MSU at ImageCLEF: Cross Language and Interactive Image Retrieval

Vineet Bansal, Chen Zhang, Joyce Y. Chai, Rong Jin

Department of Computer Science and Engineering, Michigan State University
East Lansing, MI48824, U.S.A.
{bansalvi,zhangch6,jchai,rongjin}@cse.msu.edu

Abstract. In this report, we describe our studies with cross language and interactive image retrieval in ImageCLEF 2004. Typical cross language retrieval requires special linguistic resources, such as bilingual dictionaries. In this study, we focus on the issue of how to achieve good retrieval performance given only an online translation system. We compare two approaches, i.e., a translation-based approach and a model-based approach, and find that the later one performs substantially better than the former one. For interactive image retrieval, we investigated the potential use of user relevance feedback (URF), which was designed to address the mismatch problem between user queries and system descriptions. Our strategy is to let the system select important terms for user feedback before expanding queries. However, our preliminary results appear to indicate that the URF approach developed at the current stage is not working. We report our current investigation and discuss lessons learned from this experience.

1 Introduction

Empirical studies have shown that using image features to find similar images is usually insufficient [15]. First, it is difficult for users to specify visual queries with low-level visual features. Second, low level image features cannot precisely describe user information needs. There is a gap between low-level visual descriptions and user's semantic expectation [10]. Text queries, on the other hand, are more intuitive and natural for users to specify their information needs and expectations.

In this year's ImageCLEF, we investigated two challenging tasks related to text-based image retrieval:

1) Given image descriptions in one language and user query in another language, how to effectively retrieve images using cross language retrieval? In particular, given limited bilingual resources (e.g., the online bilingual translation system), how to improve the accuracy of cross lingual information retrieval?
2) Given a target image in user's mind, how to interactively help users to find such an image. In particular, we investigated the use of user relevance feedback in such a task.

In the following sections, we devote two sections to these two tasks respectively.

2 Cross Language Retrieval Using only an Online Translation System

Cross lingual retrieval has been one of the major research areas in information retrieval during last few years [1, 2, 5-7, 9]. Most cross lingual retrieval algorithms fall into two categories: the translation-based approaches, and the approaches based on statistical models.

A simple translation-based approach will translate a query into the language of documents, and relevant documents will be found by matching the translated queries with the documents[1]. Different algorithms can be applied to translate queries, ranging from the simplest one that is based on bilingual dictionaries to the sophisticated one that is based on a full-scale machine translation system. Compared to dictionary-based translation, using a full-scale machine translation system has the advantage in that the ambiguity of a query is reduced by a full-scale translation system and only the best translation of the query is used. However, on the other hand, a cross lingual approach based on the full-scale translation system can perform poorly if a query is truly ambiguous and multiple possible translations need to be considered. In those cases, dictionary-based translation approaches for cross lingual retrieval will have advantages because it include all possible translations of query words. Thus, a good cross lingual retrieval system should be able to, on one hand, reduce the uncertainty in translating queries when possible, and on the hand, maintain the uncertainty of query translation if the query is ambiguous.

A model-based approach usually utilizes the existing statistical machine translation models that were developed by the IBM group [16]. Given a translation model θ, the relevance of a document d to a given query q is computed as $p(q \mid d;\theta)$, which is the likelihood of translating document d into query q. Compared to the translation-based approaches, the model-based approaches have advantage in that by using the translation probabilities learned from a parallel corpus, we are able to reduce translation ambiguity and yet maintain the uncertainty in translation at the same time. This is done through the adjustment of translation probabilities: an unlikely translation will be assigned with a small probability; meanwhile equally likely translations of a query will be assigned with similar translation probabilities. However, in order to build a statistical translation model, a sufficiently large bilingual parallel corpus is required. Acquiring a large parallel corpus is usually expensive and time consuming, especially for minor languages.

In this report, we study an approach that first utilizes the online translation system to create a bilingual parallel corpus and then learns a statistical translation model based on the created bilingual corpus. Unlike the translation based approaches where only the best translation is used in information retrieval, this approach maintains the uncertainty in translation and therefore will be more robust to the translation errors. On the other hand, unlike the typical model-based approaches where a large bilingual parallel corpus is required, this approach automatically creates a bilingual parallel corpus by applying the online translation system to translate test documents into the

language of queries. In the following subsections, we will first overview the statistical machine translation model, and then discuss the empirical results with the proposed approach.

2.1 A Statistical Translation Model for Cross Language Information Retrieval

For the convenience of discussion, let's assume that the language of queries is Chinese and the language of documents is English. Let the set of translation probabilities denoted by $\theta = \left\{ t(w_i^e \mid w_j^c) \right\}$. Each $t(w_i^e \mid w_j^c)$ is the probability that translates a Chinese word w_j^c into an English word w_i^e. The key to statistical translation model for cross lingual information retrieval is to automatically learn the set of word translation probabilities from a parallel corpus. Let the bilingual parallel corpus for training a statistical translation model be denoted by $\Omega = \left\{ \left(\mathbf{s}_i^c, \mathbf{s}_i^e \right) \right\}_{i=1}^{N}$. Each $\left(\mathbf{s}_i^c, \mathbf{s}_i^e \right)$ is a translation pair in which sentence $\mathbf{s}_i^c$ is the Chinese translation of sentence $\mathbf{s}_i^e$. N is the total number of translation pairs in the corpus Ω. According to the IBM translation model, $p(\mathbf{s}_i^e \mid \mathbf{s}_i^c; \theta)$, i.e., the probability of translation an English sentence $\mathbf{s}_i^e$ into a Chinese sentence $\mathbf{s}_i^c$, can be written as:

$$p(\mathbf{s}_i^e \mid \mathbf{s}_i^c; \theta) \propto \prod_{j=1}^{V_e} \left(\sum_{k=1}^{V_c} o(w_k^c, \mathbf{s}_i^c) t(w_j^e \mid w_k^c) \right)^{o(w_k^e, \mathbf{s}_i^e)}$$

where V_e and V_c stand for the size for Chinese vocabulary and English vocabulary, respectively. $o(w_k^c, \mathbf{s}_i^c)$ represents the occurrence of Chinese word w_k^c in Chinese sentence $\mathbf{s}_i^c$. So does $o(w_k^e, \mathbf{s}_i^e)$. Thus, in order to learn translation probabilities, we can maximize the log-likelihood of all translation pairs used for training, i.e.

$$\theta = \arg\max_{\theta \in \Theta} l(\Omega; \theta) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \log p(\mathbf{s}_i^e \mid \mathbf{s}_i^c; \theta)$$

A well-known Expectation Maximization algorithm can be used to efficiently learn the optimal translation probabilities. More details can be found in [3]. Finally, in order to estimate the relevancy of a document *d* to a query *q*, probability $p(q \mid d)$ is estimated using the following expression:

$$\log p(q^c \mid d^e; \theta) = \log \int dq^e \, p(q^c \mid q^e) p(q^e \mid d^e) \approx \sum_i p(w_i^e \mid q^c) \log p(w_i^e \mid d^e)$$

$$\approx \sum_i \left(\sum_j o(w_j^c, q^c) p(w_i^e \mid w_j^c) \right) \log p(w_i^e \mid d^e)$$

More details of applying statistical translation model to cross lingual information retrieval can be found in [16].

2.2 Our Approach: Training a Statistical Model Using an Online Translation System

Given the success of the statistical translation model for cross lingual information retrieval in the TREC evaluations [13, 14], we would like to apply it to the cross language image retrieval. However, the biggest problem is to acquire a bilingual parallel corpus that shares the similar content as the text collection used in the ImageCLEF evaluation. In order to acquire a bilingual corpus, we tried a simple strategy. We first applied an online translation system to translate the textual descriptions in ImageCLEF into Chinese sentences. To enhance the diversity of our translation pairs, the Chinese sentences that are generated by the online translation system are further translated back into English sentences. The final bilingual corpus is created by aggregating all the translation pairs together. The online translation system used in our experiment is Systran (http://www.systransoft.com/). With the acquired translation pairs, we now can apply the statistical translation model to automatically learn translation probabilities between Chinese words and English words. Examples of learned translation probabilities are listed in Table 1. Note that all English words are stemmed using the Porter algorithm.

Table 1. Examples of translation probabilities learned from the bilingual parallel corpus that is generated by the online translation model. All the English words are stemmed.

Chinese	English	Prob.	Chinese	English	Prob.
塔	tower	0.8692	大教堂	cathedr	0.7312
	turret	0.0200		st	0.0475
	pinnacl	0.0198		iona	0.0231
	build	0.0048		dunblan	0.0161
	clock	0.0044		eli	0.0152
	squar	0.0042		durham	0.0147
	spire	0.0042		andrew	0.0143
	church	0.0028		elgin	0.0139
	transept	0.0026		dunkeld	0.0104
	hous	0.0023		transept	0.0098

The retrieval performance using statistical translation model for cross lingual retrieval is listed in Table 2 under the column entitled as 'Model-based'. For the purpose of comparison, we also run the simple translation-based approach, which applies the online translation system to translate each Chinese query into an English query. The results of this translation-based approach are also included in Table 2 under the column entitled as 'Translation-based'.

As indicated in Table 2, the approach based on the statistical translation model performs substantially better than the simple translation-based approach in terms of almost every metric. In particular, the major difference between these two approaches lies in the region when only the top retrieved documents are examined. For example, when only the first five documents are examined, the precision for the translation-based approach is only 28.8%, while the precision for the approach based on statistical translation model is 41.6%. This fact is further confirmed by the precision results

for the low recall points. For example, when systems recall 10% of the relevant documents, the precision for the translation-based approach is only 37.4%, while the precision for the model-based approach is above 50%. Thus, we conclude that the proposed approach is a better way of utilizing the online translation system for cross lingual information retrieval than the simple translation-based approach.

Table 2. Retrieval results for both the translation-based approach and the approach based on the statistical translation model.

Recall@	**Translation-based**	**Model-based**
0.0	0.638	0.680
0.1	0.374	0.521
0.2	0.367	0.432
0.3	0.328	0.392
0.4	0.290	0.338
0.5	0.261	0.301
0.6	0.225	0.265
0.7	0.203	0.239
0.8	0.171	0.194
0.9	0.141	0.153
1.0	0.099	0.106
Avg Prec.	0.245	0.293
Prec@		
5 doc	0.288	0.416
10 doc	0.260	0.344
100 doc	0.150	0.161

3 Interactive Image Retrieval: User Relevance Feedback

Compared to example-based image retrieval, text-based image retrieval provides an intuitive and natural means for users to specify their information needs and expectations. However, text queries also face many challenges [8]. One major problem concerns both the sparsity and inconsistency of textual descriptions [12]. The words used to describe an image or a similar image vary from one user to another. Furthermore, the textual descriptions are usually short. This vocabulary variation and the conciseness of textual descriptions make it difficult for the traditional text retrieval to work effectively for image retrieval.

To address this problem, we are currently in an on-going investigation on user relevance feedback (URF) in image retrieval. Here, user relevance feedback is motivated by the success of pseudo relevance feedback (PRF) in information retrieval [10]. The difference between URF and PRF is that, in URF we introduce users in the loop to do a sanity check on potential expanded terms. Instead of automatically expanding the query as in PRF, the URF presents a list of terms to users and ask them to choose relevant terms that can describe the target image. Only those terms chosen by the user will be used in query expansion.

Our hypothesis is that this type of feedback can take advantage of the conciseness of textual descriptions and consolidate the inconsistency of user textual queries. On one hand, the concise descriptions make it possible for the system to efficiently identify potential important terms. On the other hand, the system selected terms will remedy the difference between query term and image description. Furthermore, the sanity check from the user will improve the quality of query expansion, which will ultimately result in the improvement of final retrieval results.

As a first step in our investigation, we developed several strategies to select terms and conducted simulations to evaluate different strategies. We then implemented the best strategy for the real user study. In the user study, we compared the interface using URF with a standard interface that only allows users to interactively refine or expand their queries. However, out of our expectation, the results from user studies were not able to validate our hypotheses. In fact, the results indicate the current design and implementation of URF is not working. Therefore, in this section, rather than presenting a successful story (as much as we wish), we report our current investigation and discuss lessons learned from this experience that are useful for future investigation.

3.1 Term Selection

Term selection in URF is different from that in PRF. In URF, our goal is to find terms from descriptions that are related to the initial user query terms, however with large uncertainties as to whether they are relevant. As a first step, we investigated different strategies for term selection using simulation experiments. Simulation studies are important since they can provide some insights on whether a strategy can potentially work even before the expensive user studies are conducted. In these simulated experiments, the system first selects a list of ten terms based on different strategies. To simulate human behavior in identifying relevant terms from this list, the system picks terms that occur in the description of the target images. The picked terms, together

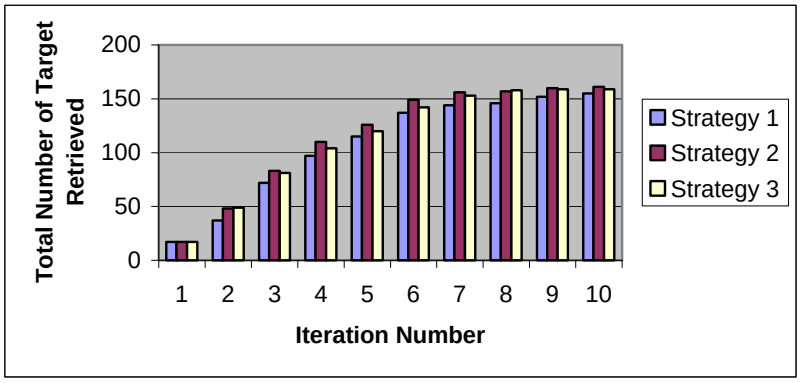

Fig. 1. Performance of three strategies at each iteration point

with the initial query terms, will be sent to the backend retrieval engine. This process repeats until either the target image is found in the top N (currently, N = 20) retrieved images or the system reaches M iterations (currently, M = 10).

To generate the terms, we have experimented with different strategies. The first strategy measures the entropy of a term based on the top *N* retrieved results (called *Top Set* later) and/or the next *100-N* retrieved results (called *Bottom Set* later). The idea is that the term with higher entropy is more uncertain in terms of whether it describes user's interest. By asking user to confirm those higher entropy terms, the system can quickly narrow down the search space. The higher the entropy is, the higher the weight is given. We tried different combinations of retrieved results to calculate the entropy for a given term, specifically the following three strategies:

- Strategy 1: Higher weights are given to terms that have higher entropy from the Bottom Set and also occur less frequently in the Top Set.
- Strategy 2: Higher weights are given to terms with higher entropy from the Bottom Set.
- Strategy 3: Higher weights are given to terms with higher entropy from the Top Set.

In the simulation experiments, we randomly picked 200 images from ImageCLEF collection [4]. For each image, we provided an initial query. Then we applied the simulation process as described above to retrieve each image. Figure 1 shows the simulation results from three different strategies as to how many out of 200 images were successfully retrieved as top 20 results at each iteration point. Results indicate that there is no significant difference between three strategies. All three strategies are more effective at earlier iterations (from 1 to 6) than later ones in the simulation. At iteration 1, since no query expansion is used, all three strategies resulted in the same number of successful retrievals only based on the initial queries. Since the strategy 2 seems slightly better, we use the strategy 2 in our user study.

We have also experimented with the inverse correlation strategy and the synonym strategy. The idea for the inverse correlation strategy is that if a term is very correlated with a query term given by the user, then that term carries less information in identifying new images that might be of user's interest. Therefore, we give a lower weight to the terms that is highly correlated with a query term using a vector space model. The idea for the synonym strategy is that if a term is the synonym of a query term, then it could be very relevant. We want to give it a higher weight since it maybe just a different vocabulary expressing the same meaning. To test this synonym strategy, we used WordNet. However, our current simulations have not shown the effectiveness of these two strategies in term selection. Therefore, we did not include these two strategies in the user study.

Once one or more prompted terms are selected (either automatically by the system in the simulation or manually by the user in the user studies), those terms will be used to expand the initial query in further retrieval cycles. The retrieval model is based on a statistical language modeling approach using textual descriptions of images [11].

3.2 User Studies

To validate our hypothesis and evaluate the effectiveness of the current URF, we conducted a comparative study following the guidelines provided by the ImageCLEF interactive track.

3.2.1 Method

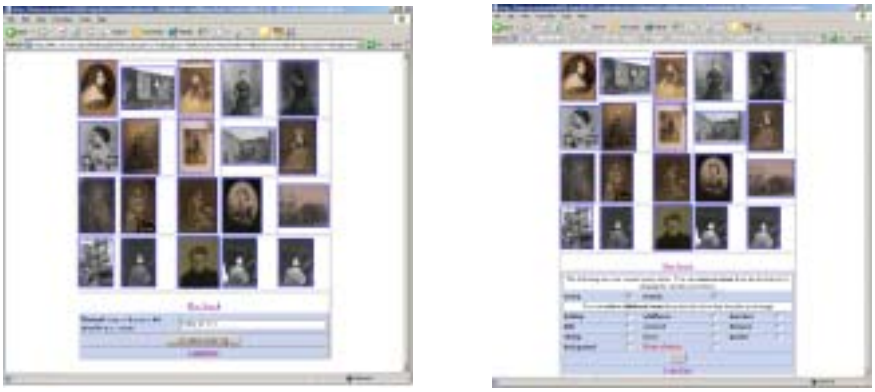

Fig. 2. (a) Standard Interface (b) Interface with URF

Eight subjects participated in the study and each of them was asked to search for 16 images from the Eurovision St. Andrew collection provided by ImageCLEF. The subjects were first asked to complete a screening questionnaire to elicit demographic data and data concerning searching experience. Then the subject was asked to use one interface to search eight images (one at a time). After using each interface, the subject was given a questionnaire to indicate how easy he feels about the search process, and how satisfied he is with a particular system. During the search, the system also automatically logged the information such as the original queries from the user, the system retrieved results, terms prompted by the system, and the time spent on searching, etc. When an image was found or when five minutes were run out, the search stopped. After searching all images using two different interfaces, each subject was asked to give an overall ranking of the two interfaces in terms of their overall satisfaction and systems' effectiveness of locating the target images.

Two interfaces used in the study are shown in Figure 2. Figure 2(a) is a standard interface to be compared with, where users could refine or expand their queries using their own terms. Figure 2(b) is the URF interface. In the URF interface, in addition to ten terms prompted by the system for user feedback, the system also shows the query terms that are used so far for the retrieval. Users have choices to revise these query terms from previous iterations by "de-select" them from the list. This feature was designed so that the URF interface is comparable to the standard interface where users can freely add/remove their query terms at each iteration.

3.2.2 Evaluation Results

Unfortunately, after three users, we found some inconsistency in the system, so we had to discard the results from those three users. The results shown here are from five out of eight users. Table 3 shows the effectiveness for the two interfaces. The successful retrieval rate is calculated by dividing the number of target images that are shown in the top 20 retrievals by the total number of target images tried for that interface. Since the success rate for the interface B (with URF) is lower than the interface A (the standard interface), we conclude that the interface B is not effective based on the current design and implementation.

Table 3. Overall performance of two interface

	Standard Interface	URF
Successful Retrieval	0.575	0.175
Average time	0:48	1:57
Average number of interactions	2.43	3.57

3.3 Discussion

The failure with the current design points to several problems that need to be addressed in our future investigation. Users were involved in the loop to provide feedback for query expansion. However, one major problem is that many of those terms do not mean much to the user. Certainly, we hope that when a prompted term appears in the description of the target image, the user would pick that term (as in our simulation). However, from our studies, we found that even those terms appear in the description, the user still could not recognize them. This caused the big performance difference between the simulated experiments and the real user study. We feel that there are different classes of terms. Some classes of terms are much easier to identify than the others. For example, “background” and “substantial”, both terms occurred in the description of a target image. However, it was very hard for users to recognize them since they did not directly match any salient features conveyed by the image. On the other, the term “bridge” would be easier for the user to recognize. It would be ideal if the system can only prompt to the users those key terms that could mean something to the user. Thus, it would be interesting to study how users respond to these different terms based on the salient features and semantic content presented in an image and how to identify those significant terms from the retrieved results. Only with such an understanding, is it possible to build a potentially effective URF.

In additional, as in the traditional text retrieval, the term mismatching is another problem for image retrieval. For example, suppose among the ten terms prompted by the system, the user chooses the term “road”. Even this term does describe some object in the target image, this term will not be effective if the term “street” is used in the description, rather than the term “road”. Therefore, in order to effectively use URF, the system needs to have a capability of handling this type of mismatching caused by variations of terms.

Because of the time limitation, here we only briefly describe some very preliminary observations and problems. We certainly need more in-depth analysis on our

collected data. Although the current experiment is not successful, what we have learned from this experience can help us focus on specific issues identified. We believe URF still has a potential in interactive image retrieval. For example, instead of only allowing URF as in our current interface, we can consider adding URF to a standard interface. However, before that happens, first of all, we need to reach a better understanding of user cognitive models on describing image content and its implication in user relevance feedback.

4. Conclusion

In this report, we examined two important issues associated with cross language image retrieval and interactive image retrieval:

1) How to improve the accuracy of information retrieval given that only an online translation system is available;
2) How to enhance text-based image retrieval using the user relevance feedback (URF).

Our empirical results with cross language retrieval have indicated that an employment of statistical translation model is effective, even when the parallel corpus is created automatically by an online translation system. Our preliminary study with interactive image retrieval has illustrated that to make user relevance feedback effective for text-based image retrieval, a carefully designed procedure of automatic term selection is critical. In particular, the selected terms should be able to not only distinguish certain images from others, but also be consistent with the users' perception of images. Thus, more in-depth investigation is needed to reach a better understanding of user cognitive models on describing image content and its implication in user relevance feedback.

References

1. Ballesteros, L. and W.B. Croft. *Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval.* in Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. 1997.
2. Ballesteros, L. and W.B. Croft. *Resolving Ambiguity for Cross-Language Retrieval.* in Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998.
3. Brown, P., et al., *The Mathematics of Statistical Machine Translation.* Computational Linguistics, 1993. **19**(2): p. 263-311.
4. Clough, P., M. Sanderson, and N. Reid. The Eurovision St Andrews Photographic Collection. http://ir.shef.ac.uk/imageclef2004/guide.pdf.
5. Federico, M. and N. Bertoldi. *Statistical Cross-Language Information Retrieval Using N-Best Query Translations.* in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2002.

6. Gao, J., et al. *Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models*. in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001.
7. Hiemstra, D. and F.M.G.d. Jong. *Disambiguation Strategies for Cross-Language Information Retrieval.* in Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL). 1999.
8. Keister, L.H., *User Types and Queries: Impact on Image Access Systems.* ASIS, 1994: p. 7-22.
9. Lavrenko, V., M. Choquette, and W.B. Croft. *Cross-Lingual Relevance Model*. in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2002.
10. Mitra, M., A. Singhal, and C. Buckley. *Improving Automatic Query Expansion.* in Proceedings of SIGIR 1998. 1998.
11. Ponte, J.M. and W.B. Croft. *A Language Modeling Approach to Information Retrieval*. in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998.
12. Seloff, G.A., *Automated Access to Nasa-Jsc Image Archives.* Library Trends, 1990. **38**(4): p. 682-696.
13. Voorhees, E.M. and D.K. Harman, eds. *Proceedings of the Ninth Text Retrieval Conference (Trec-9)*. 2000: Gaithersburgh, MD.
14. Voorhees, E.M. and D.K. Harman, eds. *Proceedings of the Ninth Text Retrieval Conference (Trec-10)*. 2001: Gaithersburgh, MD.
15. Westerveld, T. and A.P.d. Vries. *Experimental Result Analysis for a Generative Probabilistic Image Retrieval Model*. in Proceedings of the 26th ACM SIGIR. 2003.
16. Xu, J., R. Weischedel, and C. Nguyen. *Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval*. in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001.