Using Information Gain to Improve the ImageCLEF 2006 Collection

M.C. Díaz-Galiano, M.Á. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L. Alfonso Ureña-López

SINAI Research Group. Computer Science Department. University of Jaén. Spain
{mcdiaz,magc,maite,amontejo,laurena}@ujaen.es

Abstract. This paper describes the SINAI team's participation in both the ad hoc task and the medical task. For the ad hoc task we use a new Machine Translation system which works with several translators and heuristics. For the medical task, we have processed the set of collections using Information Gain (IG) to identify the best tags that should be considered in the indexing process[1].

1 The Ad Hoc Task

The goal of the ad hoc task is, given a multilingual query, to find as many relevant images as possible from a given image collection [2].

1.1 Description of Experiments

We have considered seven languages in our experiments: Dutch, English, French, German, Italian, Portuguese and Spanish. Based on our reliable 2005 results [3], this year we have used the same IR system and the same strategies, except that we used a new translation module. This module combines the following Machine Translators and heuristics:

- Epals (German and Portuguese), Prompt (Spanish), Reverso (French) and Systran (Dutch and Italian)
- Some heuristics are, for instance, the use of the translation made by the translator by default, a combination with the translations of every translator, or a combination of the words with a higher punctuation (scoring two points if it appears in the default translation, and one point if it appears in all of the other translations).

The collections have been preprocessed using stopwords and the Porter's stemmer. The collection dataset has been indexed using the LEMUR IR system[2], using two different weighting schemes (Okapi or TFIDF), and enabling or disabling PRF (pseudo-relevance feedback).

[1] This work has been supported by the Spanish Government with grant TIC2003-07158-C04-04.

[2] http://www.lemurproject.org/

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 711–714, 2007.

1.2 Results and Discussion

All the results are obtained using the title and narrative text, whenever possible. In the English monolingual task and in the German-English bilingual task, we have combined the use (or not) of pseudo-relevance feedback and the weighting function (Okapi or Tfidf). In table 1, we can see the global results: English monolingual results and bilingual results. They show that the pseudo-relevance feedback is important when Okapi is used as a weighing function. The results with Tfidf and with Okapi without PRF are very poor.

The results also show that there is a loss of MAP between the best monolingual experiment and this bilingual experiment, namely aproximately 28%. Yet, the other results in the English monolingual task are less acceptable compared to the German bilingual ones.

In general, there is a loss of precision compared to the English monolingual results. The Spanish result is aproximately 17% less acceptable. The other languages lead to a deterioration of the results.

Table 1. Summary of results for the English monolingual adhoc runs and the other bilingual runs

Language	Initial Query	Expansion	Weight	MAP	Rank
Monolingual En	title + narr	with	Okapi	0.2234	9/49
Monolingual En	title + narr	without	Okapi	0.0845	38/49
Monolingual En	title + narr	with	Tfidf	0.0846	37/49
Monolingual En	title + narr	without	Tfidf	0.0823	39/49
Bilingual DeEn	title + narr	with	Okapi	0.1602	4/8
Bilingual DeEn	title + narr	without	Okapi	0.1359	7/8
Bilingual DeEn	title + narr	with	Tfidf	0.1489	5/8
Bilingual DeEn	title + narr	without	Tfidf	0.1369	6/8
Bilingual NlEn(Dutch)	title + narr	with	Okapi	0.1261	4/4
Bilingual FrEn(French)	title + narr	with	Okapi	0.1617	5/8
Bilingual ItEn(Italian)	title + narr	with	Okapi	0.1216	13/15
Bilingual PtEn(Portuguese)	title + narr	with	Okapi	0.0728	7/7
Bilingual EsEn(Spanish)	title + narr	with	Okapi	0.1849	4/7

2 The Medical Task

This year our experiments focus on preprocessing the collection using Information Gain (IG) in order to improve the quality of results and to automate the tag selection process.

In order to generate the textual corpus we have preprocessed the collection using the ImageCLEFmed.xml file [1]. The process is the same as last year [3].

However, this year the XML tags have been selected according to the amount of information theoretically supplied. For this reason, we have used the information gain measure to select the best tags in the collection.

The method applied consists in computing the information gain for every tag at every sub-collection. Since each subcollection has a different set of tags, the information gain was calculated for each subcollection individually. Let C be the set of cases and E the value set for the E tag, then the formula that we have to compute must obey the following expression:

$$IG(C|E) = H(C) - H(C|E) \tag{1}$$

where

$IG(C|E)$ is the information gain for the E tag,
$H(C)$ is the entropy and of the set of cases C
$H(C|E)$ is the relative entropy of the set of cases C conditioned by the E tag

Both, $H(C)$ and $H(C|E)$ are calculated based on the frequencies of occurrence of tags according to the combination of words which they represent. After some basic operations, the final equation for the computation of the information gain supplied by a given tag E over the set of cases C is defined as follows:

$$IG(C|E) = -\log_2 \frac{1}{|C|} + \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} log_2 \frac{1}{|C_{e_j}|} \tag{2}$$

For every tag in every collection, its information gain is computed. Then, the tags selected to compose the final collection are those showing high values of IG. Once the document collection was generated, experiments were conducted with the LEMUR retrieval information system, applying the Kl-divergence weighting scheme.

2.1 Experiment Description and Results

Our main goal is to investigate the effectiveness of filtering tags using IG in the text collection. To that end, we have accomplished several experiments using the ImageCLEFmed2005 to identify the best tag percentage with experiments preserving 10%, 20%...100% of tags. The results were evaluated with the relevance assessments of the 2005 collection. Only runs with 20%, 30% and 40% of tags for the 2006 collection were submitted, because these settings led to the best results on the 2005 corpus.

We also wanted to compare the results obtained when using only the text associated to the query topic with the results obtained when we merge visual and textual information. For this purpose, the first experiment was conducted as a baseline case. This experiment simply consists in taking the text associated to each query as a new textual query, which is then fed into the LEMUR system. The resulting list is the baseline run.

The remaining experiments start from the ranked lists provided by the GIFT tool for each query. For each list/query we have used an automatic textual query expansion using the associated text to the first four ranked images from the GIFT

lists and the original textual query in order to generate a new textual query. The new textual query is then submitted to the LEMUR system and a new ranked list is obtained. Finally, the expanded textual list and GIFT list are merged in order to obtain one final list (FL) with relevant images ranked by relevance. The merging process was done giving different importance to the visual (VL) and textual lists (TL): $FL = TL * \alpha + VL * (1 - \alpha)$

In order to adjust these parameters some experiments were accomplished with the 2005 collection varying α in the range [0,1] with step 0.1 (i.e., 0, 0.1, 0.2,...,0.9 and 1). After analyzing the results, we have submitted runs with α set to 0.5, 0.6 and 0.7 for the 2006 collection.

These three experiments and the baseline experiment (that only uses textual information of the query) have been accomplished over the three different corpora generated with 20%, 30% and 40% of tags. All textual experiments have been carried out with LEMUR using PRF and the Kl-divergence weighting scheme, as pointed out previously. Table 2 shows the results obtained with the SINAI system.

Table 2. Performance in Medical Image Retrieval

Runs	Experiment	Precision
text only	SinaiOnlytL20	0.1955
text only	SinaiOnlytL30	0.2167
text only	SinaiOnlytL40	0.2215

Runs	Experiment	Precision
mixed	SinaiGiftT50L20	0.0495
mixed	SinaiGiftT50L30	0.0491
mixed	SinaiGiftT50L40	0.0494
mixed	SinaiGiftT60L20	0.0468
mixed	SinaiGiftT60L30	0.0465
mixed	SinaiGiftT60L40	0.0470
mixed	SinaiGiftT70L20	0.0435
mixed	SinaiGiftT70L30	0.0437
mixed	SinaiGiftT70L40	0.0441

References

1. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed, medical retrieval and annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS, Springer, Heidelberg (to appear, 2007)
2. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF, photographic retrieval and object annotation tasks. Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006 (to appear, 2006)
3. Martín-Valdivia, M.T., García-Cumbreras, M.Á., Díaz-Galiano, M.C., Ureña-López, L.A., Montejo-Ráez, A.: SINAI at ImageCLEF 2005. In: Proceedings of the Cross Language Evaluation Forum CLEF 2005 (2005)