

# 行政院國家科學委員會專題研究計畫 成果報告

## 跨語言跨媒體資訊檢索研究(2/2) 研究成果報告(完整版)

計畫類別：個別型  
計畫編號：NSC 95-2221-E-002-334-  
執行期間：95年08月01日至96年07月31日  
執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：陳信希

計畫參與人員：博士班研究生-兼任助理：張亦塵

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中華民國 96年10月03日

# Language Translation and Media Transformation in Cross-Language Image Retrieval

Hsin-Hsi Chen and Yih-Chen Chang

Department of Computer Science and Information Engineering  
National Taiwan University  
Taipei, Taiwan  
hhchen@csie.ntu.edu.tw; ycchang@nlg.csie.ntu.edu.tw

**Abstract.** Cross-language image retrieval facilitates the use of text query in one language and image query in one medium to access image collection with text description in another language/medium. The images with annotations are considered as a trans-media parallel corpus. In a media-mapping approach, we transform a query in one medium into a query in another medium by referencing to the aligned trans-media corpus. From the counterpart of results of an initial retrieval, we generate a new query in different medium. In the experiments, we adopted St. Andrews University Library's photographic collection used in ImageCLEF, and explored different models of language translation and media transformation. When both text query and image query are given together, the best MAP of a cross-lingual cross-media model **1L2M** (one language translation plus two media transformations) achieve 87.15% and 72.39% of those of mono-lingual image retrieval in the 2004 and the 2005 test sets, respectively. That demonstrates our media transformation is quite useful, and it can compensate for the errors introduced in language translation.

## 1. Introduction

For systematic construction of digital libraries and digital museums, large scale of images associated with captions, metadata, and so on, are

available. Users often use their familiar languages to annotate images and express their information needs. Cross-language image retrieval (CLMR) becomes more practical. CLMR, which is some sort of cross-language cross-media information retrieval (CL-CM-IR), allows users employ text queries (in one language) and example images (in one medium) to access image database with text descriptions (in another language/medium). Two languages, i.e., query language and document language, and two media, i.e., text and image, are adopted to express the information needs and the data collection. Both language translation and media transformation have to be dealt with.

Cross-language information retrieval (CLIR) facilitates the uses of queries in one language to access documents in another language. It touches on the multilingual aspect only. Language unification is the major issue in CLIR. Either query translation or document translation can be considered. In the past, dictionary-based, corpus-based and hybrid approaches have been proposed [3][11]. Dictionary-based approach exploits bilingual machine-readable dictionaries. Translation ambiguity, target polysemy and coverage of dictionaries are several important issues to tackle. Target term selection strategies like *select all*, *select N randomly* and *select best N*, and selection level like *words* and *phrases* have been presented. Corpus-based approach exploits a bilingual parallel corpus, which is a collection of original texts and their translations. Such a corpus may be document-aligned, sentence-aligned or word-aligned. Corpus-based approach has been employed to set up a bilingual dictionary, or to translate a source query to a target one. Dictionaries and corpora are complementary. The former provides broad and shallow coverage, while the latter provides narrow (domain-specific) but deep (more terminology) coverage of the language.

Compared with CLIR, image retrieval touches on medium aspect rather than multilingual issue. Two types of approaches, i.e., content-based and text-based approaches, are usually adopted in image retrieval [8]. Content-based image retrieval (CBIR) uses low-level visual features to retrieve images. In such a way, it is unnecessary to annotate images and transform users' queries. However, due to the semantic gap between image visual features and high-level concepts [7], it is still challenging to use a CBIR system to retrieve images with correct semantic meanings. Integrating textual information may help a CBIR system to cross the semantic gap and improve retrieval performance.

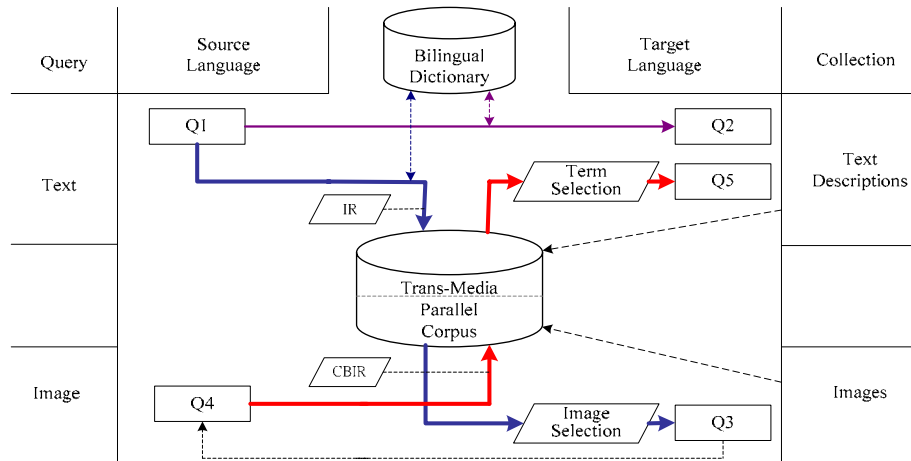
Recently several approaches tried to combine text- and content-based methods for image retrieval. A simple approach is: conducting text- and content-based retrieval separately, and merging the retrieval results of the two runs [2][9]. In contrast to the parallel approach, a pipeline approach uses textual or visual information to perform initial retrieval, and then uses the other feature to filter out irrelevant images [1]. In these two approaches, textual and visual queries are formulated by users and do not directly influence each other. Another approach, i.e., transformation-based approach, mines the relations between images and text, and uses the mined relations to transform textual information into visual one, and vice versa [10].

In this paper, we will consider how to utilize a trans-media parallel corpus to integrate textual and visual information for cross-language image retrieval. In contrast to a bilingual parallel corpus, a trans-media parallel corpus is defined to be a collection of images and their text annotations. An image is aligned to its text description. Section 2 will present a media-mapping approach. Section 3 will specify the experimental materials, i.e., St. Andrews University Library's photographic collection used in the 2004 and the 2005 ImageCLEF [4][5]. Sections 4 and 5 will show and discuss the experiments. Finally, Section 6 will conclude the remarks.

## 2. A Media-Mapping Approach

A media-mapping approach transforms a query in one medium to a query in another medium using a trans-media parallel corpus. Figure 1 sketches the concept. A visual query is submitted to a content-based information retrieval (CBIR) system. The IR system reports a set of relevant images. Since an image and its text description are aligned, the corresponding text descriptions of relevant images are also reported. Different term selection methods like high frequency terms, statistically significant terms, *etc.* proposed in multilingual text retrieval [6] can be explored.

Figure 1 shows two possible media transformations, including textual query to visual one (i.e.,  $Q1 \Rightarrow Q3$ ) and visual query to textual one (i.e.,  $Q4 \Rightarrow Q5$ ). Here,  $X \Rightarrow Y$  denotes a translation or a transformation from  $X$  to  $Y$ . Besides media transformation,  $Q1$  can also be translated into  $Q2$ , a textual query in target language.



**Figure 1.** A Media-Mapping Approach

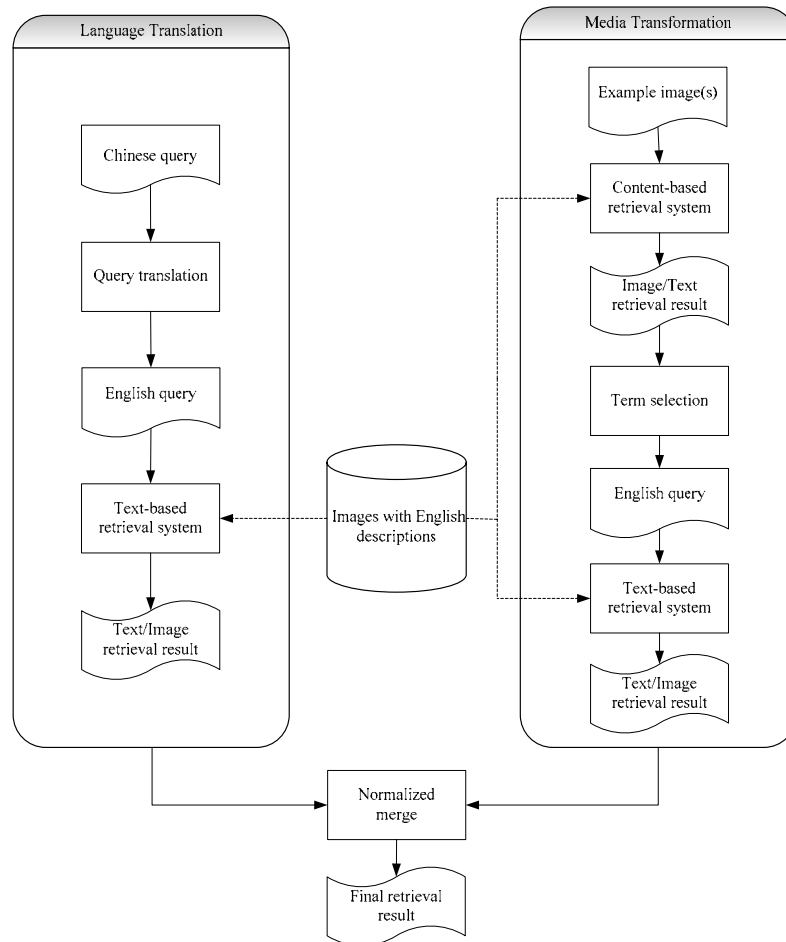
Figure 2 shows a model called **1L1M**, i.e.,  $(Q1 \Rightarrow Q2)$  and  $(Q4 \Rightarrow Q5)$ , including one language translation and one media transformation. The overall architecture consists of a textual run, an initial visual run, and a relevance feedback run. The original textual query initiates a textual run. Its procedure is the same as a traditional CLIR system. A source textual query is translated into a target textual one, and then the target query is submitted to a text-based IR system. A set of relevant text descriptions is reported along with their images.

In an initial visual run, a visual query is transformed into a textual one through the media-mapping approach. Then, the textual query is sent to a text-based IR system to retrieve the relevant text descriptions, and, at the same time, the relevant images. This procedure is similar to traditional relevant feedback except that the feedback comes from another media, i.e., text, rather than the original media, i.e., image. This is because the performance of content-based IR is usually worse than that of text-based IR.

The results generated by the textual run and the visual run are merged together. The similarity scores of images in the two runs are normalized and linearly combined by weighting.

In Section 4, we will consider an alternative model called **1L2M** which includes one language translation  $(Q1 \Rightarrow Q2)$  and two media

transformations ( $Q1 \Rightarrow Q3$ ) & ( $Q3=Q4' \Rightarrow Q5$ ). Here  $Q3=Q4'$  means the retrieval results  $Q3$  of  $Q1$  are considered as new query  $Q4'$ .



**Figure 2.** A 1L1M Cross-Language Image Retrieval System with Media Mapping

### 3. Experimental Materials

In the experiments, we adopt the 2004 and the 2005 ImageCLEF test sets [4][5]. The image collection consists of 28,133 photographs from St. Andrews University Library’s photographic collection, which is one of the largest and most important collections of historic photography in

Scotland. The majority of images (82%) in the St. Andrews image collection are in black and white. All images are accompanied by a caption written in English by librarians working at St. Andrews Library. The information in a caption ranges from specific date, location, and photographer to a more general description of an image. Figure 3 shows an example of image and its caption in the St. Andrews image collection. The text descriptions are semi-structured and consist of several fields including document number, headline, record id, description text, category, and file names of images in a 368×234 large version and 120×76 thumbnail version.



The 2004 and the 2005 test sets contain 25 topics and 28 topics, respectively. Each topic consists of a title (a short sentence or phrase describing the search request in a few words), and a narrative (a description of what constitutes a relevant or non-relevant image for that each request). In addition to the text description for each topic, one and two example images are provided for the 2004 and the 2005 topic sets, respectively. In our experiments, queries are in Chinese. Figure 4 illustrates a topic of the 2005 topic set in English and in Chinese along with two example images.

|   |  |
|---|--|
|  | <pre> &lt;DOC&gt; &lt;DOCNO&gt;stand03_1029/stand03_5473.txt   &lt;/DOCNO&gt; &lt;HEADLINE&gt; Horse and handler.   &lt;/HEADLINE&gt; &lt;TEXT&gt; &lt;RECORD_ID&gt;LHG-.000010.-.000067 &lt;/RECORD_ID&gt;   Horse Stable hand holding bridle of heavy   horse on grass slope, farm buildings, tall   chimney and trees behind wall beyond.   ca.1900 Lady Henrietta Gilmour   Scotland LHG-10-67 &lt;CATEGORIES&gt;   [chimneys - industrial],[horses &amp;   ponies],[Scotland unidentified   views],[Collection - Lady H Gilmour] &lt;/CATEGORIES&gt; &lt;/TEXT&gt; </pre> |
|---|--|

**Figure 3.** An Image and Its Description

## 4. Experiments

At first, we consider text query only. For each Chinese query term, we find its translation equivalents by using a Chinese-English bilingual dictionary. If a query term has more than one translation, the first two translations with the highest occurrences in the English image captions are considered as the target language query terms. Assume queries Q0 and Q2 are human translation and machine translation of a text query Q1, respectively, in Figure 1. Table 1 shows the mean average precision (MAP) of retrieval using Q0 and Q2. It is trivial that monolingual IR using Q0 is better than crosslingual IR using Q1 $\Rightarrow$ Q2. The MAPs of the latter are 69.72% and 60.70% of those of the former on the 2004 and the 2005 topic sets, respectively. Compared to the 2004 topic set, the MAP of using the 2005 topic set is decreased to 0.3952. It confirms that the 2005 topic set containing more general and visual queries is more challenging than the 2004 topic set [5].

|   |  |
|---|--|
|  | <pre>&lt;top&gt; &lt;num&gt; Number: 1 &lt;/num&gt; &lt;title&gt; aircraft on the ground &lt;/title&gt; &lt;narr&gt;Relevant images will show one or more airplanes positioned on the ground. Aircraft do not have to be the focus of the picture, although it should be possible to make out that the picture contains aircraft. Pictures of aircraft flying are not relevant and pictures of any other flying object (e.g. birds) are not relevant. &lt;/narr&gt; &lt;/top&gt;</pre> |
|  | <pre>&lt;top&gt; &lt;num&gt; Number: 1 &lt;/num&gt; &lt;title&gt; 地面上的飛機 &lt;/title&gt; &lt;narr&gt;相關圖片應顯示一架或多架地面上的飛 機。飛機不需要在圖片中央，但是圖片 上應該可以看到飛機。飛行中的飛機和 其他飛行物體（例如：鳥）不算相關。 &lt;/narr&gt;</pre>  |

**Figure 4.** Topic Number 1 of the 2005 Topic Set in English and in Chinese



**Table 1.** Performance of monolingual (Q0)/cross-lingual information retrieval (Q1 $\Rightarrow$ Q2)

| Model               | 2004 topic set | 2005 topic set |
|---------------------|----------------|----------------|
| Q0                  | 0.6304         | 0.3952         |
| Q1 $\Rightarrow$ Q2 | 0.4395         | 0.2399         |

Next, we consider image query only. Compared with Table 1, Table 2 shows that content-based IR is much worse than monolingual IR and crosslingual IR. Because example images (i.e., image queries) are in the data set, they are often the top-1 and the top-2 images reported by the content-based IR system for the 2004 and the 2005 topic sets, respectively. To evaluate the real performance, we consider two cases: the data sets with and without the example images. It is trivial that the MAPs of the former (0.0672 and 0.0725) are better than those of the latter (0.0149 and 0.0259).

**Table 2.** Performance of content-based information retrieval (CBIR)

|      | Keep Example Images |                | Remove Example Images |                |
|------|---------------------|----------------|-----------------------|----------------|
|      | 2004 topic set      | 2005 topic set | 2004 topic set        | 2005 topic set |
| CBIR | 0.0672              | 0.0725         | 0.0149                | 0.0259         |

With the media-mapping approach, we generate a text query Q5 from the text description counterparts of retrieved images by using Q4, shown in Figure 1. Table 3 illustrates the performance of Q5 when all the words are selected from the text descriptions of the top- $n$  retrieved images. Comparing Tables 2 and 3, media transformation from image to text is better than CBIR directly.

**Table 3.** Retrieval performance of text query transformed from image query ( $Q4 \Rightarrow Q5$ )

| Top- $n$<br>Images↓ | Keep Example Images |                   | Remove Example Images |                   |
|---------------------|---------------------|-------------------|-----------------------|-------------------|
|                     | 2004<br>topic set   | 2005<br>topic set | 2004<br>topic set     | 2005<br>topic set |
| 1                   | 0.4991              | 0.2109            | 0.0704                | 0.0582            |
| 2                   | 0.3922              | 0.3409            | 0.0486                | 0.0434            |
| 3                   | 0.2994              | 0.2912            | 0.0451                | 0.0441            |
| 4                   | 0.2380              | 0.2004            | 0.0450                | 0.0436            |
| 5                   | 0.2231              | 0.1588            | 0.0429                | 0.0450            |

Now, we consider both text and image queries. At first, we set up two baselines, i.e., merging the results of monolingual/cross-lingual IR and CBIR directly. The latter is one language translation and zero media transformation (**1L0M**). Table 4 shows the performance on two different topic sets. The weights for models  $Q0 \cup Q4$  and  $(Q1 \Rightarrow Q2) \cup Q4$  are (0.9, 0.1) and (0.7, 0.3), respectively.  $X \cup Y$  means merging results of  $X$  and  $Y$ . When example images are removed from the test collection, the naïve merging model cannot even outperform text query only model (see Table 1).

**Table 4.** Naïve merging of monolingual IR ( $Q0$ )/cross-lingual IR ( $Q1 \Rightarrow Q2$ ) and CBIR ( $Q4$ )

|                               | Keep Example Images |                   | Remove Example Images |                   |
|-------------------------------|---------------------|-------------------|-----------------------|-------------------|
|                               | 2004<br>topic set   | 2005<br>topic set | 2004<br>topic set     | 2005<br>topic set |
| $Q0 \cup Q4$                  | 0.6241              | 0.4354            | 0.5439                | 0.3770            |
| $(Q1 \Rightarrow Q2) \cup Q4$ | 0.4622              | 0.2697            | 0.4265                | 0.2365            |

We consider the 1L1M model shown in Figure 1 next. Query  $Q1$  is translated into  $Q2$  by language translation, i.e.,  $Q1 \Rightarrow Q2$ . Query  $Q4$  is transformed into  $Q5$  by media transformation, i.e.,  $Q4 \Rightarrow Q5$ . In this way, all words from the text counterparts of images retrieved by  $Q4$  form a text query  $Q5$ . Finally, we merge the retrieval results of  $Q2$  and  $Q5$  with weights 0.7 and 0.3. Table 5 depicts the experimental results of one language translation and one media transformation. No matter

whether the examples are kept in or removed from the test collection, the 1L1M model  $(Q1 \Rightarrow Q2) \cup (Q4 \Rightarrow Q5)$  is better than the 1L0M model  $(Q1 \Rightarrow Q2) \cup Q4$ .

**Table 5.** Performance of 1 Language Translation and 1 Media Transformation Using ALL

| Top- <i>n</i><br>Images↓ | Keep Example Images |                   | Remove Example Images |                   |
|--------------------------|---------------------|-------------------|-----------------------|-------------------|
|                          | 2004<br>topic set   | 2005<br>topic set | 2004<br>topic set     | 2005<br>topic set |
| 1                        | 0.5220              | 0.2930            | 0.4542                | 0.2440            |
| 2                        | 0.5243              | 0.3194            | 0.4476                | 0.2410            |
| 3                        | 0.5184              | 0.3207            | 0.4508                | 0.2497            |
| 4                        | 0.5097              | 0.3139            | 0.4484                | 0.2526            |
| 5                        | 0.4985              | 0.3030            | 0.4519                | 0.2526            |

We further introduce the 1L2M model to integrate text and image queries. This model, denoted by  $(Q1 \Rightarrow Q2) \cup ((Q1 \Rightarrow Q3 \text{ selected by } Q4) \Rightarrow Q5)$ , consists of one language translation, and two media transformations. In the first media mapping, the system employs Q1 to select 1,000 text descriptions, then Q4 to re-rank the 1,000 image counterparts, and finally reports re-ranking results Q3 as retrieval results of Q1. Because Q1 filters out most of the irrelevant images, and the search scope is narrowed down, it is more probable to select relevant images by using Q4. In the second media mapping, the re-ranking result Q3 is considered as new image query Q4', and transformed into Q5. Textual terms in Q5 are selected from the relevant text counterparts of a content-based image retrieval using Q4'. Finally, we merge the results of two monolingual text retrieval using Q2 and Q5 with weights 0.7 and 0.3.

Table 6 shows the performance of the new model when all words in the text descriptions are selected. Compared with Table 5, the performance is improved when example images are removed from test collections. We employ another alternative – say, Chi-Square, to select suitable terms to form query Q5. Table 7 shows the performance of the 1L2M model on the test collection without example images. The best MAPs are 0.4740 and 0.2729 for the 2004 and the 2005 topic sets, respectively, which are 87.15% and 72.39% of the performance of the mono-lingual image retrieval  $Q0 \cup Q4$  (refer to Table 4). Compared with 1L0M model  $(Q1 \Rightarrow Q2) \cup Q4$ , the improvement of 1L2M model

using  $\chi^2$  is verified as significant by a Wilcoxon signed-rank test with a confidence level of 95%. The boldface in Table 7 denotes the cases passing the significant test.

**Table 6.** Performance of 1 Language Translation and 2 Medial Transformations Using ALL

| Top- $n$<br>Images↓ | Keep Example Images |                   | Remove Example Images |                   |
|---------------------|---------------------|-------------------|-----------------------|-------------------|
|                     | 2004<br>topic set   | 2005<br>topic set | 2004<br>topic set     | 2005<br>topic set |
| 1                   | 0.5218              | 0.2917            | 0.4522                | 0.2628            |
| 2                   | 0.5131              | 0.3191            | 0.4552                | 0.2659            |
| 3                   | 0.5028              | 0.3210            | 0.4648                | 0.2621            |
| 4                   | 0.4921              | 0.3252            | 0.4700                | 0.2625            |
| 5                   | 0.4864              | 0.3091            | 0.4699                | 0.2654            |

**Table 7.** Performance of 1 Language Translation and 2 Medial Transformations Using  $\chi^2$  (Remove Example Images)

| $n$ ↓<br>$m$<br>→ | 2004 topic set |               |               |               | 2005 topic set |               |               |               |
|-------------------|----------------|---------------|---------------|---------------|----------------|---------------|---------------|---------------|
|                   | 10             | 20            | 30            | 40            | 10             | 20            | 30            | 40            |
| 1                 | 0.4494         | 0.4532        | 0.4519        | 0.4524        | <b>0.2586</b>  | <b>0.2628</b> | <b>0.2656</b> | <b>0.2640</b> |
| 2                 | 0.4424         | 0.4445        | 0.4493        | 0.4530        | <b>0.2659</b>  | <b>0.2625</b> | <b>0.2649</b> | <b>0.2635</b> |
| 3                 | <b>0.4737</b>  | 0.4568        | 0.4537        | 0.4607        | <b>0.2629</b>  | <b>0.2624</b> | <b>0.2610</b> | <b>0.2586</b> |
| 4                 | <b>0.4727</b>  | 0.4622        | <b>0.4740</b> | <b>0.4718</b> | <b>0.2649</b>  | <b>0.2729</b> | <b>0.2666</b> | <b>0.2636</b> |
| 5                 | <b>0.4580</b>  | <b>0.4567</b> | <b>0.4618</b> | <b>0.4600</b> | <b>0.2727</b>  | <b>0.2684</b> | <b>0.2662</b> | <b>0.2630</b> |

## 5. Discussion

We examine the retrieval results of the best model, i.e., two medial transformations, to see why the performance is improved. The following list three possible cases.

- (1) Image query compensates for the translation errors of text query. Consider a query “四輪馬車”, which is translated into “four wheel horse car” instead of “cart” or “coach”. The text retrieval returns candidate images containing “horse” at first, image retrieval then selects the most similar images from candidates by using example image, and finally the terms “cart”

or “coach” in the reported images are suggested for retrieval. Similarly, query “建築物上飄揚的旗子” is translated into “building on a flutter flag”. Image retrieval re-ranks those images containing “flags”, and contributes the concept “flying”.

- (2) The translation is correct, but the text description is not matched to the translation equivalent. For example, query “蘇格蘭的太陽” is translated into “Scotland Sun” correctly, however, the corresponding concepts in relevant images are “sunrise” or “sunset”. The first media transformation proposes those images containing “sun”, and the second media transformation suggests the relevant concepts, i.e., “sunrise” or “sunset”.
- (3) The translation is correct, and the text description of images is exactly matched to the translation equivalent. For example, query “動物雕像” is translated into “animal statue” correctly. Here, “statue” is enhanced, so that those images containing the concept are re-ranked to the top.

To sum up, in the two media transformation model, the first media mapping (i.e., text→image) derives image query to capture extra information other than text query. The second media mapping (i.e., image → text) generates text query for more reliable text-based retrieval other than content-based retrieval. These two procedures are complementary, so that the model results in good performance.

## 6. Concluding Remarks

Cross-lingual IR achieves 69.72% and 60.70% of mono-lingual IR in the two topic sets used in this paper. Content-based IR gets much less performance, i.e., it achieves only 2.36% and 6.55% of text-based IR. Naïve merging the results of text and image queries gain no benefit. It would be doubtful if integrating content-based IR and text-based IR was helpful for cross-language image retrieval under such a poor CBIR system.

Compared to content-based IR, the generated text query from the given image query using media-mapping approach improves the original performance from 0.0149 and 0.0259 to 0.0704 and 0.0582 in the best setting. When both text and image queries are considered, the best cross-lingual image retrieval model (i.e., the 1L2M model using  $\chi^2$

term selection), which achieves the MAPs of 0.4740 and 0.2729, are significantly better than the baseline cross-lingual image retrieval model (i.e., 0.4265 and 0.2365), and 87.15% and 72.39% of the baseline mono-lingual image retrieval model (i.e., 0.5439 and 0.3770).

Enhancing text-based IR with image query is more challenging than enhancing image-based IR with text query. The use of image query to re-rank the counterpart image results of a text-based IR in the first media mapping, the transformation of the re-ranked images to text queries in the second media mapping, and the employment of final text queries bring into full play of text and image queries.

In the current experimental materials, most of the images are in black and white. We will extend our models to the web, where plenty of color images are available, and various genres of annotations can be explored.

## References

1. Baan, J., van Ballegooij, A., Geusenbroek, J.M., den Hartog, J., Hiemstra, D., List, J., Patras, I., Raaijmakers, S., Snoek, C., Todoran, L., Vendrig, J., de Vries, A., Westerveld, T. and Worrying, M.: Lazy Users and Automatic Video Retrieval Tools in the Lowlands. In: Proceedings of the Tenth Text Retrieval Conference (2002) 159-168.
2. Besançon, R., Hède, P., Moellic, P.A. and Fluhr C.: Cross-Media Feedback Strategies: Merging Text and Image Information to Improve Image Retrieval. In: Proceedings of 5th Workshop of the Cross-Language Evaluation Forum LNCS 3491 (2005) 709-717.
3. Chen, H.H., Bian, G.W. and Lin, W.C.: Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In: Proceedings of 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (1999) 215-222.
4. Clough, P., Sanderson, M. and Müller, H.: The CLEF 2004 Cross Language Image Retrieval Track. In: Proceedings of 5<sup>th</sup> Workshop of the Cross-Language Evaluation Forum LNCS 3491 (2005) 597-613.
5. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J. and Hersh, W.: The CLEF 2005 Cross Language Image Retrieval Track. In: Proceedings of 6<sup>th</sup> Workshop of the Cross-Language Evaluation Forum LNCS (2006).

6. Davis, M.W. and Dunning, T.: A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval. In: Proceedings of TREC-4 (1996) 483-498.
7. Eidenberger, H. and Breiteneder, C.: Semantic Feature Layers in Content-based Image Retrieval: Implementation of Human World Features. In: Proceedings of International Conference on Control, Automation, Robotic and Vision (2002).
8. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. *Information Science* 3(2) (2000) 63-66.
9. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B. and Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St. Andrew's Collection. In: Proceedings of 5th Workshop of the Cross-Language Evaluation Forum LNCS 3491 (2005) 653-663.
10. Lin, W.C., Chang, Y.C. and Chen, H.H.: Integrating Textual and Visual Information for Cross-Language Image Retrieval. In: Proceedings of the Second Asia Information Retrieval Symposium LNCS 3689 (2005) 454-466.
11. Oard, D.W.: Alternative Approaches for Cross-Language Text Retrieval. In: Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval (1997) 131-139.

## Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval

Yih-Chen Chang and Hsin-Hsi Chen\*

Department of Computer Science and Information Engineering  
National Taiwan University  
Taipei, Taiwan  
E-mail: ycchang@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

**Abstract.** Two kinds of intermedia are explored in ImageCLEFphoto2006. The approach of using a word-image ontology maps images to fundamental concepts in an ontology and measure the similarity of two images by using the kind-of relationship of the ontology. The approach of using an annotated image corpus maps images to texts describing concepts in the images, and the similarity of two images is measured by text counterparts using BM25. The official runs show that visual query and intermedia are useful. Comparing the runs using textual query only with the runs merging textual query and visual query, the latter improved 71%~119% of the performance of the former. Even in the situation which example images were removed from the image collection beforehand, the performance was still improved about 21%~43%.

### 1. Introduction

In recent years, many methods (Clough, Sanderson, & Müller, 2005; Clough, Müller, Deselaers, Grubinger, Lehmann, Jensen, & Hersh, 2006) have been proposed to explore visual information to improve the

---

\* Corresponding author



performance of cross-language image retrieval. The challenging issue is the semantic differences among visual and textual information. For example, using the visual information “red circle” may retrieve noise images containing “red flower”, “red balloon”, “red ball”, and so on, rather than the desired ones containing “sun”. The semantic difference between visual concept “red circle” and textual symbol “sun” is called a *semantic gap*.

Some approaches conducted text- and content-based image retrieval separately and then merged the results of two runs (Besançon, et al., 2005; Jones, et al., 2005; Lin, Chang & Chen, forthcoming). Content-based image retrieval may suffer from the semantic gap problem and report noise images. That may have negative effects on the final performance. Some other approaches (Lin, Chang & Chen, forthcoming) learned the relationships between visual and textual information and used the relationships for media transformation. The final retrieval performance depends on the relationship mining.

In this paper, we use two intermedia approaches to capture the semantic gap. The main characteristic of these approaches is that human knowledge is imbedded in the intermedia and can be used to compute the semantic similarity of images. A word-image ontology and an annotated image corpus are explored and compared. Section 2 specifies how to build and use the word-image ontology. Section 3 deals with the uses of the annotated corpus. Sections 4 and 5 show and discuss the official runs in ImageCLEFphoto2006.

## **2. An Approach of Using a Word-Image Ontology**

### **2.1 Building the Ontology**

A word-image ontology is a word ontology aligned with the related images on each node. Building such an ontology manually is time consuming. In ImageCLEFphoto2006, the image collection has 20,000 colored images. There are 15,998 images containing English captions in <TITLE> and <DESCRIPTION> fields. The vocabularies include more than 8,000 different words, thus an ontology with only hundreds words is not enough.

Instead of creating a new ontology from the beginning, we extend WordNet, the well-known word ontology, to a word-image ontology. In WordNet, different senses and relations are defined for each word. For simplicity, we only consider the first two senses and kind-of relations in the ontology. Because our experiments in ImageCLEF2004 (Lin, Chang & Chen, 2006) showed that verbs and adjectives are less appropriate to be represented as visual features, we only used nouns here.

Before aligning images and words, we selected those nouns in both WordNet and image collection based on their TF-IDF scores. For each selected noun, we used Google image search to retrieval 60 images from the web. The returned images may encounter two problems: (1) they may have unrelated images, and (2) the related images may not be pure enough, i.e., the foci may be in the background or there may be some other things in the images. Zinger (2005) tried to deal with this problem by using visual features to cluster the retrieved images and filtering out those images not belonging to any clusters. Unlike Zinger (2005), we employed textual features. For each retrieved image, Google will return a short snippet. We filter out those images whose snippets do not exactly match the query terms. The basic idea is: “the more things a snippet mentions, the more complex the image is.” Finally, we get a word-image ontology with 11,723 images in 2,346 nodes.

## 2.2 Using the Ontology

### 2.2.1 Similarity Scoring

Each image contains several fundamental concepts specified in the word-image ontology. The similarity of two images is measured by the sum of the similarity of the fundamental concepts. In this paper we use kind-of relations to compute semantic distance between fundamental concepts  $A$  and  $B$  in the word-image ontology. At first, we find the least common ancestor ( $LCA$ ) of  $A$  and  $B$ . The distance between  $A$  and  $B$  is the length of the path from  $A$  through  $LCA$  to  $B$ . When computing the semantic distance of nodes  $A$  and  $B$ , the more the nodes should be traversed from  $A$  to  $B$ , the larger the distance is. In addition to the path

length, we also consider the weighting of links in a path shown as follows.

(1) When computing the semantic distance of a node  $A$  and its child  $B$ , we consider the number of children of  $A$ . The more children  $A$  has, the larger the distance between  $A$  and  $B$  is. In an extreme case, if  $A$  has only one child  $B$ , then the distance between  $A$  and  $B$  is 0. Let  $\#children(A)$  denote the number of children of  $A$ , and  $base(A)$  denote the basic distance of  $A$  and its children. We define  $base(A)$  to be  $\log(\#children(A))$ .

(2) When computing the semantic distance of a node  $A$  and its brother, we consider the level it locates. Assume  $B$  is a child of  $A$ . If  $A$  and  $B$  have the same number of brothers, then the distance between  $A$  and its brothers is larger than that between  $B$  and its brothers. Let  $level(A)$  be the depth of node  $A$ . Assume the level of root is 0. The distance between node  $A$  and its child, denoted by  $dist(A)$ , is defined to be  $C^{level(A)} \times base(A)$ . Here  $C$  is a constant between 0 and 1. In this paper,  $C$  is set to 0.9.

If the shortest path between two different nodes  $N_0$  and  $N_m$  is  $N_0, N_1, \dots, N_{LCA}, \dots, N_{m-1}, N_m$ , we define the distance between  $N_0$  and  $N_m$  to be:

$$dist(N_0, N_m) = dist(N_{LCA}) + \sum_{i=1}^{m-1} dist(N_i)$$

The larger the distance of two nodes is, the less similar the nodes are.

### 2.2.2 Mapping into the Ontology

Before counting the semantic distance between two given images, we need to map the two images into nodes of the ontology. In other words, we have to find the fundamental concepts the two images consist of. A CBIR system is adopted. It regard an image as a visual query and retrieves the top  $k$  fundamental images (i.e., fundamental concepts) in the word-image ontology. In such a case, we have two sets of nodes  $S_1 = \{n_{11}, n_{12}, n_{13}, \dots, n_{1k}\}$  and  $S_2 = \{n_{21}, n_{22}, n_{23}, \dots, n_{2k}\}$ , which correspond to the two images, respectively. We define the following formula to compute the semantic distance:

$$SemanticDistance(S_1, S_2) = \sum_{i=1}^k \min(dist(n_{1i}, n_{2j})), \quad where \quad j = 1, \dots, k$$

Given a query with  $m$  example images, we regard each example image  $Q$  as an independent visual query, and compute the semantic distance between  $Q$  and images in the collection. Note that we determine what concepts are composed of an image in the collection before retrieval. After  $m$  image retrievals, each image in the collection has been assigned  $m$  scores based on the above formula. We choose the best score for each image and sort all the images to create a rank list. Finally, the top 1000 images in the rank list will be reported.

### **3. An Approach of Using an Annotated Image Corpus**

An annotated image corpus is a collection of images along with their text annotations. The text annotation specifies the concepts and their relationships in the images. To measure the similarity of two images, we have to know how many common concepts there are in the two images. An annotated image corpus can be regarded as a reference corpus. We submit two CBIRs to the reference corpus for the two images to be compared. The corresponding text annotations of the retrieved images are postulated to contain the concepts embedded in the two images. The similarity of text annotations measures the similarity of the two images indirectly.

The image collection in ImageCLEFphoto2006 can be considered as a reference annotated image corpus. Using image collection itself as intermedia has some advantages. It is not necessary to map images in the image collection to the intermedia. Besides, the domain can be more restricted to peregrine pictures. In the experiments, the <DESCRIPTION>, <NOTE>, and <LOCATION> fields in English form the annotated corpus.

To use the annotated image corpus as intermedia to compute similarity between example images and images in image collection, we need to map these images into intermedia. Since we use the image collection itself as intermedia, we only need to map example images in this work. An example image is considered as a visual query and submitted to retrieve images in intermedia by a CBIR system. The corresponding text counterparts of the top returned  $k$  images form a long text query and it is submitted to an Okapi system to retrieval

images in the image collection. BM25 formula measures the similarity between example images and images in image collection.

#### 4. Experiments

In the formal runs, we submitted 25 cross-lingual runs for eight different query languages. All the queries with different source languages were translated into target language (i.e., English) using SYSTRAN system. We considered several issues, including (1) using different intermedia approaches (i.e., the text-image ontology and the annotated image corpus), and (2) with/without using visual query. In addition, we also submitted 4 monolingual runs which compared (1) the annotation in English and in German, and (2) using or not using visual query and intermedia. At last, we submitted a run using visual query and intermedia only. The details of our runs are described as follows:

- (1) 8 cross-lingual and text query only runs:

NTU-PT-EN-AUTO-NOFB-TXT,  
 NTU-RU-EN-AUTO-NOFB-TXT,  
 NTU-ES-EN-AUTO-NOFB-TXT,  
 NTU-ZHT-EN-AUTO-NOFB-TXT,  
 NTU-FR-EN-AUTO-NOFB-TXT,  
 NTU-JA-EN-AUTO-NOFB-TXT,  
 NTU-IT-EN-AUTO-NOFB-TXT, and  
 NTU-ZHS-EN-AUTO-NOFB-TXT.

These runs are regarded as baselines and are compared with the runs using both textual and visual information.

- (2) 2 monolingual and text query only runs:

NTU-EN-EN-AUTO-NOFB-TXT, and  
 NTU-DE-DE-AUTO-NOFB-TXT.

These two runs serve as the baselines to compare with cross-lingual runs with text query only, and to compare with the runs using both textual and visual information.

- (3) 1 visual query only run with the approach of using an annotated image corpus:

NTU-AUTO-FB-TXTIMG-WEprf.

This run will be merged with the runs using textual query only, and is also a baseline to compare with the runs using both visual and textual queries.

(4) 8 cross-lingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-PT-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-RU-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-ES-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-FR-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-ZHS-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-JA-EN-AUTO-FB-TXTIMG-T-WEprf,  
NTU-ZHT-EN-AUTO-FB-TXTIMG-T-WEprf, and  
NTU-IT-EN-AUTO-FB-TXTIMG-T-WEprf.

These runs merge the textual query only runs in (1) and visual query only run in (3) with equal weight.

(5) 8 cross-lingual runs, using both textual and visual queries with the approach of using word-image ontology:

NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-FR-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-ZHS-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-JA-EN-AUTO-NOFB-TXTIMG-T-IOntology,  
NTU-ZHT-EN-AUTO-NOFB-TXTIMG-T-IOntology, and  
NTU-IT-EN-AUTO-NOFB-TXTIMG-T-IOntology.

These runs merge textual query only runs in (1), and visual query runs with weights 0.9 and 0.1.

(6) 2 monolingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-EN-EN-AUTO-FB-TXTIMG, and  
NTU-DE-DE-AUTO-FB-TXTIMG

These two runs using both textual and visual queries. The monolingual run in (2) and the visual run in (3) are merged with equal weight.

## 5. Results and Discussions

Table 1 shows experimental results of official runs in ImageCLEFphoto2006. We compare performance of the runs using textual query only, and the runs using both textual and visual queries

(i.e., Text Only vs. Text + Annotation and Text + Ontology). In addition, we also compare the runs using word-image ontology and the runs using annotated image corpus (i.e., Text + Ontology vs. Text + Annotation). The runs whose performance is better than that of baseline (i.e., Text Only) will be marked in bold. The results show all runs using annotated image corpus are better than the baseline. In contrast, only two runs using word-image ontology are better.

**Table 1.** Performance of Official Runs (T=text only, A=annotated image corpus, O=word-image ontology)

| Query Language | MAP           | Description | Runs                                   |
|----------------|---------------|-------------|--|
| Portuguese     | 0.1630        | T           | NTU-PT-EN-AUTO-NOFB-TXT                |
|                | <b>0.2854</b> | T+A         | NTU-PT-EN-AUTO-FB-TXTIMG-T-WEprf       |
|                | 0.1580        | T+O         | NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology |
| Russian        | 0.1630        | T           | NTU-RU-EN-AUTO-NOFB-TXT                |
|                | <b>0.2789</b> | T+A         | NTU-RU-EN-AUTO-FB-TXTIMG-T-Weprf       |
|                | 0.1591        | T+O         | NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology |
| Spanish        | 0.1595        | T           | NTU-ES-EN-AUTO-NOFB-TXT                |
|                | <b>0.2775</b> | T+A         | NTU-ES-EN-AUTO-FB-TXTIMG-T-Weprf       |
|                | 0.1554        | T+O         | NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology |
| French         | 0.1548        | T           | NTU-FR-EN-AUTO-NOFB-TXT                |
|                | <b>0.2758</b> | T+A         | NTU-FR-EN-AUTO-FB-TXTIMG-T-WEprf       |
|                | 0.1525        | T+O         | NTU-FR-EN-AUTO-                        |

|                        |               |     |   |
|------------------------|---------------|-----|---|
|                        |               |     | NOFB-TXTIMG-T-<br>IOntology                     |
| Simplified             | 0.1248        | T   | NTU-ZHS-EN-AUTO-<br>NOFB-TXT                    |
| Chinese                | <b>0.2715</b> | T+A | NTU-ZHS-EN-AUTO-FB-<br>TXTIMG-T-Weprf           |
|                        | <b>0.1262</b> | T+O | NTU-ZHS-EN-AUTO-<br>NOFB-TXTIMG-T-<br>IOntology |
| Japanese               | 0.1431        | T   | NTU-JA-EN-AUTO-NOFB-<br>TXT                     |
|                        | <b>0.2705</b> | T+A | NTU-JA-EN-AUTO-FB-<br>TXTIMG-T-Weprf            |
|                        | 0.1396        | T+O | NTU-JA-EN-AUTO-NOFB-<br>TXTIMG-T-IOntology      |
| Traditional<br>Chinese | 0.1228        | T   | NTU-ZHT-EN-AUTO-<br>NOFB-TXT                    |
|                        | <b>0.2700</b> | T+A | NTU-ZHT-EN-AUTO-FB-<br>TXTIMG-T-Weprf           |
|                        | <b>0.1239</b> | T+O | NTU-ZHT-EN-AUTO-<br>NOFB-TXTIMG-T-<br>IOntology |
| Italian                | 0.1340        | T   | NTU-IT-EN-AUTO-NOFB-<br>TXT                     |
|                        | <b>0.2616</b> | T+A | NTU-IT-EN-AUTO-FB-<br>TXTIMG-T-Weprf            |
|                        | 0.1287        | T+O | NTU-IT-EN-AUTO-NOFB-<br>TXTIMG-T-IOntology      |

The reason why the word-image ontology does not perform as our expectation may be that the images in the word-image ontology come from the web and the images in the web still contain much noise even after filtering. To deal with this problem, a better method in the image filtering is indispensable.

Since the example images in this task are in the image collection, the CBIR system always correctly maps the example images into themselves at mapping step. We made some extra experiments to



examine the performance of our intermedia approach. In the experiments, we took out the example images from the image collection when mapping example images into intermedia. Table 2 shows the experiment results. Comparing Table 1 and Table 2 we find the performance of Table 2 is lower than that of Table 1. It shows the performance of CBIR in mapping stage will influence the final result and that is very critical. From Table 2, we also find that the approaches of annotated image corpus are better than the runs using textual query only. It shows even mapping stage have some errors, the annotated image corpus can still work well.

**Table 2.** MAP of Runs by Removing Example Images from the Collection

| Query Language      | Text Only | Text + Annotated image corpus |
|---------------------|-----------|-------------------------------|
| Portuguese          | 0.1630    | <b>0.1992</b>                 |
| Russian             | 0.1630    | <b>0.1880</b>                 |
| Spanish             | 0.1595    | <b>0.1928</b>                 |
| French              | 0.1548    | <b>0.1848</b>                 |
| Simplified Chinese  | 0.1248    | <b>0.1779</b>                 |
| Japanese            | 0.1431    | <b>0.1702</b>                 |
| Traditional Chinese | 0.1228    | <b>0.1757</b>                 |
| Italian             | 0.1340    | <b>0.1694</b>                 |

Table 3 shows the experiment results of monolingual runs. Using both textual and visual queries are still better than runs using textual query only. The performance of the runs by taking out the example images from collection beforehand is still better than the runs use textual query only. From this table, we also find the runs using textual query only does not perform well even in monolingual runs. This may be because the image captions of this year are shorter and we do not have enough information when we use textual information only. In addition, when image captions are short too, the little differences in vocabularies between query and document may influence the results a lot. Therefore, German monolingual run and English monolingual run perform so different.

Table 4 shows the experiment of runs that using visual query and annotated image corpus only, i.e., the textual query is not used. When example images were kept in the image collection, we can always map the example images into the right images. Therefore, the translation from visual information into textual information will be more correctly. The experiment shows the performance of visual query runs is better than that of textual query runs when the transformation is correct.

**Table 3.** Performance of Monolingual Image Retrieval

| Query Language    | MAP           | Description | Runs                         |
|-------------------|---------------|-------------|------------------------------|
| English           | 0.1787        | T           | NTU-EN-EN-AUTO-NOFB-TXT      |
| (+example images) | <b>0.2950</b> | T+A         | NTU-EN-EN-AUTO-FB-TXTIMG     |
| (-example images) | <b>0.2027</b> | T+A         | NTU-EN-EN-AUTO-FB-TXTIMG-NoE |
| German            | 0.1294        | T           | NTU-DE-DE-AUTO-NOFB-TXT      |
| (+example images) | <b>0.3109</b> | T+A         | NTU-DE-DE-AUTO-FB-TXTIMG     |
| (-example images) | <b>0.1608</b> | T+A         | NTU-DE-DE-AUTO-FB-TXTIMG-NoE |

**Table 4.** Performance of Visual Query

| MAP           | Description           | Runs                         |
|---------------|-----------------------|------------------------------|
| 0.1787        | T (monolingual)       | NTU-EN-EN-AUTO-NOFB-TXT      |
| <b>0.2757</b> | V+A (+example images) | NTU-AUTO-FB-TXTIMG-Weprf     |
| 0.1174        | V+A (-example images) | NTU-AUTO-FB-TXTIMG-Weprf-NoE |

## 6. Conclusion

The experiments show visual query and intermedia approaches are useful. Comparing the runs using textual query only with the runs merging textual query and visual query, the latter improved 71%~119% of performance of the former. Even in the situation which example images are removed from the image collection, the performance can still be improved about 21%~43%. We find visual query in image retrieval is important. The performance of the runs using visual query only can be even better than the runs using textual only if we translate visual information into textual one correctly. In this year the word-image ontology built automatically still contain much noise. We will investigate how to filter out the noise and explore different methods.

## References

1. Besançon, R., Hède, P., Moellic, P.A., & Fluhr, C. (2005). Cross-media feedback strategies: Merging text and image information to improve image retrieval. In Peters, C.; Clough, P.; Gonzalo, J.; Jones, G.J.F.; Kluck, M.; Magnini, B. (Eds.), *Proceedings of 5th Workshop of the Cross-Language Evaluation Forum*, LNCS 3491, (pp. 709-717). Berlin: Springer.
2. Clough, P., Sanderson, M. & Müller, H. (2005). The CLEF 2004 cross language image retrieval track. In Peters, C.; Clough, P.; Gonzalo, J.; Jones, G.J.F.; Kluck, M.; Magnini, B. (Eds.), *Proceedings of 5th Workshop of the Cross-Language Evaluation Forum*, LNCS 3491, (pp. 597-613). Berlin: Springer.
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., & Hersh, W. (2006). The CLEF 2005 cross-language image retrieval track, *Proceedings of 6<sup>th</sup> Workshop of the Cross Language Evaluation Forum*, Lecture Notes in Computer Science, 2006.
4. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., & Way, A. (2005). Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection. In Peters, C.; Clough, P.; Gonzalo, J.; Jones, G.J.F.; Kluck, M.; Magnini, B. (Eds.), *Proceedings of 5th Workshop of the Cross-*

*Language Evaluation Forum*, LNCS 3491, (pp. 653-663). Berlin: Springer.

5. Lin, W.C., Chang, Y.C. and Chen, H.H. (forthcoming). "Integrating Textual and Visual Information for Cross-Language Image Retrieval: A Trans-Media Dictionary Approach." *Information Processing and Management*, Special Issue on Asia Information Retrieval Research.
6. Zinger, S. (2005). "Extracting an Ontology of Portable Objects from WordNet." *Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*.

# 行政院國家科學委員會補助國內專家學者出席國際學術會議報告

96年5月21日

附件三

|                |  |              |                        |
|----------------|--|--------------|------------------------|
| 報告人姓名          | 陳信希  | 服務機構<br>及職稱  | 國立台灣大學<br>資訊工程學系<br>教授 |
| 時間<br>會議<br>地點 | 96年5月14日-5月19日<br>日本東京   | 本會核定<br>補助文號 | NSC 95-2221-E-002-334  |
| 會議<br>名稱       | (中文)<br>(英文) The 6th NTCIR Workshop on Evaluation of Information Access Technologies   |              |                        |
| 發表<br>論文<br>題目 | (英文)<br>(1) Overview of CLIR Task at the Sixth NTCIR Workshop<br>(2) Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task<br>(3) Overview of Opinion Analysis Pilot Task at NTCIR-6<br>(4) Using Opinion Scores of Words for Sentence-Level Opinion Extraction |              |                        |

報告內容應包括下列各項：

### 一、參加會議經過

The 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (第六屆 NTCIR 資訊存取評估會議, NTCIR-6), 5月15日-5月18日在日本東京 National Institute of Informatics 大會堂舉行, 會議前後(5月14日和5月19日)也舉行 NTCIR-7 圓桌會議, 討論 NTCIR-7 相關事宜。筆者於5月13日搭乘中華航空 CI 104 班機, 抵達東京, 5月20日 NTCIR-6 國際會議結束後, 搭乘中華航空 CI 101 班機返回台北。

### 二、與會心得

EVIA 2007 (1<sup>st</sup> International Workshop on Evaluating Information Access)是 NTCIR-6 會前會, 本次會議邀請亞洲中日韓語之外的語言檢索研究人員參與, 探討簡體中文資訊檢索評比、越南文件檢索、印度文資訊檢索評比規劃、和泰文搜索引擎評比, 以及新的評估方法研究。

正式議程包括 QAC, CLIR, CLQA, PATENT, Opinion, 和 MuST 等六個評比項目的整體報告(評比過程、使用測試集、參與的研究團隊、採用的方法、和效能分析), 以及參與評比研究團隊的系統和技術報告。此外, 會議單位也邀請另外兩個資訊檢索國際評比: TREC 和 CLEF 做專題報告, 以經驗分享。會議中也舉行數次圓桌會議, 針對評比內容、程序、評估方式等方面, 由 task organizers 和 task participants 進行面對面交流, 以作為下次會議的參考。

下年度共有 8 個規劃提案: Complex CLQA (CCLQA)、CLIR For Blog (CLIRB)、Multilingual Opinion Analysis (MOAT)、Multimodal Summarization for Trend Information (MuST)、Patent Processing (translation, mining) (PAT)、Question Answering Challenge (QAC5)、Simplified Chinese IR (CLIR-SC)、和 User Satisfaction Task (USAT)中, 在 PC meeting 中, 選出 Complex CLQA、Multilingual Opinion Analysis (MOAT)、和 Patent Processing (PAT), 作為 NTCIR-7 的評比項目。其中 Multilingual Opinion Analysis (MOAT) 為筆者、Yohei Seki、和 David Kirk Evans 所合作提出。

比較可惜的是舉辦數年的 CLIR 評比, 下年度不再為獨立項目, 而是併入到 MOAT 中。主要的原因之一是: CLIR 近年來在技術層面上, 並沒有明顯的進展, 雖然我們擬引進部落格語料, 但是語料的特徵並沒有很突顯出來。另一個可惜的是 CLQA 過去兩屆是由林川傑教授和 Yutaka Sasaki 博士共同舉辦, 由於 Sasaki 博士轉到曼徹斯特大學任教, 工作繁忙, 下年度轉由 CMU 的 Mitamura 和 Nyberg 舉辦。

### 三、建議

NTCIR 是國際三大資訊檢索評比, 台灣大學過去 5 年參與規劃舉辦 CLIR、CLQA、和 Opinion Analysis 三項主題, 有很多國際研究團隊參加評比, 實驗系統的效能, 未來持續的參與, 才能發揮影響力。

### 四、攜回資料名稱及內容

Proceedings of The 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, and CD-ROM