A Textual Approach Based on Passages Using IR-n in WikipediaMM Task 2008

Sergio Navarro, Rafael Muñoz, Fernando Llopis
Natural Language Processing and Information Systems Group
University of Alicante, Spain
snavarro,rafael,llopis@dlsi.ua.es

Abstract

In this paper we have focused our efforts on comparing the behaviour of two relevance feedback methods in this task - LCA and PRF - and in checking if our passage based information rerieval (IR) system is useful in a competition with small sized documents. Furthermore we have added an adaptation to this domain based on decompound in single terms those file names which use a Camel Case notation. We base our decision on the belief that the most meaningful information of an image file appointed by a human is on the file name itself. Thus, it is important to make visible this terms when they are hidden in a compounded file name. Finally we have added a geographical query expansion and a visual concept expansion. We have obtained a 29th place within a total of 77 runs with our baseline run - which only used the passage IR system -, and a 3rd place obtained with our best run - which used the passage IR system with Camel Case decompounding -. It shows us on one hand the usefulness of our passage based IR system in this domain, and on the other hand it confirms our belief in the existence of specially meaningful information within the file names. In the the relevance feedback respect, we have obtained contradictory results about the suitability of LCA or PRF to the task, but we have found that LCA has a more robust behavior than PRF.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2 [**Database Managment**]: H.2.5 Heterogenous Databases

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Image Retrieval, Geographic Expansion, Relevance Feedback, PRF, LCA, Camel Case decompounding

1 Introduction

This is the first time we are participating in the WikipediaMM task. We had experience of participation in the photo retrieval task of 2007 [8]. Our participation in ImageCLEFphoto [1]

[1] *http://www.imageclef.org*

involved the use of an information retrieval system based on passages. We analyzed the suitability of our system for the short text annotations related to the images in the collection. We concluded that our system improved the results more in comparison to other systems, which were similar to our system except the fact that they did not use passages. The experiments also showed that relevance feedback is a good tool for improving results.

Since our last participation in ImageCLEF our efforts has been focused on finding an alternative strategy for the relevance feedback. So far, the most common strategy between the participants of ImageCLEFphoto task on last year edition was to use PRF [9]. Thus, we are comparing in this CLEF edition PRF with Local Context Analisy (LCA) [10], as alternate strategy.

Thus, in our participation in the WikipediaMM task we wanted to figure out if good results can be also achieved with our passages based system for this domain.

Furthermore, in order to adapt our system to this task we have added a module to preprocess the wikipedia documents. It skips the useless tags and decompose the file names with camel case notation in single terms.

Finally, due to it is so common to see queries that are looking for an image that involve a geographical term. We have added a geographical query expansion module in order to experiment its behavior in this task.

This paper is structured as follows: Firstly, it presents the main characteristics of the IR-n system focusing on the relevance feedback strategies, the geographical query expansion and the Wikipedia processing module, then it moves on to explain the experiments we have made to evaluate the system, and finally it describes the results and conclusions.

2 The IR-n System

In our approach, we used IR-n - an information retrieval system based on passages -. Passage-based IR systems treat each document as a set of passages, with each passage defining a portion of text or contiguous block of text. Unlike document-based systems, these systems can consider the proximity of words with each other, that appear in a document in order to evaluate their relevance [6].

The IR-n passage-based system differs from other systems of the same category with regard to the method proposed for defining the passage - that is - using sentences as unit. Thus, passages are defined by a number of consecutive sentences in a document [6].

IR-n uses stemmer and stopword lists to determine which information in a document will be used for retrieval. For a list of stemmers and stopwords used by IR-n, see www.unine.ch/infor/clef.

IR-n uses several weighting models. Weighting models allow the quantification of the similarity between a text (a complete document or a passage in a document) and a query. Values are based on the terms that are shared by the text and query and on the discriminatory importance of each term.

2.1 Relevance Feedback

Most IR systems use relevance feedback techniques [1]. These systems usually employ local feedback. The local feedback assumes that top-ranked documents are relevant. The added terms are, therefore, common terms from the top-ranked documents. Local feedback has become a widely used relevance feedback technique. Although, it can deter retrieval, in case most of the top-ranked documents are not relevant, results in TREC an CLEF conferences show that is an effective technique [10]. In fact, almost all the systems that participated at ImageCLEF 2007 used Probabilistic Relevance Feedback (PRF) - the most common relevance feedback method - [4] [5] [3] [2].

In the selection of terms, PRF gives more importance to those terms which have a higher frequency in the top relevant documents than in the whole collection. An alternative query expansion method relies on the Local Context Analysis (LCA), based on the hypothesis that a common term from the top-ranked relevant documents will tend to co-occur with all query terms within the top-ranked documents. That is an attempt to avoid including terms from top-ranked,

non-relevant documents in the expansion. Furthermore, in the case of polysemus words, this method will help to retrieve documents more related to the sense of the query, since it is logical to think that the user will use words from the domain associated with this sense to complete the query.

The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents.

2.2 Geographical Query Expansion

To carry out the geographical terms expansion, our system first determines which are the geographical terms of the query. And afterwards generates the terms expansion based on the features of the term found.

2.2.1 Selecting Geographical Terms

During the search phase using Freeling [2] a language analysis tool suite the system extracts from the query the names and adjectives with their Wordnet [3] most frequent sense. With this information, the Semantic Domains [7] associated to the sense of each term are obtained.

The system consider as geographical terms of the query the terms that pertain to the 'administration' or 'geography' dominions and that are hyponym of one of these concepts: 'location' or 'landmass' for toponyms, 'nationality', 'asian', 'european', 'australian', 'american', 'african', 'person of color' for demonyms 'language' for names for languages.

For instance, consider the following query from the ImageCLEF07: "Asian women and/or girls". 'Asian' term pertains to the Semantic Domain 'geography', and it is an hyponym of 'person of color', our system classify it as a demonym.

2.2.2 Terms Expansion

The toponyms are expanded using WordNet with their synonyms, their direct holonyms and their hierarchical meronyms. And the names of inhabitants or languages are expanded with WordNet synonyms and hierarchical meronyms of their direct pertainyms concept.

The system uses a little stopword list for the geographical expansion to determine which of the generated terms by the expansion will be skipped. These terms are so common between locations and are not relevant to distinguish different locations.

Following the sample of the query "Asian women and/or girls". How 'Asian' term has been classified as a demonym, the system obtain its pertainyms concept, which is 'Asia'. The expansion terms are the hierarchical meronyms of the 'Asia'term.

2.3 Wikipedia Preprocessing Module

In order to use significant information in the IR-n indexing phase of the collection, our preprocessing module skips the useless information and optionally it can decompose the compounded file names in single terms with meaning.

Specifically, for the IMAGE tag the preprocessing module skips the data which it contain, and for the other tags it only uses their text nodes. Optionally, the decompounding function can be activated. It decompose the file names which use camel case notation in single terms. An example of a document before and after the preprocessing phase can be seen at Figure 2 and Figure 1 respectively. We can observe in this example how the terms "EgiptyanDesert" that are joined in a file name are decomposed, making this document more visible to a textual query based on natural language.

[2] available at *http://garraf.epsevg.upc.es/freeling*
[3] available at *http://wordnet.princeton.edu*

3 Training

IR-n is a parameterizable system, which means that it can be adapted in line with the concrete characteristics of the task at hand. The parameters for this configuration are the number of sentences that form a passage, the weighting model to use, the type of expansion, the number of documents/passages on which the expansion is based, the average number of words per document, and the use of geographical query expansion. This section describes the training process that was carried out in order to obtain the best possible features for improving the performance of the system. The collections and resources are described first, and the next section describes specific experiments.

3.1 Data Collection

In 2008, ImageCLEF wikipediaMM have used the image collection created and employed by the INEX Multimedia (MM) Track (2006-2007). This collection contains approximately 150,000 images that cover diverse topics of interest. These images are associated with unstructured and noisy textual annotations in English. Table 1 shows the characteristics extracted from the textual annotations in the collection using IR-n splitter after preprocessing it with the wikipedia preprocessing module - without Camel Case decompounding -.

Table 1: Data Collection

NoDocs	**WDAvg**	**WD Max**	**SentAvg**	**Sent Max**	**Language**
151.519	20,03	3.101	1,7	784	English

Bellow is a descriptions of Table 1 columns:

- **NoDocs**: is the number of documents.
- **WDAvg**: is the average of words by document.
- **WD Max**: is the maximum number of words in a document.
- **SentAvg**: is the average of sentences by document.
- **Sent Max**: is the maximum number of sentences in a document.
- **Language**: is the language of the collection.

Each image is associated with user-generated alphanumeric, unstructured metadata in English. These metadata usually contain a brief caption or description of the image,the Wikipedia user who uploaded the image, and the copyright information. These descriptions are highly heterogeneous and of varying length. The Figure 1 provides a metadata example associated to an image.

The topics for the 2008 ImageCLEF wikipediaMM task on one hand have included the topics previously used in INEX MM and ImageCLEF photo tasks, and on the other hand topics created by this year's task participants.

The topics are multimedia queries that can consist of a textual, a visual and a conceptual part, with the latter two parts being optional. The tags that compound a topic are the TITLE, which contains the query by keywords, the CONCEPT, which contains query by one or more visual concepts - optional -, and the IMAGE, which contains query by one or more images - optional -.

For the training of systems, the organization distributed between the participants the topics used in INEX MM 2006 and 2007 editions and their relevance assessments.

Figure 1: Metadata File Example

```
<?xml version="1.0"?>
<article><name id="103136">EgyptianDesert.JPG</name>
<image xmlns:xlink="http://www.w3.org/1999/xlink"
xlink:type="simple" xlink:actuate="onLoad"
xlink:show="embed" xlink:href="../pictures/EgyptianDesert.JPG"
id="103136" part="images-20000">
EgyptianDesert.JPG</image>
<text><wikitemplate parameters="2">
<wikiparameter number="0"><value>PD-user</value></wikiparameter>
<wikiparameter number="1" last="1"><value>Manos</value>
</wikiparameter></wikitemplate>
brrrrrrrrrrrrrrkjfhdikfhas
</text></article>
```

Figure 2: Processed Metadata File Example

```
<DOC><DOCNO> 103136 </DOCNO>
Egyptian Desert .
PD user Manos brrrrrrrrrrrrrrkjfhdikfhas</DOC>
```

3.2 Experiments

The experiment phase aims to establish the optimum values for the configuration of the system. Below is a description of the input parameters of the system used in this task:

- **Passage size (PS)**: Number of sentences in a passage.
- **Weight model**: We used DFR weighting model. We have based this decision on the good results obtained with this model for English language on our last participation in ImageCLEF.
- **Query expansion parameters**: we can use relevance feedback based on passages or based on documents. Moreover we have to select the number of passages or documents that the expansion will use, and indicate the k terms extracted from the best ranked passages or documents from the original query.
- **Geographical query expansion (Geog)**: Indicate if the system use the geographical query expansion module.
- **Camel Case**: Indicate if the system use the collection with camel case terms decompounded.
- **Concept**: Indicate if the system has to join the terms that represent the visual concepts of the topic to the topic title in order to compound the query. Or if it only has to use the TITLE keywords for the query.

For this phase we have carried out all the experiments separately for the two training query sets - INEX MM 2006 and 2007 -. Our objective was to find which parameter values are better independent of the query set used. Thus, we have looked for common parameter values between best runs which have used 2006 query set and best runs which have used 2007 query set.

Due to the great number of configurations used for this training. We have had to made an effort for synthesizing the results obtained, in order to find common issues of the best runs. We

have focused on the comparison of the best results obtained for each combination of Camel Case, Concept and Geog parameters. Moreover, in order to analyze the suitability of LCA an PRF to the task, we have showed the best results for each Camel Case, Concept and Geog combination using LCA or PRF as relevance feedback strategy.

Next, there is a description of the fields compounding the result tables:

- **rk**: Shows the ranking position of a combination of Camel Case, Concept and Geog parameter values. It is based on the maximum MAP measure obtained.
- **rk [no fb|prf|lca]**: Shows the ranking position based on MAP for the best run that have used LCA/PRF or that not have used relevance feedback (NOFB).
- **ps [no fb|prf|lca]**: Shows the passage size of the best run that that have used LCA/PRF or that not have used relevance feedback (NOFB).
- **map [no fb|prf|lca]**: MAP or Mean Average Precision of the best run that have used LCA/PRF or that not have used relevance feedback (NOFB).
- **[lca|prf] impro**: Shows the percentage of improvement using LCA or PRF regarding not use relevance feedback.
- **maIm**: Mean average of percentage of improvement - for $lcaimpro$ and $prfimpro$ -.

The following tables - Table 2 and Table 3 - shows the best results obtained in the training phase, their data is presented in the increasing order of rk value.

Table 2: Query Set 2006 - Best Results

rk	**cam case**	**vis con**	**geo**	**rk no fb**	**ps no fb**	**map no fb**	**rk lca**	**ps lca**	**map lca**	**rk prf**	**ps prf**	**map prf**	**% lca impro**	**% prf impro**
1	no	yes	no	8	4	0.2627	4	2	0.4493	1	3	**0.4709**	71.03	79.25
2	yes	yes	no	1	4	0.4608	1	2	0.4616	2	5	**0.4696**	0.17	1.91
3	yes	yes	yes	3	7	0.4409	8	5	0.4343	3	5	**0.4681**	-1.49	6.17
4	yes	no	no	2	4	0.4459	2	4	**0.4589**	4	4	0.4289	2.92	-3.81
5	yes	no	yes	6	4	0.3950	3	4	**0.4539**	7	4	0.4158	14.91	5.27
6	no	no	no	4	1	0.4065	5	1	0.4471	5	3	**0.4534**	9.99	11.54
7	no	no	yes	5	1	0.4046	6	1	**0.4463**	6	3	0.4268	10.31	5.49
8	no	yes	yes	7	4	0.3749	7	2	**0.4407**	8	3	0.4085	17.55	8.96
												maIm:	**15,67**	14.35

Table 2 shows us that for 2006 query set the best run has been obtained with the configuration that use only the visual concept and PRF as relevance feedback strategy. However this configuration without relevance feedback (NOFB) obtains the worst MAP value. Moreover, we can see that almost all the runs with and without relevance feedback which use the camel case decompounding, improve the baseline result - MAP 0.4065 -. And therefore they are top ranked in the table. We also can observe that geographical expansion always get worse results than the results obtained with the same configuration but without the geographical expansion.

In relevance feedback respect we observe that PRF has obtained the best result but the worst too. Showing LCA a better behaviour if we attend to the mean average improvement (MaImpr).

And finally in passage size respect we see that there is not an ideal size to use with this query set. Since each configuration has its own suitable passage size.

The experiments with the 2007 query set show us that camel case decompounding obtains the best results. We also can observe that visual concept expansion always get worse results than

Table 3: Query Set 2007 - Best Results

rk	**cam case**	**vis con**	**geo**	**rk no fb**	**ps no fb**	**map no fb**	**rk lca**	**ps lca**	**map lca**	**rk prf**	**ps prf**	**map prf**	**% lca impro**	**% lca impro**
1	yes	no	no	1	5	0.3052	1	5	0.3201	1	5	**0.3245**	4.88	6.32
2	yes	no	yes	2	5	0.3032	2	5	0.3198	2	5	**0.3227**	5.47	6.43
3	yes	yes	yes	3	5	0.2940	3	5	0.3138	3	5	**0.3151**	6.73	7.18
4	yes	yes	no	4	5	0.2937	4	5	0.3135	4	5	**0.3140**	6.74	6.91
5	no	no	no	5	3	0.2803	5	1	**0.2872**	5	2	0.2847	2.46	1.57
6	no	no	yes	6	3	0.2787	6	1	**0.2871**	6	2	0.2835	3.01	1.72
7	no	yes	yes	7	5	0.2607	7	5	**0.2720**	7	5	0.2656	4.33	1.88
8	no	yes	no	8	5	0.2605	8	5	**0.2718**	8	5	0.2654	4.34	1.88
												maIm:	**4.75**	4.24

the results obtained with the same configuration but without the visual concept. In geographical expansion respect, we observe that it does not meaningfully change the results.

Furthermore we saw that relevance feedback always improves the results. On one hand PRF as well as for 2006 query set experiments obtains the best and the worst results of the relevance feedback runs and on the other hand LCA shows another time a good performance in a mean average improvement way.

In passage size respect we see that for almost all the runs the best passage size has a value of 5.

Finally, for our participation in WikipediaMM08, we have sent 24 runs corresponding with the configurations that better results have given with the 2007 query set. This is because we have found very few common values between the parameters of the best runs of the 2006 and 2007 query sets experiments. Thus, we decided to use the parameter values which best results obtained for 2007 query set since those parameter values were more homogeneous than the parameter values which gave best results for the 2006 query set - it can be observed seeing the passage sizes but it is extensible to other parameters as the number of documents and terms used for the relevance feedback that in order to summarise we have not shown in the data tables -.

4 Results in WikipediaMM08

Table 4 shows the results for each Camel Case, Concept and Geog combination in the WikipediaMM task of this year. It also shows us the ranking position for each run in the WikipediaMM08 task ranking (RK CLEF).

As we expected the best results has been obtained using camel case decompounding. However we have obtained unexpected negative results with the runs using relevance feedback. Even so it is observable that LCA always has better MAP results than PRF, and therefore a great difference for the MAImpr.

Table 4 shows together a comparison of the MAP results of our baseline run and the 5 best MAP results in the competition, along with an average for the 77 runs - which have been sent by a total number of 12 participants - (Avg MAP CLEF). Moreover it is showed the improvement of each run respect the Avg MAP CLEF.

It is important to highlight that our baseline run using only a passage based system - with a passage size of 3 sentences - has obtained the 29th position in the ranking of 77 runs submitted.

Table 4: Results in WikipediaMM08

rk	**rk clef**	**cam case**	**vis con**	**geo**	**rk no fb**	**map no fb**	**rk lca**	**map lca**	**rk prf**	**map prf**	**% lca impro**	**% prf impro**
1	3	yes	no	no	1	**0.2700**	1	0.2614	2	0.2321	-3.19	-14,04
2	6	yes	no	yes	2	**0.2605**	3	0.2583	3	0.2287	-0.84	-12,21
3	8	yes	yes	no	3	0.2587	2	**0.2593**	1	0.2326	0.23	-10.09
4	17	yes	yes	yes	4	0.2509	4	**0.2537**	4	0.2238	1,11	-10.80
5	30	no	yes	no	5	**0.2183**	5	0.2158	5	0.2083	-1.15	-4.58
6	31	no	no	no	6	**0.2178**	7	0.2053	6	0.2033	-5,74	-6.66
7	35	no	yes	yes	7	**0.2091**	6	0.2067	8	0.1997	-1.15	-4.50
8	36	no	no	yes	8	**0.2091**	8	0.2010	7	0.2003	-3,87	-4.21
										maIm:	**-1.83**	-8.39

Table 5: Comparation with Other Participants WikipediaMM08

rk	**run**	**cam case**	**vis con**	**geo**	**rel fb**	**map**	**avg map**	**% run impro**
1	(upeking)zzhou3					0.3444	0.1756	0,9613
2	ceaTxtCon					0.2735		0,5575
3	**IRnNoCamel**	**yes**	**no**	**no**	**no**	**0.2700**		**0,5376**
4	ceaTxt					0.2632		0,4989
5	IRnNoCamelLca	yes	no	no	lca	0.2614		0,4886
29	**IRn**	**no**	**no**	**no**	**no**	**0.2178**		**0,2403**

5 Conclusion and Future Work

A major finding of these results is that, camel case decompounding has been decisive for the success in this task. It makes us to believe that there are two main reasons for it. The first and most obvious one is that with this procedure we have accessed the terms, not visible for other systems and therefore we have increased our chances to perform a good retrieval. The other one - more intuitive - is that the terms used for a file name usually are the most meaningful terms that exist within the metadata of an image. The reason is that the user tries to synthesize the image content from a few words. It would explain the boost with this procedure experimented by our system.

The problem of mismatch between a concept in a query and in a document when it is expressed - with different terms than found in the collection - is aggravated in collections with small sized documents. Hitherto relevance feedback always has been a good tool for improving the results in our experiments with image annotations. However unexpected results has been obtained with relevance feedback strategies. Even though we have seen that LCA has showed a more robust behaviour, while PRF shows more unpredictable results.

As future work study we will research the causes of the relevance feedback negative results. and if it is possible, we will try to develop a method to forecast when a relevance feedback is going to be negative for the retrieval, in order to refrain from using it.

Furthermore, in order to achieve the continuing improvement of the system, we shall attempt to include an efficient method for the decompounding file names which does not use a standard notation. Moreover, in accordance with our intuition with respect to the significance of file name, we will research the effect of giving more weight in the retrieval phase to the terms that compound

the file names.

6 Acknowledgement

This research has been partially funded by the Spanish Government within the framework of the TEXT-MESS (TIN-2006-15265-C06-01) project and by European Union (EU) within the framework of the QALL-ME project (FP6-IST-033860).

References

[1] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Lecture notes in Computer Science*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[2] M.C. Daz-Galiano, M.A. Garca-Cumbreras, M.T. Martn-Valdivia, A. Montejo-Raez, and L.A. Urea-Lpez. Sinai at imageclef 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[3] H. Jair Escalante, Carlos A. Hernndez, Aurelio Lpez, Heidy M. Marn, Manuel Montes, Eduardo Morales, Luis E. Sucar, and Luis Villaseor. Tia-inaoes participation at imageclef 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[4] Sheng Gao, Jean-Pierre Chevallet, Thi Hoang Diem Le, Trong Ton Pham, and Joo Hwee Lim. Ipal at imageclef 2007 mixing features, models and knowledge. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[5] Anni Jrvelin, Peter Wilkins, Tomasz Adamek, Eija Airio, Gareth J. F. Jones, Alan F. Smeaton, and Eero Sormunen. Dcu and uta at imageclefphoto 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[6] Fernando Llopis. *IR-n: Un Sistema de Recuperacin de Informacin Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[7] Bernardo Magnini Luisa Bentivogli, Pamela Forner and Emanuele Pianta. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING 2004 Workshop on 'Multilingual LinguisticResources'*, pages 101–108, Geneva, Switzerland, August 2004.

[8] Sergio Navarro, Fernando Llopis, Rafael Muñoz, and Elisa Noguera. Information Retrieval of Visual Descriptions with IR-n System based on Passages. In *In on-line Working Notes, CLEF 2007*, 2007.

[9] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–46, 1976.

[10] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.