UB at CLEF2004 Cross Language Medical Image Retrieval

Miguel E. Ruiz[1] and Munirathnam Srikanth[2]

[1] State University of New York at Buffalo,
School of Informatics, Dept. of Library and Information Studies,
534 Baldy Hall, Buffalo, NY 14260-1020 USA
meruiz@buffalo.edu
http://www.informatics.buffalo.edu/faculty/ruiz

[2] Language Computer Corporation Richardson, TX, 75080, USA
srikanth@languagecomputer.com

Abstract. This paper presents the results of the State University of New York at Buffalo in the cross-language medical image retrieval task at CLEF 2004. Our work in image retrieval explores the combination of image and text retrieval using automatic query expansion. The system uses pseudo relevance feedback on the case descriptions associated with the top 10 images to improve ranking of images retrieved by a CBIR system. The results show significant improvements with respect to a base line that uses only image retrieval.

1 Introduction

The cross language medical image retrieval task requires participants to retrieve information from a collection that includes medical images and physician's annotations, given an initial query that consists only of an image [1]. In this track our goal is to improve image retrieval by using retrieval feedback on the related case descriptions of the top n retrieved images to re-rank the final list of retrieved images. Because our statistical language model system (TAPIR) did not support retrieval feedback (which is a feature that was still under development by the time we worked on this task) we decided to use a version of the SMART retrieval system that we used in our participation in CLEF2003 [3].

Section 2 presents a description of our system for the medical image retrieval task. Section 3 discusses the details about document and image processing, indexing and query expansion. Section 4 presents our experimental results and analysis. The last section presents our conclusion and future work.

2 Combining Retrieval of Medical Images and Case Descriptions

Our goal in this task is to explore ways to expand the initial image retrieval with the multilingual text of the case descriptions associated to each image. For this

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 773–780, 2005.

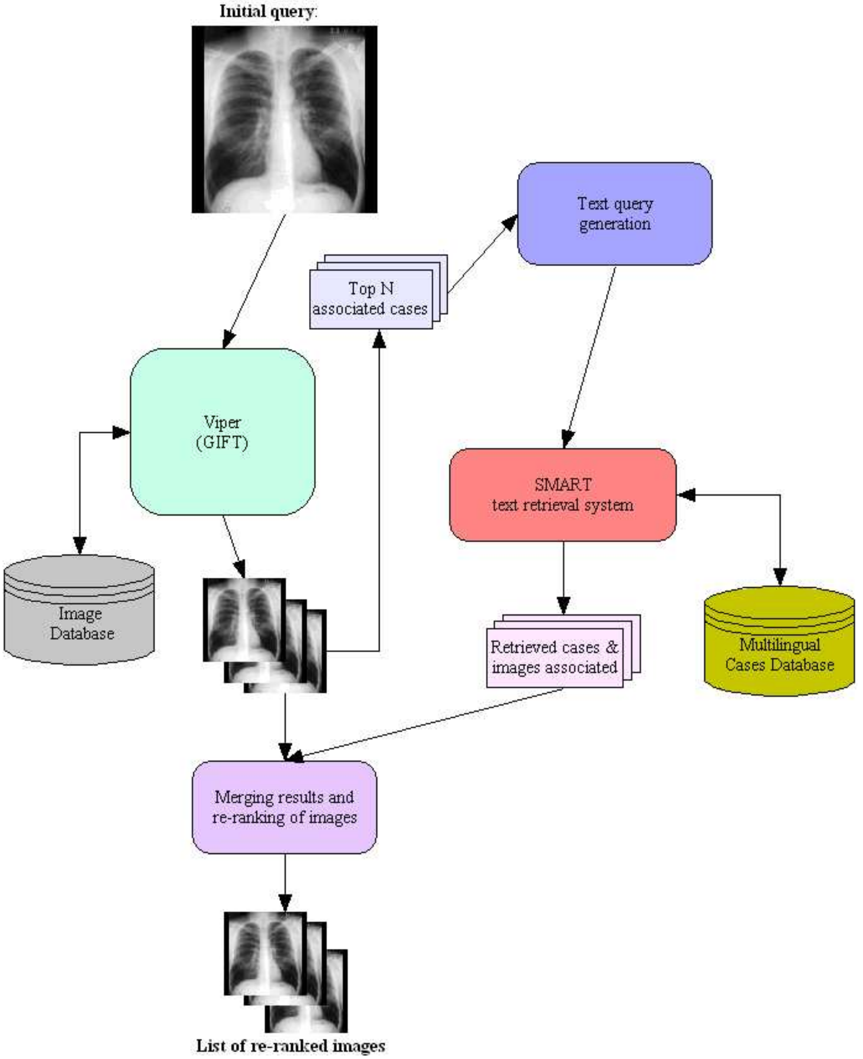
Initial query:
Text query generation
Top N associated cases
Viper (GIFT)
SMART text retrieval system
Image Database
Retrieved cases & images associated
Multilingual Cases Database
Merging results and re-ranking of images
List of re-ranked images

Fig. 1. Diagram of our text and image retrieval system

purpose we use a pseudo relevance feedback mechanism. The first step consists in performing retrieval using the database of images indexed. The top n images are used to locate the corresponding case descriptions. These case descriptions are used to build a query that is submitted to the text retrieval system to obtain other related case descriptions. Figure 1 presents an schematic design of this system.

Our system combines the Viper system [5], which is a publicly available image retrieval system developed at University of Geneva, and the well known SMART system [4], which is an information retrieval system deloped by G. Salton and his collaborators.

3 Collection Preparation and Indexing

The collection consists of $8,726$ images and $2,081$ cases descriptions that contain clinical information (details about the collection can be found in [1]. Our initial inspection of the data revealed that there were 209 cases that have images associated with them but no textual information. We discarded these cases from our experiments because they would not be suited for our evaluation. In consequence, our text collection consists of $1,872$ cases.

We used the list of images retrieved by *Viper*, which was supplied by the organizers of this track. Details about the preparation of this list of images can be found in [1]. For this reason, our efforts in pre-processing concentrated on manipulating the text descriptions associated with these images.

We decided to use almost all tags included in the documents with the exception of dates, URLs, and personal information from the patients (i.e. birth date, age, etc). The tags were classified and grouped into 9 types:

- Textual description: this includes fields such as title, description, commentary, questions, and answers.
- Diagnosis: The actual diagnosis associated to each case.
- Keywords and codes: This type includes keywords assigned to the case and radiology classification codes (ACR).
- Authors and organizations: Author, reviewer, hospital, department.
- Language.
- Orthopedic information: This includes all tags related to orthopedic annotations.
- Images: We added the list of image ids associated with each case.

Each of these types of information has its own characteristics that merit a different treatment during text processing and indexing. Our initial design creates a separate index for each type. The final score for ranking the retrieved cases is a weighted linear combination of each index score. Ideally, the weight of the contribution of each type should be determined experimentally. However, because we did not have a reliable way to estimate the contribution of each type to the final score of the document we decided to use the same weight for all parts.

Most of the case descriptions have a language field that indicates the language used in them. However, some case descriptions have no language specified in this filed or text in both languages (French and English) and we were not sure how often this occurred in the collection. We use a simple algorithm to estimate whether the actual language used in the document corresponds with the language assigned to the case in the language field. This algorithm identifies stop words in English and French and computes a score for each language based on the proportion of English and French stopwords present in the document. Through this process we found that $1,693$ cases were in French, 177 were in English and 16 cases have text in both languages (i.e. French description with English comments). Given the nature of these bilingual texts we decided to build a single retrieval index for all documents instead of separating them into two subcollections.

Our previous experience with medical documents has shown that using an aggressive stemming such as Porter's stemmer could reduce terms to roots that are actually quite different from their intended meaning. For example, "organization" is stemmed to "organ", which has a very different meaning from the original word. For this reason we use a simple stemming strategy that takes care only of plurals (in both English and French). We also used a stopword list that combined English and French stopwords and was manually reviewed to assure that it did not contain stop words that could have medical meaning (for example, the original stopword list from SMART includes "B" and "E" as a stop words, but if we discard this words it would be difficult for the system to distinguish between articles that talk about "vitamin B" and "vitamin E").

Indexing of the case descriptions was performed using a version of the SMART system adapted to handle the ISO-latin-1 encoding in our CLEF 2003 work [3]. The documents were indexed using *atc* weighting (augmented term frequency, idf, and cosine normalization) while the queries used *atn* weighting (augmented term frequency, idf, no normalization).

3.1 Query Expansion

Our retrieval approach follows a classical pseudo relevance feedback method. The initial image is send as a query to *Viper* and the top ten images retrieved are used to build a query for the textual database. Our initial text query consists of the image ids of the top ten images retrieved (Note that we have added the list of image ids related to each case). We perform an initial retrieval step using these queries and retrieve the top 1000 cases. The top n cases are marked as relevant while the bottom 100 cases are marked as non relevant. This information is used to obtain terms to expand the original query. The query expansion step uses Rocchio's formula to compute the weight of each of the terms as follows:

$$Q_{new} = \alpha \times Q_{orig} + \beta \times \frac{\sum_{D \in Rel} D}{R} + \gamma \times \frac{\sum_{D \ni Rel} D}{N - R} \tag{1}$$

Terms are ranked according to Rocchio's score and the top m terms are selected for expansion. We tried several values for the number of cases assumed to be relevant after the initial retrieval (n = 5, 10, 20) and for the number of terms used to build the expanded query (m = 20, 50, 100). Since we were not sure whether the usage of the original image ids would be important or not to the final retrieval we decided to use two different values for the coefficient α: 0 (don't take into account these original terms) and 1. The second coefficient (β) of the Rocchio's formula controls the contribution of the relevant documents. We set it to 64 because this is the most important information that will allow us to expand the query. The third coefficient γ controls the penalty assigned to terms that appear in the "non-relevant" documents (bottom 100 cases retrieved in the initial retrieval) and was set to 16. In summary, we tried two different sets of coefficients for the Rocchio expansion formula ($\alpha = 1$, $\beta = 64$ and $\gamma = 16$)

Table 1. Top 30 terms generated by the query expansion method for the first image query

Weight	Term	Weight	Term
0.28935	im10654	0.08574	iliite
0.27833	im10361	0.08092	pied
0.26395	im11040	0.07697	acr33.3320
0.26294	im11114	0.07697	acr44.3320
0.25794	im10945	0.07697	im10362
0.25652	im10170	0.07212	l'èvolution
0.25585	im9832	0.07180	dèmasquage
0.25585	im9833	0.06919	sènile
0.25585	im9835	0.06819	kindyni
0.23936	im10916	0.06573	psoriasi
0.13769	sacro	0.06572	patiente
0.11502	bassin	0.06379	toutefoi
0.09297	iliaque	0.06258	im11042
0.08798	acr44.562	0.06258	im11041
0.08798	im10655	0.06169	collection

and ($\alpha = 0$, $\beta = 64$ and $\gamma = 16$). An example of the expanded query is shown in Table 1.

The expanded query is then submitted to the text retrieval system and the score of each retrieved case is assigned to the images associated with it. A final score for each image was computed by combining the scores obtained from the image retrieval system and the text retrieval system. We use a linear combination of the scores to compute the final image score:

$$W_k = \lambda Iscore_k + \delta Tscore_k \tag{2}$$

where $Iscore_k$ and $Tscore_k$ are the scores assigned to the image k by the image retrieval system (Viper) and text retrieval system (SMART) respectively, λ and δ are coefficients that weight the contribution of each score. Usually the coefficients are estimated from experimental results. However, due to the lack of training data we decided tu use $\lambda = \delta = 1$ (observe that this simple addition of scores is possible due to the fact that both scores are scaled between 0 and 1).

4 Analysis of Results

We submitted three runs. The first run (UBMedImTxt01) used the top 10 documents to expand the query with the top 100 terms ranked by Rocchio's formula with coefficients $\alpha = 1$, $\beta = 64$ and $\gamma = 16$. This is a run that uses an aggressive expansion strategy and takes into account the image ids of the top ten images retrieved by *Viper* as actual terms. The second run (UBMedImTxt02) differs from the first run in the fact that the coefficient $\alpha = 0$ disregards the image ids as actual query terms. The third run (UBMedImTxt03) uses a more conservative strategy for expansion with only the top 5 cases and coefficients $\alpha = 1$, $\beta = 64$ and $\gamma = 16$.

Our official results are presented in Table 2. The performance of the system shows a positive impact in improving relevance of the images retrieved. The best run UBMedTxt01 performed above the median in all queries and obtained the best performance of all official runs in automatic query construction (note that the difference between the top 5 systems is very small and would not be statistically significant). Our second run (UBMedImTxt02) performs 5.3% below our best run and performs above the median in 20 queries. Observe that the only difference between these two runs is that we use the ids of the images as actual terms for query expansion. These image ids work as anchors that reinforce the notion that cases that those cases, which have images associated with the assumed top 10 retrieved images, are regarded as relevant in our initial retrieval. Our third run (UBMedImTxt03) performs 4.7% below the best run and performs above the median systems in 20 queries. This third run uses a more conservative query expansion assuming that only the top 5 retrieved cases are relevant and perform query expansion.

Table 2 also includes the performance for our baseline system that corresponds to the list of images retrieved by Viper. The performance of our baseline is 0.3502. Our best run performs 11.5% above the baseline and this difference is statistically significant.

Table 2. Performance of official runs in Medical Image Retrieval

		diff with	number of queries	
	AvgP	Baseline	> median	> baseline
UBMedImTxt01	0.3904	11.5% **	26	19
UBMedImTxt02	0.3696	5.5%	20	14
UBMedImTxt03	0.3722	6.2%	20	16
Baseline	0.3502	–	–	–

We have to note that improvements to the final performance of the expanded queries are highly dependent of the quality of the initial set of images retrieved by the CBIR system. Figure 2 shows that there is a strong correlation between the performance of the expanded queries and the original initial retrieval using only the image retrieval system. This figure also shows a line that represents the performance of the baseline system. The points above this line are queries that have improved performance after the pseudo-relevance feedback of image and text.

Observe that because we have indexed French and English documents as a single collection the expanded query actually includes terms in both languages. A different approach could be to perform query expansion in two separate collections and then merge the results in a single list. Another approach could identify the language of a terms and add the corresponding translation. However, this will require the use of a specialized bilingual lexicon.

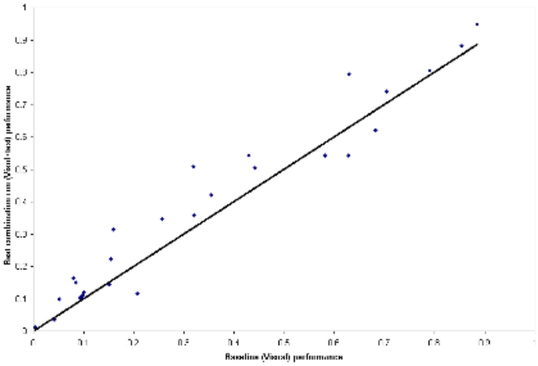

Fig. 2. Query by query comparison of best visual+text run vs visual baseline

We asked two physicians (an specialist in pneumonology and an urologist) to help us validate the results of the retrieved images (although we did not use this feedback to change the ranking of the images or the way the system processed the queries). We asked them to give general feedback to understand whether the results retrieved by the system would make sense to a medical professional. This helped us to realize that some of the aspects of how a medical professional could use this type of system in their daily work. They also emphasized that the actual diagnostic of a patient is usually a complex process that includes not only the review of images but also the analysis of the clinical data that in many cases is more indicative of a specific diagnosis than the image itself. This seems to be corroborated by the fact that adding the text description of the actual cases associated to the image makes a significant difference.

5 Conclusions and Future Work

We can conclude that our experiments confirm the hypothesis that pseudo-relevance feedback on the case descriptions associated with the medical images can be used to improve performance of a CBIR system.

Our method for preprocessing the actual structure of the cases have to be refined but it seems to work well for retrieval purposes.

We plan to add some extra query expansion using the UMLS Metathesaurus produced by NLM to add related medical phrases to the cases and verify whether this would actually improve performance.

Acknowledgement

We would like to thank Doctors Freddy and Adriana Karausch for their help in validating the results retrieved by the system and giving positive feedback that we will use for making general improvements to our system.

References

1. Clough, P. and Müller, H. and Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In Peters, C. and Clough, P. and Gonzalo, J. and Jones, G. and Kluck, M. and Magnini, B. (Eds.) Fifth Workshop of the Cross Language Evaluation Forum (CLEF 2004) Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print), 2005.
2. Rocchio, J. J.: Relevance feedback in information retrieval. In G. Salton (Ed.) The SMART Retrieval System: Experiments in Automatic Document Processing (pp.313.323). Englewood Cliff, NJ, Prentice Hall, 1971.
3. Ruiz, M. E.: Automatically generated phrases and relevance feedback for improving Cross-Language Information Retrieval. In Proceedings of the Cross Language Evaluation Forum (CLEF) 2003, August 21st . 22nd, Trondheim, Norway, 2003.
4. Salton, G. (Ed.): The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliff, NJ, Prentice Hall, 1971.
5. Viper Research Group URL: viper.unige.ch