REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval

Angel Zazo, Carlos G. Figuerola, and José Luis A. Berrocal
REINA Research Group - Universidad de Salamanca
C/ Francisco de Vitoria 6-16, 37008 Salamanca, SPAIN
http://reina.usal.es

Abstract

This paper describes our work at CLEF 2006 Robust task. This task is an ad-hoc task that explores methods for stable retrieval by focusing on poorly performing topics. We have realized experiments for all subtask: monolingual (EN, ES, FR and IT), bilingual (IT→ES) and multilingual (ES→[EN ES FR IT]) retrieval.

For monolingual retrieval we have focused our work on local query expansion, i.e. using only the information from retrieved documents. External corpora, such as the Web, were not used. Our document retrieval system is simple; it is based on vector space model. Some local expansion techniques were applied for training topics. The best improvement was achieved using association thesauri, which were constructed employing co-occurrence relations in term windows, not in complete document. This technique is effective and can be easily implemented without tuning some parameters. Our mandatory runs (title+description topic fields) have obtained good positions in all monolingual subtasks we participate.

For bilingual retrieval two machine translation programs were used to translate the topics from Italian into Spanish. Both translations were joined before searching. The same expansion technique was also applied. Our mandatory run has got the top rank in the bilingual subtask. For multilingual research we used the same procedure to obtain the retrieval list for each target language, and we combined them with the MAX-MIN data fusion method. In this subtask, our mandatory run has been in the lower part of the ranking of runs.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: *Indexing methods, Thesauruses*; H.3.3 [**Information Search and Retrieval**]: *Query formulation, Relevance feedback*; H.3.4 [**Systems and Software**]: *Performance evaluation*; I.2.7 [**Natural Language Processing**]: *Machine Translation*

General Terms

Measurement, Performance, Experimentation

Keywords

Robust Retrieval, Query Expansion, Term Windows, Association Thesauri, CLIR, Machine Translation

1 Introduction

Robust retrieval tries to obtain stable performance over all topics by focusing on poorly performing topics. Robust tracks were carried out in TREC 2003, 2004 and 2005 (Voorhees, 2003, 2004, 2005) for monolingual retrieval, but not for cross-language information retrieval.

The users of a information retrieval system don't know concepts such as average precision, recall, etc. They only use it, and they usually remember better failures than success. Failures decide if a system will be used again. The robustness ensures that all topics obtain minimum effectiveness levels. In information retrieval the mean of the average precision (MAP) has been used to measure the systems' performance. But, poorly performing topics have little influence on MAP. At TREC, geometric average (rather than MAP) turned out to be the most stable evaluation method for robustness (Voorhees, 2004). The geometric average (GMAP) has the desired effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between larger scores.

In CLEF 2006 Ad-hoc track a new robust task was introduced. Three subtask were designed for robust task:

- **Monolingual**: for all six document languages: Dutch (NL), English (EN), German (DE), French (FR), Italian (IT) and Spanish (ES).
- **Three bilingual**: Italian→Spanish, French→Dutch and English→German.
- **Multilingual**: All six languages are allowed as topic language.

Our research group has participated in all subtasks. We have carried out monolingual (EN, ES, FR, IT), bilingual (IT→ES) and multilingual (ES→[EN ES FR IT]) experiments. For each subtask two runs was submitted, one with `title` and `description` topic fields (mandatory) and one with only the `title` field. All experiments were run with the same setup (except for language specific resources).

2 Experiments

We have focused our work on local query expansion, i.e. using only the information from retrieved documents. In CLEF 2002 we used association and similarity thesauri to expand sort queries: all documents of the collection (i.e. global query expansion) was used to construct the thesauri (Zazo et al., 2003). In later works (Zazo et al., 2002, 2005; Zazo, 2003) we have studied in depth several query expansion techniques: local vs. global analysis, term reweighting, coefficients for expansion, etc. Some conclusions we have taken out:

- Query expansion depends on the technique using to obtain relations between terms.
- Performance improves if terms added to the original query have high relation value with all terms of the original query, not with only one separately.
- Expansion depends on the importance (weight) of the terms added to the original query.
- Performance is higher for sort queries than long queries. Long queries usually have well defined the user information need, and frequently several additional terms are not necessary to improve performance.
- In most cases the expansion techniques are based on local analysis, using the retrieved documents to obtain relations between terms. The performance of the first retrieval is fundamental to obtain high improvement with the expansion: a good retrieval system (term weighting) is better than a good expansion technique.

Considering these items, a lot of experiments have been carried out, only with training topics (mandatory). One observes that the topic collection of robust task came from CLEF 2001 through CLEF 2003, but the document collections came from CLEF 2003, and they were different than CLEF 2001 and 2002 collections. It's known that retrieval performance depends not only in term weighting, but topic and document collections; for the same document collection and weighting schema, two different topic collections obtain different performance. So, we take a daring decision: for our experiments we have only used the training topics of CLEF 2003 topic collection.

Our primary effort was monolingual retrieval. The steps in monolingual subtask will be explained bellow. For bilingual and multilingual experiments we have used machine translation (MT) programs to translate the topics into document language, and then performing monolingual retrieval. The MAX-MIN data fusion method was used to joining lists in multilingual retrieval.

2.1 Monolingual Experiments

Our document retrieval system is simple. It is based on vector space model. Not additional plugins for word sense disambiguation nor other linguistic techniques were used. We have focused our work on local query expansion, i.e. using only the information from retrieved documents. Complete document collection or external corpora, such as the Web, were not used. First, it is necessary to have a good term weighting schema to take as the base, and to check if stop words removing or stemming processes improve robustness. Second, we have applied some local query expansion techniques to see which had better improvement over the least effective topics.

For each test we realized, each topic was classified into three category: "OK" if its average precision was >MAP; "bad" if it was only >MAP/2, and "hard" if it was <MAP/2. Our effort was to improve hard topics, taking account that MAP and GMAP of all topics would be the measures of robustness in the task. We have had in mind sort and long queries too.

Term weighting

For our experiments, we converted all word to lowercase, suppressed the stress signs, and included number as index terms. We used the training topics from CLEF 2003. A lot of tests were carried out to obtain the best document-query weighting schema. The best was **dnu-ntc** schema (using SMART letters conventions) for all document collections we partipated (EN, ES, FR, IT). For documents, letter u stands for the pivoted document normalization (Singhal et al., 1996). In this normalization is usual setting up *pivot* to the average document length and tuning *slope*. How odd! For all collections the best performance was reached with $slope = 0.1$.

We have tried to find some heuristic to predict topic difficulty without relevance information, and use it to make determinations at run time. We used term frequency and inverse document frequency of the topic terms, but no heuristic was found.

Stop words removing

Stops words are too frequent words or words with slight semantics. They are removed in indexing process, but is not easy to build a good set of stop words. For each document collection, we sorted all words according to their document frequency and we decided to probe with three thresholds: words that appear in more than 50, 25 and 15 percent of the documents. We inspect the set to remove some important words, for example de word "angeles" –from Los Angeles– in the English document collection.

There was hardly any difference in MAP and GMAP when stop words were removed. But it's important to note that for English and Italian collections, removing stop words improve slightly the performance of hard topics. For French and Spanish no improvement is observed. We decide to use the better situation, with threshold of 25%, for the rest of the experiments.

Stemming

Stemming can be thought of as a mechanism for query expansion, since each word is expanded with other words having the same stem. For each language we used a different stemmer: for English the well-known Porter stemmer, and for French, Italian and Spanish the stemmers provide by Jacques Savoy in the web page `http://www.unine.ch/info/clef/`.

Stemming improved MAP and GMAP for all collections, but hard topics had different behavior. For the English collection the hard topics hardly improved performance. For the Spanish collection hard topics had a little improvement. For French and Italian the improvement was important.

Blind relevance feedback

BRF is a common used technique for local query expansion. The Rocchio formula is the most frequent for this purpose. In (1) γ was fixed to 0.0 and was necessary tuning α and β. A lot of tests were carried out to obtain the better performance, using the first 5, 10, 15 and 20 retrieved documents for expansion.

$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_r}\sum_{i \in rel}^{n_r} \vec{d_i} - \frac{\gamma}{n_{nr}}\sum_{j \in norel}^{n_{nr}} \vec{d_j} \tag{1}$$

For the English collection BRF deteriorated performance for the experiment with title and description topic fields. For the experiment with the title field the improvement was only near 3% for MAP and GMAP. Hard topics had no improvement. For the rest collections BRF improved MAP and GMAP about 7%, and also hard topics. The best test for all languages and experiments was using 5 or 10 documents for the expansion, with $\alpha = 1$, $\beta \sim 2.5$ and using an expanded query with about 50 terms.

Local association thesauri

Term co-occurrence has been frequently used in IR to identify some of the semantic relationships that exist among terms. In fact, this idea is based on the Association Hypothesis (van Rijsbergen, 1979, p.104). If query terms are useful to identify relevant and non relevant documents, then their associated terms will also be useful, and can be added to the original query.

Several coefficients have been used to calculate the degree of relationship between two terms. All of them measure the number of documents in which they occur separately, in comparison with the number of documents in which they co-occur. In our tests three well-known coefficients have been used (Salton and McGill, 1983):

$$\begin{aligned}\text{Tanimoto}(t_i, t_j) &= \frac{c_{ij}}{c_i + c_j - c_{ij}}\\ \text{Cosine}(t_i, t_j) &= \frac{c_{ij}}{\sqrt{c_i \cdot c_j}}\\ \text{Dice}(t_i, t_j) &= \frac{2 \cdot c_{ij}}{c_i + c_j}\end{aligned}$$

where c_i and c_j are the number of documents in which terms t_i and t_j occur, respectively, and c_{ij} is the number of documents in which t_i and t_j co-occur. Computing co-occurrence values for all terms we obtain a matrix, i.e. the association thesaurus. To construct the matrix we only used the first retrieved documents.

The aim of using the association matrix is to expand the entire query, not only separate terms. To expand the original query, terms with a high association value with all terms of the query must be selected. In the vector space model query q is represented with a vector, $\vec{q} = (q_1, q_2, \ldots, q_m)$, where q_i is the weight of the term t_i in the query. We use the measurement of scalar product with all query terms to obtain the terms with highest potential to add to the original query:

$$\mathrm{rel}(q,t_e) = \vec{q}^T * \vec{t_e} = (\sum_{t_i \in q} q_i \cdot \vec{t_i})^T * \vec{t_e} = \sum_{t_i \in q} q_i \cdot (\vec{t_i}^T * \vec{t_e}) = \sum_{t_i \in q} q_i \cdot \mathrm{ASS}(t_i, t_e) \tag{2}$$

The terms t_e with highest values of association obtained from (2) are added to the original query. Finally, only the weight of each term t_e remains to be determined. It seems natural to consider it according to (2), i.e. employing the sum of the weight of the original terms:

$$q_e = \frac{\mathrm{rel}(q,t_e)}{\sum_{t_i \in q} q_i}$$

A lot of tests were realized to obtain the better improvement, using the first 2, 5, 10, 25, 50 and 100 retrieved documents and all association coefficients. The results were discouraged. In all tests this expansion technique deteriorated retrieval performance. The reason is the weighting schema for topics: **ntc**, i.e., the vector of the query was normalized, and the added terms obtain higher weight than original terms.

We repeated the tests using **ntn** schema for topics. In this case we obtained better results. For French experiments we obtained improvement about 12% in GMAP, and about 4% in MAP. Hard topics had important improvements for the experiment with title and descriptions topic fields, but worsening for the experiment with title field. For Spanish experiments the improvement was about 8% in both MAP and GMAP, and also hard topics achieved effectiveness. For Italian we obtained little improvement, about 4% in both MAP and GMAP, but hard topics was made worse. For all these collections the best test was using 5 or 10 documents for the expansion, with Tanimoto (Jaccard) coefficient and adding between 5 to 40 terms to the original query. For the English collection no improvement was obtained in any test.

Local association thesauri with term windows

In essence this technique is identical to preceding one, but using term windows to obtain the co-occurrence relations, instead of complete document. In a document terms close to query terms must have higher relation value than other terms. In our tests we use several distances to setup the windows. If distance is 0 between two terms, both are adjacent. To compute the distance stop words are removed and sentence limits have no effects. A new term window is created when a query term appears in the document: around terms are included in the window depending on distance. If another query term appears inside the window, the window enlarges its right limit to contain other terms within distance. In this way some terms will co-occur with more query terms.

For all collections we have realized a lot of tests to obtain the best setups. For all collections the highest improvement using this technique was achieved using a distance value of 1 or 2, using the first 10 retrieved documents for local expansion, and adding about 40 terms to original query. With these setup, the English and Italian experiments had a improvement in retrieval performance about 4% in both MAP and GMAP. The improvement for the Spanish collection was about 10% in both measurements. For the French collection the improvement in MAP was only about 4%, but about 16% in GMAP. In all experiments the hard topics achieved improvements.

This technique was the best for our experiments, so we only submitted the corresponding runs.

2.2 Bilingual Retrieval IT→ES

Our CLIR system was the one used in monolingual retrieval. A previous step was carried out before searching to translate into Spanish the Italian topics. We use two MT programs: Power Translation Pro 7.0 and Wordlingo (`http:\\www.wordlingo.com`). The resulting translations were not post-edited. For each topic we joined the terms of the translations in a single topic: this is another expansion process, although in most cases the two translations were identical. Finally a monolingual retrieval was done. The local expansion with term windows was also applied.

2.3 Multilingual Retrieval ES→(EN ES FR IT)

The Spanish topics were translated into the language of each document collection, using these MT programs:

- English: Power Translator Pro 7.0, Systrans and Reverso.
- French: Systrans and Reverso.
- Italian: Power Translator Pro 7.0 and Wordlingo.

The resulting translations were not post-edited. For each topic and target language we joined the translation to obtain a single topic, and performed monolingual retrieval. The local expansion with term windows was also applied.

We also used the retrieval list of the monolingual Spanish experiment. This list and the ones for all target languages were joined using the MAX-MIN data fusion method. We consider a weighting factor of 1.02 for the Spanish list, and 1 for the rest ones.

3 Results

We only analyze results of our TEST runs, i.e., for the test topics of the robust task.

Monolingual Retrieval

English: Our 'reinaENtd' run (td stands for title and description topic fields) obtains a good position in the English subtask. The MAP for this run is over the mean and over the third quartile too. Our 'reinaENt' run (only title topic field) has the MAP measurement over the mean and between first and second quartile.

French: Our experiments ('reinaFRtd' and 'reinaFRt') achieve good positions in the ranking. The MAP for both runs are over the mean and about the third quartile.

Italian: The experiments for Italian had have apposite behavior. For 'reinaITtd' run the MAP is over the mean and the third quartile. The MAP of 'reinaITt' run is under the mean and is about the first quartile.

Spanish: Our 'reinaEStd' run gets the intermediate position of the subtask, with MAP over the mean. The other run 'reinaESt' is the worst run of the subtask, but is the only one with solely the title field.

Bilingual Retrieval

Our 'reinaIT2EStd' run is the top ranked run in this subtask. The run 'reinaIT2ESt' obtains worse position: MAP is under the mean and near the first quartile.

Multilingual Retrieval

In this subtask our mandatory title+description run, 'reinaES2mtd', is in the lower part of the ranking of runs. The other run, 'reinaES2mt' is the worst run of the subtask.

4 Conclusions

Local query expansion using association thesauri constructed with term windows is an effective and simple expansion technique, and can be easily implemented without tuning some parameters such as BRF. It is important to note that for thesauri contruction only the information from retrieved documents was used. We also used a simple document retrieval system based on vector space model, and we looked for a good document-query weigthing schema as the basis for the next expansion experiments. With these setups, our mandatory runs using local query expansion with term windows have obtained good positions in the monolingual subtask for all language we participate.

For the bilingual Italian to Spanish subtask, collecting terms from some translations of a topic is another query expansion technique that, combining with the previous we used in monolingual experiments, obtains performance improvement. Our mandatory run was the best run in the subtask.

For multilingual research we used the same procedure to obtain the retrieval list for each target language, and we combined them with the MAX-MIN data fusion procedure. In this subtask our mandatory run had not obtained a good position. We think that the reason is the ineffectiveness of the employed procedure for data fusion.

5 Acknowledgement

This work was partially supported by a Spanish Science and Education Ministery project grand, number EA2006-0080.

References

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New-York.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In Frei, H.-P., Harman, D. K., Schäuble, P., and Wilkinson, R., editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96. August 18–22, 1996, Zurich, Switzerland*, pages 21–29. ACM.

van Rijsbergen, C. J. (1979). *Information Retrieval.* Dept. of Computer Science, University of Glasgow.

Voorhees, E. M. (2003). Overview of the TREC 2003 robust retrieval track. In *The Twelfth Text REtrieval Conference (TREC 2003)*, pages 69–77. NIST Special Publication 500-255.

Voorhees, E. M. (2004). Overview of the TREC 2004 robust retrieval track. In *The Thirteen Text REtrieval Conference (TREC 2004), Gaithersburg, Maryland, November 16-19.* NIST Special Publication 500-261.

Voorhees, E. M. (2005). Overview of the TREC 2005 robust retrieval track. In *The Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, November 15-18.* NIST.

Zazo, Á. F. (2003). *Técnicas de Expansión en los Sistemas de Recuperación de Información.* PhD thesis, Departamento de Informática y Automática. Universidad de Salamanca.

Zazo, Á. F., Figuerola, C. G., Alonso Berrocal, J. L., and Rodríguez, E. (2005). Reformulation of queries using similarity thesauri. *Information Processing & Management*, 41(5):1163–1173.

Zazo, Á. F., Figuerola, C. G., Alonso Berrocal, J. L., and Rodríguez Vázquez de Aldana, E. (2002). Tesauros de asociación y similitud para la expansión automática de consultas. Algunos resultados experimentales. Technical Report DPTOIA-IT-2002-007, Departamento de Informática y Automática - Universidad de Salamanca.

Zazo, Á. F., Figuerola, C. G., Berrocal, J. L. A., Rodríguez, E., and Gómez, R. (2003). Experiments in term expansion using thesauri in Spanish. In *Advances in Cross-Language Information Retrieval. Third Workshop of the Cross-Languge Evaluation Forum, CLEF 2002, Rome, Italy. September, 2002 Revised Papers*, volume 2785 of *Lecture Notes in Computer Science*, pages 301–310. Springer.