Cross-Language Image Retrieval via Spoken Query

Wen-Cheng Lin, Ming-Shun Lin and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN
{denislin, mslin}@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

This paper studies cross-language cross-medium information retrieval. We introduce several approaches to unify the languages and media of queries and documents. We experiment on cross-language image retrieval via spoken query. Two approaches are proposed to recognize and translate spoken queries. We also propose a similarity-based approach to identify and backward transliterate named entities in a spoken query.

1. Introduction

Cross language information retrieval (CLIR) (Oard & Diekema, 1998) facilitates using one language (source language) to access documents in another language (target language). The major argument of this approach is: users that are not familiar with the target language still cannot afford to understand the retrieved documents. Images, which are neutral to different language users, are considered as alternative visualization media by CLIR community (Sanderson & Clough, 2002). People with no strong language skills can easily understand and judge the relevance of retrieved images.

There are two types of approaches, i.e., content-based and text-based approaches, to retrieve multimedia data. 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 images or videos look like, but cannot tell us what exactly they are. On the other hand, text can describe the content of multimedia objects. Several hybrid approaches (The Lowlands Team, 2001; Westerveld, 2000, 2002) 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 focused on monolingual retrieval. Little work has been done on cross-language tasks. To encourage the research in cross-language retrieval of images via captions, a new track, i.e. ImageCLEF, was organized in Cross Language Evaluation Forum (CLEF) 2003 (Clough & Sanderson, 2003).

Compared to conventional text input, speech is a more natural way to express users' information needs. In the meantime, spoken access to image databases also introduces some challenging research issues, including how to identify named entities like person names, location names, *etc*. from spoken utterances; how to translate/transliterate information needs from source query to target query; how to retrieve images satisfying users' needs. These issues will affect the performance of spoken cross-language retrieval of images. Chen (2003) proposed a framework of using Chinese speech to access images via English captions. In his study, named entity transliteration/translation problem was dealt with.

This paper studies how to retrieve images based on annotated captions via spoken query in different language. Section 2 discusses cross-lingual cross-medium research issues. Section 3 deals with spoken query recognition and translation. Unknown word detection and translation is discussed in Section 4. Section 5 shows the experimental results of Chinese-English image retrieval. Finally, Section 6 concludes the remarks.

2. Cross-Lingual and Cross-Medium Issues

How to unify the languages in documents and queries is an important issue in cross-language information retrieval. It can be done by translating queries into the language that documents are written in, translating documents into the language that queries are written in, or transforming both queries and documents into an intermediate representation. Query translation is commonly adopted in cross-language text retrieval. This is because translating a large number of documents costs too much time and computing resources. When translating queries, translation ambiguity, target polysemy problem and unknown word handling have to be faced. Dictionary-based, corpus-based and integrated approaches have been proposed to deal with query translation problem (Oard & Diekema, 1998; Chen, Bian, & Lin, 1999).

When queries and documents are in different types of media, translation problem becomes more complicated. In addition to language translation, media transformation is needed. Media transformation combining with language translation problem is shown in Figure 1. Horizontal direction indicates language translation, and vertical direction means media transformation. There are several alternatives to unify the media forms and languages of queries and documents. Similar to cross-language text retrieval, we can translate/transform queries, documents or both of them. Take spoken cross-language access to image collection via captions as an example. The data that users request is images while query is in terms of speech. Images are represented by captions in a language different from that of query. In this way, the medium of target document is transformed into text. In principle, spoken queries in source language and textual documents in target language can be transformed in either of the following ways.

(1) We can transform spoken query in source language into text using a speech recognition system, then translate the textual query into target language and retrieve documents in target language.

(2) In contrast to query translation, we can translate textual documents in target language into source language by a machine translation system. In this model, text retrieval is taken in source language side.

(3) Theoretically, we can transform both source language spoken query and target language textual documents into target language spoken data. Target language textual documents can be transformed into spoken documents by a text-to-speech synthesizer. Comparing to text-to-speech transformation, translating source language spoken queries into target language spoken queries directly is harder. It needs a spoken-to-spoken machine translation system.

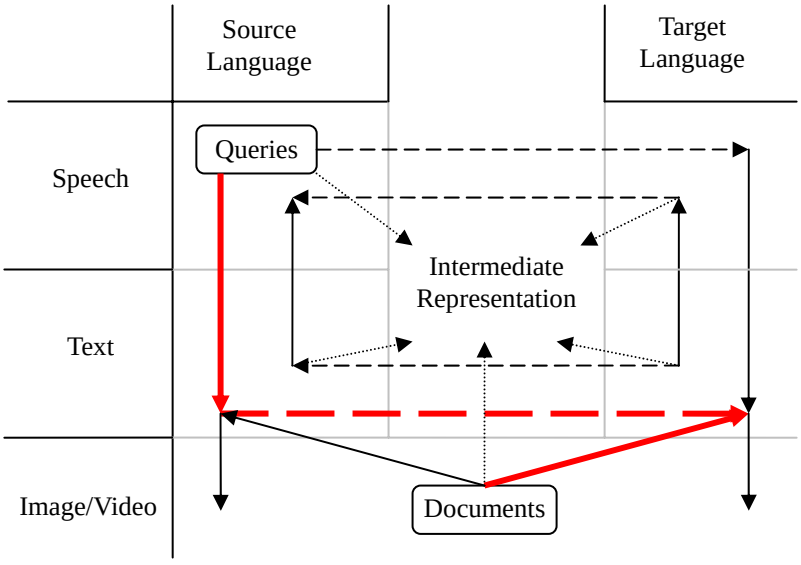

Figure 1: Media Transformation and Language Translation

Furthermore, retrieving spoken data on speech level is not an easy task.

(4) Another alternative is transforming both source language spoken query and target language textual documents into an intermediate form, e.g., International Phonetic Alphabet (IPA).

Transforming spoken query into text is a straightforward and commonly adopted approach. A cross-lingual text retrieval system can be used to retrieve documents after textual query is recognized. In text retrieval stage, we can translate query, document, or both of them. Since image captions are usually short, translating all captions into source language would not cost too much. Therefore, document translation is feasible in cross-language image retrieval.

In translation/transformation process, ambiguity problem occurs in several stages and has to be deal with. In speech recognition, a speech pattern may have several syllable candidates, and a syllable sequence may have many word candidates. How to select the best word is an important issue. In text translation, translation ambiguity and target polysemy problem occur. All the ambiguity problems have to be resolved in cross-lingual cross-medium applications.

In addition to ambiguity problem, unknown word, named entities in particular, is another problem. Thompson and Dozier (1997) reported an experiment over periods of several days in 1995. It showed 67.8%, 83.4%, and 38.8% of queries to Wall Street Journal, Los Angeles Times, and Washington Post, respectively, involve name searching. However, named entities are often not listed in lexicons. If dictionary-based approach is adopted, it is hard to recognize or translate the named entities that are not included in dictionary, and therefore affect retrieval performance.

In this paper, we experiment on Chinese-English image retrieval. Queries are Chinese speech signal, while image captions are text in English. Model (1) mentioned above, which is shown as the red bolder lines in Figure 1, is adopted to unify the forms of queries and documents. Source language spoken query is transformed into source language text, and then is translated into target language. We propose several approaches, which will be discussed in the next two sections, to deal with ambiguity and unknown word problems.

3. Spoken Query Recognition and Translation

The speech recognition tool kit followed (Wang & Chen, 2000) is used to process acoustic signal. The acoustic models are 39 dimensional feature vectors, and syllable bi-gram language model is employed for speech recognition. The 39 dimensions are 12 melfrequency cepstral coefficients (MFCCs) and the logarithmic energy, and their first and second time derivatives. There are total 112 context-dependent initials HMM with 3 states and 38 context-independent finals HMM with 4 states for syllable recognition. Each HMM has 32 Gaussian mixtures. Viterbi search is adopted to find the best 9 syllable candidates on each word.

After syllable candidates are proposed, possible words are extracted from a lexicon according to the syllable sequences. Since the recognized query terms are used to retrieve image captions, if a query term does not appear in captions, it is useless when retrieving images. In this way, we restrict the terms in the lexicon to the terms used in the captions. Because queries and image captions are in different languages, caption terms are translated into source language to construct source language lexicon. We translate the words in English captions into Chinese by looking up an English-Chinese bilingual dictionary. Translation ambiguity appears here, i.e., an English word may have several Chinese translations. Two alternatives may be considered, i.e., all translation equivalents are kept, or the senses are disambiguated and the best one is kept. Since our goal is to construct a lexicon for speech recognition, we keep all translation equivalents in our experiments. The senses of an English word in different captions may be different, and all these translations will be included in the lexicon. That is, disambiguation is not performed when constructing lexicon.

Due to the segmentation problem, there may have (1+2+…+k) possible syllable sequences for k syllables. A syllable sequence may generate several candidate words. How to select the correct

words from candidates is a problem. Selection may be taken in source language side or target language side. Conventionally, term selection is taken is source language side, and then the selected terms are translated into target language. When translating source language terms, translation ambiguity occurs. Thus, another selection is needed. In order to simplify process, term selection may be postponed to translation stage. We proposed two approaches shown as follows to select words in source language and target language sides, respectively.

(1) Selection in source language side

In this approach, appropriate Chinese words are selected, and then translated into English. When selecting Chinese words, we try to find useful words for retrieval. Thus, more than one term may be selected for a speech pattern. Mutual Information (MI) (Church, *et al.*, 1989) is adopted to do the selection. Term selection is determined by the following steps.

Let C be the set of all the candidate words. Let C_b be the set of bi-character words in C, and C_l the set of words in C which are composed of more than two Chinese characters. Initially, the output set O is set to empty. Define pos(*w*,*k*) as the position of the *k*-th character of word *w*.

(i) For each *w* in C_l, if *w* has positive MI score with another candidate word, it is selected. The word w_1 which has the highest MI score with *w* is also selected. All the positions which *w* and w_1 cover are marked and will not be considered later. But if their MI score is smaller than 4, we unmark the position pos(w_1,1). If w_1 is in C_b, find the word w_2 in C_b which has the highest MI score with w_1. Select w_2 and mark pos(w_2,2).

Take Topic 9 in CLEF image track, i.e., “攝影師亞當森拍攝的漁夫” (Fishermen by the photographer Adamson), as an example. Some Chinese candidates of Topic 9 are showed in Table 1. Since “攝影師 (she ying shi)” has positive MI with “拍攝 (pai she)”, both words are selected. Positions 1, 2, 3 and 8 are marked.

(ii) Next, we try to identify the single character word “的 (de)”. We mark the rightmost position which has syllable “ㄉㄜ” in the top three candidates. In Table 1, position 9 is identified as “的 (de)” and is marked.

(iii) We select the word *w* in C_l that has no positive MI score with the other candidate words and all the positions that *w* covers are not marked. In our example, because position 8 has been marked in step (i), “試著去 (shi zhe qu)” is not selected in this step.

Chinese query	攝(she)	影(ying)	師(shi)	亞(ya)	當(dang)	森(sen)	拍(pai)	攝(she)	的(de)	漁(yu)	夫(fu)
Position	1	2	3	4	5	6	7	8	9	10	11
Chinese candidate words	色(se)	應(ying)	思(si)	要(yao)	當(dang)	森(sen)	拍(pai)	設(she)	**的(de)**	魚(yu)	赴(fu)
	瑟(se)	英(ying)	使(shi)	亞(ya)	放(fang)	生(sheng)	牌(pai)	生(sheng)	德(de)	女(nv)	鋪(pu)
	:	:	:	:	:	:	:	:	:	:	:
	攝影(she ying)		資料(zi liao)		**方針(fang zhen)**		**拍攝(pai she)**		爭取(zheng qu)		
	索引(suo yin)		自然(zi ran)		**豐盛(feng sheng)**		快速(kuai su)		等於(deng yu)		
	社評(she ping)		治療(zhi liao)		**放生(fang sheng)**		探視(tan shi)		證據(zheng ju)		
	:		:		:		:		:		
		平時(ping shi)		**亞當(ya dang)**		**生態(sheng tai)**		使得(shi de)		**旅途(lv tu)**	
		飲食(yin shi)		**飄蕩(piao dang)**		神態(shen tai)		升等(sheng deng)		**局部(ju bu)**	
		臨時(lin shi)		**藥房(yao fang)**		審判(shen pan)		色澤(se ze)		**音符(yin fu)**	
		:		:		:		:		:	
	攝影師(she ying shi)							試著去(shi zhe qu)			

Table 1: The Chinese candidates of Topic 9.

(iv) Let P_C={(c_1,c_2)| c_1,c_2 in C_b, both two positions of c_1 are not marked, and c_1 and c_2 do not overlap in position}. Check the pair (c_1,c_2) in P_C with the highest MI score. If their MI score is larger than a threshold (4 in our experiment), c_1 is selected and pos(c_1,2) is marked. In such a case, if both two positions of c_2 are not marked, c_2 is also selected and pos(c_2,2) is marked. All the pairs (c_3,c_4) in P_C which any position of c_3 is marked in this step are removed from P_C. This step repeats until P_C is empty.

In Table 1, the MI score between word pair (生態(sheng tai), 自然(zi ran)) exceeds the threshold, thus "生態 (sheng tai)" is selected and position 7 is marked. Because position 3 is marked in step (i), "自然 (zi ran)" is not selected. In second iteration, "方針 (fang zhen)" is selected.

(v) Finally, if a position is not marked yet, we select the first three most frequent bi-character words starting in this position. In Table 1, "亞當(ya dang)", "飄蕩(piao dang)", "藥房(yao fang)", "豐盛(feng sheng)", "放生(fang sheng)", "旅途(lv tu)", "局部(ju bu)" and "音符 (yin fu)" are selected in this step.

After selection, the output set O contains the words selected from C_b and C_l. Finally, the recognized Chinese words are translated into English to retrieve English captions.

(2) Selection in target language side

The recognition and translation method described above needs two selections, i.e., one for speech recognition, and the other one for query translation. An alternative way is to select once only in target language side. All Chinese candidate words are retained during speech recognition. For each Chinese candidate word, we find its English translations by looking up the lexicon constructed from captions. Then the appropriate translations are selected as the English query terms by using co-occurrence information. The selection method is described as follows.

(i) For each c_1 in C_l, we select the English words in the top three word pairs (e_1,e_2) with the highest MI scores which e_1 is a translation of c_1, e_2 is a translation of another word c_2 in C_b, and c_1 and c_2 do not overlap in position. The positions that c_1 covers and pos(c_2,2) are marked.

Table 2 lists some English candidate words of Topic 9. The top three words that have highest positive MI scores with the translation of "攝影師 (she ying shi)" are "Photographic", "actually" and "local". Thus "photographer", "Photographic", "actually" and "local" are selected, and positions 1, 2, 3, 8, 9, and 11 are marked.

Position	1	2	3	4	5	6	7	8	9	10	11
English candidate words	photograph		Information		**plentiful**		**Photographic**		woo		
	index		Natural		**reigned**		**rising**		equal		
	catalogue		Cure		**abundance**		**Photograph**		evidence		
	:		:		:		:		:		
		accordingly		**crank**		**proper**		**actually**		**local**	
		lever		**drift**		**posed**		promote		fisherman	
		grooming		**Adam**		**pose**		coloration		catch	
		Shadow		:		:		:		:	
	photographer							**try to**			

Table 2: The English candidates of Topic 9.

(ii) If all translations of a word c in C_l have no positive MI score with the translations of all words in C_b, the first three translations with the highest frequency in the English image captions are selected. The positions that c covers are marked. In Table 2, "try to" is selected in this step, and positions 8, 9, and 10 are marked.

(iii) Let P_E={(e_1,e_2)| e_1 and e_2 are the translations of c_1 and c_2 in C_b, respectively, pos(c_1,1) is not marked, and the MI score between e_1 and e_2 is larger than 10}. Define count(i) as the number of selected words for a position i that is not marked. Count(i) is initially set to be 0. Check the pair (e_1,e_2) in P_E with the highest MI score. We select e_1 and increase count(pos(c_1,1)) by 1. If pos(c_2,1) is not marked, e_2 is also selected and count(pos(c_2,1)) is increased. The pair (e_1,e_2) is removed from P_E. All the pairs (e_3,e_4) in P_E which count(pos(c_3,1)) equal to 3 are removed from P_E. This step repeats until P_E is empty.

In Table 2, the MI scores of word pairs (accordingly, plentiful), (lever, crank), (grooming, reigned), (abundance, catch), (therefore, proper) and (Shadow, drift) are exceed the threshold, thus "plentiful", "crank", "reigned", "abundance", "proper" and "drift" are selected.

(iv) For a position i which count(i) is less than 3, we select the most frequent word e which is the translation of c staring at position i, and increase count(i). This step repeats until all the count(i)'s are equal to 3. Since count(4), count(6) and count(7) are less than 3, "Adam", "posed", "pose", "rising" and "Photograph" are selected in this step.

4. Named Entity Backward Transliteration

In the stage of speech recognition, candidate words are supported by a lexicon constructed from image captions. When constructing lexicon, some English words are not included in bilingual dictionary, thus no translations can be included in the special lexicon. In this way, these unknown words will not be included in the final English queries when the approaches in Section 3 are used to recognize and translate queries. Many unknown words are named entities which are important terms for retrieval. If a named entity in a query is not translated in CLIR, the retrieval performance might be decreased. In CLEF 2003 image track, Lin, Yang and Chen (2003) used similarity-based backward transliteration to transliterate named entities that are not included in dictionary. The similarities of a Chinese named entity and English candidate words are computed, and the candidate words with the higher similarity are chosen as the translations of the Chinese word. The similarities are measured at phoneme level. Experimental results showed that the retrieval performances were improved after the unknown named entities were transliterated.

The backward transliteration problem becomes more complicated for spoken query. In textual query, named entities are identified first. Then transliteration and translation parts of the named entities are recognized. Finally, the similarities between the transliteration part and candidates words are measured to find its original word. In spoken query, the unknown words may be recognized as some words in the lexicon, thus they could not be considered as named entities correctly. The pronunciations of a transliterated word and its original word are in IPA, thus we may recognize transliterated name at phoneme level rather than text level. In this way, the problem becomes how to find a segment of speech that is a transliterated name from a spoken query. The concept of similarity-based backward transliteration is adopted. The similarities of a spoken query and English candidate words are computed. In the meantime, the boundaries of the segment of a query that is similar to a candidate word are also identified. The candidate word with the highest similarity is chosen as the transliteration of the segment of the query.

The similarities are measured at phoneme level. First, spoken query and candidate words are converted to IPA representations. Then we find a segment of IPA string of a query that is most similar to the IPA string of a candidate word. For a spoken query, the speech signals are recognized as syllables, and then the syllables are converted into IPA. As described in Section 2, a speech pattern may have several syllable candidates. If more than one syllable candidate of a speech pattern

are converted to IPA, the spoken query is converted to an IPA matrix instead of an IPA string. Take Topic 9 as an example. The top two syllable candidates of Topic 9 are showed in Table 3. Table 4 shows the IPA representation of each syllable in Table 3. A Chinese syllable is converted to one or two IPA symbols. We divide the IPA symbols of a Chinese syllable into two parts. If a Chinese syllable maps to only one IPA symbol, the IPA symbol is duplicated. In Table 4, the symbols in italic type are duplicated one. Given an English word, we will find an IPA string in Table 4 that is most similar to the IPA string of this English word.

The similarity between two IPA strings depends on their alignment. String 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. Figure 2 is an alignment of IPA of "亞當森 (ya dang sen)" and "Adamson". A similarity score which is determined by Formula 1 is given to an alignment. The similarity score of an alignment is the summation of the similarity scores of the aligned IPA symbol pairs. The similarity score between two IPA symbols is learned from 1,574 pairs of English names and their transliterated Chinese words (Lin & Chen, 2002).

$$Sim(S_1, S_2) = \sum s(a,b) \qquad (1)$$

where S_1 and S_2 are IPA strings,
a and *b* are aligned IPA symbols, and
s(a,b) is the similarity score between two IPA symbols.

Chinese Word	攝(she)	影(ying)	師(shi)	亞(ya)	當(dang)	森(sen)	拍(pai)	攝(she)	的(de)	漁(yu)	夫(fu)
Syllable candidates	ㄕㄜ (she)	ㄧㄥ (ying)	ㄙ (si)	ㄧㄠ (yao)	ㄉㄤ (dang)	ㄕㄣ (shen)	ㄆㄞ (pai)	ㄕㄜ (she)	ㄉㄜ (de)	ㄩ (yu)	ㄈㄨ (fu)
	ㄙㄜ (se)	ㄧㄣ (yin)	ㄕ (shi)	ㄧㄚ (ya)	ㄍㄤ (gang)	ㄕㄥ (sheng)	ㄊㄞ (tai)	ㄙㄜ (se)	ㄓㄜ (zhe)	ㄌㄩ (lv)	ㄆㄨ (pu)

Table 3: The top two syllable candidates of Topic 9.

Chinese Word	攝(she)		影(ying)		師(shi)		亞(ya)		當(dang)		森(sen)		拍(pai)		攝(she)		的(de)		漁(yu)		夫(fu)	
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
IPA of syllable	sr	e	*ing*	ing	s	i	*iao*	iao	t[	ang	sr	en	ph	ai	sr	e	t[	e	*v*	v	f	u
	s	e	*in*	in	sr	i	*ia*	ia	k	ang	sr	eng	t[h	ai	s	e	tsr	e	l	v	ph	u

Table 4: IPA representations of syllables in Table 3.

亞當森	ia t[ang s en 10 10 2.72 10 10	Sim(S_1,S_2)=42.72
Adamson	@ d & m s & n	

Figure 2: Alignment of IPAs of "亞當森 (ya dang sen)" and "Adamson"

When computing similarity between a query and a candidate word, how to find the segment that is most similar to the candidate word is a research issue. We can measure the similarities between all possible segments and the candidate word, and then select the segment with the highest similar score. Since there are many possible segments and combinations of IPA symbols, computing the similarity score of all segments costs too much time. We use a greedy method to find the segment that is most similar to a candidate word. First, we find the possible start position that is similar to the first IPA symbol of candidate word. From the start position, we try to find an IPA string aligning to the IPA string of candidate word. The details of the approach are described as follows.

(1) First, the segments that may be general words are deleted. In the selection method described in Section 3, the Chinese long word c_i that has positive MI score with other candidate word is selected at first. The word c_j which has the highest MI score with c_i is also selected. Then the Chinese word "的 (de)" is identified. Because the correct rate of selecting long word and "的 (de)" is high, we can treat these words are correctly recognized. The speech segments of these words are excluded in the following steps.

(2) Define a table *s* in which the value $s_{i,j}$ of a cell in row *i* and column *j* is the maximum similarity score between the *i*-th IPA symbol of English word and the IPA symbols in position *j* of Chinese spoken query. For example, the maximum similarity score between the first IPA symbol of "Adamson", i.e., "@", and the IPA symbols in position 7 in Table 4 is about 10 (*s(@, ia)*= 9.999845868). Thus $s_{1,7}$ is set to be 10 in Table 5.

(3) From table *s*, we try to find the possible paths from start points. Start points are the cells that have positive value in rows 1 and 2. A reasonable path is defined as: the distance between two adjacent aligned IPA pairs is not greater than one in both Chinese and English words. We compute the similarity scores of possible paths from $s_{1,1}$ in *s* column by column. The maximum score of the paths ending at cell (i,j) is determined by Formula 2. Table 6 is the scores of paths computed from Table 5.

$$m_{i,j} = \max_{1 \le p,q \le 2} s_{i-p,\, j-q}$$

$$s'_{i,j} = \begin{cases} 0 & \text{if } s_{i,j} = 0, \text{ or } m_{i,j} = 0 \text{ and } i > 2 \\ s_{i,j} + m_{i,j} & \text{otherwise} \end{cases} \qquad (2)$$

where $s'_{i,j}$ is the maximum score of the paths ending at cell (i,j).

E-IPA		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
n	7							0	0	0	9.85	0	10										
&	6							8.39	8.39	0	0	0.46	3.28										
s	5							0	0	0	0	10	0										
m	4							0	0	0	2.72	0	2.04										
&	3							8.39	8.39	0	0	0.46	3.28										
d	2							0	0	10	0	0	0										
@	1							10	10	0	4.11	0	1.37										

Table 5: The maximum similarity score between an IPA symbol of English word "Adamson" and an IPA symbol in each position (a column in Table 4)

E-IPA		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
n	7							0	0	0	0	0	42.72				
&	6							0	0	0	0	23.18	36				
s	5							0	0	0	0	32.72	0				
m	4							0	0	0	22.72	0	22.51				
&	3							0	18.39	0	0	20.46	7.39				
d	2							0	0	20	0	0	0				
@	1							10	10	0	4.11	0	1.37				

Table 6: The maximum similarity score between a segment of IPA string of English word "Adamson" and a segment of IPA string of Chinese Topic 9

Run	# correctly recognized term	# proposed term
Speaker_01	92	581
Speaker_02	84	559
Speaker_03	82	551
Speaker_04	88	540

Table 7: The performance of speech recognition

(4) Finally we find the best path that ends at an end point, and back trace to its start point. An end point is a cell with positive score in the last two rows. In Table 6, cell (7, 12) has the largest score in the last two rows. Thus, it is the end point of the best path. After back tracing to the start point, we find a segment of IPA string, from position 7 to 12, of Chinese Topic 9 that is most similar to "Adamson".

5. Experiments

CLEF 2003 image test collection and topics are adopted to evaluate our system. The image collection contains 28,133 images and collateral English captions. Okapi IR system (Robertson, *et al.*, 1998) is adopted to index and retrieve the image captions. The weighting function is BM25. For each image, the caption text, <HEADLINE> and <CATEGORIES> sections are used for indexing. The words in these sections are stemmed, and stopwords are removed. The test bed contains 50 English topics composed of title and narrative fields. The titles of each topic have been translated into Spanish, Italian, German, French, Dutch and Chinese. In our experiments, Chinese queries are used as source language queries. The 50 Chinese queries are read by four users individually as the spoken queries for testing. Thus we have four evaluation runs for each approach.

In the experiments, English textual captions are indexed and retrieved. Spoken queries are transformed to Chinese words and then translated into English to retrieve English captions. The selection method described in Section 3 was used to select source language, i.e., Chinese, terms. The MI scores trained from Academia Sinica Balance Corpus (ASBC) (Huang & Chen, 1995) are used to select Chinese words. There are total 195 Chinese words in 50 queries. The numbers of correctly recognized terms and total proposed terms of each run are shown in Table 7. The average recall is about 44.35%. Because at most four Chinese terms are proposed for each speech pattern, the number of total proposed term is larger than that of the original queries.

After Chinese terms are recognized, the proposed Chinese words are translated into English. For each Chinese query term, we found its translations by looking up the lexicon constructed from image captions. Then we use the following approaches to select appropriate English terms.

(1) CO model (Chen, Bian, & Lin, 1999)
The CO model employs word co-occurrence information extracted from a target language text collection to disambiguate query term translations. We adopt MI to measure the co-occurrence strength between words. The MI scores of English words are trained from the English image captions. For a query term, we compare the MI scores of all the translation pairs (x, y), where x is the translation of this term, and y is the translation of another query term. The word pair (x_i, y_j) with the highest MI value is extracted, and the translation x_i is regarded as the best translation of this query term.

(2) First two highest frequency (F2HF)
The first two translations with the highest frequency of occurrence in the English image captions are considered as the target language query terms. If a Chinese word has only one English translation, the English translation is duplicated.

The selected English terms were submitted to Okapi IR system to retrieve English captions. The retrieval performances are shown in Table 8. In order to compare the performances of spoken query with textual query, we conduct two runs that using original Chinese textual queries. CO model and F2HF model are used to translate Chinese queries into English. The retrieval performances are shown in Table 9. The performance of spoken query is about 45% and 40% of textual query in CO model and F2HF model, respectively. From Table 8 and 9, F2HF model performs better than CO model. In F2FH model, the top two most frequent translations are selected. In some cases, these two words are synonyms. It is like query expansion and helps to find more relevant images.

Translation model	Run	Average precision	Average of 4 runs
CO model	CO_01	0.0802	0.0857
	CO_02	0.0797	
	CO_03	0.0803	
	CO_04	0.1027	
First two highest frequency	F2HF_01	0.0782	0.0949
	F2HF_02	0.1026	
	F2HF_03	0.0898	
	F2HF_04	0.1089	

Table 8: The performances of image retrieval

Translation model	Run	Average precision
CO model	Text_CO	0.1894
First two highest frequency	Text_F2HF	0.2402

Table 9: IR performances when using textual queries

Translation model	Run	Average precision	Average of 4 runs
One-selection	1selection_01	0.0332	0.0434
	1selection_02	0.0612	
	1selection_03	0.0275	
	1selection_04	0.0517	

Table 10: IR performances of one-selection model

Translation model	Run	Average precision	Average of 4 runs
CO model + name	CO_name_01	0.0762	0.0696
	CO_name_02	0.0681	
	CO_name_03	0.0565	
	CO_name_04	0.0774	
First two highest frequency + name	F2HF_name_01	0.0752	0.0890
	F2HF_name_02	0.0887	
	F2HF_name_03	0.0873	
	F2HF_name_04	0.1048	
One-selection + name	1selection_name_01	0.0299	0.0388
	1selection_name_02	0.0509	
	1selection_name_03	0.0260	
	1selection_name_04	0.0485	

Table 11: IR performances after names are transliterated

We conduct another experiment that uses the selection method in English side. Here, we do not select Chinese terms in speech recognition stage. All Chinese word candidates are retained. English selection method is used to select English terms from the translations of all Chinese terms. We call this one-selection approach. Only one selection stage is taken. Table 10 is the IR performance of one-selection model. The performance of one-selection approach is worse than two-selection approaches, both CO model and F2HF model. Although one-selection model reduces the number of times of selection process, the number of English candidate words is larger than that of Chinese candidate words. Selecting appropriate terms from so many English candidate words is difficult.

Among the 50 queries, a total of 14 queries contain proper nouns which are not in our dictionary. Without the correct translations of these names, the retrieval performances of the 14 queries are poor. The backward transliteration method described in Section 4 is adopted to identify and translate names in spoken queries. English name candidates are extracted from English captions. We collect a list of English names that contained 50,979 person names and 19,340 location names. If a term in the captions can be found in the name list, it is extracted. Total 3,599 names are extracted from the image captions. For each query, we compute the similarity scores between it and the 3,599 English names. The top 10 English names with the highest similarity score are selected as query terms and added to the translated English query. The similarity scores of the selected English names must exceed a threshold. The threshold is set to be 30.

The retrieval performances after names are transliterated are shown in Table 11. Unfortunately, average precisions are decreased. Since we do not know which query contains named entities, all queries are processed by the name transliteration module. The queries that have no named entities are also expanded with proposed names, which become noises and decrease IR performance. When measuring similarity, the segments that may be general words are deleted at first. In some queries, some parts of a name are also deleted. Therefore, we cannot compute similarity score using the right segment. Which syllable should be retained for similarity measurement is an important issue. If the correct syllables of a Chinese name are not retained, the IPA symbols of other syllables might not be aligned to the original English name. If many syllable candidates are kept, other name candidates could gain high similarity score by aligning to the wrong syllables. The coverage of English name candidates is another problem. There are 14 named entities that are not included in the lexicon. Among the 14 names, four terms are not in the name list, thus they cannot be translated correctly. These four names are “Tay”, “Yarmouth”, “Culross”, and “Henrietta” in Topic 15, 17, 19, and 21, respectively.

6. Concluding Remarks

In this paper, we study cross-language cross-medium information retrieval problem. Queries and documents are in different languages and media. Queries and documents are needed to be transformed into the same representation for retrieval. We introduce several approaches to unify the languages and media of queries and documents. In this paper, we experiment on cross-language image retrieval via spoken query. The textual captions are used to represent images. Spoken queries are transformed into text and translated into target language that image captions are written in. Thus images and queries are in the same representation, i.e., text in target language. A monolingual text retrieval system is adopted to retrieve relevant image captions. We proposed two approaches to recognize and translate spoken queries. The first approach uses co-occurrence information to select appropriate Chinese words in speech recognition stage, and then translates the selected Chinese words into English. The second approach postpones term selection to translation stage. All Chinese word candidates are retained. Term selection is taken in English side. The results show that selecting appropriate Chinese terms at first is better than postpone approach. The performances are compared to textual queries. Experimental results showed that the performance of spoken query is about 40% of textual query.

We proposed a similarity-based approach to identify and backward transliterate named entity in a spoken query. The similarity is measured on phoneme level. However, this approach does not meet our expectation. Many factors affect the performance. Which syllable should be retained for similarity measurement, how to find the correct segment of a transliterated name, and the coverage of candidates are important issues. We will study further in the future.

Acknowledgements

We would like to thank Professor Berlin Chen (berlin@csie.ntnu.edu.tw) for providing their speech recognition module and language model for us.

References

Chen, H.H. (2003). Spoken Cross-Language Access to Image Collection via Captions. In *Proceedings of Eurospeech 2003.*

Chen, H.H., Bian, G.W., & Lin, W.C. (1999). Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics* (pp. 215--222).

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). Parsing, Word Associations and Typical Predicate-Argument Relations. In *Proceedings of International Workshop on Parsing Technologies* (pp. 389--398).

Clough, P. & Sanderson, M. (2003). The CLEF 2003 Cross Language Image Retrieval Task. In *Working Notes of CLEF 2003*.

Huang, C.R. & Chen, K.J. (1995). Academic Sinica Balanced Corpus. Technical Report 95-02/98-04. Academic Sinica, Taipei, Taiwan.

Lin, W.H. & Chen, H.H. (2002). Backward Machine Transliteration by Learning Phonetic Similarity. In *Proceedings of 6th Conference on Natural Language Learning* (pp. 139--145).

Lin, W.C., Yang, C., & Chen, H.H. (2003). Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval. In *Working notes of CLEF 2003*.

Oard, D. & Diekema, A. (1998). Cross-Language Information Retrieval. *Annual Review of Information Science and Technology*, 33, 223--256.

Robertson, S.E., Walker, S., & Beaulieu, M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* (pp. 253--264).

Sanderson, M. & Clough, P. (2002). Eurovision-An Image-Based CLIR System. In *Proceedings of Workshop at SIGIR2002, Cross-Language Information Retrieval: A Research Roadmap.*

The Lowlands Team (2001). Lazy Users and Automatic Video Retrieval Tools in (the) Lowlands. In *Proceedings of The Tenth Text REtrieval Conference (TREC 2001)* (pp. 159--168).

Thompson, P. & Dozier, C. (1997). Name Searching and Information Retrieval. In *Proceedings of Second Conference on Empirical Methods in Natural Language Processing* (pp. 134--140).

Wang, H.M. & Chen, B. (2000). Content-based Language Models for Spoken Document Retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL 2000)* (pp. 149--155).

Westerveld, T. (2000). Image Retrieval: Content versus Context. In *Proceedings of RIAO 2000*, 1, (pp. 276--284).

Westerveld, T. (2002). Probabilistic Multimedia Retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)* (pp. 437--438).