UNT at ImageCLEF 2010: CLIR for Wikipedia Images

Miguel E. Ruiz, Jiangping Chen, Karthikeyan Pasupathy, Pok Chin and Ryan Knudson

University of North Texas, College on Information, Department of Library and Information Sciences, 1155 Union Circle 311068
Denton, Texas 76203-1068, USA
{Miguel.Ruiz, Jiangping.Chen}@unt.edu

Abstract. This paper presents the results of the team of the University of North Texas in the Wikipedia image retrieval track of Image-CLEF-2010. Our approach is based on performing translation of the French and German image captions to English and using of Language Models for generating our runs. We also explore the use of complex queries by asking two users to manually build queries based on the original topics distributed. Our results indicate that the approach of translating the image captions is feasible and yields results that are quite competitive with other teams that participated in the same track.

1 Introduction

This paper presents the results of the UNT team participation in the Wikipedia retrieval task. Traditionally, the most common approach to solve the cross language retrieval problem is to perform automatic translation of the user queries into the language of the document to be retrieved. However, in the presence of short queries the automatic translation might not have enough context to generate an appropriate translation. Our main goal was to explore the efficacy of using the captions associated with the Wikipedia images and providing automatic translations of them in English. We also address the effectiveness of using this approach using automatic queries as well as manual queries constructed by real users.

Section 2 of this paper presents a short background of the CLIR retrieval problem in image retrieval. Section 3 presents the methods used to conduct our experiments. Section 4 presents our results and preliminary analysis of results. The last section of this paper presents our conclusion and plans for future work.

2 Background

Retrieval of images in multilingual collections is a task that has been studied in CLEF since 2003 (Peters, 2009). Previous research in CLEF addressing this problem have explored the use of different resources for translation and for most part concentrated on combining visual and textual features automatically extracted from images

(Müller, et al., 2009). There has been a lot on emphasis on trying to improve the current Content-Based Image Retrieval (CBIR) using automatically extracted visual features to match sample images given in the official topics. However, our own research as well as the results from other participants has shown that the most successful approach to solve the image retrieval problem relies on high quality text retrieval (Müller, et al., 2009; Ruiz M. E., 2006; Ruiz & Névéol, 2007). The combination of visual and textual features has proven to contribute to small improvements in mean average precision (MAP) which is the standard measure that is used in CLEF to compare system performance.

One of the key issues that needs to be explored is to find approaches that can contribute more to solve the retrieval problem when the given data collection contains annotations generated in multiple languages. For the current Wikipedia retrieval task this is an important issue since the collection has a relatively even distribution of annotations in three languages (English, French and German). The CLIR problem has been solved using several methods but the most commonly used approach consists of translating the text of the user query to the language of the document, performing monolingual retrieval in each language and then combining the results of several monolingual runs. However, this approach has two main potential challenges:

1. Use of machine translation on short queries can be difficult due to the loss of context and finding appropriate disambiguation for automatic translation.
2. Finding optimal parameters to adjust the mechanism to merge results from multiple monolingual runs is challenging. Moreover, these optimal parameters can change from one collection to another making it hard to find a general optimal set of parameters.

We decided to explore a solution that translates all the documents to a single language and perform the retrieval in that language only. We recognize that this is an expensive solution that might not work for a general CLIR problem. However, for image retrieval it is a viable option due to the relatively short length of image captions (compared to the full text associated with the images in an article in Wikipedia). Also, image captions usually contain enough contextual information to allow appropriate translation disambiguation. The translation of captions can be achieved relatively fast using MT translation systems that are freely available on the Internet such as Google Translation. This also reduces the CLIR problem to a simple monolingual translation for which the technology is more stable, and there is no need to deal with merging the results that come from different collections.

3 Methodology

We used the Indri/Lemur Retrieval system to index our collection using standard Krovetz stemming and the standard Language Model implemented in Lemur (Lemur Project, 2001-2008).

Data Collection preparation:

For our experiments we translated all the French and German texts in the captions associated with the images into English using the Google Translation service. This translation was added to the caption using a new field that was indexed together with the original English caption (if it was available).

Topics preparation:

For our runs we used just one language at a time from the three provided in the original ImageCLEFwiki Topics and built topics automatically using a simple strategy that converted all the words to a "#combine" statement in lemur. We used first the English topics as our base line. A second run with French topics that were translated into English was created to measure the effect on query translation. We also asked two of the members of our group to use the Indri Query Language and create manual queries that could take advantage of the advanced option of the more advanced operators in Lemur. For this purpose we made available for these users the Indri web search engine (which is based on Lemur) and asked them to conduct searches with the system until they were satisfied with the results that were retrieved. Each user learned the syntax of the Indri Query Language and then created queries that tried to use the capabilities of the query language. For example, for our first user the procedure followed to build the query is described below:

All query statements used to perform manual image retrieval were built based on the Indri Query Language. The user developed all seventy manual query statements using the following methods:

1. The user tried different combinations of keywords to retrieve images from a sample database.
2. Based on the returned images, the user refined the query statements based on the following criteria:
    a. Incorporate those observable objects within images that can match the question topics into the query keywords using Indri Query Language operators such as #combine and #filreq.
    b. Reject those observable objects within images that cannot match the question topics using Indri Query Language operators such as #filrej.
    c. The user reviewed the first 50 images returned and reiterated the two steps mentioned above until the precision of the first 50 images reached at least 80%.

For example, for topic number seven. In order to find most images representing "striking lighting in the sky", the user tried the method mentioned in 2a to incorporate all potential keywords that could imply "striking lighting in sky" such as lightning, day, night, strike, struck, striking, and sky. The user also rejected the keyword "fighter" using method mentioned in 2b so that the aircraft fighter "lighting" would not be selected for this question. The final query submitted in the official run for this topic is:

#filrej(fighter #filreq(lightning #combine(day night strike struck striking sky)))

4 Results and Analysis

We submitted three official runs and have an unofficial manual run

- untaTxEn: This run uses automatic query construction using the portion of the original topic.
- untaTxFr: This run uses automatic query construction using the original French portion of the text which was translated to English using Google Translation. (This can be considered a standard CLIR scenario)
- untMan1En: This run uses the final version of the queries created by our first user for each of the 70 topics.
- untMan2En: This run uses the final version of the queries created by our second user for each of the 70 topics.

Table 1 shows the retrieval performance of each of our runs. As expected the manual runs had the best performance on P@5, P@10 and P@20 retrieved documents which is consistent with the procedures that the users followed to build their queries. The best MAP for our official runs was the automatically generated English run. However, the manual run generated by the second user did perform better than all our runs. Our manual run 1 achieved a pretty high performance in terms of P@5, P@10 and P@20. When we compared this run with all runs of other participant teams is among the top 10 runs in these measures. However, if we use MAP the best scoring of our official runs is the Automatic English run followed by our unofficial manual run 2.

Table 1 Performance of Official and Unofficial Runs

	untaTxEn	untaTxFr	untMan1En	untMan2En
#ret	55647	58476	11779	20255
#Relev	17660	17660	17660	17660
#relret	7840	7641	4584	5768
Avg-P	0.2251	0.22	0.2064	**0.2349**
exact-P	0.3025	0.2855	0.2603	0.3002
P@5	0.4857	0.46	**0.6314**	0.6171
P@10	0.4314	0.4229	0.5886	**0.5914**
P@20	0.3871	0.3986	0.5021	**0.5521**

Comparing the runs using a standard Recall-Precision graph gives a better picture of the performance of the manual and automatic runs (see Figure 1). The manual runs in general perform better on the early R-P levels (0-0.2) while the automatic runs perform better on the higher levels of recall. This seems to be correlated to the amount of images retrieved which basically indicate that the manual queries are

optimized to generate high precision but low recall (this can also be appreciated in the total number of retrieved documents in Table 1).

Regarding the CLIR run that uses French as original language and English as target (or collection language) we can see that the performance is very close to approach that translates the documents instead of the queries.

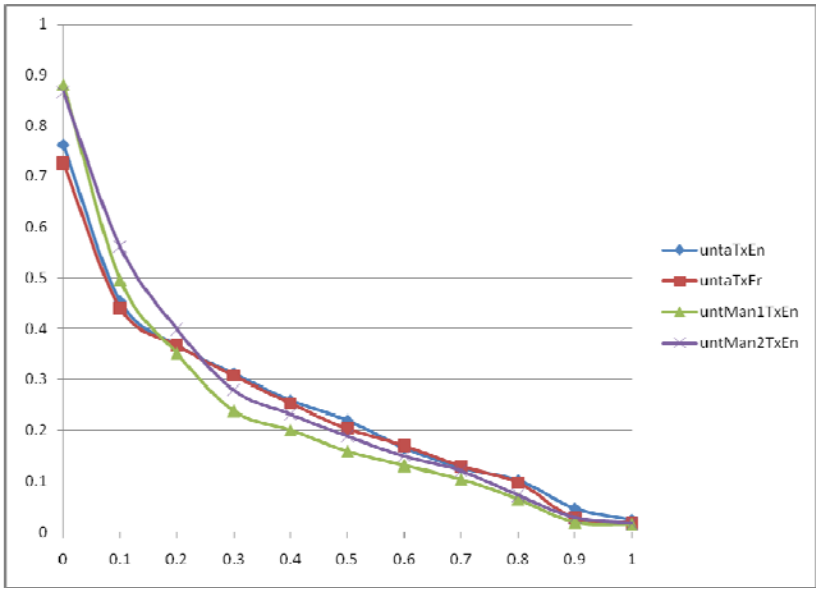

Figure 1 R-P Graph comparing UNT's runs

5 Conclusions

We conclude that the approach of translating the captions instead of the queries is feasible for image collection and competitive with other more complex approaches that use more complex algorithms for performing CLIR.
The manually created queries allowed us to explore potential strategies that could help in future research that can take advantage of the complex query language available in Lemur. We still have to do more analysis on the official runs as well as other unofficial runs that will be included for the extended paper of the proceedings.
We also plan to explore the use of visual features with these queries but need to get a better understanding of the way users would interact with the system using an appropriate interface that combines CBIR and CLIR.

References

Gonzalo, J., Clough, P., & Karlgren, J. (2008). Overview of iCLEF 2008: Search Log Analysis for Multilingual Image Retrieval. *Working Notes for the CLEF 2008 Workshop.* Aarhus, Denmark.

Lemur Project (2001-2008). *The Lemur Project*. (University of Massachusetts and Carnegie Mellon University) Retrieved August 15, 2010, from http://www.lemurproject.org

Müller, H., Kalpathy-Crame, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., et al. (2009). Overview of the CLEF 2009 medical image retrieval track, CLEF working notes 2009. Corfu, Greece.

Peters, C. (2009). What happened in CLEF 2009: Introduction to the Working Notes. *Working Notes for the CLEF 2009 Workshop.* Corfu, Greece.

Ruiz, M. E. (2006). Combining Image Features, Case Descriptions and UMLS Concepts to Improve Retrieval of Medical Images. *Proceedings of the Symposium of the American medical Informatics Association*, (pp. 674-8). Washington, D.C.

Ruiz, M., & Névéol, A. (2007). Evaluation of Automatically Assigned MeSH Terms for Retrieval of Medical Images. *In Revised and selected papers of the 2006 Cross Language Evaluation Forum.* Springer.