Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora

Pu-Jen Cheng[*], Yi-Cheng Pan[*], Wen-Hsiang Lu[+], and Lee-Feng Chien[*†]
[*] Institute of Information Science, Academia Sinica, Taiwan
[+] Dept. of Computer Science and Information Engineering, National Cheng Kung Univ., Taiwan
[†] Dept. of Information Management, National Taiwan University, Taiwan
{pjcheng, thomas02, whlu, lfchien}@iis.sinica.edu.tw

Abstract

The purpose of this paper is to automatically create multilingual translation lexicons with regional variations. We propose a transitive translation approach to determine translation variations across languages that have insufficient corpora for translation via the mining of bilingual search-result pages and clues of geographic information obtained from Web search engines. The experimental results have shown the feasibility of the proposed approach in efficiently generating translation equivalents of various terms not covered by general translation dictionaries. It also revealed that the created translation lexicons can reflect different cultural aspects across regions such as Taiwan, Hong Kong and mainland China.

1 Introduction

Compilation of translation lexicons is a crucial process for machine translation (MT) (Brown et al., 1990) and cross-language information retrieval (CLIR) systems (Nie et al., 1999). A lot of effort has been spent on constructing translation lexicons from domain-specific corpora in an automatic way (Melamed, 2000; Smadja et al., 1996; Kupiec, 1993). However, such methods encounter two fundamental problems: *translation of regional variations* and *the lack of up-to-date and high-lexical-coverage corpus source*, which are worthy of further investigation.

The first problem is resulted from the fact that the translations of a term may have variations in different dialectal regions. Translation lexicons constructed with conventional methods may not adapt to regional usages. For example, a Chinese-English lexicon constructed using a Hong Kong corpus cannot be directly adapted to the use in mainland China and Taiwan. An obvious example is that the word "taxi" is normally translated into "的士" (Chinese transliteration of taxi) in Hong Kong, which is completely different from the translated Chinese words of "出租车" (rental cars) in mainland China and "計程車" (cars with meters) in Taiwan. Besides, transliterations of a term are often pronounced differently across regions. For example, the company name "Sony" is transliterated into "新力" (xinli) in Taiwan and "索尼" (suoni) in mainland China. Such terms, in today's increasingly internationalized world, are appearing more and more often. It is believed that their translations should reflect the cultural aspects across different dialectal regions. Translations without consideration of the regional usages will lead to many serious misunderstandings, especially if the context to the original terms is not available.

Halpern (2000) discussed the importance of translating simplified and traditional Chinese lexemes that are semantically, not orthographically, equivalent in various regions. However, previous work on constructing translation lexicons for use in different regions was limited. That might be resulted from the other problem that most of the conventional approaches are based heavily on domain-specific corpora. Such corpora may be insufficient, or unavailable, for certain domains.

The Web is becoming the largest data repository in the world. A number of studies have been reported on experiments in the use of the Web to complement insufficient corpora. Most of them (Kilgarriff et al., 2003) tried to automatically collect parallel texts of different language versions (e.g. English and Chinese), instead of different regional versions (e.g. Chinese in Hong Kong and Taiwan), from the Web. These methods are feasible but only certain pairs of languages and subject domains can extract sufficient parallel texts as corpora. Different from the previous work, Lu et al. (2002) utilized Web anchor texts as a comparable bilingual corpus source to extract translations for out-of-vocabulary terms (OOV), the terms not covered by general translation dictionaries. This approach is applicable to the compilation of translation lexicons in diverse domains but requires powerful crawlers and high network bandwidth to gather Web data.

It is fortunate that the Web contains rich pages in a mixture of two or more languages for some lan-

guage pairs such as Asian languages and English. Many of them contain bilingual translations of terms, including OOV terms, e.g. companies', personal and technical names. In addition, geographic information about Web pages also provides useful clues to the regions where translations appear. We are, therefore, interested in realizing whether these nice characteristics make it possible to automatically construct multilingual translation lexicons with regional variations. Real search engines, such as *Google* (http://www.google.com) and *AltaVista* (http://www.altavista.com), allow us to search English terms only for pages in a certain language, e.g. Chinese or Japanese. This motivates us to investigate how to construct translation lexicons from bilingual search-result pages (as the corpus), which are normally returned in a long ordered list of *snippets* of summaries (including titles and page descriptions) to help users locate interesting pages.

The purpose of this paper is trying to propose a systematic approach to create multilingual translation lexicons with regional variations through mining of bilingual search-result pages. The bilingual pages retrieved by a term in one language are adopted as the corpus for extracting its translations in another language. Three major problems are found and have to be dealt with, including: (1) *extracting translations for unknown terms* – how to extract translations with correct lexical boundaries from noisy bilingual search-result pages, and how to estimate term similarity for determining correct translations from the extracted candidates; (2) *finding translations with regional variations* – how to find regional translation variations that seldom co-occur in the same Web pages, and how to identify the corresponding languages of the retrieved search-result pages once if the location clues (e.g. URLs) in them might not imply the language they are written in; and (3) *translation with limited corpora* – how to translate terms with insufficient search-result pages for particular pairs of languages such as Chinese and Japanese, and simplified Chinese and traditional Chinese.

The goal of this paper is to deal with the three problems. Given a term in one language, all possible translations will be extracted from the obtained bilingual search-result pages based on their similarity to the term. For those language pairs with unavailable corpora, a *transitive translation model* is proposed, by which the source term is translated into the target language through an intermediate language. The transitive translation model is further enhanced by a competitive linking algorithm. The algorithm can effectively alleviate the problem of error propagation in the process of translation, where translation errors may occur due to incorrect identification of the ambiguous terms in the intermediate language. In addition, because the search-result pages might contain snippets that do not be really written in the target language, a filtering process is further performed to eliminate the translation variations not of interest.

Several experiments have been conducted to examine the performance of the proposed approach. The experimental results have shown that the approach can generate effective translation equivalents of various terms – especially for OOV terms such as proper nouns and technical names, which can be used to enrich general translation dictionaries. The results also revealed that the created translation lexicons can reflect different cultural aspects across regions such as Taiwan, Hong Kong and mainland China.

In the rest of this paper, we review related work in translation extraction in Section 2. We present the transitive model and describe the direct translation process in Sections 3 and 4, respectively. The conducted experiments and their results are described in Section 5. Finally, in Section 6, some concluding remarks are given.

2 Related Work

In this section, we review some research in generating translation equivalents for automatic construction of translational lexicons.

Transitive translation: Several transitive translation techniques have been developed to deal with the unreliable direct translation problem. Borin (2000) used various sources to improve the alignment of word translation and proposed the pivot alignment, which combined direct translation and indirect translation via a third language. Gollins et al. (2001) proposed a feasible method that translated terms in parallel across multiple intermediate languages to eliminate errors. In addition, Simard (2000) exploited the transitive properties of translations to improve the quality of multilingual text alignment.

Corpus-based translation: To automatically construct translation lexicons, conventional research in MT has generally used statistical techniques to extract translations from domain-specific sentence-aligned parallel bilingual corpora. Kupiec (1993) attempted to find noun phrase correspondences in parallel corpora using part-of-speech tagging and noun phrase recognition methods. Smadja et al. (1996) proposed a statistical association measure of the Dice coefficient to deal with the problem of collocation translation. Melamed (2000) proposed statistical translation models to improve the techniques of word alignment by taking advantage of pre-existing knowledge, which was more effective than a knowledge-free model. Although high accuracy of translation extraction can be easily achieved by these techniques, sufficiently large parallel corpora for

(a) Taiwan (Traditional Chinese) (b) Mainland China (Simplified Chinese) (c) Hong Kong (Traditional Chinese)

Figure 1: Examples of the search-result pages in different Chinese regions that were obtained via the English query term "George Bush" from Google.

various subject domains and language pairs are not always available.

Some attention has been devoted to automatic extraction of term translations from comparable or even unrelated texts. Such methods encounter more difficulties due to the lack of parallel correlations aligned between documents or sentence pairs. Rapp (1999) utilized non-parallel corpora based on the assumption that the contexts of a term should be similar to the contexts of its translation in any language pairs. Fung et al. (1998) also proposed a similar approach that used a vector-space model and took a bilingual lexicon (called seed words) as a feature set to estimate the similarity between a word and its translation candidates.

Web-based translation: Collecting parallel texts of different language versions from the Web has recently received much attention (Kilgarriff et al., 2003). Nie et al. (1999) tried to automatically discover parallel Web documents. They assumed a Web page's parents might contain the links to different versions of it and Web pages with the same content might have similar structures and lengths. Resnik (1999) addressed the issue of language identification for finding Web pages in the languages of interest. Yang et al. (2003) presented an alignment method to identify one-to-one Chinese and English title pairs based on dynamic programming. These methods often require powerful crawlers to gather sufficient Web data, as well as more network bandwidth and storage. On the other hand, Cao et al. (2002) used the Web to examine if the arbitrary combination of translations of a noun phrase was statistically important.

3 Construction of Translation Lexicons

To construct translation lexicons with regional variations, we propose a transitive translation model $S_{trans}(s,t)$ to estimate the degree of possibility of the translation of a term s in one (source) language l_s into a term t in another (target) language l_t. Given the term s in l_s, we first extract a set of terms $C=\{t_j\}$, where t_j in l_t acts as a translation candidate of s, from a corpus. In this case, the corpus consists of a set of search-result pages retrieved from search engines using term s as a query. Based on our previous work (Cheng et al., 2004), we can efficiently extract term t_j by calculating the association measurement of every character or word n-gram in the corpus and applying the local maxima algorithm. The association measurement is determined by the degree of cohesion holding the words together within a word n-gram, and enhanced by examining if a word n-gram has complete lexical boundaries. Next, we rank the extracted candidates C as a list T in a decreasing order by the model $S_{trans}(s,t)$ as the result.

3.1 Bilingual Search-Result Pages

The Web contains rich texts in a mixture of multiple languages and in different regions. For example, Chinese pages on the Web may be written in traditional or simplified Chinese as a principle language and in English as an auxiliary language. According to our observations, translated terms frequently occur together with a term in mixed-language texts. For example, Figure 1 illustrates the search-result pages of the English term "George Bush," which was submitted to Google for searching Chinese pages in different regions. In Figure 1 (a) it contains the translations "喬治布希" (George Bush) and "布希" (Bush) obtained from the pages in Taiwan. In Figures 1 (b) and (c) the term "George Bush" is translated into "布什"(busir) or "布甚"(buson) in mainland China and "布殊"(busu) in Hong Kong. This characteristic of bilingual search-result pages is also useful for other language pairs such as other Asian languages mixed with English.

For each term to be translated in one (source) language, we first submit it to a search engine for locating the bilingual Web documents containing the term and written in another (target) language from a specified region. The returned search-result pages containing snippets (illustrated in Figure 1), instead of the documents themselves, are collected as a corpus from which translation candidates are extracted and correct translations are then selected.

Compared with parallel corpora and anchor texts, bilingual search-result pages are easier to collect and can promptly reflect the dynamic content of the Web.

In addition, geographic information about Web pages such as URLs also provides useful clues to the regions where translations appear.

3.2 The Transitive Translation Model

Transitive translation is particularly necessary for the translation of terms with regional variations because the variations seldom co-occur in the same bilingual pages. To estimate the possibility of being the translation $t \in T$ of term s, the transitive translation model first performs so-called *direct translation*, which attempts to learn translational equivalents directly from the corpus. The direct translation method is simple, but strongly affected by the quality of the adopted corpus. (Detailed description of the direct translation method will be given in Section 4.)

If the term s and its translation t appear infrequently, the statistical information obtained from the corpus might not be reliable. For example, a term in simplified Chinese, e.g. 互联网 (Internet) does not usually co-occur together with its variation in traditional Chinese, e.g. 網際網路 (Internet). To deal with this problem, our idea is that the term s can be first translated into an intermediate translation m, which might co-occur with s, via a third (or intermediate) language l_m. The correct translation t can then be extracted if it can be found as a translation of m. The transitive translation model, therefore, combines the processes of both *direct translation* and *indirect translation,* and is defined as:

$$S_{trans}(s,t) = \begin{cases} S_{direct}(s,t), & \text{if } S_{direct}(s,t) > \theta \\ S_{indirect}(s,t) = \sum_{\forall m} S_{direct}(s,m) \times S_{direct}(m,t) \times \varpi(m), & \text{otherwise} \end{cases}$$

where m is one of the top k most probable intermediate translations of s in language l_m, and ϖ is the confidence value of m's accuracy, which can be estimated based on m's probability of occurring in the corpus, and θ is a predefined threshold value.

3.3 The Competitive Linking Algorithm

One major challenge of the transitive translation model is the propagation of translation errors. That is, incorrect m will significantly reduce the accuracy of the translation of s into t. A typical case is the *indirect association problem* (Melamed, 2000), as shown in Figure 2 in which we want to translate the term s_1 (s=s_1). Assume that t_1 is s_1's corresponding translation, but appears infrequently with s_1. An indirect association error might arise when t_2, the translation of s_1's highly relevant term s_2, co-occurs often with s_1. This problem is very important for the situation in which translation is a many-to-many mapping. To reduce such errors and enhance the reliability of the estimation, a *competitive linking* algorithm, which is extended from Melamed's work (Melamed, 2000), is developed to determine the most probable translations.

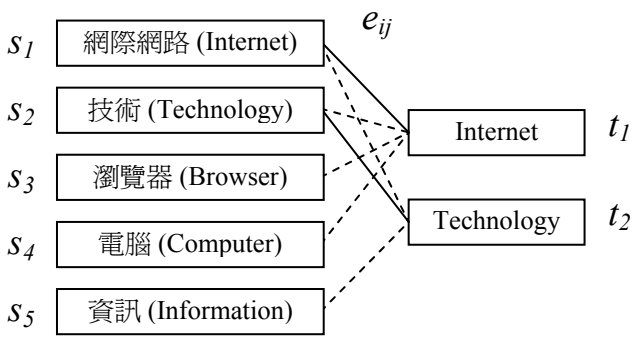

Figure 2: An illustration of a bipartite graph.

The idea of the algorithm is described below. For each translated term $t_j \in T$ in l_t, we translate it back into original language l_s and then model the translation mappings as a bipartite graph, as shown in Figure 2, where the vertices on one side correspond to the terms $\{s_i\}$ or $\{t_j\}$ in one language. An edge e_{ij} indicates the corresponding two terms s_i and t_j might be the translations of each other, and is weighted by the sum of $S_{direct}(s_i,t_j)$ and $S_{direct}(t_j,s_i)$. Based on the weighted values, we can examine if each translated term $t_j \in T$ in l_t can be correctly translated into the original term s_1. If term t_j has any translations better than term s_1 in l_s, term t_j might be a so-called indirect association error and should be eliminated from T. In the above example, if the weight of e_{22} is larger than that of e_{12}, the term "Technology" will be not considered as the translation of "網際網路" (Internet). Finally, for all translated terms $\{t_j\} \subseteq T$ that are not eliminated, we re-rank them by the weights of the edges $\{e_{ij}\}$ and the top k ones are then taken as the translations. More detailed description of the algorithm could be referred to Lu et al. (2004).

4 Direct Translation

In this section, we will describe the details of the direct translation process, i.e. the way to compute $S_{direct}(s,t)$. Three methods will be presented to estimate the similarity between a source term and each of its translation candidates. Moreover, because the search-result pages of the term might contain snippets that do not actually be written in the target language, we will introduce a filtering method to eliminate the translation variations not of interest.

4.1 Translation Extraction

The Chi-square Method: A number of statistical measures have been proposed for estimating term association based on co-occurrence analysis, including mutual information, DICE coefficient, chi-square test, and log-likelihood ratio (Rapp, 1999). Chi-square test (χ^2) is adopted in our study because the required parameters for it can be obtained by submit-

ting Boolean queries to search engines and utilizing the returned page counts (number of pages). Given a term s and a translation candidate t, suppose the total number of Web pages is N; the number of pages containing both s and t, $n(s,t)$, is a; the number of pages containing s but not t, $n(s,\neg t)$, is b; the number of pages containing t but not s, $n(\neg s,t)$, is c; and the number of pages containing neither s nor t, $n(\neg s, \neg t)$, is d. (Although d is not provided by search engines, it can be computed by $d=N-a-b-c$.) Assume s and t are independent. Then, the expected frequency of (s,t), $E(s,t)$, is $(a+c)(a+b)/N$; the expected frequency of $(s,\neg t)$, $E(s,\neg t)$, is $(b+d)(a+b)/N$; the expected frequency of $(\neg s,t)$, $E(\neg s,t)$, is $(a+c)(c+d)/N$; and the expected frequency of $(\neg s,\neg t)$, $E(\neg s,\neg t)$, is $(b+d)(c+d)/N$.

Hence, the conventional chi-square test can be computed as:

$$S^{\chi 2}_{direct}(s,t) = \sum_{\forall X\in\{s,\neg s\},\forall Y\in\{t,\neg t\}} \frac{[n(X,Y)-E(X,Y)]^2}{E(X,Y)} = \frac{N\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)}.$$

Although the chi-square method is simple to compute, it is more applicable to high-frequency terms than low-frequency terms since the former are more likely to appear with their candidates. Moreover, certain candidates that frequently co-occur with term s may not imply that they are appropriate translations. Thus, another method is presented.

The Context-Vector Method: The basic idea of this method is that the term s's translation equivalents may share common contextual terms with s in the search-result pages, similar to Rapp (1999). For both s and its candidates C, we take their contextual terms constituting the search-result pages as their features. The similarity between s and each candidate in C will be computed based on their feature vectors in the vector-space model.

Herein, we adopt the conventional *tf-idf* weighting scheme to estimate the significance of features and define it as:

$$w_{t_i} = \frac{f(t_i,p)}{\max_j f(t_j,p)} \times \log(\frac{N}{n}) \ ,$$

where $f(t_i,p)$ is the frequency of term t_i in search-result page p, N is the total number of Web pages, and n is the number of the pages containing t_i. Finally, the similarity between term s and its translation candidate t can be estimated with the cosine measure, i.e. $S^{CV}_{direct}(s,t)=cos(cv_s, cv_t)$, where cv_s and cv_t are the context vectors of s and t, respectively.

In the context-vector method, a low-frequency term still has a chance of extracting correct translations, if it shares common contexts with its translations in the search-result pages. Although the method provides an effective way to overcome the chi-square method's problem, its performance depends heavily on the quality of the retrieved search-result pages, such as the sizes and amounts of snippets. Also, feature selection needs to be carefully handled in some cases.

The Combined Method: The context-vector and chi-square methods are basically complementary. Intuitively, a more complete solution is to integrate the two methods. Considering the various ranges of similarity values between the two methods, we compute the similarity between term s and its translation candidate t by the weighted sum of $1/R_{\chi 2}(s,t)$ and $1/R_{CV}(s,t)$. $R_{\chi 2}(s,t)$ (or $R_{CV}(s,t)$) represents the similarity ranking of each translation candidate t with respect to s and is assigned to be from 1 to k (number of output) in decreasing order of similarity measure $S^{\chi 2}_{direct}(s,t)$ (or $S^{CV}_{direct}(s,t)$). That is, if the similarity rankings of t are high in both of the context-vector and chi-square methods, it will be also ranked high in the combined method.

4.2 Translation Filtering

The direct translation process assumes that the retrieved search-result pages of a term exactly contain snippets from a certain region (e.g. Hong Kong) and written in the target language (e.g. traditional Chinese). However, the assumption might not be reliable because the location (e.g. URL) of a Web page may not imply that it is written by the principle language used in that region. Also, we cannot identify the language of a snippet simply using its character encoding scheme, because different regions may use the same character encoding schemes (e.g. Taiwan and Hong Kong mainly use the same traditional Chinese encoding scheme).

From previous work (Tsou et al., 2004) we know that word entropies significantly reflect language differences in Hong Kong, Taiwan and China. Herein, we propose another method for dealing with the above problem. Since our goal is trying to eliminate the translation candidates $\{t_j\}$ that are not from the snippets in language l_t, for each candidate t_j we merge all of the snippets that contain t_j into a document and then identify the corresponding language of t_j based on the document. We train a uni-gram language model for each language of concern and perform language identification based on a discrimination function, which locates maximum character or word entropy and is defined as:

$$lang(t_j) = \arg\max_{l\in L}\left\{\sum_{w\in N(tj)} p(w\mid l)\ln p(w\mid l)\right\},$$

where $N(t_j)$ is the collection of the snippets containing t_j and L is a set of languages to be identified. The candidate t_j will be eliminated if $lang(t_j)\neq l_t$.

To examine the feasibility of the proposed method in identifying Chinese in Taiwan, mainland China and Hong Kong, we conducted a preliminary experiment. To avoid the data sparseness of using a tri-gram language model, we simply use the above unigram model to perform language identification. Even so, the experimental result has shown that very high identification accuracy can be achieved. Some Web portals contain different versions for specific regions such as *Yahoo! Taiwan* (http://tw.yahoo.com) and *Yahoo! Hong Kong* (http://hk.yahoo.com). This allows us to collect regional training data for constructing language models. In the task of translating English terms into traditional Chinese in Taiwan, the extracted candidates for "laser" contained "雷射" (translation of laser mainly used in Taiwan) and "激光" (translation of laser mainly used in mainland China). Based on the merged snippets, we found that "激光" had higher entropy value for the language model of mainland China while "雷射" had higher entropy value for the language models of Taiwan and Hong Kong.

5 Performance Evaluation

We conducted extensive experiments to examine the performance of the proposed approach. We obtained the search-result pages of a term by submitting it to the real-world search engines, including *Google* and *Openfind* (http://www.openfind.com.tw). Only the first 100 snippets received were used as the corpus.

Performance Metric: The average top-*n* inclusion rate was adopted as a metric on the extraction of translation equivalents. For a set of terms to be translated, its top-*n* inclusion rate was defined as the percentage of the terms whose translations could be found in the first *n* extracted translations. The experiments were categorized into direct translation and transitive translation.

5.1 Direct Translation

Data set: We collected English terms from two real-world Chinese search engine logs in Taiwan, i.e. *Dreamer* (http://www.dreamer.com.tw) and *GAIS* (http://gais.cs.ccu.edu.tw). These English terms were potential ones in the Chinese logs that needed correct translations. The Dreamer log contained 228,566 unique query terms from a period of over 3 months in 1998, while the GAIS log contained 114,182 unique query terms from a period of two weeks in 1999. The collection contained a set of 430 frequent English terms, which were obtained from the 1,230 English terms out of the most popular 9,709 ones (with frequencies above 10 in both logs). About 36% (156/430) of the collection could be found in the LDC (*Linguistic Data Consortium*, http://www.ldc.upenn.edu/Projects/Chinese) English-to-Chinese lexicon with 120K entries, while about 64% (274/430) were not covered by the lexicon.

English-to-Chinese Translation: In this experiment, we tried to directly translate the collected 430 English terms into traditional Chinese. Table 1 shows the results in terms of the top 1-5 inclusion rates for the translation of the collected English terms. "χ^2", "*CV*", and "χ^2*+CV"* represent the methods based on the chi-square, context-vector, and chi-square plus context-vector methods, respectively. Although either the chi-square or context-vector method was effective, the method based on both of them (χ2+CV) achieved the best performance in maximizing the inclusion rates in every case because they looked complementary. The proposed approach was found to be effective in finding translations of proper names, e.g. personal names "Jordan" (喬丹, 喬登), "Keanu Reeves" (基努李維, 基諾李維), companies' names "TOYOTA" (豐田), "EPSON" (愛普生), and technical terms "EDI" (電子資料交換), "Ethernet" (乙太網路), etc.

English-to-Chinese Translation for Mainland China, Taiwan and Hong Kong: Chinese can be classified into *simplified Chinese* (SC) and *traditional Chinese* (TC) based on its writing form or character encoding scheme. SC is mainly used in mainland China while TC is mainly used in Taiwan and Hong Kong (HK). In this experiment, we further investigated the effectiveness of the proposed approach in English-to-Chinese translation for the three different regions. The collected 430 English terms were classified into five types: *people*, *organization*, *place*, *computer and network*, and *others*.

Tables 2 and 3 show the statistical results and some examples, respectively. In Table 3, the number stands for a translated term's ranking. The underlined terms were correct translations and the others were relevant translations. These translations might benefit the CLIR tasks, whose performance could be referred to our earlier work which emphasized on translating unknown queries (Cheng et al., 2004). The results in Table 2 show that the translations for mainland China and HK were not reliable enough in the top-1, compared with the translations for Taiwan. One possible reason was that the test terms were collected from Taiwan's search engine logs. Most of them were popular in Taiwan but not in the others. Only 100 snippets retrieved might not balance or be sufficient for translation extraction. However, the inclusion rates for the three regions were close in the top-5. Observing the five types, we could find that type *place* containing the names of well-known countries and cities achieved the best performance in maximizing the inclusion rates in every case and almost had no regional variations (9%, 1/11) except

Table 1: Inclusion rates for Web query terms using various similarity measurements

Method	Dic			OOV			All		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
χ^2	42.1%	57.9%	62.1%	40.2%	53.8%	56.2%	41.4%	56.3%	59.8%
CV	51.7%	59.8%	62.5%	45.0%	55.6%	57.4%	49.1%	58.1%	60.5%
χ^2+ CV	52.5%	60.4%	63.1%	46.1%	56.2%	58.0%	50.7%	58.8%	61.4%

Table 2: Inclusion rates for different types of Web query terms

Type	Extracted Translations								
	Taiwan (Big5)			Mainland China (GB)			Hong Kong (Big5)		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
People (14)	57.1%	64.3%	64.3%	35.7%	57.1%	64.3%	21.4%	57.1%	57.1%
Organization (147)	44.9%	55.1%	56.5%	47.6%	58.5%	62.6%	37.4%	46.3%	53.1%
Place (11)	90.9%	90.9%	90.9%	63.6%	100.0%	100.0%	81.8%	81.8%	81.8%
Computer & Network (115)	55.8%	59.3%	63.7%	32.7%	59.3%	64.6%	42.5%	65.5%	68.1%
Others (143)	49.0%	58.7%	62.2%	30.8%	49.7%	58.7%	28.7%	50.3%	60.8%
Total (430)	50.7%	58.8%	61.4%	38.1%	56.7%	62.8%	36.5%	54.0%	60.5%

Table 3: Examples of extracted correct/relevant translations of English terms in three Chinese regions

English Terms	Extracted Correct or Relevant Target Translations		
	Taiwan (Traditional Chinese)	Mainland China (Simplified Chinese)	Hong Kong (Traditional Chinese)
Police	警察 (1) 警察隊 (2) 警察局 (4)	警察 (1) 警务 (2) 公安 (4)	警務處 (1) 警察 (3) 警司 (5)
Taxi	計程車 (1) 交通 (3)	出租车 (1) 的士 (4)	的士 (1) 的士司機 (2) 收費表 (15)
Laser	雷射 (1) 雷射光源 (3) 測距槍(4)	激光 (1) 中国 (2) 激光器 (3) 雷射 (4)	激光 (1) 雷射 (2) 激光的 (3) 鐳射 (4)
Hacker	駭客 (1) 網路 (2) 軟體 (7)	黑客 (1) 网络安全 (5) 防火墙 (6)	駭客 (1) 黑客 (2) 互聯網 (9)
Database	資料庫 (1) 中文資料庫 (3)	数据库 (1) 数据库维护 (9)	資料庫 (1) 數據庫 (3) 資料 (5)
Information	資訊 (1) 新聞 (3) 資訊網 (4)	信息 (1) 信息网 (3) 资讯 (7)	資料 (1) 資訊 (6)
Internet café	網路咖啡 (3) 網路 (4) 網咖 (5)	网络咖啡 (1) 网络咖啡屋 (2) 网吧 (6)	網吧 (1) 香港 (3) 網站 (4)
Search Engine	搜尋器 (2) 搜尋引擎 (5)	搜索引擎工厂 (1) 搜索引擎 (3)	搜索器 (1) 搜尋器 (8)
Digital Camera	相機 (1) 數位相機 (2)	数码相机 (1) 数码影像 (6)	像素 (1) 數碼相機 (2) 相機 (3)

Table 4: Inclusion rates of transitive translations of proper names and technical terms

Type	Source Language	Target Language	Intermediate Language	Top-1	Top-3	Top5
Scientist Name	Chinese	English	None	70.0%	84.0%	86.0%
	English	Japanese	None	32.0%	56.0%	64.0%
	English	Korean	None	34.0%	58.0%	68.0%
	Chinese	Japanese	English	26.0%	40.0%	48.0%
	Chinese	Korean	English	30.0%	42.0%	50.0%
Disease Name	Chinese	English	None	50.0%	74.0%	74.0%
	English	Japanese	None	38.0%	48.0%	62.0%
	English	Korean	None	30.0%	50.0%	58.0%
	Chinese	Japanese	English	32.0%	44.0%	50.0%
	Chinese	Korean	English	24.0%	38.0%	44.0%

that the city "Sydney" was translated into 悉尼 (Sydney) in SC for mainland China and HK and 雪梨 (Sydney) in TC for Taiwan. Type *computer and network* containing technical terms had the most regional variations (41%, 47/115) and type *people* had 36% (5/14). In general, the translations in the two types were adapted to the use in different regions. On the other hand, 10% (15/147) and 8% (12/143) of the translations in types *organization* and *others*, respectively, had regional variations, because most of the terms in type *others* were general terms such as "bank" and "movies" and in type *organization* many local companies in Taiwan had no translation variations in mainland China and HK.

Moreover, many translations in the types of *people*, *organization*, and *computer and network* were quite different in Taiwan and mainland China such as the personal name "Bred Pitt" was translated into "毕彼特" in SC and "布萊德彼特" in TC, the company name "Ericsson" into "爱立信" in SC and "易利信" in TC, and the computer-related term "EDI" into "電子數據聯通" in SC and "電子資料交換" in TC. In general, the translations in HK had a higher chance to cover both of the translations in mainland China and Taiwan.

5.2 Multilingual & Transitive Translation

Data set: Since technical terms had the most region variations among the five types as mentioned in the previous subsection, we collected two other data sets for examining the performance of the proposed approach in multilingual and transitive translation. The data sets contained 50 scientists' names and 50 disease names in English, which were randomly selected from 256 scientists (Science/People) and 664 diseases (Health/Diseases) in the *Yahoo! Directory* (http://www.yahoo.com), respectively.

English-to-Japanese/Korean Translation: In this experiment, the collected scientists' and disease names in English were translated into Japanese and Korean to examine if the proposed approach could be applicable to other Asian languages. As the result in Table 4 shows, for the English-to-Japanese translation, the top-1, top-3, and top-5 inclusion rates were 35%, 52%, and 63%, respectively; for the English-to-Korean translation, the top-1, top-3, and top-5 inclusion rates were 32%, 54%, and 63%, respectively, on average.

Chinese-to-Japanese/Korean Translation via English: To further investigate if the proposed transitive approach can be applicable to other language pairs that are not frequently mixed in documents such as Chinese and Japanese (or Korean), we did transitive translation via English. In this experiment, we first manually translated the collected data sets in English into traditional Chinese and then did the Chinese-to-Japanese/Korean translation via the third language English.

The results in Table 4 show that the propagation of translation errors reduced the translation accuracy. For example, the inclusion rates of the Chinese-to-Japanese translation were lower than those of the English-to-Japanese translation since only 70%-86% inclusion rates were reached in the Chinese-to-English translation in the top 1-5. Although transitive translation might produce more noisy translations, it still produced acceptable translation candidates for human verification. In Table 4, 45%-50% of the extracted top 5 Japanese or Korean terms might have correct translations.

6 Conclusion

It is important that the translation of a term can be automatically adapted to its usage in different dialectal regions. We have proposed a Web-based translation approach that takes into account limited bilingual search-result pages from real search engines as comparable corpora. The experimental results have shown the feasibility of the automatic approach in generation of effective translation equivalents of various terms and construction of multilingual translation lexicons that reflect regional translation variations.

References

L. Borin. 2000. You'll take the high road and I'll take the low road: using a third language to improve bilingual word alignment. In *Proc. of COLING-2000*, pp. 97-103.

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

Y.-B. Cao and H. Li. 2002. Base noun phrase translation using Web data the EM algorithm. In *Proc. of COLING-2002*, pp. 127-133.

P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. 2004. Translating unknown queries with Web corpora for cross-language information retrieval. In *Proc. of ACM SIGIR-2004*.

P. Fung and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of ACL-98*, pp. 414-420.

T. Gollins and M. Sanderson. 2001. Improving cross language information with triangulated translation. In *Proc. of ACM SIGIR-2001*, pp. 90-95.

J. Halpern. 2000. Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In *Proc. of Workshop on Asian Language Resources and International Standardization*.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3): 333-348.

J. M. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of ACL-93*, pp. 17-22.

W.-H. Lu, L.-F. Chien, and H.-J. Lee. 2004. Anchor text mining for translation of web queries: a transitive translation Approach. *ACM TOIS* 22(2): 242-269.

W.-H. Lu, L.-F. Chien, and H.-J. Lee. 2002. Translation of Web queries using anchor text mining. *ACM TALIP*: 159-172.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2): 221-249.

J.-Y. Nie, P. Isabelle, M. Simard, and R. Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proc. of ACM SIGIR-99*, pp. 74-81.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora, In *Proc. of ACL-99,* pp. 519-526.

P. Resnik. 1999. Mining the Web for bilingual text. In *Proc. of ACL-99*, pp. 527-534.

M. Simard. 2000. *Multilingual Text Alignment*. In "Parallel Text Processing", J. Veronis, ed., pages 49-67, Kluwer Academic Publishers, Netherlands.

F. Smadja, K. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1): 1-38.

B. K. Tsou, T. B. Y. Lai, and K. Chow. 2004. Comparing entropies within the Chinese language. In *Proc. of IJCNLP-2004.*

C. C. Yang and K.-W. Li. 2003. Automatic construction of English/Chinese parallel corpora. *JASIST* 54(8): 730-742.