Spanish. The pre-processing of the collections is the usual onw in CLIR, taking into account the lexicographical and morphological idiosyncracy of every language. The pre-processing is summarized in table 1.

- English has been pre-processed as in past years. Stop-words have been eliminated and we have used the Porter algorithm[7] as implemented in the ZPrise system.
- Dutch, German and Swedish are agglutinating languages. Thus, we have used the decompounding algorithm depicted in [6]. The stopwords list and the stemmer algorithm have been both obtained in the Snowball site [2].

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

– The resources for French and Spanish have been updated using the stopword lists and stemmers from http://www.unine.ch/info/clef. The translation from English has been done using Reverso[3] software.
– Dutch and Swedish translations have been done using on-line FreeTrans service[4].

Table 1. Language preprocessing and translation approach

	Dutch	English	French	German	Spanish	Italian
Preprocessing	stop words removed and stemming					
Decompounding	yes	no	no	yes	no	yes
Translation approach	FreeTrans		Reverso	Reverso	Reverso	FreeTrans

Once the collections have been pre-processed, they are indexed with the IR-N [10] system, an IR system based on passage retrieval. The OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV, using the OKAPI probabilistic model (fixed empirically at $b = 0.75$ and $k1 = 1.2$) [8]. As usual, we have not used blind feedback, because the improvement is unsubstantial for these collections, and the precision is even worse for some languages (English and Swedish).

2.2 Merging strategies

This year we have selected the following merging algorithms: round-Robin, raw scoring [11, 14], normalized raw scoring [13], logistic regression [12], raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression [6] and mixed 2-step RSV based on bayesian logistic regression as implemented in the BBR package[5].

3 Query expansion using the Internet as a resource

Expanding user queries by using web search engines such as Google has been successfully used for improving robustness of retrieval systems over collections in English. Due to the multilinguality of the web, we have assumed that this could be extended to additional languages, though the smaller amount of web pages not in English could be a major obstacle. The process is splitted into the following sequential steps:

1. **Web query generation**. The process varies depending on whether we consider the title field or the description field of the original query:

[3] Reverso is available on-line at www.reverso.net
[4] FreeTrans is available on-line at www.freetranslation.com
[5] BBR software available at http://www.stat.rutgers.edu/m̃adigan/BBR.

- *From title*. Experiments expanding queries based just on the title field take all the terms in the field in lower case joined by the AND operator.
- *From description*. Here, terms have to be selected. To that end, stop words are removed (using a different list according to the language which the description is written in) and the top 5 ranked terms are taken to compose, as for the title field, an AND query. The score computed for each term to rank the same formula used by Kwok [3].

2. **Web search**. Once the web query has been composed, the web search engine is called to retrieve relevant documents. We can automate the process of query expansion through Google using its Java API. This web search is done specifying the language of the documents expected for the retrieval. Therefore, a filter on the language is set on the Google engine.
3. **Term selection from web results**. The 20 top ranked items returned by Google are taken. Each item points at an URL but also contains the so-called snippet, which is a selection of text fragments from the original pages containing the terms used in the web search (i.e. the query terms) as a sort of summary intended to let the user decide if the link is worth following. We have performed experiments using just the snippets as retrieved text in order to propose new query terms, and also experiments where terms are selected from full web page content (downloading the document from the returned URL).
   In both cases (selection of terms from snippets or selection from full web pages), the final set of terms is the composite of those 60 terms with the highest frequency after discarding stop words. Of course, in the case of full web pages, the HTML tags are also conveniently eliminated. To generate the final expanded query, terms are repeated according to their frequencies (normalized to that of the least frequent term in the group of 60 selected terms).

As an example of the queries generated by this process, for a title containing words "`inondation pays bas allemagne`" the resulting expansion would produce the text:

```
pays pays pays pays pays pays pays pays pays pays pays pays pays pays
pays pays pays bas bas bas bas bas bas bas bas bas bas bas bas bas
allemagne allemagne allemagne allemagne allemagne allemagne allemagne
inondations inondations inondations france france france inondation
inondation inondation sud sud cles cles belgique belgique grandes
grandes histoire middot montagne delta savent fluviales visiteurs
exportateur engag morts pend rares projet quart amont voisins ouest
suite originaires huiti royaume velopp protection luxembourg convaincues
galement taient dues domination franque xiii tre rent commenc temp
monarchie xii maritime xive proviennent date xiiie klaas xiie ques
```

3.1 Experiments and results

We have generated four different sets of queries for every language, one without expansion and three with web-based expansion:

1. `base` – No expansion, the original query is used and its results taken as base case
2. `sd-esnp` – Expansion using the original description field for web query generation and final terms selected from snippets
3. `st-esnp` – Expansion using the original title field for web query generation and final terms selected from snippets
4. `st-efpg` – Expansion using the original title field for web query generation and final terms selected from full web pages

The results obtained are discouraging, as all our expansions lead to worse measurements of both *R-precision* and *average precision*. Figures 1 a and 1 b show graphically the values obtained when evaluating these measures. For technical reasons, the expansion of type `st-efpg` for Dutch was not generated.

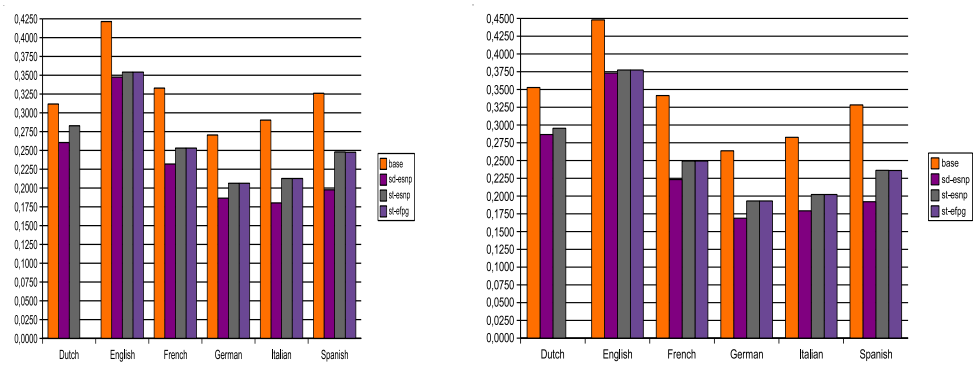

Fig. 1. R-precision (a) and average precision (b) measurements

Some conclusions can be drawn from these results. The main one is that the title field is a much more suitable source of items for a web-based expansion. Indeed, for many authors the title can be considered as the set of query terms that the users should pass to a search engine. Thus, web query generation from the description field, even using sophisticated formulae, is, as the results show, a worse choice when a title field is available.

We can check that very similar results are obtained, independently of the final selection of terms, that is, it seems that the decision of taking final terms either from snippets or from full web pages text does not determine significant differences in the results obtained. This issue needs further analysis, because expanded queries are quite different on the last half of the series of selected terms (those that are less frequent). These results hint at the system not profiting from the full set of terms passed.

Finally, we find that the results depend on the language under study. We think that this is due to differences on the size of existing collections of pages

for each language on the web. That could explain the slightly better results in the case of English compared to the rest of languages.

4 Multilingual experiments

As the commented formerly, the merging algorithm is the only difference between all our multilingual experiments. Table 2 shows the results obtained in terms of 11-pt average precision, R-precision and geometric precision. From the point of view of the average precision, the most interesting finding is the relatively poor result obtained withthe methods based on machine learning. Thus, mixed 2-step RSV-LR and mixed 2-step RSV-BLR perform slightly worse than raw mixed 2-step RSV in spite that the latter approach does not use any training data. As usual, logistic regression performs better than round-Robin and raw scoring, but the difference is not as relevant as other years. Thus, we think that difficult queries are not learned as fine as usual queries. This is probably because, given a hard query, the relation between score, ranking and relevance of a document is not clear at all, and therefore, machine learning approaches are not capable to learn a good enough prediction function. Similarly, this year there are not only hard queries, but also very heterogeneous queries from the point of view of average precision. Thus, the distribution of average precision is very smooth and it makes more difficult extracting useful information from the training data.

Table 2. Multilingual results

Merging approach	11Pt-AvgP	R-precision	Geometric Precision
round-robin	23.20	25.21	10.12
raw scoring	22.12	24.67	10.01
normalized Raw scoring	22.84	23.52	10.52
logistic regression	25.07	27.43	12.32
raw mixed 2-step RSV	27.84	32.70	15.70
mixed 2-step RSV based on LR	26.91	30.50	15.13
mixed 2-step RSV based on BLR	26.04	31.05	14.93

Since the 2-step RSV largely overcomes the rest of tested merging algorithms when they are evaluated using geometric precision measure, we think that a 2-step RSV merging algorithm is better suited than other merging algorithms to improve the robustness of CLIR systems. Thus, if we use geometric precision to evaluate the CLIR system, the difference of performance between results obtained using 2-step RSV and the rest of merging algorithms is higher than using traditional 11Pt-AvP or R-precision measures.

5 Conclusions

We have reported on our experimentation for Ad-Hoc Robust Multilingual track CLEF task about web-based query expansion for languages other than English.

First, we try to apply the expansion of queries using a web search engine such as Google. This approach has obtained remarkable results in monolingual IR systems evaluated in TREC conferences. But in a multilingual scenario, the results obtained are very poor and we think that our implementation of the approach must be tuned for CLEF queries, regardless of our belief that an intensive tuning work is unrealistic for real-world systems. In a robust evaluation the key measure should be the *geometric average precision*, also at monolingual tasks, because it emphasizes the effect of improving retrieved documents on weak queries, as the task itself defines. Future research includes to study the value obtained on this measure when using expanded queries and when merging retrieved items in a multilingual retrieval, as it is difficult to explain the goodness of our approach to the robust task without it.

In addition, as was expected, the quality of the expanded terms depends on the language selected. The second issue reported is about the robustness of several widespread merging algorithms. We have found that Round-Robin, raw scoring and methods based on logistic regression perform worst from the point of view of robustness. On the other hand, 2-step RSV merging algorithms perform better than the other algorithms when geometric precision is applied. In any case, we think that the development of a robust CLIR system does not require special merging approaches: it "only" requires good merging approaches. It may be that other CLIR problems such as translation strategies or the development of an effective multilingual query expansion should be revisited in order to obtain such a robust CLIR model.

References

1. E. M. Voorhees: *The TREC Robust Retrieval Track*, TREC Report 2005
2. K.L. Kwok, L. Grunfeld, P. Deng: *Improving Weak Ad-Hoc Retrieval by Web Assistance and Data Fusion*, AIRS 2005, LNCS 3689, pp. 17–30, 2005
3. K.L. Kwok, L. Grunfeld, H.L. Sun, P. Deng: *TREC2004 Robust Track Experiments using PIRCS, 2004*, 2005
4. L. Grunfeld, K.L. Kwok, N. Dinstl, P. Deng, 2003, *TREC2003 Robust, HARD and QA Track Experiments using PIRCS*, 2003
5. S.T. Dumais. Latent Semantic Indexing (LSI) and TREC-2. *Proceedings of TREC'2, volume 500-215, pages 105-115, Gaithersburg*, 1994. NIST, D. K. Harman.
6. F. Martinez-Santiago, L.A. Ureña, and M. Martin. A merging strategy proposal: two step retrieval status value method. *Information Retrieval, vol. 9, issue 1, 71-93*, Jan 2006.
7. M.F. Porter. An algorithm for suffix stripping. *Program 14, pages 130-137*, 1980.
8. S. E Robertson, S. Walker., and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management, vol. 1, 95-108*, 2000.
9. J. Savoy. Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management, vol 39, 75-115*, 2003.
10. F. Llopis, H. Garcia Puigcerver, Mariano Cano, Antonio Toral, Hector Espi. *IR-n System, a Passage Retrieval Architecture. TSD, 57-64*, 2004.
11. J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. *Proceedings of the 18th International Conference of the ACM SIGIR'95, pages 21-28, New York*, 1995. The ACM Press.

12. A. Calve and J. Savoy. Database merging strategy based on logistic regression. *Information Processing and Management, 36:341-359*, 2000.
13. A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selection on distributed searching. *The ACM Press., Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, pages 232-239, New York.* 2000.
14. E. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. *D. K. Harman, editor, Proceedings of the 3rd Text Retrieval Conference TREC-3, volume 500-225, pages 95-104, Gaithersburg*, 1995. National Institute of Standards and Technology, Special Publication.