Pattern-Based Image Retrieval with Constraints and Preferences on ImageCLEF 2004*

Maximiliano Saiz-Noeda, José Luis Vicedo, and Rubén Izquierdo

Departamento de Lenguajes y Sistemas Informáticos,
University of Alicante, Spain
{max, vicedo, ruben}@dlsi.ua.es

Abstract. This paper presents the approach used by the University of Alicante in the ImageCLEF 2004 adhoc retrieval task. This task is performed through multilingual search requests (topics) against an historic photographic collection in which images are accompanied with English captions. This approach uses these captions to perform retrieval and is based on a set of constraints and preferences that allow the rejection or scoring of images for the retrieval task. The constraints are implemented through a set of co-occurrence patterns based on regular expressions and the approach is extended in one of the experiments with the use of WordNet synonyms.

1 Introduction

Bilingual ad hoc retrieval is one of the tasks defined within the ImageCLEF 2004 campaign [1] as part of the Cross Language Evaluation Forum (2004). The objective of this task, celebrated since last 2003 campaign [2], is to retrieve relevant photographic documents belonging to a historic photographic collection in which images are accompanied with English captions. These photographs integrate the *St Andrews photographic archive* consisting of 28,133 (approximately 10% of the total) photographs from *St Andrews University Library photographic collection* [3].

The method followed to retrieve relevant images is based on three experiments where a set of preferences and constraints are applied. The constraints, based on a set of co-occurrence patterns will reject potentially incompatible (non-relevant) images related to the query. Preferences will score the images in order to give a list according to their degree of relevance. Furthermore, a Wordnet-based query expansion is tested.

This is the first time that the University of Alicante has participated in this specific task and the main objective in the starting premise is to make a simple and low cost approach for this kind of search task.

* This work has been partially supported by the Spanish Government (CICYT) with grant TIC2003-07158-C04-01.

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 626–632, 2005.

The next sections describe specific characteristics of the dataset, relevant for the retrieval process, and the strategy used by the University of Alicante's team in order participate in the forum. Finally, some evaluation results will be discussed and some future improvements to the system will be presented.

2 Photographic Dataset

As mentioned, the photographic dataset used for the ImageCLEF 2004 ad hoc evaluation is a collection of 28,133 historical images from *St Andrews University Library photographic collection*. Photographs are primarily historic in nature from areas in and around Scotland; although pictures of other locations also exist.

All images have an accompanying textual description consisting of a set of fields. In this approach, we have used a file containing all image captions in a TREC-style format as detailed below:

```
<DOC>
 <DOCNO>stand03_2096/stand03_10695.txt</DOCNO>
 <HEADLINE>Departed glories - Falls of Cruachan Station above Loch
 Awe on the Oban line.</HEADLINE>
 <TEXT>
  <RECORD_ID>HMBR-.000273</RECORD_ID>
  <SHTITLE>Falls of Cruachan Station.</SHTITLE>
  <DESCRIPTION>Sheltie dog by single track railway below embankment,
  with wooden ticket office, and signals; gnarled trees lining
  banks.</DESCRIPTION>
  <DATE>ca.1990</DATE>
  <PHOTOGRAPHER>Hamish Macmillan Brown</PHOTOGRAPHER>
  <LOCATION>Argyllshire, Scotland</LOCATION>
  <NOTES>HMBR-273 pc/ADD: The photographer's pet Shetland collie
  dog, 'Storm'.</NOTES>
  <CATEGORIES>[tigers],[Fife all views],[gamekeepers],[identified
  male],[dress - national],[dogs]</CATEGORIES>
  <SMALL_IMG>stand03_2096/stand03_10695.jpg</SMALL_IMG>
  <LARGE_IMG>stand03_2096/stand03_10695_big.jpg</LARGE_IMG>
 </TEXT>
</DOC>
```

The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English, although the language also contains colloquial expressions. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are in black and white, although colour images are also present in the collection.

The type of information that people typically look for in this collection include the following: Social history, e.g. old towns and villages, children at play and work. Environmental concerns, e.g. landscapes and wild plants. History of photography, e.g. particular photographers. Architecture, e.g. specific or general places or buildings. Golf, e.g. individual golfers or tournaments. Events, e.g. historic, war related. Transport, e.g. general or specific roads, bridges etc. Ships and shipping, e.g. particular vessels or fishermen.

Although all these fields can be used individually or collectively to facilitate image retrieval, in this approach only a few of them have been used. In particular, fields related to the photographer, location and date (apart from the headline) have been selected for the retrieval.

3 A Description of Our Technique

As it is the first time this group has participated in this task, we decided to make use of a naive approach with the smallest possible quantity of resources and implementation-time required. So, this technique does not use any kind of indexing, dictionary or entity recognition and makes use of a single POS tagging approach. Nevertheless, within the three experiments, improvements of the method includes the use of co-occurrence patterns and WordNet for query expansion.

Figure 1 shows the process followed by the system. This figure includes three steps related to the three experiments carried out for the evaluation that will be detailed below. To apply this basic strategy to retrieval, it is necessary to create files with questions and images. As mentioned, the file with the whole set of images in TREC format has been used for retrieval.

Constraints and preferences applied to the retrieval process make use of morphological information. Furthermore, the retrieval process is based on word lemmas. This means POS tagging of both the question and the image files is necessary. This POS tagging has been performed using the TreeTagger analyzer [4]. For the retrieval process itself, a file of stop words have been used in order to eliminate unhelpful words and improve speed of the system.

In order to cope with multilingual retrieval, we use a translation method to perform query translation. In concrete terms, the Babelfish [5] Machine Translation (MT) tool is used. This resource has allowed us to test the system with topics in German, Chinese, Spanish, French, Dutch, Italian, Japanese and Russian. Following translation into English, all languages are treated equally thereafter. According to the information required for retrieval, three different experiments have been carried out:

1. Preferences-based retrieval
2. Constraints and Preferences-based retrieval
3. Constraints and Preferences-based retrieval with question expansion

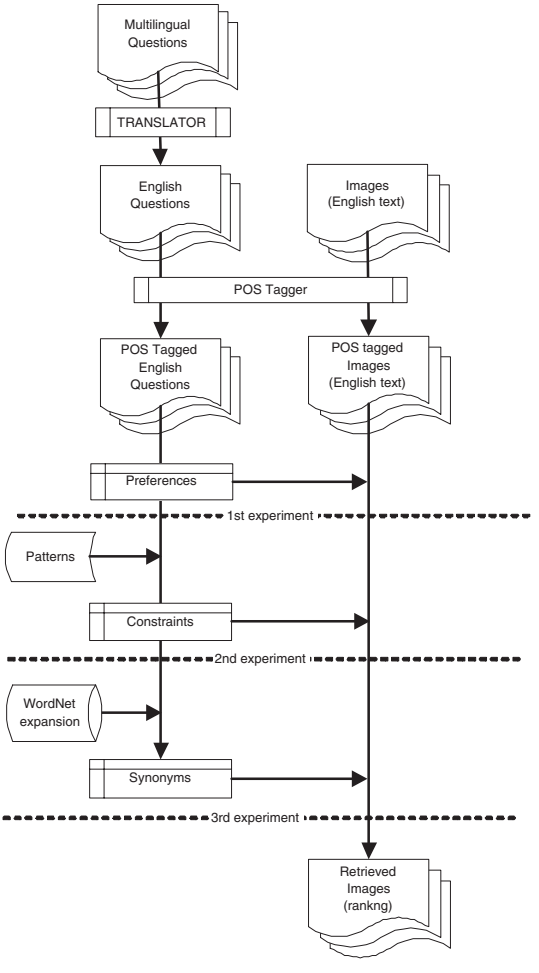

Fig. 1. Image retrieval process for different experiments

At the beginning of the process, all the images are suggested as a solution for each question[1]. From this scoring, some images will be added.

3.1 Experiment 1 - Preferences

For applying preferences, a single word matching between the question and the *HEADLINE* field of the image is used. This experiment is used as a baseline and its main function, as discussed later, is to determine the effects of additional information added in other experiments on the retrieval process.

[1] This condition is guided by the idea of giving 1000 images for each question, what constitutes a misunderstanding of the evaluation process and will be discussed bellow as an evaluation handicap.

For scoring in this experiment, we have assumed the relevance of proper nouns, nouns and verbs. So we have scored following this order when matching is related to these elements. Furthermore, if applicable, relations between nouns and verbs with the same lemma are also scored (if we are looking for "golfers swinging their clubs" probably we are interested in a "golfer's swing").

It seems almost obvious that good performance of this technique should be based on good entity recognition that ensures the correct detection of proper nouns in the question and in the image text. Probably, as it will be discussed later, using a better named-entity recognizer would improve the overall performance of this experiment.

3.2 Experiment 2 - Constraints and Preferences (The Patterns)

This experiment makes use of the previously described preferences and integrates constraints as a new selection criterion. The main aspect of the constraint is that it should be a rule strong enough to reject an image based on a compatibility guideline. This rule is built through the definition of a set of co-occurrence patterns that establish rejecting rules related to three of the fields contained in the image information: *DATE*, *PHOTOGRAPHER* and *LOCATION*.

These patterns are applied to the question (topic) and generate an XML-style file with information provided by the patterns. For example, topic:

```
1. Portrait pictures of church ministers by Thomas Rodger
```

is converted into the file:

```
<PREG>
 <PREGNO>1</PREGNO>
 <HEADLINE> Portrait pictures of church ministers by Thomas
  Rodger</HEADLINE>
 <DATE> </DATE>
 <PHOTOGRAPHER> by Thomas Rodger </PHOTOGRAPHER>
 <LOCATION> </LOCATION>
</PREG>
```

where labels $< DATE >$, $< PHOTOGRAPHER >$ and $< LOCATION >$ contain all information extracted about these information items.

The patterns are built over regular expressions that allow the extraction of a string contained in any of the mentioned labels. *DATE* constraints try to reject images by comparing not only question and image years, but applying extra information such as months or quantifiers. This way, if the topic asks for "Pictures of Edinburgh Castle taken before 1900", all the photos taken after 1900 will be discarded. *PHOTOGRAPHER* constraints are based in the whole name of the photographer. *LOCATION* constraints use not only the location itself but also, if applicable, possible relations with other locations (city, country, ...).

As can be seen, this technique is very general and, therefore, the possibility of error is also high. To reduce the possibility of errors, the strategy also uses

statistical information from the image corpus. This way, matches that are incorrectly treated by the pattern as photographers or locations are considered as noise and rejected because of their low or null appearance frequency in the corresponding field in the image caption. In fact, we can use the same pattern for both location and photographer and then decide what to apply depending on the image. For example, according to the image captions, a capitalized word after a comma can be consider both a photographer or a location (as shown in topics "Men in military uniform, *George Middlemass Cowie*" and "College or university buildings, *Cambridge*"). After including the extracted string in both fields of the topic generated, the statistical information will determine what is a photographer and what is a location (unless there is a town called George Middlemass or a photographer called Cambridge).

Once the constraint features are determined and included in the topic through their corresponding labels, the system makes a matching task to reject non-compatible images. For example, if it determines that the photographer of the searched pictures is "Thomas Rodger", all the images that don't contain "Thomas Rodger" (or any of its parts) in the *PHOTOGRAPHER* label are rejected.

3.3 Experiment 3 - Query Expansion Using Wordnet

The last experiment has been designed to incorporate extra information regarding potential synonym relations between terms in the query and image. In this case, the system expands the topic with all noun and verb synonyms contained in WordNet [6].

Using this the scoring for each image is increased not only if a lemma of a word appears in the topic, but also if its synonyms from WordNet also appear in the image *HEADLINE* text. Due to there being no lexical disambiguation in the process, noun synonyms are best scored than verb synonyms assuming that the former tend to be less generic than the latter. If the synonym is found but with different POS tag, a smaller score is added.

4 Evaluation

Although we knew this to be a very general approach to this task, the results obtained after the evaluation of the system are not as successful as desired. At the moment of the writing of this paper we are trying to determine if there is any kind of computing processing mistake that has affected the final scoring. Anyway, there are some considerations extracted from the results.

For the evaluation results, the system was prepared to always provide 1000 images as output. This is an error because some images given by the system are not relevant at all (they have no specific nor score).

Another problematic issue is the way the system scores the images. This scoring is also very general and often generates the same score for a large number of images (in fact, all the images can be grouped in four or five different scores). All the images that are equally scored have, for the system, the same order in

the final evaluation scored list. For the evaluation, comparing results we see that there are big differences depending on the order of the images.

Related to the results of the three experiments, one of the most "eye-catching" things is that, in general, the preferences-baseline experiment gives the best result or is improved in a very small degree by the rest of experiments. This situation can be put down to the lack of additional information regarding named-entity and recognition of proper nouns.

Another interesting observation from the evaluation is that although there are not large differences between the monolingual and the bilingual results, it is clear that automatic translation (such as the method used in these experiments) introduces errors and noise which ultimately decreases system performance. Furthermore, basic techniques of lexical disambiguation and restricted-domain ontologies could improve the use of WordNet.

In summary, although the results are not very good, the system itself presents many possibilities for improvement through the refinement of the scoring system, the addition of new techniques based on named-entity recognition, the use of better translation resources and dictionaries and the incorporation of new semantic and ontological information that enforces WordNet access.

5 Conclusions

In this paper we have described the system carried out by the University of Alicante in the ImageCLEF 2004 adhoc retrieval task. Information about the process itself, the strategies and experiments developed for the retrieval task have been given. The results of the evaluation have been justified and different solutions to improve these results have been outlined in order to define future improvements to obtain a better retrieval system.

References

1. www: Image CLEF in the Cross Language Evaluation Forum 2004. http://ir.shef.ac.uk/imageclef2004/index.html (2004) Last visited 2-Aug-2004.
2. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Working Notes for the CLEF 2003 WorkShop, Trondheim, Norway (2003) 379–388
3. www: St Andrews University Library photographic collection. http://special collections.st-and.ac.uk/photcol.htm (2004) Last visited 2-Aug-2004.
4. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing, Manchester, UK (1994) 44–49
5. www: BabelFish translator. http://world.altavista.com/ (2004) Last visited 2-Aug-2004.
6. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Five Papers on WordNet. Special Issue of the International Journal of Lexicography **3** (1993) 235–312