

Spoken Cross-Language Access to Image Collection via Captions

Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University, Taipei, Taiwan

hh_chen@csie.ntu.edu.tw

Abstract

This paper presents a framework of using Chinese speech to access images via English captions. The formulation and the structure mapping rules of Chinese and English named entities are extracted from an NICT foreign location name corpus. For a named location, name part and keyword part are usually transliterated and translated, respectively. Keyword spotting identifies the keyword from speech queries and narrows down the search space of image collections. A scoring function is proposed to compute the similarity between speech query and annotated captions in terms of International Phonetic Alphabets. The experimental results show that the average rank and the mean reciprocal rank are 2.04 and 0.8322, respectively, which is very close to the best performance, i.e., 1, for both average rank and mean reciprocal rank.

1. Introduction

Cross language information retrieval (CLIR) [1] 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 [2]. That is also realized in real world. Many image collections are available on the Internet, and image search is a basic capability provided by search portals. In such a situation, cross-language retrieval of images via captions is promoted and considered as one of evaluation tasks in Cross Language Evaluation Forum in Europe [3].

Compared to conventional text input, speech is a more natural way to express users' information need. In the meantime, spoken access to image databases also introduces some challenging research issues, including how to identify named entities like personal names, location names, *etc.* from spoken utterances; how to translate/transliterate information need 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.

Query translation is a well-known methodology to unify the language usages between source query and target document collection in cross language text retrieval. Dictionary-based, corpus-based and integrated approaches have been proposed to deal with query translation problem [1, 4]. Named entities are common targets that users are interested in. Thompson and Dozier [5] 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. The papers [6-8] touched on extraction and translation/transliteration of named entities in CLIR. Two approaches, i.e., similarity scoring on

grapheme [6] and phoneme levels [7, 8], were proposed. Besides that, the papers [9, 10] dealt with named entity extraction from speech. Named entities are often not listed in lexicons, so that a name may be converted to different character strings during speech recognition. In this way, it is not feasible to transform speech to text and do query translation similar to conventional text retrieval.

This paper will study how to retrieve relevant images based on annotated captions in different languages. Section 2 sketches the architecture of this work. Section 3 deals with the Chinese spoken named entity recognition, named locations in particular. The transcribed results are in terms of International Phonetic Alphabet (IPA). Section 4 shows how to retrieve relevant captions and images at the same time using IPAs. Finally, Section 5 concludes the remarks.

2. Architecture

Figure 1 demonstrates architecture of spoken cross language access to an image database with English captions. Users express their information needs with Chinese speech. A Chinese named entity recognizer extracts the named entity from speech, and tells out its type (e.g., named locations, named people, *etc.*) and lexical structure (i.e., the name part and the keyword part in named locations). For location names, the name part is usually transliterated and the keyword part is translated. For example, a Chinese speech query “雷尼爾山” (lei ni er shan) is composed of name (“雷尼爾”, lei ni er, Rainier) and keyword (“山”, shan, mountain). The former is transliterated from “Rainier” and the latter is translated from “Mountain”. The speech query can be converted to a text query “雷尼爾” (lei ni er), “累倪耳” (lei ni er) or something else, but it is not necessary for spoken cross language retrieval. Here, spoken named entity recognition will convert the phonetic string of name part to a canonical form in terms of IPA. In the previous example, “l ei n i er” will be produced. Name transliteration/translation module will rank the annotated captions by IPA similarity matching and term matching, and the most similar image will be retrieved from the image database.

In the initial study, we did not extract features from images. Instead, English captions are regarded as metadata to describe each image. A metadata manager extracts the possible named entities from captions, converts the name parts to canonical forms (i.e., IPAs), and labels on each image. For example, Rainier is converted to “9 ei n I 9”. The transformation of grapheme to phoneme is shown as follows. At first, we look up the CMU pronouncing dictionary [11]. If the name part is in the dictionary, then we take the pronunciation and transform it into IPA. Otherwise, we apply a speech synthesis system, MBRDICO [12], to generate the pronunciation of the name part.

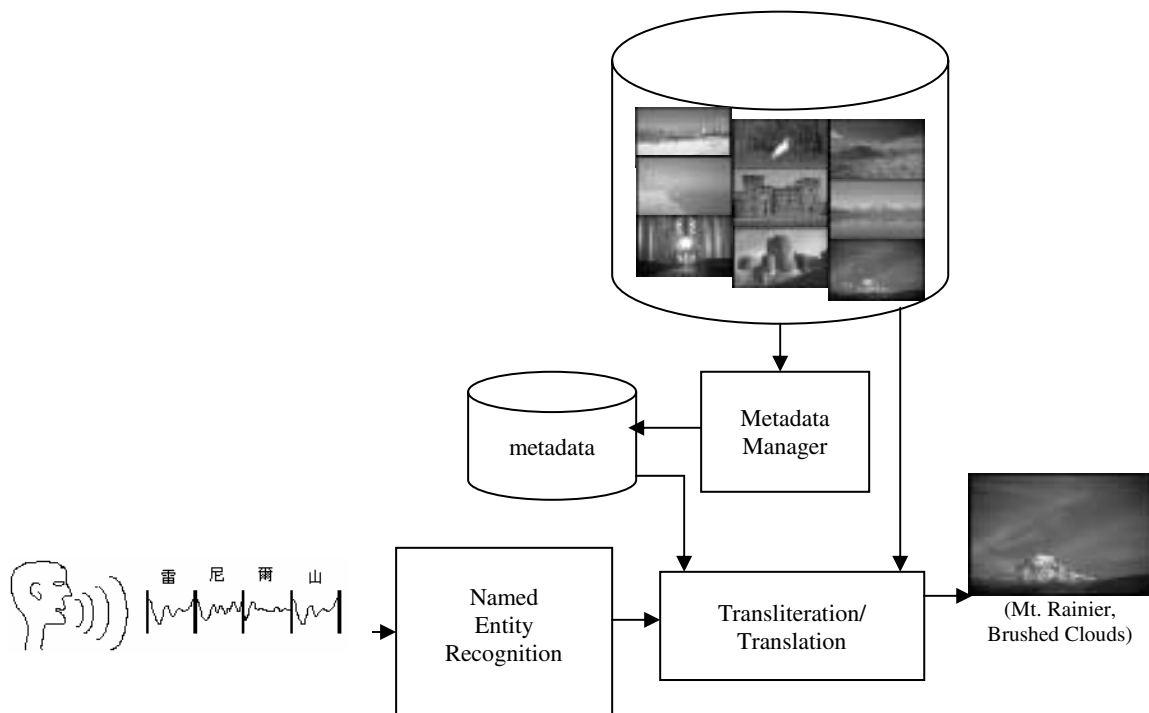


Figure 1: Flow of Spoken Cross Language Access to Image Collection

Take an image collection - Corel (<http://elib.cs.berkeley.edu/photos/corel/>) as an example. The images are divided into several predefined categories such as AirSnow, Animal, ..., Mountain, ..., and so on. Each image is annotated with English captions. The following is the English annotation for image "Mt. Rainier" in Corel.

Caption: Mt. Rainier, Brushed Clouds

Keywords: mountain clouds snow rock

Not all the captions include named entities. Metadata manager maintains an index consisting of named entities and/or general content words. In the following sections, we will focus on spoken named entity recognition and name transliteration/translation.

3. Spoken Named Entity Recognition

In the famous message understanding competition MUC [13], there are eight types of named entities, including named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions. This paper focuses on named locations only. A named location is usually composed of a name part and a keyword part. Both name and keyword may consist of more than one word. During conversion among English and Chinese, some part may be translated and some part may be transliterated. We try to find Chinese and English keywords used to formulate location names, the formulation rules, and the conversion rules among Chinese and English location names from a Chinese-English foreign location name corpus compiled by National Institute for Compilation and Translation of Taiwan [14].

This foreign location corpus abbreviated NICT corpus hereafter consists of 42,500 entries. Each entry has three parts including foreign name, translation/transliteration name and country name. For example, (Elephant Island, "愛麗芬島" (ai li fen dao), UK), (Little Colorado River, "小科羅拉多河"

(xiao ke luo la duo he), USA), (Great Salt Lake, "大鹽湖" (da yan hu), USA), and (Edinburgh, "愛丁堡" (ai ding bao), UK). From country name, we know the language used. Different languages have different formulation rules. In the first example, the name part is transliterated and the keyword part is translated. That is, Elephant and Island correspond to "愛麗芬" (ai li fen) and "島" (dao), respectively. Comparatively, name part is both translated and transliterated in the second example. In this case, Little and Colorado correspond to "小" (xiao) and "科羅拉多" (ke luo la duo), respectively. In the third example, both the name part (i.e., "大鹽" (da yan)) and keyword part ("湖") are translated. In other words, the three characters "大" (da), "鹽" (yan), and "湖" (hu) correspond to Great, Salt, and Lake, respectively, which are meaning translation instead of phoneme transliteration. There is no keyword part in the last example, and the name part is transliterated.

Because Chinese has segmentation problem, we start the keyword extraction from the English part of the NICT location name corpus. The following shows a keyword extraction algorithm.

- Compute the word frequency of each word in the English location list.
- Regard those words that appear more than a threshold as English keywords.
- Because the NICT corpus is English-Chinese aligned, we can cluster the Chinese name list based on English keywords. For each Chinese name cluster, we identify the Chinese keyword sets by frequency.

The following keywords were extracted from the NICT location name corpus: (River, 河 (he), 677), (Island, 島 (dao), 544), (Lake, 湖 (hu), 320), (Mountain, 山 (shan), 230), (Bay, 灣 (wan), 166), (Mountain, 峰 (feng), 108), (Peak, 峰 (feng),

99), (Island, 群島 (qun dao), 93), (Mountains, 山脈 (shan mo), 90), (Corner, 角 (jiao), 90), (City, 城 (cheng), 52), and so on. Each tuple (E, C, F) consists of three components, i.e., English name E, Chinese name C and Frequency F, respectively. For example, “Mountain”, which is an English location keyword, corresponds to two Chinese location keywords, i.e., “山” (shan) and “峰” (feng), whose frequencies are 230 and 108, respectively. On the other hand, the Chinese location keyword “峰” (feng) can be translated into two English location keywords “Mountain” and “Peak”.

After keyword sets are extracted from the NICT corpus, the formulation rules for English and Chinese are learned in the meantime. A complete English (or Chinese) location name is in terms of name-keyword combination. The transliteration/translation rules between Chinese and English location names are also determined from this corpus.

A spoken named entity recognizer is composed of traditional acoustic and lexical models. Because named entities are usually unknown words, i.e., not listed in a lexicon, the lexical model is trained by using different named entity corpora. Take named locations as an example. We removed the keywords and some words, such as “大” (da) and “小” (xiao), which are usually translated, from the NICT corpus. The remaining characters, which belong to the name part of named locations, are transformed into phonemes by table lookup. N-gram model captures the patterns embedded in the phoneme sequence. Keyword spotting will recognize the type of the named entities according to the precompiled keyword sets, and suitable lexical model will be selected. In this paper, only named locations are considered, thus the problem is simplified.

4. Name Translation/Transliteration

After spoken name recognition, a complete location name is converted to two parts, i.e., name part in terms of IPA and keyword part. Because images are partitioned into several predefined categories, the keyword part will narrow down the search scope. Edit distance is regarded as a metric to compute the similarity between IPAs of a spoken named entity C and an annotated caption of an image E . The distance is in terms of various costs of insertions, deletions, and substitutions required to transforming IPA_C into IPA_E . The first step to compute the similarity score of these two phonetic strings, IPA_C and IPA_E , is alignment. We try to insert spaces to IPA_C and IPA_E such that the resulting strings are of equal length. Consider two phonetic strings $\langle h j u g oU \rangle$ and $\langle v k uo \rangle$. There are two possible alignments – say, $\langle h j u g oU \rangle \Leftrightarrow \langle _ _ v k uo \rangle$ and $\langle h j _ u g oU \rangle \Leftrightarrow \langle _ v k _ uo _ \rangle$, where $_$ denotes a space.

Each character in a string corresponds to a character in another string. The similarity score of an alignment is the summation of similarity scores of all corresponding character pairs. The similarity scores of the above two alignments are computed as follows.

$$\begin{aligned} \text{Score1} &= s(h, _) + s(j, _) + s(u, v) + s(g, k) + s(oU, uo) \\ \text{Score2} &= s(h, _) + s(j, v) + s(_, k) + s(u, _) + s(g, uo) + \\ &\quad s(oU, _) \end{aligned}$$

The function s defines the similarity score between two phonetic characters. Table 1 shows a sample matrix of similarity scores. Score1 and Score2 are equal to 16 and -47 by

Table 1: Sample matrix of similarity scores

s(a,b)	h	j	u	v	g	k	oU	uo	_
h	10	0	-8	0	0	-9	0	-4	-10
j	0	10	-1	0	0	-1	0	-1	3
u	0	0	10	3	0	-4	0	-2	-10
v	0	0	-6	10	0	-6	0	-5	-10
g	0	0	-10	0	10	10	0	-7	-10
k	0	0	-10	-1	0	10	0	-10	-10
oU	0	0	2	4	0	-4	10	10	-10
uo	0	0	0	0	0	0	0	10	-10
_	-10	-10	-10	-5	-10	-10	-10	-10	

using this matrix. It indicates the Alignment 1 is more preferable to Alignment 2.

The similarity scores among phonetic characters can be coded manually [7] or learned automatically from a training corpus [8]. The higher the score of two phonemes is, the more similar they are. The similarity score of two IPA strings is the score of the optimal alignment of these two strings. The optimal alignment of two strings can be computed using dynamic programming. A matrix T of $n+1$ rows and $m+1$ columns, where n and m are the lengths of IPA_C and IPA_E , will be defined as follows.

$$T(i,0) = \sum_{1 \leq k \leq i} s(IPA_C(k), _) \quad (1)$$

$$T(0, j) = \sum_{1 \leq k \leq j} s(_, IPA_E(k)) \quad (2)$$

$$T(i, j) = \max \begin{pmatrix} T(i-1, j-1) + s(IPA_C(i), IPA_E(j)), \\ T(i-1, j) + s(IPA_C(i), _), \\ T(i, j-1) + s(_, IPA_E(j)) \end{pmatrix} \quad (3)$$

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

$T(n, m)$ will be the similarity score of the optimal alignment of IPA_C and IPA_E .

To evaluate the performance of transliteration, which is regarded as a similarity measurement in the above model, a corpus consisting of 1,574 pairs of English name and Chinese transliterated names is prepared. Total 97 phonemes are used to represent these names. Ten-fold cross validation methodology is adopted. That is, 8/10 and 1/10 of the corpus are used as training and validation, and the remaining 1/10 of the corpus is used for testing. The 9/10 of the corpus develops the score matrix s . Average score of ten folds are computed as the final performance. Two metrics, i.e., average rank and mean reciprocal rank, are proposed. Given a Chinese transliteration name in the test corpus, we try to find the corresponding English name. The rank of the correct English name in a list sorted by similarity score indicates the effect of the proposed method. The average rank is defined as follows.

$$AvgR = \frac{1}{N} \sum_{i=1}^N R_i \quad (4)$$

where N is total number of Chinese-English name pairs,

R_i is the rank of the i^{th} Chinese-English similarity matching.

The smaller the average rank is, the better the performance is. The mean reciprocal rank (MRR) proposed by TREC QA task [15] is also adopted.

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{R_i} \quad (5)$$

The value of *MRR* is between 0 and 1. The higher the *MRR* is, the better the performance is. In our test, the average rank is 2.04, and the mean reciprocal rank is 0.8322, which are very close to the best performance, i.e., 1, for both the two metrics.

5. Concluding Remarks

This paper proposes a framework to retrieve images with captions using speech. The speech and the annotated captions are in different languages. We tell out which part should be translated and which part should be transliterated. A scoring mechanism is presented to compute the similarity between speech query and annotated captions in terms of IPAs. In current work, only named location is allowed in speech query. Extending to other named entities and combining with general content words will be studied in the future. When more candidates have to be considered in similarity ranking, the speed issue should be addressed. Although optimal alignment of two phonetic strings requires $O(n \times m)$ time, where n and m are length of two strings, the overall time complexity depends on the number of candidates to be ranking. Indexing and filtering by phonemes can decrease the number of candidates before similarity computation. That will also be investigated.

6. References

- [1] Oard, D. and Diekema, A. "Cross-Language Information Retrieval", *Annual Review of Information Science and Technology*, vol. 33, pp. 223-256, 1998.
- [2] Sanderson, M. and Clough, P. "Eurovision-An Image-Based CLIR System", *Proceedings of Workshop at SIGIR2002, Cross-Language Information Retrieval: A Research Roadmap*, 2002.
- [3] CLEF *Cross Language Evaluation Forum*, <http://clef.iei.pi.cnr.it:2002/>, 2003.
- [4] Chen, H.H.; Bian, G.W. and Lin, W.C. "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval", *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pp. 215-222, 1999.
- [5] Thompson, P. and Dozier, C. "Name searching and information retrieval", *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.
- [6] Chen, H.H. *et al.* "Proper Name Translation in Cross-Language Information Retrieval", *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada, August 10-14 1998, pp. 232-236, 1998.
- [7] Lin, W.H. and Chen, H.H. "Similarity Measure in Backward Transliteration between Different Character Sets and Its Application to CLIR", *Proceedings of Research on Computational Linguistics Conference*, Taipei, Taiwan, pp. 97-113, 2000.
- [8] Lin, W.H. and Chen, H.H. "Backward Machine Transliteration by Learning Phonetic Similarity", *Proceedings of 6th Conference on Natural Language Learning: A COLING 2002 Workshop*, Taipei, Taiwan, 2002.
- [9] Appelt, D.E. and Martin, D. "Named Entity Extraction from Speech: Approach and Results using the TextPro System", *Proceedings of DARPA Broadcast News Workshop*, pp. 51-54, 1999.
- [10] Kubala, F.; Schwartz, R.; Stone, R. and Weischedel, R. "Named Entity Extraction from Speech", *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [11] CmuDict, *The CMU Pronouncing Dictionary* 0.6, <ftp://ftp.cs.cmu.edu/project/speech/dict>, 1995
- [12] Pagel, V. *et al.* "Letter to Sound Rules for Accented Lexicon Compression", *Proceedings of the 1998 International Conference on Spoken Language Processing*, 1998.
- [13] MUC (1998) Message Understanding Competition, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.
- [14] NICT (1995) *Translation of Foreign Location Names*, National Institute for Compilation and Translation, Taipei, Taiwan.
- [15] Voorhees, E.M. and Tice, D.M. (2000) "The TREC-9 Question Answering Track Report," *Proceedings of The Ninth Text REtrieval Conference (TREC 9)*, NIST Special Publication 500-249, Gaithersburg.