

Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval

Wen-Cheng Lin, Changhua Yang, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
denislin@nlg.csie.ntu.edu.tw
{d91013, hh_chen}@csie.ntu.edu.tw

Abstract. In this paper we propose an approach to deal with the Chinese-English cross-language image retrieval problem. Text-based image retrieval and query translation methods were adopted in the experiments. A similarity-based backward transliteration model with candidate filter was proposed to translate proper nouns that have no entries in a bilingual dictionary. The experimental results showed that using similarity-based backward transliteration increased retrieval performances.

1 Introduction

Multimedia data has an explosive growth nowadays, and more and more people are searching and using it. Searching in a large amount of data is not easy, thus the efficient retrieval of multimedia data becomes an important research issue. Two types of approaches, i.e., content-based and text-based approaches, are usually adopted. Content-based approaches use low-level visual features such as color, texture and shape to represent multimedia objects. Text-based approaches use collateral texts to describe the objects. Low-level visual features only show what the images or videos look like, but cannot tell us exactly what they are. On the other hand, text can describe the content of multimedia objects. Several hybrid approaches [12, 14, 15] that integrate visual and textual information have been proposed. Experimental results showed that the optimal technique depends on the query. The combined approach could outperform text- and content-based approaches in some cases.

Most of the previous work in image retrieval has focused on monolingual retrieval. Little work has been done on cross-language tasks. Sanderson and Clough [11] pointed out the need for cross-language image retrieval and discussed some issues in image CLIR. Images are good media in the context of cross language. People with no strong language skills can easily understand and judge the relevance of retrieved images. In this paper, we adopt a text-based approach to deal with the Chinese-English cross-language image retrieval problem. Query translation is adopted to unify the languages in queries and image captions. Proper noun processing plays an important role in query translation [1, 9]. IR systems must handle proper noun translation appropriately to achieve better performance. Bilingual dictionaries are usually

adopted to provide translations when translating query terms. However, proper nouns are often not listed in lexicons. We propose a similarity-based backward transliteration model to translate proper nouns not included in dictionaries.

The rest of this paper is organized as follows. Section 2 describes similarity-based backward transliteration. Section 3 shows the query translation methods. Section 4 discusses the experimental results. Finally, Section 5 provides concluding remarks.

2 Backward Transliteration

2.1 Similarity-Based Backward Transliteration

Lin and Chen [7, 8] proposed a similarity-based framework to model backward transliteration. In the similarity-based framework, the similarities of a transliterated word and candidate words are computed, and the candidate word with the highest similarity is chosen as the original word. The similarities can be measured at three levels, i.e. physical sounds, graphemes and phonemes. Comparing similarities at phoneme level has been shown to outperform the grapheme level [7, 8]. When comparing similarities at the phoneme level, the transliterated word and candidate words are first transformed into phonetic representations, i.e., International Phonetic Alphabet (IPA), and then the similarities between the IPA strings are measured.

The similarity score of two strings is the score of the optimal alignment. Given two strings S_1 and S_2 , let Σ be the alphabet of S_1 and S_2 , $\Sigma' = \{\Sigma, \text{'_'}\}$, where '_' stands for space. Space could be inserted into S_1 and S_2 such that they are of equal length and denoted as S_1' and S_2' . S_1' and S_2' are aligned when every character in either string has a one-to-one mapping to a character or space in the other string. The similarity score of an alignment is measured by the following formula.

$$Score = \sum_{i=1}^l s(S_1'(i), S_2'(i)) \quad (1)$$

where $s(a, b)$ is the similarity score between the character a and b in Σ' ,

$S'(i)$ is the i^{th} character in the string S' , and

l is the length of S_1' and S_2' .

The similarity score $s(a, b)$ can be manually assigned or automatically learned. Lin and Chen [8] proposed a learning approach based on the Widrow-Hoff rule [6] to acquire phonetic similarities from a training corpus. The learning algorithm can capture subtle similarities that cannot easily be manually assigned based on phonological knowledge. The experimental results showed that learned similarities are more discriminative than manually assigned one.

The optimal alignment of two strings S_1 and S_2 can be computed efficiently using dynamic programming. Let T be an $n+1$ by $m+1$ table, where n and m are the lengths of S_1 and S_2 respectively. By filling table T row by row, we can obtain the optimal alignment and the similarity score of S_1 and S_2 . The base condition is defined as follows.

$$T(i,0) = \sum_{1 \leq k \leq i} s(S_1(k), ' _') . \quad (2)$$

$$T(0, j) = \sum_{1 \leq k \leq j} s(' _', S_2(k)) . \quad (3)$$

The recurrence formula is defined as follows.

$$T(i, j) = \max \begin{bmatrix} T(i-1, j-1) + s(S_1(i), S_2(j)), \\ T(i-1, j) + s(S_1(i), ' _'), \\ T(i, j-1) + s(' _', S_2(j)) \end{bmatrix} . \quad (4)$$

where $1 \leq i \leq n$, $1 \leq j \leq m$.

2.2 Candidate Filter

Similarity based backward transliteration with automatically learned phonetic similarities works well, but will cost too much time if there are a lot of candidate words. This is not suitable for some applications like online IR systems. To reduce processing time, we use a pre-processing to decrease the number of candidates. A transliterated word and its original word should contain the same or similar phonemes, and the order of the phonemes should remain the same. In other words, if two IPA strings contain several identical or similar characters, their similarity may be higher. A vector space IR model is adopted to select the appropriate candidates for a transliterated word. The document set is the set of IPA strings of a list of proper nouns in the source language. Each proper noun is treated as one document. The query is the IPA string of the transliterated word. After retrieval, the top ranked documents (i.e., candidate words) are selected as the appropriate candidates of the transliterated word.

The transliterated word and its original word do not always contain the same phonemes due to the different pronunciation of different languages. For example, the English phoneme 'g' is usually transliterated into the Chinese phoneme 'k'¹. If only the phonemes of the transliterated word are used as the query terms, the original word may not be retrieved. Thus, the query has to be expanded with the most co-transliterated phonemes. The co-transliterated Chinese-English phoneme pairs are trained from a Chinese-English personal name corpus, which has 51,114 pairs of Chinese transliterated names and the corresponding English original names. A variant Mutual Information is adopted to measure the strength of co-transliteration of two phonemes. The variant Mutual Information can solve the preference for rare terms problem of traditional Mutual Information [5]. Let x be a Chinese phoneme and y an English phoneme. The Mutual Information of x and y is defined as follows:

¹ All phonemes are represented in SAMPA, which can represent IPA in ASCII.

$$MI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \times \log(f(x, y)) \quad (5)$$

where $p(x)$ is the occurrence probability of phoneme x in Chinese names,
 $p(y)$ is the occurrence probability of phoneme y in English names,
 $p(x, y)$ is the probability that x and y occur in a pair of transliterated and original names, and
 $f(x, y)$ is the frequency of x and y occurring in a pair of transliterated and original names.

A phoneme x in a transliterated word will be expanded with the phonemes that have positive MI values with x . The augmented phonemes are weighted by $MI(x, y)/\text{the number of augmented terms}$.

3 Query Translation

In our experiments, Chinese queries were used as the source language queries. The Chinese queries are translated from English by native speakers. We adopted query translation to unify the languages used in queries and documents. First, the Chinese queries were segmented by a word recognition system, and tagged by a POS tagger. Name entities were then identified [3]. For each Chinese query term, we found its translation equivalents by looking up a Chinese-English bilingual dictionary. The bilingual dictionary is integrated from four resources, including the LDC Chinese-English dictionary, Denisowski's CEDICT², BDC Chinese-English dictionary v2.2³ and a dictionary used in query translation in the MTIR project [1]. The dictionary contains 200,037 words, where a word may have more than one translation. We adopted the following two methods to select appropriate translations.

(1) CO model [2]

The CO model employed word co-occurrence information extracted from a target language text collection to disambiguate query term translations. We adopted traditional Mutual Information to measure the co-occurrence strength between words. The MI values of English words were trained on the TREC6 text collection [13]. For a query term, we compared the MI values of all the translation equivalent pairs (x, y) , where x is the translation equivalent of this term, and y is the translation equivalent of another query term within a sentence. The word pair (x_i, y_j) with the highest MI value is extracted, and the translation equivalent x_i is regarded as the best translation equivalent of this query term. Selection is carried out on the basis of the order of the query terms.

² The dictionary is available at <http://www.mandarintools.com/cedict.html>

³ The BDC dictionary is developed by the Behavior Design Corporation (<http://www.bdc.com.tw>).

(2) First-two-highest-frequency

The first two translation equivalents with the highest frequency of occurrence in the English image captions were considered as the target language query terms.

There are 150 distinct Chinese query terms in 50 topics. A total of 16 query terms could not be found in our dictionary. Of these 16 terms, 7 terms were tagged as personal names, and 5 terms were location names. These names are Chinese translations of foreign names. We can use a backward transliteration scheme to translate these names. First, we adopted the transformation rules [4] to identify the name part and keyword part of a name. The keyword parts are general nouns, e.g., “湖” (lake), “河” (river) and “橋” (bridge), and can be translated by looking up in the dictionary. We used the first-two-highest-frequency method to translate keywords. The name parts are transliterations of foreign names, and were transliterated into English in the following way.

- (1) The personal names and the location names in the English image captions were extracted. We collected a list of English names that contained 50,979 personal names and 19,340 location names. If a term in the captions can be found in the name list, it was extracted. Total 3,599 names were extracted from the image captions.
- (2) For each Chinese name, 300 candidates were selected from the 3,599 English names by using the candidate filter described in Section 2.2.
- (3) The similarity-based backward transliteration approach described in Section 2.1 was adopted to translate the Chinese name. The top 6 candidates with the highest similarities were considered as the translations of the Chinese name.

In the segmentation and name identification stage, some terms were segmented or tagged incorrectly. These errors propagated to the translation stage and affected the performance of backward transliteration. In order to evaluate the real performance of similarity-based backward transliteration, we conducted manual runs in which the Chinese queries were segmented and tagged manually. In the manual runs, there are 136 distinct query terms and 18 terms have no translations. Among the 18 terms, 5 terms were tagged as person names, 9 terms were location names and 1 term was an organization name.

4 Experiments

In the experiments, we adopted a text-based approach. The captions were used to represent the images. The Okapi IR system [10] was adopted to index and retrieve the image captions. The weighting function was BM25. For each image, the caption text, <HEADLINE> and <CATEGORIES> sections were used for indexing. The words in these sections were stemmed, and stopwords were removed. The translated English queries were used to retrieve the image captions. Only the title sections of topics were used to construct queries.

We submitted eight runs in CLEF 2003 image track. The performances of two query translation methods with or without similarity-based backward transliteration were compared. The details of the submitted runs are shown in Table 1. The performances are shown in Table 2.

Table 1. Configurations of official runs

Run	Segmentation and Tagging	Query Translation	Backward Transliteration
NTUiaCo	Automatically	CO	No
NTUiaCoP	Automatically	CO	Yes
NTUiaF2hf	Automatically	First-two-highest-frequency	No
NTUiaF2hfP	Automatically	First-two-highest-frequency	Yes
NTUimCo	Manually	CO	No
NTUimCoP	Manually	CO	Yes
NTUimF2hf	Manually	First-two-highest-frequency	No
NTUimF2hfP	Manually	First-two-highest-frequency	Yes

Table 2. Results of official runs (Average Precision)

Run	Intersection Strict	Intersection Relaxed	Union Strict	Union Relaxed
NTUiaCo	0.1712	0.1876	0.1921	0.1869
NTUiaCoP	0.1892	0.2054	0.2103	0.2060
NTUiaF2hf	0.2635	0.2754	-	0.2496
NTUiaF2hfP	0.2888	0.3004	0.2852	0.2785
NTUimCo	0.1985	0.2210	0.2233	0.2219
NTUimCoP	0.2241	0.2459	0.2483	0.2475
NTUimF2hf	0.2821	0.3042	0.2808	0.2814
NTUimF2hfP	0.3143	0.3359	0.3148	0.3193

The results show that using similarity-based backward transliteration to translate proper nouns improves performance. In the automatic segmentation runs, twelve topics had proper nouns not contained in our dictionary. After applying the similarity-based backward transliteration model, the proper nouns in six topics were translated correctly, and the average precision for these topics increased dramatically. The performance of the twelve topics is shown in Table 3. Terms in square brackets are keywords extracted by transformation rules. In Topic 16, “丹地” (dan di), transliterated from “Dundee”, was not translated correctly due to an error of keyword extraction. “丹地” was tagged as a location name, and “地” (di) was identified as a keyword according to the transformation rules. By dictionary look-up, “地” (di) was translated into “field” and “ground”. Only “丹” (dan) was transliterated by similarity-based backward transliteration and the similarity between “丹” (dan) and “Dundee” is low. Five terms were segmented incorrectly, and thus were transliterated incorrectly.

Table 3. Performance of similarity-based backward transliteration in automatic runs (Intersection Strict)

Topic	Proper Noun	Translation Result		NTUiaCo	NTUiaCoP	NTUiaF2hf	NTUiaF2hfP
		Name	Keyword				
3	安德魯斯 (Andrews)	Correct (rank 1)	-	0.0012	0.0216	0.0000	0.0123
9	亞當森 (Adamson)	Correct (rank 2)	-	0.0017	0.0271	0.0072	0.0477
10	克萊德 (Clyde)	Correct (rank 6)	-	0.0001	0.0870	0.0027	0.1019
12	安德魯斯 (Andrews)	Correct (rank 1)	-	0.0000	0.4314	0.0000	0.4293
	[北][街] (North street)	-	Correct				
14	蒙湖	Segmentation error	-	0.0188	0.0131	0.0025	0.0017
16	丹[地] (Dun- dee)	Keyword ex- traction error	-	0.0786	0.0573	0.1156	0.1165
17	茅斯	Segmentation error	-	0.0038	0.0034	0.0011	0.0011
19	卡羅斯大	Segmentation error	-	0.0000	0.0000	0.0063	0.0053
21	吉爾摩 (Gilmour)	Correct (rank 2)	-	0.0110	0.3227	0.0088	0.2416
27	灘上	Segmentation error	-	0.1425	0.0716	0.1020	0.1297
38	湯普森 (Thompson)	Correct (rank 1)	-	0.0104	0.1468	0.0444	0.4704
40	瓦倫	Segmentation error	-	0.0341	0.0185	0.0184	0.0165

In manual segmentation runs, the segmentation error problem is excluded. A total of 14 topics contain proper nouns which were not in our dictionary. The performances of seven and ten topics were increased after applying similarity-based backward transliteration model followed CO and first-two-highest-frequency models, respectively. The performances are shown in Table 4. Examining the results of backward transliteration, the original English words of about 50% Chinese proper nouns were in the top 6 ranks. Remember that the top 6 ranked terms were added to the query. One term at most was correct, the others were noise. Although noise was introduced, the performance of the topics in which the proper nouns were backward transliterated correctly improved. If the performance of backward transliteration is improved and fewer incorrect terms are added to the queries, the retrieval performance should be better. We also found that the original words of four Chinese transliterated words were not included in the name list. Thus, these original English words were not contained in the candidate lists. How to enlarge the coverage of the name list is also an important issue.

Table 4. Performances of similarity-based backward transliteration in manual runs (Intersection Strict)

Topic	Proper Noun	Translation Result		NTUimCo	NTUimCoP	NTUimF2hf	NTUimF2hFP
		Name	Keyword				
3	聖安德魯斯 (St Andrews)	Keyword extraction error	-	0.0012	0.0004	0.0002	0.0001
9	亞當森 (Adamson)	Correct (rank 2)	-	0.0017	0.0271	0.0072	0.0477
10	克萊德[河] (Clyde river)	Correct (rank 6)	Correct	0.0003	0.0974	0.0031	0.1019
12	聖安德魯斯 (St Andrews)	Keyword extraction error	-	0.0000	0.0678	0.0000	0.0678
	[北][街] (North street)	-	Correct				
14	洛蒙[湖] (Lomond loch)	Correct (rank 4)	Correct	0.0206	0.7295	0.0048	0.4463
15	泰[橋] (Tay bridge)	Has no en- try in the name list	Correct	0.0210	0.0433	0.1119	0.1335
16	丹[地] (Dundee)	Keyword extraction error	-	0.0786	0.0573	0.1156	0.1165
17	[大]亞茅斯 (Great Yar- mouth)	Has no en- try in the name list	Wrong	0.0057	0.0025	0.0036	0.0011
19	卡羅斯 (Cul- ross)	Has no en- try in the name list	-	0.0072	0.0055	0.0044	0.0038
21	亨利耶塔 (Henrietta)	Has no en- try in the name list	-	0.0233	0.3281	0.0133	0.2996
	吉爾摩 (Gilmour)	Correct (rank 2)	-				
26	勃恩斯 (Burns)	Correct (rank 3)	-	0.3571	0.3022	0.0119	0.2386
35	尼維[峰] (Nevis ben)	Wrong (rank 129)	Wrong	0.0000	0.0091	0.0000	0.0091
38	湯普森 (Thompson)	Correct (rank 1)	-	0.0104	0.1468	0.0444	0.4704
40	瓦倫坦 (Valentine)	Correct (rank 1)	-	0.0362	0.0253	0.0236	0.0175

Comparing the performances of two query translation models, surprisingly, the CO model was worse than the first-two-highest-frequency model. In the CO model, only one translation equivalent is selected for a query term. Since the image captions are very short, the suggested English translation may be not used in the captions. If we expand queries or captions, the performance may be better. On the other hand, the first-two-highest-frequency model selects the translations with the highest frequency in the target documents. Most of the English translated query terms present in the captions. The term usage is more consistent in the first-two-highest-frequency model.

5 Conclusion

In this paper we proposed an approach to deal with the Chinese-English cross-language image retrieval problem. Text-based image retrieval and query translation were adopted in the experiments. A similarity-based backward transliteration model with candidate filter was proposed to translate the proper nouns. The experimental results showed that using similarity-based backward transliteration increased the retrieval performance. The average precision of about 50% of topics containing proper nouns that are not included in dictionary were increased. The performance of the remaining topics decreased due to the failure of backward transliteration. The errors come from segmentation errors, named entity identification errors, keyword extraction errors, and the coverage of the list of names. Several methods such as learning phoneme similarity from larger data, extracting more named entities from the target document set, and improving the performance of candidate filter and keyword extraction will be further investigated to improve the performance of the similarity-based backward transliteration model.

The consistency of term usage is also an important issue. The image captions are usually short and the words used in captions are limited. Query expansion or document expansion could resolve this problem. We will experiment with various expansion approaches in the future.

References

1. Bian, G.W. and Chen, H.H.: Cross Language Information Access to Multilingual Collections on the Internet. *Journal of American Society for Information Science*, 51(3). (2000) 281-296.
2. Chen, H.H., Bian, G.W. and Lin, W.C.: Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In: *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (1999) 215-222.
3. Chen, H.H., Ding, Y.W, Tsai, S.C. and Bian, G.W.: Description of the NTU System Used for MET2. In: *Proceedings of 7th Message Understanding Conference*. (1998).
4. Chen, H.H., Yang, C. and Lin, Y.: Learning Formulation and Transformation Rules for Multilingual Named Entities. In: *Proceedings of ACL 2003 Workshop on Multilingual and Mixed Language Named Entity Recognition: Combining Statistical and Symbolic Models*. Association for Computational Linguistics (2003) 1-8.

5. Church, K., Gale, W., Hanks, P. and Hindle, D.: Parsing, Word Associations and Typical Predicate-Argument Relations. In: Proceedings of International Workshop on Parsing Technologies. (1989) 389-398.
6. Duda, R.O., Hart, P.E. and Stork, D.G.: Pattern Classification. 2nd edition. Wiley-Interscience Publication (2001).
7. Lin, W.H. and Chen, H.H.: Similarity Measure in Backward Transliteration Between Different Character Sets and Its Application to CLIR. In: Proceedings of the Thirteenth Conference of Research on Computational Linguistics (ROCLING XIII). ROCLING (2000) 97-113.
8. Lin, W.H. and Chen, H.H.: Backward Machine Transliteration by Learning Phonetic Similarity. In: Proceedings of 6th Conference on Natural Language Learning. Association for Computational Linguistics (2002) 139-145.
9. Oard, D.W.: Issues in Cross-Language Retrieval from Document Image Collection. In: 1999 Symposium on Document Image Understanding Technology. (1999).
10. Robertson, S.E., Walker, S. and Beaulieu, M.: Okapi at TREC-7: Automatic ad hoc, Filtering, VLC and Interactive. In: Voorhees, E.M. and Harman, D.K. (Eds.): Proceedings of the Seventh Text REtrieval Conference (TREC-7). National Institute of Standards and Technology (1998) 253-264.
11. Sanderson, M. and Clough, P.: EuroVision - An Image-Based CLIR System. In: Proceedings of Cross-Language Information Retrieval: A Research Roadmap (Workshop at SIGIR 2002). (2002).
12. The Lowlands Team: Lazy Users and Automatic Video Retrieval Tools in (the) Lowlands. In: Voorhees, E.M. and Harman, D.K. (Eds.): Proceedings of The Tenth Text REtrieval Conference (TREC 2001). National Institute of Standards and Technology (2002) 159-168.
13. Voorhees, E.M. and Harman, D.K. (Eds.): Proceedings of the Sixth Text REtrieval Conference (TREC-6). National Institute of Standards and Technology (1997).
14. Westerveld, T.: Image Retrieval: Content versus Context. In: Proceedings of RIAO 2000, Vol. 1. (2000) 276-284.
15. Westerveld, T.: Probabilistic Multimedia Retrieval. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002). ACM Press (2002) 437-438.