Journal of Chinese Language and Computing 15 (1): (33-44)

Incorporating Pronunciation Variation into Extraction of Transliterated-term Pairs from Web Corpora

Jin-Shea Kuo [1, 2]

[1]Chung-Hwa Telecommunication Laboratories, Taoyuan, Taiwan

jskuo@cht.com.tw

Ying-Kuei Yang [2]

[2]Electrical Eng. Dept.,
National Taiwan University of Science and Technology, Taipei, Taiwan

ykyang@mouse.ee.ntust.edu.tw

Abstract

A novel approach to automatically extracting transliterated-term pairs from Web corpora is proposed in this paper. One of the most important issues addressed is that of taking pronunciation variation into account. Pronunciation variation is a phenomenon of pronunciation ambiguity that seriously affects the term transliteration and hence affects those results produced by transliteration processes. Extracting transliterated-term pairs is a fundamental yet important task in natural language processing to collect large enough paired cognates for further studies on transliteration. To mitigate the problem of pronunciation variation in extracting paired cognates is not an easy task. The proposed method successfully exploits ASR (automated speech recognition)-generated confusion matrices as a basis for both alleviating pronunciation variation and constructing cross-linguistic syllable-and-phoneme conversions and it improves the extraction performance gradually by using cross-linguistic syllable-phoneme confusion matrices trained and refined progressively from extracted term pairs. Many terms extracted in the experiment are new to the existing lexicons. Experiments on mining information from the extracted pairs also have been conducted. From the experimental results showed that taking pronunciation variation into account did make extraction of paired cognates more effective

Keywords

Machine translation, machine transliteration, transliteration term extraction, pronunciation variation, Web corpora, data mining, text mining, Web mining, resource acquisition

1. Introduction

Many transliterated-term pairs have been requested to train various models and to learn rules in studying machine transliteration (Al-Onaizan 2002; Knight 2000; Lin 2002), cross-language information retrieval (Qu 2002; Virga 2000) and cross-language spoken document retrieval (Meng 2001). Most of the transliterated-term lists used in these papers are small in scale and/or compiled manually. A transliteration lexicon composed of many transliterated-term pairs is an important resource to researches on machine transliteration. However, it is time- and labor-consuming to prepare such a lexicon.

Transliterated-term extraction using parallel corpora has been conducted (Lee 2003). Generally speaking, parallel corpora are smaller in scale and less versatile in coverage as compared to non-parallel corpora. Query logs recorded by Internet search engines reveal users' intentions and contain much information about users' behaviors. An iterative process, which extracted Japanese-English cognate pairs from query logs, has been proposed. There are several problems associated with this process. First, the resource used is not publicly accessible. Second, a large English lexicon is required and each collected katakana term has to compare with each term in the English lexicon to calculate the term similarity until a threshold is reached in order to extract possible cognate pairs. This process not only resulted in high computing overheads but also degraded the extraction performance. If a term is not in the English lexicon, it is not possible to extract an English-katakana pair by this approach. However, this paper revealed an idea of mining transliterated-term pairs from a special Web resource when dealing with transliterated-term extraction.

The Internet is one of the largest distributed databases in the world. It comprises various kinds of data and at the same time is growing rapidly. Though the World Wide Web is not systematically organized, much invaluable information can still be obtained from this large text corpus, which can be accessed publicly. Constructing an English-Chinese transliteration lexicon automatically from Web corpora is the most important goal of this paper.

One of the most important factors that affect constructing a transliteration lexicon is pronunciation variation. Pronunciation variation is a problem of pronunciation ambiguity. Some phonemes in source language terms may be pronounced swiftly, quietly or strongly in many different situations according to speakers' speaking conventions. For example, different translators may transliterate "Disney" and "Honeywell" into different transliterations shown in Table 1in Chinese.

Disney	迪士尼 /DI-SHI-NI/[1]	迪斯耐 /DI-SI-NI	狄斯耐 /DI-SI-NI/
Honeywell	漢尼威 /HAI-NI-WEI/	霍尼威 /HUO-NI-WEI/	霍尼偉 /HUO-NI-WEI/

Table 1. Transliteration variation on the Web.

There are two kinds of pronunciation variations, namely lexical variation and allophonic variation (Jurafsky 2000). Dialect variation is one source of lexical variation and allophonic variation has to do with the phonemes changed in different contexts. For example, elision is quite common in English speech. /t/ and /d/ are often elided before consonants or when they

[1] Both English and Chinese pronunciations are referred to in / /. The English ones are in lower case, whereas, the Chinese ones represented in Hanyu are in capital.

are parts of a sequence of two or three consonants. Another type of isolated but not always elided pronunciation units, such as /l/ of "polder", may or may not be transliterated into one in Chinese depending on the translators. Two transliterated terms, "波德/BO-DE/" and "波爾德/BO-ER-DE/," can be generated for "polder" depending on whether /l/ is mapped to "爾/ER/" or null; however, both terms are correct for term transliteration. These two term pairs, "polder" and "波德/BO-DE/" and "polder" and "波爾德/BO-ER-DE/", should be able to extract when deal with transliterated-term extraction. Pronunciation variation has often been encountered in daily conversations and, therefore, in transliteration. This issue has not previously been discussed extensively with respect to the extraction of transliterated-term pairs. To model transliterated-term extraction effectively, pronunciation variation has to be taken into account.

English and Chinese are two languages with different alphabets and phoneme inventories. Each word in Chinese is monosyllabic. On the other hand, most of words in English are polysyllabic. Converting phonemes rendered from source- and target- language terms cross-linguistically between the languages that belong to different language families statistically is not an easy task. It is even more difficult if the pronunciation variation issue is taken into account.

In this paper, an approach, which takes pronunciation variation into consideration, is proposed for transliterated-term extraction from Web corpora. First, by using confusion matrices generated by a speech recognition process as a basis for both alleviating pronunciation variation and constructing phoneme conversion, one can extract paired transliterated-terms from the training text corpora. Then, a cross-linguistic syllable-and-phoneme conversion is trained using the extracted term pairs, which reflects the real cases of term transliteration. The generated conversion then provides a more rigid basis for extraction in next rounds. The process iterates until a criteria reached.

The remainder of the paper is organized as follows: Section 2 describes how English-Chinese transliterated term pairs can be extracted automatically in a bootstrapping manner. Experimental results obtained using Web corpora are presented in section 3. Section 4 provides an extensive discussion of transliterated-term extraction Conclusions are drawn in section 5.

2. The proposed approach

A new approach using different confusion matrices to boost the performance of extracting transliterated-term pairs is described in this section. Initially, confusion matrices produced by a speech recognition process act as a basis and then the progressively refined cross-linguistic syllable-and-phoneme conversion is used in paired cognate extraction. These conversions are used not only to construct the relation of phoneme mapping between two different languages, but also to alleviate the pronunciation variation occurred during transliterated-term extraction.

Generally, a sentence separated by a punctuation mark is selected from the training text corpora when extracting paired cognates. Then the candidates in the target language (Chinese in this paper) are obtained from the contexts of the located source-language (English in this paper) string in the selected sentence. Both strings in the source language and target language are converted into phonemes of the same representation in order to calculate the degree of similarity between these two terms. English phonemes are then syllabified into consonant-vowel pairs. The converted English syllables are transformed into Chinese syllables by using a basic English-to-Chinese phoneme conversion with handcoded rules initially when ASR (Automated Speech Recognition)-generated confusion matrices are used. Then the similarity degree between syllables is calculated, and a pair of transliterated terms can be extracted, depending on whether the similarity degree is larger than a predefined threshold or not.

Figure 1 shows the system diagram of the transliterated-term extraction. There are six steps in the proposed approach to transliterated-term extraction, namely, locating transliterated-term candidates, phoneme conversion, calculating the degree of similarity between paired terms, training cross-linguistic syllable-phoneme conversion and stop criterion evaluation. Each step is described in the following sections.

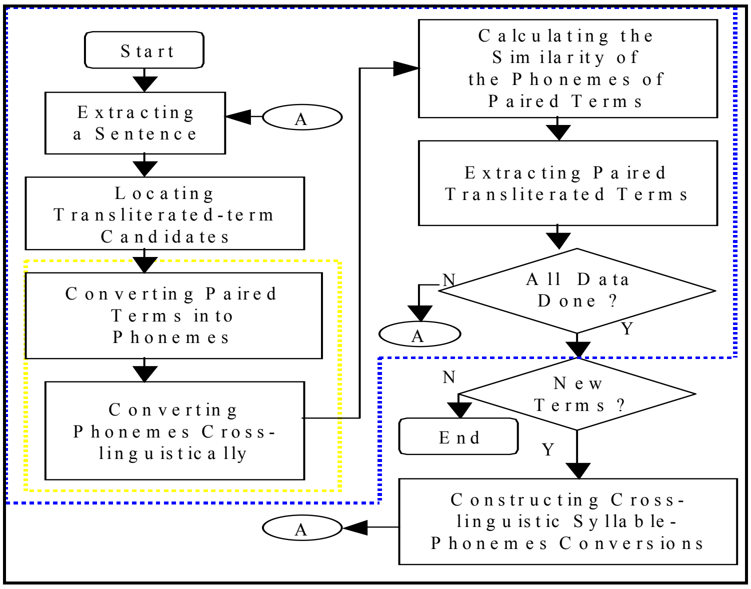
Start
Extracting a Sentence
A
Locating Transliterated-term Candidates
Converting Paired Terms into Phonemes
Converting Phonemes Cross-linguistically
Calculating the Similarity of the Phonemes of Paired Terms
Extracting Paired Transliterated Terms
All Data Done ?
N
A
Y
New Terms ?
N
End
Y
Constructing Cross-linguistic Syllable-Phonemes Conversions
A

Figure 1. The system diagram of extracting transliterated-term pairs.

2.1. Locating Transliterated-Term Candidates Using Local Context Analysis Algorithm

In order to perform term-to-term similarity calculation described later, candidate terms in source language and target language have to be aligned first; however, term alignment is not easy in processing non-parallel corpora. A local context analysis algorithm is used to locate the source-language string first and then to find the target-language strings within the context of the source-language string. These candidate strings in target language are not divided into terms using word sense disambiguation because most the terms are out-of-vocabulary. The actual boundaries of the transliterated equivalents in target language extracted from candidate strings are determined by calculating the similarity degree of the phonetic features between the selected candidate strings and their source-language string.

An approach was proposed for using partially parallel corpora on the Web to extract translation terms (Nagata 2001). The researchers observed that there are many translated-terms in English-Japanese mixed texts. These terms are located closely and possible translation candidates are always surrounded by parentheses in a sentence. Actually, this phenomenon has been observed not only in Japanese articles but also frequently in articles published in other oriental languages such as Korean and Chinese, and it can be adapted and applied to perform transliterated-term extraction. However, there are exceptions, as shown for example in the following text.

『...經營 Kuro 庫洛 P2P 音樂交換軟體的飛行網，3 日發表 P2P 與版權爭議的解決方案—C2C (Content to Community)。...』

In the above passage, C2C is not a transliterated-term meaning "Content to Community" at all and vice versa. On the other hand, "庫洛/KU-LUO/", which is not encompassed in parentheses, is indeed a transliterated term for "Kuro". The most important point observed here is that source language terms and their transliterated candidates are frequently closely related in mixed-language texts. To obtain paired cognates according to this point of view from non-parallel corpora, a more general algorithm, that employs local context analysis, is proposed here to set the ranges of transliterated-term candidates.

A sentence, $\boldsymbol{S}=\{\boldsymbol{s}_1\boldsymbol{s}_2...\boldsymbol{s}_m\}$, where each $\boldsymbol{s}_{i,}$ $i=1..m$ is a character, is extracted when any punctuation symbol is encountered in an article. In order to extract English-Chinese cognates, an attempt is made to locate a string in the source language. When such a string, $\boldsymbol{W}$, is found, it is decomposed into a set of tokens, i.e., $\boldsymbol{W}\in\boldsymbol{S}$, $\boldsymbol{W}=\{\boldsymbol{t}_1\boldsymbol{t}_2...\boldsymbol{t}_n\}$, where each token, $\boldsymbol{t}_i$, $i=1..n$, is composed of one or more characters. Possible strings in the target language can be selected within the left and right contexts of $\boldsymbol{W}$. Suppose that $\boldsymbol{W}$ is found at in position $[\boldsymbol{l}_{el}, \boldsymbol{l}_{er}]$, where $\boldsymbol{l}_{el}$ and $\boldsymbol{l}_{er}$ are the starting and ending positions of $\boldsymbol{W}$, respectively. The ranges of the left and right contexts can be expressed in terms of position by $[\max(\boldsymbol{l}_{ss}, \boldsymbol{l}_{cl}), \boldsymbol{l}_{el})$ and $(\boldsymbol{l}_{er}, \min(\boldsymbol{l}_{se}, \boldsymbol{l}_{cr})]$, respectively, where $\boldsymbol{l}_{ss}$ and $\boldsymbol{l}_{se}$ are the starting and the ending boundaries of the selected sentence, respectively, and $\boldsymbol{l}_{cl}$ and $\boldsymbol{l}_{cr}$ are the positions where an English character or a Chinese symbol is encountered next on the left-hand side and the right-hand side of $\boldsymbol{W}$, respectively.

2.2. English Phoneme Syllabification and Conversion

In order to determine the similarity between transliterated-term candidates, a common representation of English phonemes and Chinese phonemes is selected. Initially, confusion matrices, which have relations between syllables and phonemes are generated by a Chinese speech recognition system, are used to perform similarity calculation. These confusion matrices are used for phoneme conversion; however, it does not bear a cross-linguistic relation between English and Chinese phonemes. Using handcrafted rules initially and then using the trained cross-linguistic syllable-and-phoneme conversion directly in the rest of the process, a phoneme conversion can be conducted.

Once the phoneme conversion is performed, the degree of similarity between terms in different languages can be calculated. Before doing similarity calculation, both source language and target language terms are converted into phonemes using letter-to-sound systems and then these phonemes are syllabified and converted into a common representation.

2.3. Candidate Matrix Construction for Consonant-Vowel Pairs

Speech is introduced as an intermediate medium in transliteration process. Similar or confused sounds may be produced according to different tongue hump positions and tongue hump heights. In order to extract term pairs from a large text corpus and tackle the problem of pronunciation variation, confusion matrices are used. Confusion matrices have been generated and used mainly for error analysis to improve the performance of the recognition system. However, such data are invaluable to term extraction especially for extracting terms from scratch when no information about cross-linguistic phoneme conversion is available initially. Each row in the confusion matrix consists of a set of syllables that are correctly or erroneously recognized statistically and used in the recognition process according to the acoustic features.

A candidate matrix is constructed for consonant-vowel pairs by referring to confusion matrices generated by a speech recognition system initially, or the confusion matrices trained using extracted cognates. Only qualified syllables that are larger than the predefined threshold are kept in the candidate matrix. If qualified syllables are not found, sub-syllables are combined to form possible candidates. Once the cross-linguistic phoneme relationship is created, the similarity degree between the members of each term pair can be estimated.

2.4. Similarity Calculation between Paired Terms

Calculating the degree of similarity between candidate pairs and taking pronunciation variation into account by means of a model deduced from the source channel model, the transliterated cognates may be extracted. Statistical models for machine translation have been proposed in Brown (1993), and a noisy channel model for spelling correction has been proposed in Brill (2001). These models can be adapted so as to take pronunciation variation into consideration and applied to term extraction. Because MBRDICO (Pagel 1998) is used for English letter-to-phoneme transformation and according to the modular learning algorithm (Knight 2000), only cross-linguistic phoneme modeling is of interest in term extraction. Cross- linguistic phoneme similarity can be estimated by $p(PH_t \mid PH_s)$, where PH_t and PH_s are phoneme sequences of the target language and source language, respectively. The similarity degree between candidate string pairs can be estimated by

$$\boldsymbol{p(W_t \mid W_s) = \sum_{W_s^i} p(W_t \mid W_s{}^i)}\ , \tag{1}$$

where $\boldsymbol{W_s}$ is the source-language term used to examine whether a transliterated term exists in $\boldsymbol{W_t}$ or not, and $\boldsymbol{W_t}$ is the target-language candidate string. $\boldsymbol{W_s}$ can be decomposed further into tokens and can be expressed as $\boldsymbol{W_s = W_s^1 W_s^2 W_s^3,...,W_s^m}$, where each $\boldsymbol{W_s^i}$ is a token, which is the most elementary unit of source-language term in extraction.

In equation (1), there is a transliteration equivalent, Ť, in target language with the largest probability can be selected for each token Š in source language. Suppose that the possible transliterated-token pair is denoted by $\boldsymbol{\hat{J}}$ =(Š, Ť). Each $\boldsymbol{\hat{J}}$ can be determined by means of equation (2). In order to compare the similarity degree at the syllable level, each source-language token is converted into syllables, which are syllabified to obtain syllables. Equation (2) can be expressed as:

$$\begin{aligned}\hat{J} &= \mathop{\arg\max}_{W_t} p(W_t \mid W_s^i) \approx \mathop{\arg\max}_{W_t} p(W_t \mid H_s^i) \\ &\approx \mathop{\arg\max}_{H_t^j\ H_s^{iu}} p(H_t^1 H_t^2,\ldots,H_t^n \mid H_s^{i1} H_s^{i2},\ldots,H_s^{ik}),\end{aligned} \tag{2}$$

where $\boldsymbol{H_s^i}$, which is equal to $\boldsymbol{H_s^{i1} H_s^{i2},\ldots,H_s^{ik}}$, is converted from the English syllable sequence into the Chinese syllables. The target-language candidate string, which needs to be converted into the same representation as $\boldsymbol{H_s^i}$, is transformed into syllables with the help of Chinese homograph disambiguation and is denoted as $\boldsymbol{H_t}$, which is equal to $\boldsymbol{H_t^1 H_t^2,\ldots,H_t^n}$. When pronunciation variation is taken into consideration, $\boldsymbol{H_s^i}$ may contain many different combinations in which syllables with an isolated consonant may or may not be sounded in transliteration and can be expressed as $\boldsymbol{H_s^i} = \{(\boldsymbol{H_s^{11},\ldots,H_s^{1n_1}}),\ldots,(\boldsymbol{H_s^{U1},\ldots,H_s^{Un_U}})\}$, and U sub-sets of syllables in $\boldsymbol{H_s^i}$ in total. Each sub-set of the source-language syllables is a basic unit in examining whether there is a transliterated-term in the target-language candidate string or not. The window size of the target-language syllables varies dynamically according to the size of the selected basic unit. Therefore, equation (2) is changed into equation (3) under the assumption of independence:

$$\boldsymbol{\hat{J} \approx \mathop{\arg\max}_{H_t^j\ H_s^{iu}} \prod_{w=1}^{M_{iu}} p(H_t^{jw} \mid H_s^{iuw}) \approx \mathop{\arg\max}_{R_t^j\ R_s^{iu}} p(R_t^j \mid R_s^{iu})}, \tag{3}$$

where $\boldsymbol{M_{iu}} = \left|\boldsymbol{H_s^{iu}}\right|$ is the window size of each sub-set in $\boldsymbol{H_s^i}$. $\boldsymbol{R_t^j}$ and $\boldsymbol{R_s^{iu}}$ are the syllable sub-sets of the target-language term and the source-language term, respectively. Equation (3) determines the goodness of transliterated-token pairs based on the degree of syllable similarity. The similarity score for each syllable pair can be calculated using the information available in the syllable-based confusion matrix directly and can be expressed as $\boldsymbol{p(R_t^j \mid R_s^{iu}) = p(ST_t^j \mid ST_s^{iu})}$, where $\boldsymbol{ST_t^j}$ and $\boldsymbol{ST_s^{iu}}$ are units of syllables of the target language and source language, respectively.

The phoneme-based confusion matrix also has a fine-grained control that is used when the quality of the syllable-based confusion matrix is not good enough at the very beginning of the extraction process. $\boldsymbol{p(R_t^j \mid R_s^{iu})}$ can be estimated using these low-level primitives. A Chinese syllable can be divided into the initial and final parts. A Chinese initial is almost the same as a consonant cluster in English, and a Chinese final is also analogous to an English vowel or a combination of a vowel and a final consonant cluster in terms of functionality. Suppose that the initial part and the final part are generated independently. $\boldsymbol{p(R_t^j \mid R_s^{iu})}$ can also be estimated by equation (4):

$$\boldsymbol{p(R_t^j \mid R_s^{iu}) = p(PT_t^j \mid PT_s^{iu}) \approx \prod_{w=1}^{M_{iu}} p(I_t^{jw} \mid I_s^{iuw})\, p(F_t^{jw} \mid F_s^{iuw})}, \tag{4}$$

where $\boldsymbol{I_t^{jw}}$ and $\boldsymbol{F_t^{jw}}$ are sets of initial and final parts of a sliding window of target-language syllables, respectively, whereas, $\boldsymbol{I}_s^{iuw}$ and $\boldsymbol{F}_s^{iuw}$ are sets of the initial and the final parts of a sliding window of source-language syllables, respectively.

Equations (3) and (4) can be combined to form an equation with different weights for syllable-based and phoneme-based confusion matrices, respectively. By changing the weights, we can control the effects of different confusion matrices on transliterated-term extraction.

2.5. Training a Model for Cross-linguistic Syllable-Phoneme Conversion

From error analysis of extraction results, one of the most important issues that affect the performance of extraction is the conversion of English phonemes into Chinese syllables using the rules defined manually in the first iteration. If many term pairs have been extracted, a cross-linguistic syllable-phoneme conversion that will make the conversion fit better can be obtained for transliterated-term extraction in next iterations.

2.6. Stop Criteria Evaluation

To acquire enough term pairs for both transliteration and training, a refined cross-linguistic phoneme conversion is used in each transliterated-term extraction. The extraction process continues until no new term or sufficiently large term pairs are generated.

3. Experimental results

There are two phases in this experiment. First, confusion matrices generated by a speech recognition were used a basis to extract transliterated-term pairs which reflected the real cases of term transliteration. A progressively refined cross-linguistic conversion could be obtained by exploring from the extracted pairs. The precision and recall of the term extraction on training corpus were also estimated in order to estimate the qualified terms extracted in the large-scale term extraction described later. Second, the final cross-linguistic mapping obtained in the training phase was used to extract paired transliterated terms from a large corpus.

Initially, a mixed (English-Chinese) text corpus of 500MB with 15,822,984 pages, which were collected from the Internet using a Web spider and converted into plain text, was used as a training set. This corpus is called SET1. From SET1, 80,094 qualified sentences that occupied 5MB were extracted. A qualified sentence was one that was composed of at least one English string.

	NO_CM_NO_ELISION	NO_CM_WITH_ELISION
DQTP	1073	1177

Table 2. The results obtained without using confusion matrices.

The results obtained by running transliterated-term pairs extraction with the options of without using confusion matrix and with or without taking elision into account were shown in Table 2. All the distinct qualified term pairs (DQTP) reported in this paper were verified manually.

In order to improve the performance of transliterated-term extraction, confusion matrices (AGCM) produced by a speech recognition system were applied to this corpus. The results are depicted in Table 3. Those pairs produced by using both CASCM (the syllable part of AGCM) and CAPCM (the phoneme part of AGCM), which was called CACM for abbreviation, got the best results in generating term pairs.

	CASCM	CAPCM	CACM
DQTP	1,971	3,353	3,831

Table 3. The results obtained by using confusion matrices produced by a speech recognition system.

The collection of extracted term pairs produced by using AGCM was a "parallel" corpus and reflected the real cases of term transliteration. A CLSPC (cross-linguistic syllable-phoneme conversion) could be explored using this collection. The syllable and phoneme confusion matrices trained progressively by using extracted term pairs were called TCSCM and TCPCM, respectively.

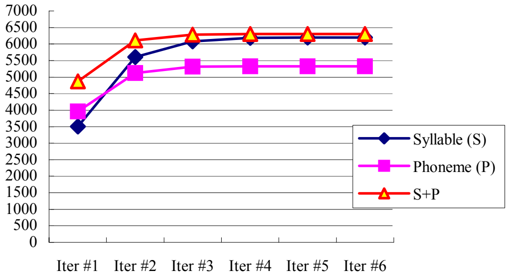

Figure 2. Extracting transliterated-term pairs using trained cross-linguistic syllable-phoneme conversions.

An algorithm for word alignment based on minimizing the edit distance between words with the same representation has been proposed (Brill 2001). However, the mapping between cross-linguistic phonemes is obtained only after the cross-linguistic relation is constructed. Such a relation is not available at the very beginning. A simple and fast approach that adopts equal syllable numbers is used to align syllables and phonemes cross-linguistically (Kuo 2004). The progress of extracting transliterated-term pairs using trained cross-linguistic syllable-phoneme conversions is shown in Figure 2.

The results obtained by using TCSCM, TCPCM and both TCSCM and TCPCM (TCCM for abbreviation) are displayed in Table 4. The results generated by using trained CLSPC were better than those produced by using AGCM, respectively.

	TCSCM	TCPCM	TCCM
DQTP	6,174	5,221	6,475

Table 4. The final result obtained by using cross-linguistic syllable-phoneme conversion.

To estimate the recall, precision and F-measures, which is equal to 2*recall*precision/(recall + precision), achieved by using the two proposed methods, 200 qualified sentences were randomly selected from the training corpus. The results shown in Table 5 reveal the achieved improvements. The improved recall was helpful to transliterated-term extraction.

	Precision	Recall	F-Measure
CACM	95.238%	32.258%	0.482
TCCM	71.429%	72.581%	0.720

Table 5. Estimated precisions, recalls and F-measures achieved by the proposed methods.

In order to provide a solid ground for term transliteration, a large-scale transliterated-term extraction conducted on the SET2 text corpus using the finally trained TCCM directly. The SET2 corpus was also collected from the Internet and was composed of 1,260,154 Web pages and occupied 3,336,303,998 bytes. About 72,329 qualifying term pairs were obtained.

Some examples of qualified transliterated-term pairs using the proposed approach are shown in Table 6. One important point worth of noting is that some newly transliterated terms such as “homework” and “fans” are found. These terms are out-of-vocabulary in existing dictionaries.

Homework (洪沃客)	Style (史黛爾)	Fans (粉絲)	House (浩室)	Oxygen (歐思淨)	Wild (王爾德)	Robert (蘿蔔(特))
Logo (漏狗)	Lightning (雷霆)	Model (魔豆)	Togo (土狗)	Order (歐德)	Short (蕭特)	Richard (瑞麒)

Table 6. Newly Transliterated pairs extracted using the proposed approach.

3.1. Mining information from the extracted pairs

A large quantity of transliterated-term pairs were extracted successfully using the proposed approach in this experiment. Several phenomena can be observed from the extracted pairs. Comparing those source language terms of the paired cognates with the existing lexicons available in the Internet, 31.080% and 47.768% of the extracted terms were not found in CMU Pronunciation Dictionary and Shorter Oxford English Dictionary, respectively. This means that the results produced by our approach achieved good performance as a supplement to the available dictionaries. It is more scalable and cost-effective as compared to the corpus[2], which was prepared manually over a period of several years.

[2] http://client.cna.com.tw/name/

In order to realize the characteristics of pronunciation variation in the extracted transliteration lexicon, some information was obtained by mining from the extracted pairs. All transliterated-term pairs used in the mining process were manually verified correct. We grouped the positions where an isolated syllable located in a word into two categories, namely, middle and final parts in analyzing the elision rates. Table 7 shows the elision rates of the top-six isolated syllables explored from the extracted transliteration lexicon in different cases including the whole lexicon, the whole lexicon in different positions and only elided cases in different positions. From the Table 7, it reveals that /r/ has been always elided in a word especially in the middle of syllables rendered from a word. It also means that incorporating this kind of information into the extraction of paired cognates can improve the extraction performance.

Isolated Syllables	Elision Rates of Whole Transliteration Lexicon	Elision Rates of Whole Transliteration Lexicon in Different Positions		Elision Rates of Elided Cases in Different Positions	
		Middle	Final	Middle	Final
/r/	61.132%	66.891%	33.706%	90.798%	9.202%
/l/	38.215%	47.529%	25.860%	72.293%	27.707%
/d/	29.219%	27.159%	31.860%	287761%	71.239%
/t/	28.088%	43.524%	21.438%	49,023%	50.997%
/z/	9.791%	7.516%	10.258%	11.165%	88.835%
/s/	5.732%	5.699%	8.037%	45.643%	54.357%

Table 7. Elision rates of top-six isolated syllables in the extracted transliteration lexicon.

4. Conclusions

In this paper, a novel approach by taking pronunciation variation for transliterated-term extraction has been proposed. Initially, using confusion matrices produced by a speech recognition system and finally a cross-linguistic syllable-and-phoneme conversion explored from the real cases of term transliteration, many transliterated-term pairs were extracted. The final conversion was used not only to construct the relation of phoneme mapping between two different languages, but also to alleviate the pronunciation variation occurred during transliterated-term extraction. By taking pronunciation variation into account, our approach was able to successfully extract transliterated-term pairs from Web pages, and in turn provides a solid ground for term transliteration. Experiments on mining information from the extracted transliteration lexicon also were conducted, from the experimental results showed that taking pronunciation variation into account did make extraction of paired cognates more effective.

5. Acknowledgement

The authors would like to thank Prof. Lee-Feng Chien at Academia Sinica for providing valuable suggestions. Many thanks are given to Prof. Wen-Hsiang Lu at National Cheng-Kung University, Dr. Wern-Jun Wang and Dr. Wen-Hsiang Lai for their very helpful discussions and supports in preparing the experiments.

References

Al-Onaizan, Y. and Knight, K., 2002, Machine Transliteration of Names in Arabic Text, In *Proceedings of 40th ACL Workshop on Computational Approaches to Semitic Languages*, pp. 34-46.

Brill, E., Kacmarcik, G., Brockett, C., 2001, Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs, In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pp. 393-399.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L., 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, vol. 19, no. 2, pp. 263-311

Jurafsky, D. and Martin, J. H., 2000, Speech and Language Processing, pp. 91-188, Prentice-Hall, New Jersey.

Knight, K. and Graehl, J., 2000, Machine Transliteration, Computational Linguistics, vol. 24, no. 4, pp. 599-612.

Kuo, J. S. and Yang, Y. K., 2004, Constructing Transliterations Lexicons from Web Corpora, In *the Companion Volume to the Proceedings of 42nd ACL,* pp. 102-105.

Lee, C. J. and Chang, J. S., 2003, Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model, In *Proceedings of HLT-NAACL*, Edmonton, Canada, pp. 96-103.

Lin, W. H. and Chen, H. H., 2002, Backward Machine Transliteration by Learning Phonetic Similarity, In *Proceedings of Sixth Conference on Natural Language Learning*, pp. 139-145

Meng, H., Lo, W. K., Chen, B. and Tang, K., 2001, Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval, In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, pp. 311-314.

Nagata, M., Saito, T., and Suzuki, K., 2001, Using the Web as a Bilingual Dictionary, In *Proceedings of 39th ACL Workshop on Data-Driven Methods in Machine Translation*, pp. 95-102.

Pagel, V., Lenzo, K., and Black, A., 1998, Letter to Sound Rules for Accented Lexicon Compression, In *Proceedings of ICSLP*, pp. 2015-2020.

Qu, Y., Grefenstette, G., and Evans, D., 2003, Automatic Transliteration for Japanese-to-English Text Retrieval, In *Proceedings of the 26th Annual International ACM SIGIR*, pp. 353-360.

Virga, P. and Khudanpur, S., 2003, Transliteration of Proper Names in Cross-Lingual Information Retrieval, In *Proceedings of 41st ACL Workshop on Multilingual and Mixed Language Named Entity Recognition*, pp. 57-64.