國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩　士　論　文

利用網路探勘之中英專名萃取研究

BILINGUAL PROPER NOUNS
EXTRACTION THROUGH WEB MINING

研 究 生　：蘇傳堯

指導教授　：梁　婷 博士

中華民國　九十五　年　六　月

利用網路探勘之中英專名萃取研究

BILINGUAL PROPER NOUNS EXTRACTION THROUGH WEB MINING

研 究 生：蘇傳堯　　Student: Chuan-Yao Su

指導教授：梁　婷　　Advisor: Tyne Liang

國立交通大學

資訊科學與工程研究所

碩士論文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao-Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master

in

Computer Science and Engineering

June 2006

Hsinchu, Taiwan, Republic of China

中華民國 九十五 年 六 月

利用網路探勘之中英專名萃取研究

研究生：蘇傳堯　　指導教授：梁婷　博士

國立交通大學資訊科學與工程研究所

摘　　要

專名翻譯的研究可以幫助解決許多自然語言領域的問題，如自動問答系統、機器翻譯、以及跨語言資訊擷取。以往研究著重在利用平衡語料庫或字典來完成，而隨著網路資源的普及，利用網路資源的研究也越來越多。本論文提出了一套整合性的方法，利用網頁資源當作語料庫來完成中英專名翻譯，其中包括搜尋詞擴展和利用事先蒐集好的表面樣式來幫助擷取翻譯候選詞。最後再用我們提出的公式排序翻譯候選詞並得到最後的翻譯結果。在實驗中，我們測試了 1376 筆專有名詞，在英翻中部分，當名次第一的翻譯候選詞即是正確翻譯的機率可達到 87%。在中翻英的部份，當名次第一的翻譯候選詞即是正確翻譯的機率可達到 83%。

BILINGUAL PROPER NOUNS EXTRACTION THROUGH WEB MINING

Student: Chuan-Yao Su Advisor: Dr. Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung University

Abstract

Proper noun translation plays significant role in many natural language applications, such as question answering, machine translation, cross-language information retrieval. Traditional researches of bilingual term extraction focus on utilizing parallel/comparable texts or general dictionaries. Today the Web becomes the largest resource and is utilized in recent researches. This thesis proposes an integrated extraction method to employ query expansion, surface-patterns mined from web corpus, and new ranking scheme to improve bilingual term extraction. Experimental results on 1376 proper nouns show that the presented extraction can achieve 87% accuracy for English-to-Chinese extraction, and 83% for Chinese-to- English extraction.

Acknowledgement

本篇論文能順利完成，首先得感謝我的指導教授梁婷老師。在老師的細心指導下，使學生對資訊擷取與自然語言的相關研究產生濃厚的興趣，且老師也提供了優良的研究環境，使學生可以順利做研究，再次感謝老師的敦敦教誨。

其次我要感謝所有口試委員：張俊盛教授、胡毓志教授、宋定懿教授給我許多的寶貴意見與指正。另外，我還要感謝實驗室的所有學長與同學給予我許多的關愛與協助。

最後，我要感謝我的家人們，他們永遠對我保持著信心，謝謝大家。

Table of Contents

List of Tables

List of Figures

Chapter1 Introduction

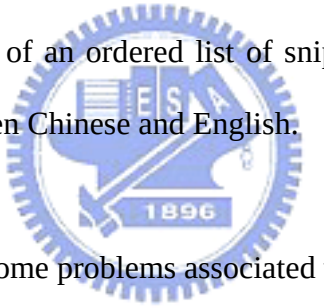
1.1 Background

Proper nouns such as person names, movie titles, company names, medical terms, science terms, and others, are usually referred in our daily life. Most of these proper nouns are out-of-vocabulary (OOV) terms, becoming a bottleneck for some natural language processing applications such as reading/writing assistant, machine translation, question answering, and cross-language information retrieval.

Past term translation researches focus on utilizing parallel/comparable corpus [Wu et al. 1994; Xu et al. 2000; Rapp 1995] or general dictionaries [Gao et al. 2001; Liu et al. 2005]. Parallel corpus contains bilingual sentences, from which translations can be extracted by using appropriate words or sentences alignment methods. Although researches by utilizing parallel corpus can get good translation accuracy, but it is difficult to get sufficient parallel corpora in various domains and languages. On the other hand, comparable corpus consists of documents in one language aligned with documents in another language, where each pair of documents are considered to cover a similar topic. Though, it is easier to collect comparable corpus than to collect parallel corpus, yet the approach using comparable corpus is more difficult to get good performance because of lack of parallel correlation between word pairs. Dictionary-based methods are widely used for their simplicity, but there are multiple translation equivalents in a bilingual dictionary. So how to select appropriate translations is the major task. However both methods encounter some problems like lack of up-to-date data resources and out-of-vocabulary terms problem. Therefore, we propose a Web-based term translation approach to deal with these problems.

Today, the Web is considered as the largest database in the world. Many researches have been developed by exploiting three kinds of web resources, namely parallel webpages [Nie et al. 1999], anchor texts [Lu et al. 2001], and search-result pages [Cheng et al. 2004; Zhang et al. 2004]. However, the approaches based on parallel webpages or anchor texts face the insufficiency of useful corpora [Huang et al. 2005].

On the other hand, real search engine like Google[1] allows us to search terms in one language and get result pages in another language, so we can obtain enough resources in certain language pairs easily. The Web contains huge amounts of data resources in various kinds of subject domains in the world. In this thesis, we exploit search-result-pages consisting of an ordered list of snippets as our corpus to extract proper noun translation between Chinese and English.



However, there are also some problems associated with Web corpus. For instance, the Web contains noise and insufficiency of snippets for some queries. Therefore, we need to exploit strategies such as query expansion or surface patterns to deal these problems.

[1] Google Search Engine: http://www.google.com.tw

Figure 1. Search-result pages from Google by querying "GONE WITH THE WIND"

1.2 Overview of Search-Result-Based Method

Most of the search-result-based methods [Cheng et al. 2004; Zhang et al. 2004; Huang et al. 2005] involve at least three processing phases as follows:

1. Web data collection: collect bilingual search-result pages from search engine by source query.
2. Candidate collection: collect translation candidates from bilingual search-result pages.
3. Translation collection: rank every translation candidates and extract top ranked ones as target translation.

In this thesis, we aim to mine translation of proper noun from these search-result pages. There are three major issues to be concerned:

1. How to crawl the web resources which are more relevant to source query in web data collection phase,
2. How to filter irrelevant information of the web resources, in order to obtain more accurate translation candidates in candidate collection phase,
3. How to rank these translation candidates and get correct translation of source query in translation collection phase.

1.3 Motivation

In order to effectively overcome three major issues and enhance extraction performance, we propose an integrated method to improve each phase in search-result-based method.

1. In the phase of web data connection, we exploit query expansion in order to get more relevant search-result pages.
2. In the phase of candidate collection, we exploit surface patterns proposed by [Wu et al. 2005] to filter miscellaneous information and extract more exact translation candidates.
3. In the phase of translation collection, we utilize statistical information like word length, occurrence frequency, and position between source query and translation candidates in bilingual search-result pages, so as to select appropriate candidates.

The remainder of this thesis is organized as follows. In Chapter 2, we survey the related work. In Chapter 3, we describe baseline method and our proposed method.

Chapter 4 describes the experimental setup, experimental results, and analysis. Chapter 5 is conclusions and future work.

Chapter 2 Related Work

In this chapter, we will briefly describe some researches of automatic term translation. The methods are classified into three categories according to the corpus they used:

1. Parallel/comparable corpus-based method [Nie et al. 1999; Shao et al. 2004; Lee et al. 2005],
2. Bilingual dictionary-based method [Gao et al. 2001; Seo et al. 2005],
3. Web-based method [Lu et al. 2001; Lu et al. 2004; Zhang et al. 2004; Huang et al. 2005; Wu et al. 2005; Fang et al. 2005; Wang et al. 2006]

2.1 Parallel/Comparable Corpus-Based Method

A parallel corpus is a collection of sentence pairs with the same meaning but in different languages. Nie et al. [1999] proposed a method to automatically gather parallel texts from the Web based on anchor texts, hypertexts, webpage names, and HTML structure. They used a probabilistic model to extract translations from parallel texts they gathered. The core of the model is the probability p(t|s), the probability of having a word t in the translation of a sentence containing a word s. However, for language pairs other than English-French in their case, the amount of parallel documents on the Web might not always be enough. Lee et al. [2005] proposed a model for extracting proper names and corresponding translations from parallel corpus. They proposed statistical transliteration model $P(C|E)$ to calculate the probability between English proper name and Romanized transliteration of Chinese terms. The parameters of the model are automatically learned from a bilingual proper name list using the EM algorithm. Experimental results show that the average rates of word and character precision are 93.8% and 97.8%, respectively

A comparable corpus consists of a first-language corpus and a second-language corpus of the same domain. Shao et al. [2004] proposed a method to mine new word translations from comparable corpora, by combining context and transliteration information. They exploit language modeling approach *P(Q|D)* to extract translation on the basis of context information. They experimented six month of Chinese and English Gigaword corpora. They got about 78% precision and about 32% recall.

2.2 Bilingual Dictionary-Based Method

Dictionary-based method is a widely used approach in term translation, because of its simplicity and the increasing availability of readable dictionaries. In this method, the major task is word sense disambiguation, because one query term maybe has multiple translation equivalents in the bilingual dictionary. Gao et al. [2001] used statistical models to overcome this problem. First, they recognized and translated the noun phrases by using statistically models and phrase translation patterns. Second, they selected the best translations based on the cohesion between translation words. The cohesion is term similarity measured by EMMI proposed by [Van Rijsbergen 1979]. However, it is difficult to obtain sufficient amount of word/phrase-aligned parallel corpus so as to extract phrase translation patterns is difficult. Seo et al. [2005] proposed new translation selecting model, they first generated all possible candidate translation queries, and then calculated similarity scores among the terms in each translation candidate query respectively. This method attempts to get target query in which translation equivalents have strong relations with each other. However, proper nouns are not often included in bilingual dictionaries. Thus, it is difficult to handle translation only via dictionaries.

2.3 Web-Based Method

The researches based on Web resources focus on two parts, anchor texts and search-result pages. Lu et al. [2001a, 2001b] extracted translation pairs from anchor texts pointing to the same webpage. They first collected anchor-text-set of a Web page. For a query term, they found its translation $term_a$ if $term_a$ is written in the target language and frequently co-occurs with the source term in the same anchor-text sets. They employed Probabilistic Inference Model to extract translation of query term. They experimented 622 English query terms, and get about 57% accuracy. However, not every pair of languages contains sufficient anchor texts for effective extraction of translations for Web queries. To deal with this problem, Lu et al. [2004] proposed transitive translation model, the translations of a query term can be extracted via its translation in an intermediate language. They further exploit Competitive Linking Algorithm to reduce interference from translation errors. The experiments showed that the approach is particularly useful when the considered language pair lack of sufficient anchor texts.

There are many researches focus on search-result pages. Zhang et al. [2004] extracted translation of query term from search-result pages. First, they detected potential Chinese out-of-vocabulary terms based on Hidden Markov Model and term co-occurrence. First, they submitted Chinese out-of-vocabulary terms to search engine, and get top-100 Chinese snippets. Second, they extracted translation candidates that occurred immediately proceeding/succeeding the Chinese out-of-vocabulary. Final, they ranked translation candidates by their lengths, and frequencies. Wang et al. [2006] proposed a Web-based approach for dealing with the translation of unknown query terms for cross-language information retrieval in digital libraries. The proposed new

association measurement, called SCPCD, combines the symmetric conditional probability [Silva et al. 1999] with the concept of context dependency [Chien 1997] of the n-gram. They use the new formula to extract translation candidates based on the frequencies of its substrings and the number of its unique left and right adjacent words or characters. Finally, they linear combine the Chi-Square Test [Gale et al. 1991] and Context Vector Analysis to rank translation candidates. The experiments showed that they can effectively translate unknown terms.

In order to improve performance of translation, a number of effective techniques have been proposed. Fei Huang et al. [2005] used query expansion phase in order to get more related snippets and used combination of transliteration, translation, and frequency-distance models to rank translation candidates. First, they extracted expansion candidates from returned snippets by querying source query terms. They prepared a dictionary to translate expansion candidates and used rules to filter out some irrelevant terms. Finally, they extracted top frequency terms as expansion terms. In experiments, they achieve 80% accuracy with 165 snippets. Fang et al. [2005] used character-based string frequency estimation to gather translation candidates. They defined two kinds of candidate noises: subset redundancy information and prefix/suffix redundancy information. The subset redundancy information is that the $term_a$ is a subset of another $term_b$, but the rank of $term_a$ is lower than $term_b$. The prefix/suffix redundancy information is $term_a$ is the prefix or suffix of $term_b$, but rank of $term_a$ is greater than $term_b$. They proposed sort-based subset deletion and mutual information methods to deal with these two noise information respectively. After removing candidate noise, we can rank remain candidates and get better results. They experimented 401 English terms, and get about 72% accuracy.

Additionally, Wu et al. [2005] proposed a *TermMine* system. In this system, they used surface patterns which are learned by a list of bilingual terms to extract translation candidates more exact. Surface pattern means the co-occurring format between source query and its translation. For example, we submit "*Picasso*" and "*畢卡索*" to search engine, and we get some texts as follow:

"…*Picasso* (*畢卡索*)…" and "…*畢卡索* Pablo *Picasso*…".

We can extract surface patterns "*E*(*C*" and "*CwE*", in which *E* is source English word from bilingual list, *C* is translation of *E*, *w* is any other English word, and others are punctuations.

They are first submitted bilingual pairs to search engine and extracted surface patterns from the search-result pages. Translation candidates are extracted if they matched the surface patterns. Finally, they rank these translation candidates based on frequencies or probability calculated by transliteration model. They experimented 300 English terms, and get 86 % accuracy.

Chapter 3 Extract Translation from Web Snippets

In this chapter, we will describe how to extract translation candidates for an unknown source term from a set of snippets by the proposed formula. The extraction is aimed for both of “English to Chinese” and “Chinese to English”. The following description, we focus on “English to Chinese” direction.

3.1 Baseline Method

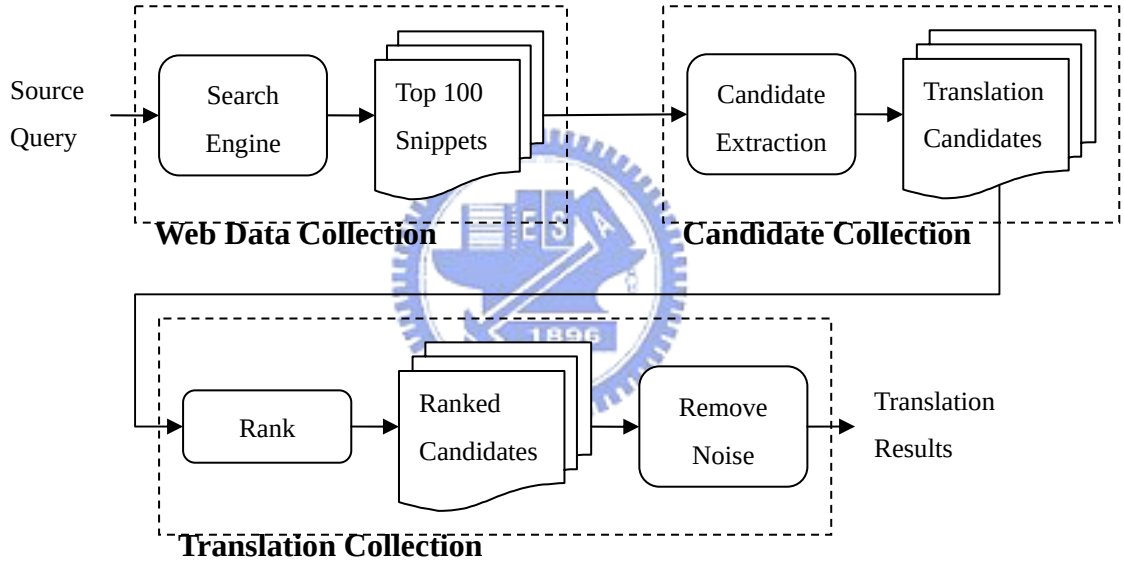


Figure 2. The Flow Chart of Baseline Method

Figure 2 shows flow chart of the baseline method. The method has three phases: Web data collection, candidate collection and translation collection. First, we submit source query to Google search engine and collect top 100 snippets, from which we extract and collect possible translation candidates. Finally, we get translation results by a ranking formula

3.1.1 Search Engine Module

First, we crawl web pages that contain English source query and Chinese language characters. We describe the procedure as follows:

1. We submit source English query terms to Google[2] search engine, and then we collect top 100 snippets of Chinese documents.
2. We remove HTML tag of snippets and leave raw text only.

3.1.2 Candidate Extraction Module

In our observation, when English unknown words appear in Chinese text, their translations probably appear nearby. We collect co-occurring of Chinese characters and English source query within a predefined window size, and extract translation candidates as follows:

1. Scan raw text for English source query and collect Chinese characters which appear immediately proceeding/succeeding of English source query within window size as translation candidate string (TCS). We define window size as 15 characters based on analyzing the length of answer translation in the answer set. And we define punctuation and English word as one character size. For example, we query "**Atomic Physics**" and get one snippet as follow:

   "美國田納西大學物理系博士後副研究員. 研究領域：, 非線性光學 (Nonlinear optics). 雷射物理 (Laser Physics). 雷射光譜學 (Laser Spectroscopy). 原子物理(**Atomic Physics**). 個人興趣：, 遊山玩水、美食、球類運動、科學與真理的探索",

   we can get five translation candidate strings,

[2] Google Search Engine: http://www.google.com.tw

i. Proceeding: { “原子物理”, “雷射光譜學” }

ii. Succeeding: { “個人興趣”, “遊山玩水”, “美食” }

2. We generate all substring of each translation candidate strings with length greater than 1 as translation candidates. From above example, we can generate { “原子”, “子物”, “物理” , “原子物” , “子物理” , “原子物理” } as translation candidates from translation candidate string { “原子物理” }.

3.1.3 The Rank Module

The rank module is to rank every translation candidate by the following equation, that

$$r(x) = freq(x) \times \log(length(x)) \quad (1)$$

where *freq(x)* is the frequency of *x* and *length(x)* is the string length of *x*. We treat those candidates with the highest value to be the translation result.

3.1.4 Noise Removing Module

In the candidate collection step we generate all substring of translation candidate string, so we may get many redundancy noises in ranked result list. In the noise removing module, we prefer those words with longer string size since they contain more information. We remove the lower rank translation result items if they are the substring of the higher translation result items. For example, English source query is “Ford Motor” and we get ranked result list {“福特汽車公司”, “福特”, “福特汽車”, “美國福特汽車公司”}. We show that “福特” and “福特汽車” are substrings of “福特汽車公司” and their ranks are lower than “福特汽車公司”, so we remove “福特” and “福特汽車”.

3.2 The Chinese-Translation Extraction

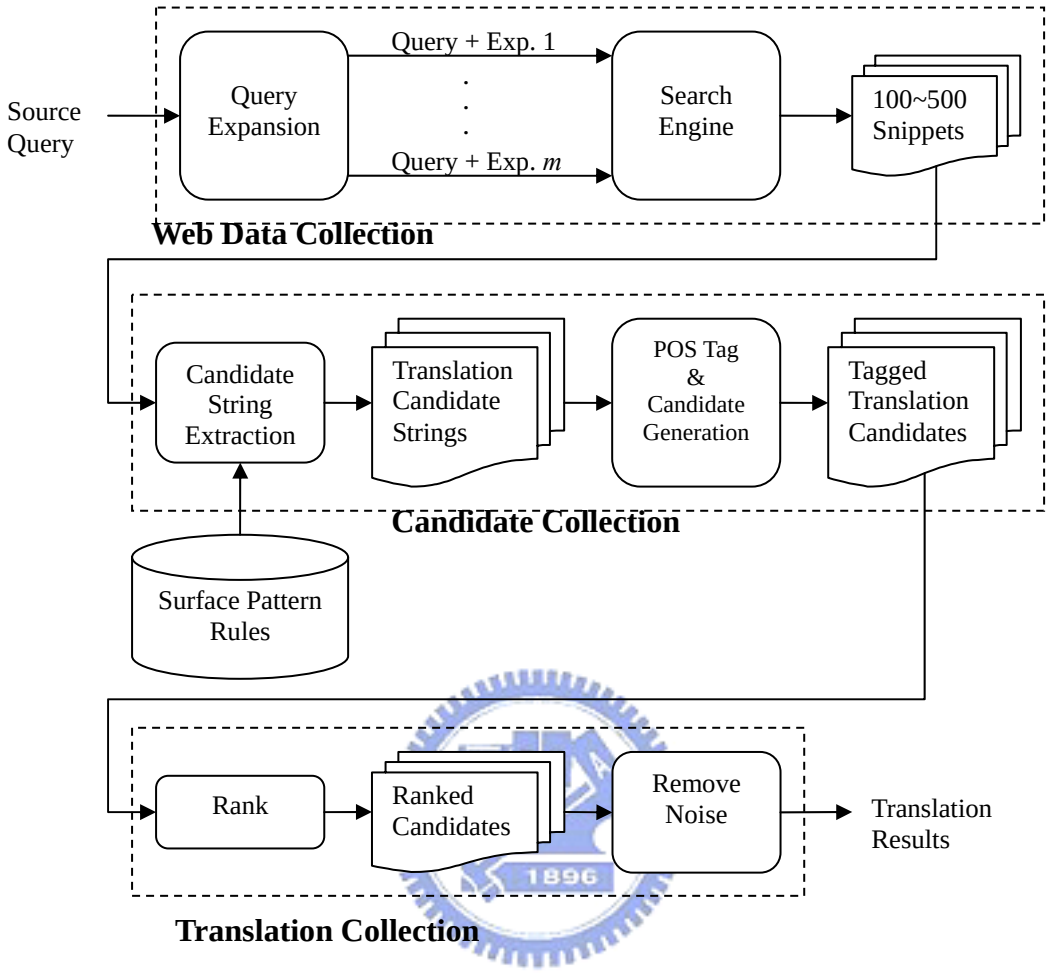


Figure 3. The Flow Chart of Our Method

Figure 3 shows flow chart of our proposed method, which contains three phrases namely, Web data collection, candidate collection and translation collection. At web data collection, a query expansion module is used to get more related snippets. At candidate collection, first, the candidate extraction module is implemented to extract translation candidate strings (TCSs) on the basis of surface pattern rules. Second, the POS module is used to identify translation candidates with more exact segment boundary. At translation collection, a rank module is presented by considering the minimum distance between English source query and translation candidate.

3.2.1 Query Expansion Module

In this module, we assume that there is a Chinese term that is relevant to and possibly co-occurs with English source query. For example, we submit English source query “All Saints” and its translation is “聖女合唱團”. Then there are only two answers in the top five snippets as shown in Figure 4. However, if we add a Chinese term “音樂” to English source query, we can get more answers from return snippets as shown in Figure 5. That is because “音樂” is relevant to source query.

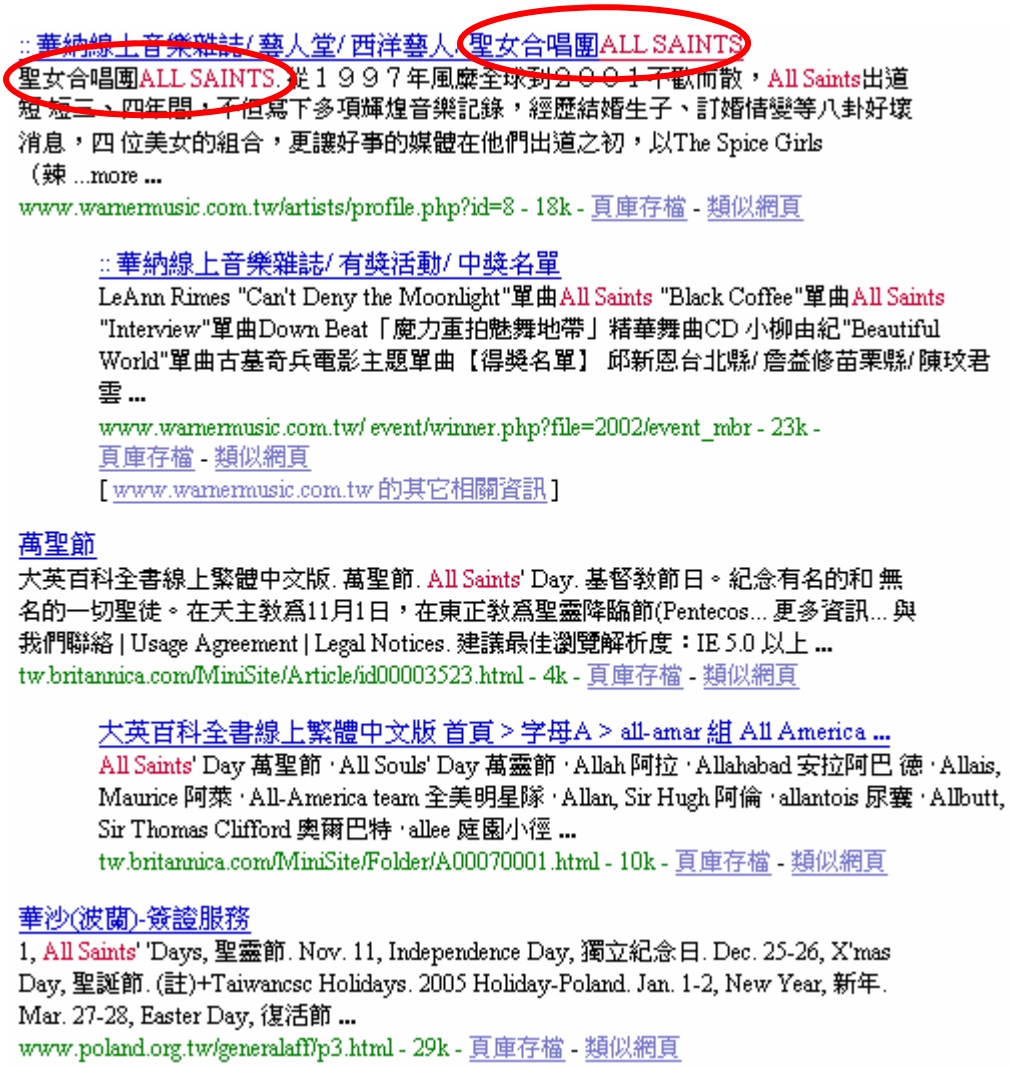


Figure 4. Returned top five snippets by submitting “All Saints”

Figure 5. Returned top five snippets by submitting “All Saints”+“音樂”

Because most proper nouns are out-of-vocabulary terms, it is unlikely to obtain much information from the existing corpus. We proposed a web-based method and exploit statistical model to extract expansion terms from returned snippets of source English query terms. The query expansion is implemented as follows:

1. We submit source English query terms to Google search engine and collect top 100 snippets of Chinese documents.
2. After removing HTML tag, we do part of speech[3] (POS) tagging for raw text.

[3] CKIP Chinese Segment Tagger: http://ckipsvr.iis.sinica.edu.tw/

3. We extract Chinese words with “Na”, “Nb” or “Nc” tags as expansion candidates. We use POS tag defined by CKIP, in which “Na” is the generic noun, “Nb” is the proper noun, and “Nc” is the toponym,
4. For each expansion candidates we compute its association score $s_{u,v}$ with respect to source query on the basis of association clusters proposed in [R. Baeza-Yates et al. 1999]. We describe at below.
5. Finally, we return the top *m* Chinese word as expansions terms.

The association score $s_{u,v}$ is computed by Equation (2),

$$s_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}} \tag{2}$$

where *u* is English source query, *v* is expansion candidate, and $c_{u,v}$ is computed as follows,

$$c_{u,v} = \sum_{d_j \in D_l} f_{s_u,j} \times f_{s_v,j} \tag{3}$$

where d_j is the j_{th} snippets, D_l is all snippets, $f_{s_u,j}$ is the frequency of English source query in the j_{th} snippet, $f_{s_v,j}$ is the frequency of expansion candidate in the j_{th} snippet.

We add each expansion term to each query. Then, we get top 100 snippets for each expanded query from Google. For example, source query is “Clinton” and expansion terms are “美國”, “總統”, and “山屋”. We expand source query to “Clinton+美國”, “Clinton+總統”, and “Clinton+山屋”.

3.2.2 Search Engine Module

The function of this module is the same as function of the baseline method. We submit expanded query command to Google search engine and get top 100 snippets respectively. We have *m* expansion terms, so we can get *m**100 snippets ideally.

3.2.3 Candidate Extraction by Surface Pattern

To filter miscellaneous information and extract more exact translation candidates, we exploit surface patterns proposed by [Wu et al. 2005] to help us extract translation candidates. We describe procedure as follows:

1. Scan raw text for English source query and collect Chinese characters matched surface patterns as translation candidate string (TCS). For example, we query "**Atomic Physics**" and get one snippet as follow:

   "美國田納西大學物理系博士後副研究員. 研究領域：, 非線性光學 (Nonlinear optics). 雷射物理 (Laser Physics). 雷射光譜學 (Laser Spectroscopy). 原子物理(**Atomic Physics**). 個人興趣：, 遊山玩水、美食、球類運動、科學與真理的探索…",

   suppose we had surface pattern C(E, but did not have surface pattern E).C. We get one translation candidate string "原子物理" matched by surface pattern "C(E".
2. Segment all translation candidate strings.
3. Generate all substring of each tagged translation candidate strings with length greater than 1 as translation candidates. From above example, we get translate candidate string as "原子(Na)　物理(Na)", and we generate { "原子", "物理", "原子物理" } as translation candidates.

3.2.4 The Proposed Rank Module

Based on the statistical data about distribution of distances between source terms and target terms in web pages proposed in [Wu et al. 2005], we know that if the distance is shorter, then the co-occurrence frequency is higher.

We proposed a new formula on the basis of occurrence frequency, word length, and distribution of distance. For every instance of translation candidate, we must record its minimum distance to source queries. We describe the procedure as follows:

1. Scan for the instance of translation candidate
2. Count number of token that occurred immediately preceding/succeeding the instance until meet the source query. We define Chinese character, English word, and punctuation as one token size.
3. Select minimum number as distance

For example, Figure 6 shows the web texts we retrieve when submit source query E to search engine, and C is instance of one translation candidate. Table 1 is the information of frequency with all distances.

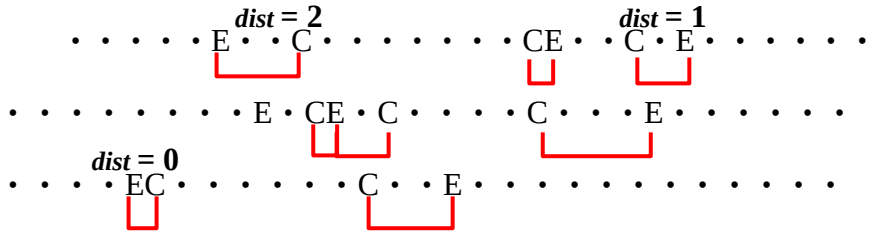


Figure 6. Example of Web Text

Table 1. Frequency of translation candidate with the j_{th} distance

Distance	Frequency
0	3
1	2
2	2
3	1

The proposed formula is shown following:

$$r'(x_i) = \sum_{j=1}^{k} \lambda^{dist_j} \times freq(x_{i,j}) \times \log(length(x_i)) \quad (4)$$

where x_i is the i_{th} translation candidate, $x_{i,j}$ is instance of with the j_{th} distance, λ is penalty weight, we define the value of λ from 0.5 to 0.9, $dist_j$ is the j_{th} distance, and k means kinds of distances. For above example, the $r'(E) = (\lambda^0 \times 3 + \lambda^1 \times 2 + \lambda^2 \times 2 + \lambda^3 \times 1) \times \log(length(C))$.

Chapter 4 Experiments and Analysis

4.1 Experimental Setup

We collected 1376 English-Chinese term pairs from 7 domains as test set, including 269 person names, 140 school names, 161 movie titles, 129 company names, 257 location names, 156 medical terms, 264 science and technology terms. Table 2 shows some English-Chinese term pairs from 7 domains.

Table 2. Example of term pairs from 7 domains

Domain	English Term	Chinese Term
Person name	Galileo	伽利略
	Vincent van Gogh	梵谷
School name	Harvard University	哈佛大學
	Carnegie Mellon University	卡內基美隆大學
Movie title	The Sound Of Music	真善美
	The Godfather	教父
Company name	General Motors	通用汽車
	Starbucks	星巴克
Location name	California	加利福尼亞
	Chicago	芝加哥
Medical term	Mediterranean anemia	地中海型貧血
	Down's syndrome	唐氏症
Sci & Tech term	Fibonacci Number	費氏數
	Kinetic Theory of Gases	氣體動力論

4.1.2 The Extracted Surface Patters

We randomly selected 750 English-Chinese term pairs from the Encyclopedia Britannica Online[4] as training data. We got 184 surface patterns from 46537 instances which . Table 3 shows the top-13 frequent surface patterns, and they cover 90.70% of all instances. In the presented candidate extraction module, we employ these 13 surface patterns, in which *E* is source English word from bilingual list, *C* is translation of *E*, *w* is any other English word, and others are punctuations.

Table 3. Top 13 Frequent Surface Patterns

Surface Pattern	Frequency	Acc. Rate
CE	17135	36.82%
C(E	6804	51.44%
CwE	5667	63.62%
EC	2798	69.63%
CwwE	2166	74.28%
E(C	1345	77.18%
Cw(E	1131	79.61%
C.E	1063	81.89%
EwC	1024	84.09%
C,E	983	86.20%
E,C	806	87.93%
C/E	751	89.55%
E.C	537	90.70%

4.1.3 Experimental Comparison Setup

The translation extraction is implemented with query expansion module, pos module, candidate extraction with surface pattern, and the proposed formula. The extraction performance is verified by checking each component respectively. We define notation as follow:

[4] Encyclopedia Britannica Online: http://tw.britannica.com/MiniSite/B00000000.html

(1) qe: Query expansion module.

(2) seg: Segmentation of translation candidate dtring.

(3) pat: Candidate extraction with surface patterns.

(4) dist: Distance-based rank formula.

(5) Base: Baseline Method.

4.1.4 Performance Metric

We utilized the *average top-n inclusion rate* as a metric on the extraction of translation equivalents. We defined *average top-n inclusion rate* as the percentage of terms whose translations could be found in the first *n* extracted translations.

If the extracted translation is substring of the correct answer, we judged it is correct. For example, the translation of "Puff Daddy" is "吹牛老爹". We judge "美國歌手吹牛老爹" is also the correct answer, but "牛老爹" is not the correct one.

4.2 Experiments and Analysis of English-to-Chinese Translation Extraction

The overall translation accuracies are shown in Table 4. We define the parameter *m* is 5, and parameter λ is 0.9. We can show that if we utilized more modules, we can get better efficiency. When we consider all modules, we can enhance about 17% accuracy than Baseline method with Top-1 inclusion rate.

Table 4. Top-5 Inclusion Rates of All Models for English-to-Chinese Extraction

	Top1		Top2		Top3		Top4		Top5	
SEG+/-	Seg+	Seg-	Seg+	Seg-	Seg+	Seg-	Seg+	Seg-	Seg+	Seg-
Base	74.1%	70.5%	87.5%	85.7%	92.0%	90.3%	93.4%	92.7%	94.4%	94.2%
Base + qe	75.4%	75.2%	88.1%	87.4%	92.0%	91.1%	93.4%	93.4%	94.5%	94.3%
Base + dist	80.9%	77.4%	91.0%	88.6%	93.8%	92.6%	95.1%	94.1%	95.9%	95.3%
Base + pat	82.2%	75.9%	90.6%	87.7%	93.8%	92.0%	94.8%	93.3%	95.5%	94.3%
Base + qe + dist	82.5%	81.0%	91.7%	90.0%	93.8%	93.5%	94.9%	94.6%	95.6%	95.2%
Base + pat + dist	84.2%	78.4%	92.2%	89.2%	94.2%	92.7%	95.1%	94.0%	95.6%	94.8%
Base + pat + qe	85.5%	79.2%	92.3%	89.8%	94.2%	93.1%	95.0%	94.4%	95.5%	95.2%
Base +qe + pat + dis	87.2%	82.3%	93.1%	91.3%	94.8%	94.0%	95.4%	94.8%	96.1%	95.3%

We have shown that when we add all modules (Base + qe + pat + dist + seg) to baseline method, we will get the best results. Table 5 shows translation accuracy of each domain separately.

1. In “Company name” domain, we often get “公司” as translation leads to degression of performance.
2. In “Sci & Tech term” and “Medical term” domains, some queries have quite few or zero number of snippets returned from search engine, so we have not enough bilingual information. For example, we can’t get any Chinese terms from snippets when we query “theory of repression”.
3. Some source queries have many translations in snippets lead to translation results are substring of correct answers or incorrect. For example, “Imperial College London” have two translations “倫敦帝國學院” and “倫敦大學帝國學院” result in “帝國學院” be the best result because of it has higher frequency. In “Person name” domain, the translation of “Jewel” is "珠兒”, but “Jewel” also has another translation “寶石” that cause us to failure.

4. Some miscellaneous information is related to source query and match surface patterns. For example of “Intel”, we get final results are “處理器”, “晶片組”,and “電腦”, because these terms occur frequently and are related to source query.

Table 5. Top-5 Inclusion Rate of Each Domain for English-to-Chinese

	Number	Ave. Len. (words)	Top1	Top2	Top3	Top4	Top5
Company name	129	1.88	76.70%	87.60%	89.90%	92.20%	93.80%
Sci & Tech term	264	1.64	80.70%	88.60%	91.70%	93.20%	95.10%
Medical term	156	1.33	85.30%	91.00%	92.90%	92.90%	92.90%
School name	140	3.32	86.40%	97.10%	98.60%	99.30%	99.30%
Person name	269	1.80	90.00%	92.60%	94.40%	94.40%	94.80%
Location name	257	1.18	92.60%	97.70%	98.10%	98.10%	98.10%
Movie title	161	2.64	96.30%	97.50%	98.80%	98.80%	99.40%

Finally, we want to define the parameters m and λ in the model “Base + qe + pat + dist + seg” which has the best performance. m is the number of expansion terms we used, and λ is the distance penalty weight. We experimented 250 English terms chosen randomly from test set. Figure 7 shows the average top-1 inclusion rate when we consider different value of m and λ. It is clear that when we exploit more expansion terms, we can get better performance. We also find that when we define distance penalty weight λ as 0.5 or 0.6, we can get the best performance.

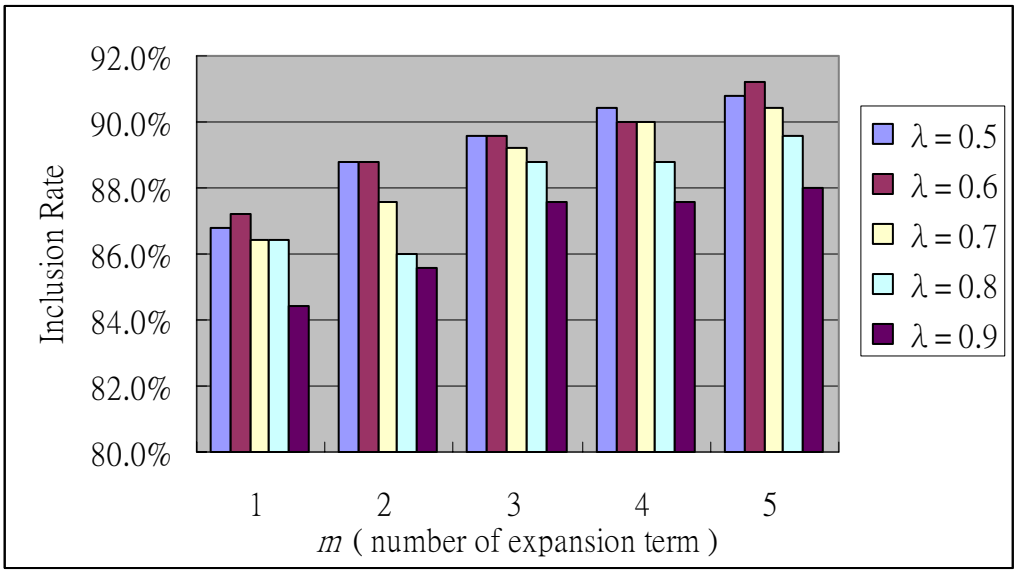


Figure 7. Average Top-1 Inclusion Rate Based on *m* and λ

4.3 Experiments and Analysis of Chinese-to-English Translation Extraction

The overall translation accuracies are shown in Table 6. We define the parameter *m* is 5, and parameter λ is 0.7. The results also show that if we utilized more modules, we can get better efficiency. When we consider all modules, we can enhance 10.6% accuracy than Baseline method with Top-1 inclusion rate.

Table 6. Top-5 Inclusion Rates of All Models for Chinese-to-English Extraction

	Top1	Top2	Top3	Top4	Top5
Base	72.5%	85.8%	91.3%	93.0%	93.8%
Base + qe	72.0%	85.2%	90.3%	92.1%	92.5%
Base + dist	76.0%	87.8%	91.5%	93.9%	94.8%
Base + pat	76.7%	88.1%	91.6%	93.0%	93.2%
Base + qe + dist	81.0%	89.3%	92.5%	94.8%	95.8%
Base + pat + dist	79.6%	89.1%	92.0%	93.0%	93.5%
Base + pat + qe	81.3%	89.4%	92.8%	95.3%	96.1%
Base +qe + pat + dist	83.1%	91.6%	94.4%	95.7%	96.7%

Table 7 shows model "Base + qe + pat + dist + seg" translation accuracy of each domain separately.

1. In domain of "Company Name", we will not encounter the problems like English-to-Chinese, because "company" is not often the contained in company name.
2. In domain of "School Name", we often get "University" as translation leads to degression of performance.
3. In domain of "Medical Term", some source queries are too short or too general, so the translations we get are not correct, but they relate to source queries. For example, the source query is "氣管炎", and its translation is "tracheitis", but we get the translation is "infectious laryngotracheitis (傳染性喉頭氣管炎)".
4. Others are described in section 4.2.

Table 7. Top-5 Inclusion Rate of Each Domain for Chinese-to-English Extraction

	Number	Ave. Len. (characters)	Top1	Top2	Top3	Top4	Top5
School name	140	5.66	65.70%	82.10%	89.30%	92.10%	93.60%
Medical term	156	3.17	68.60%	87.80%	91.00%	92.90%	93.60%
Sci & Tech term	264	3.83	76.50%	85.60%	90.50%	91.70%	93.90%
Movie title	161	4.16	85.10%	92.50%	94.40%	96.30%	97.50%
Location name	257	3.16	89.90%	96.50%	98.80%	99.60%	99.60%
Company name	129	4.99	93.00%	96.90%	96.90%	96.90%	98.40%
Person name	269	4.40	94.80%	97.00%	97.40%	98.50%	98.50%

Finally, we define the parameters m and λ in the model "Base + qe + pat + dist" which has the best performance. We experimented 250 Chinese terms chosen randomly from test set. Figure 8 shows the average top-1 inclusion rate when we consider different value of m and λ. In most case, we find that if we use more

expansion terms, we can get better performance. We can define distance penalty weight λ from 0.6 to 0.8. Table 8 is a summary of web-based approaches.

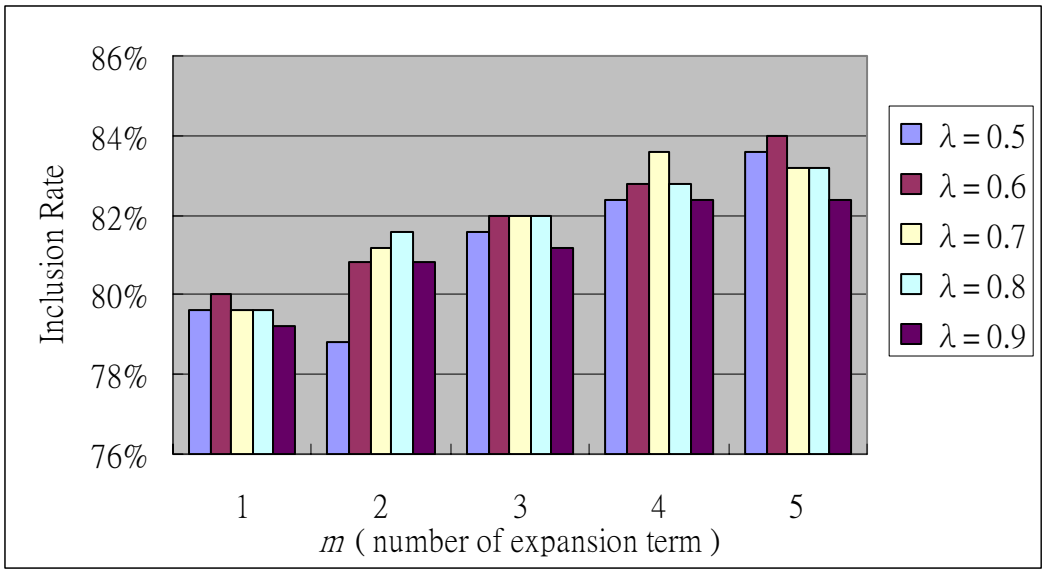


Figure 8. Average Top-1 Inclusion Rate Based on m and λ

Table 8. A Summary of Web-Based Approaches

	Cheng et al. 2004	**Fang et al. 2005**	**Wu et al. 2005**	**Huang et al. 2005**	**Our Method**
Resource	➢ Search result page ➢ Anchor text	➢ Search result page	➢ Search result page	➢ Search result page	➢ Search result page
Method	➢ Anchor text ➢ Chi-square ➢ Context-Vector	➢ Distribution forms ➢ Noise deletion	➢ Surface Pattern ➢ Transliteration ➢ Expanding Tentative Translation	➢ Query expansion ➢ Transliteration	➢ Query expansion ➢ Surface Pattern ➢ Noise deletion ➢ Distance distribution
Test Data	➢ 50	➢ 401	➢ 300	➢ 310	➢ 1376
Performance	➢ Eng-to-Ch: 61.2%	➢ Eng-to-Ch: 71.8%	➢ Eng-to-Ch: 86%	➢ Ch-to-Eng: 80%	➢ Eng-to-Ch: 87.2% ➢ Ch-to-Eng: 83.1%

Chapter 5 Conclusion and Future Work

5.1 Conclusion

In this thesis, we describe a web-based approach to deal with proper noun translation by mining from search-result pages. First, we add expansion terms to source query, and then retrieval snippets by expanded query. Second, translation candidate strings are extracted if they matched the surface patterns, and then we generate translation candidates from translation candidate strings. Finally, the proposed formula is used to rank translation candidates. From the experiments, our approach has a good performance for finding translations of proper nouns through Web resources. We summarize the contributions as follows:

1. We integrate some improved ways to enhance efficiency of proper noun translation. From the experiments show that our proposed method has good work.
2. We proposed a statically web-based query expansion method. Most of proper nouns are out-of-vocabulary terms, so we proposed web-based method can overcome the lack of resources to generate expansion terms.
3. We proposed a new formula on the basis of word length, word frequency, and distance distribution.

5.2 Future Work

Future work we will focus on sub-query translation. This approach may deals with the error cause from few numbers of returned snippets or not enough bilingual information, especially for long queries. Form example, we submit source query “李登輝學校” to search engine, but we can not get any information of its translation “Lee Teng-Hui Academy” from returned snippets. Therefore, we will segment the

source query into “李登輝” and “學校” and translate them respectively. Finally, we can exploit word sense disambiguation technique and some composition rules to merge them and get the final answer.

References

Cheng, P. J., Teng, J. W., Chen, R. C., Wang J. H., and Lu, W. H., “Translating unknown queries with web corpora for cross-language information retrieval.” *In the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 146-153, Sheffield, United Kingdom, 2004.

Fang, G., Yu, H., and Nishino, F., “Web-Based Terminology Translation Mining.” *In the International Joint Conference on Natural Language Processing*, pp. 1004-1016, Jeju Island, Korea, 2005.

Gao, J., Nie, J. Y., Xun, E., Zhang, J., Zhou, M., and Huang, C., “Improving query translation for cross-language information retrieval using statistical models.“ *In the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96-104, New Orleans, Louisiana, United States, 2001.

Huang, F., Zhang, Y. and Vogel, S., “Mining Key Phrase Translations from Web Corpora.” *In the Proceedings of the Human Language Technologies Conference (HLT-EMNLP 2005)*, pp. 483–490, Vancouver, BC, Canada, October 2005.

Kishida, K., Kando, N., and Chen, K. H., “Two-Stage Refinement of Transitive Query Translation with English Disambiguation for Cross-Language Information Retrieval: An Experiment at CLEF 2004”. *In the CLEF 2004*, pp. 135-142, Bath, UK, 2004.

Lam, W., Chan, K., Radev, D., Saggion, H., and Teufel, S., “Context-based generic cross-lingual retrieval of documents and automated summaries: Research Articles.” *In the Journal of the American Society for Information Science and Technology*, pp. 129-139, January 2005.

Lee, C. J., Chang, J. S., and Jyh-Shing Roger Jang, "Extraction of Transliteration Pairs from Parallel Corpora Using a Statistical Transliteration Model." *In the Information Sciences*, 2005.

Li, S. and Ng, H. T., “Mining New Word Translations from Comparable Corpora.” *In the Proceedings of the 20th International Conference on Computational Linguistics*, pp. 618-624, University of Geneva, Geneva, Switzerland, 2004.

Lin, F. and Mitamura, T., “Keyword Translation from English to Chinese for Multilingual QA.” *In the AMTA 2004*, pp. 164-176, 2004.

Liu, Y., Jin, R., and Chai, J. Y., “A maximum coherence model for dictionary-based cross-language information retrieval.” *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 536-543, Salvador, Brazil, 2005.

Lu, W. H., Chien, L. F., Lee, H. J., "Anchor Text Mining for Translation of Web Queries." *In the Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, pp. 401-408, 2001.

Lu, W. H., Lee, H. J., Chien, L. F., “Anchor Text Mining for Translation Extraction of Query Terms.” *In the SIGIR*, pp. 388-389, 2001.

Nie, Yun, J., Simard, M., Isabelle, P., Durand, R., “Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web.” *In the proceedings of ACM-SIGIR*, pp. 74—81, 1999

R. Baeza-Yates and B. Ribeiro-Neto, “Modern Information Retrieval.” *Addison-Wesley & ACM Press*, Harlow, UK, 1999.

Rapp, R., “Identifying word translations in non-parallel texts.” *In the Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 320-322, Cambridge, Massachusetts. 1995.

Seo, H. C., Kim, S. B., Rim, H. C., Myaeng S. H., “Improving query translation in English-Korean cross-language information retrieval.” *In the Information Processing and Management: an International Journal*, pp. 507-522, 2005.

Van Rijsbergen, D. J, “Information retrieval, 2nd. ed.Butterworths.” London, 1979.

Wang, J. H., Teng, J. W., Lu, W. H., and Chien, L. F., “Exploiting the Web as the multilingual corpus for unknown query translation.” *In the Journal of the American Society for Information Science and Technology*, pp. 660-670, 2006.

Wu, D. and Xia, X., "Learning an English-Chinese lexicon from a parallel corpus". *In the AMTA-94: Assoc. for Machine Translation*, pp. 206-213. Columbia, MD, October 1994.

Wu, J. C., Lin, T., and Chang, J. S., "Learning Source-Target Surface Patterns for Web-Based Terminology Translation." *In proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 37-40, Ann Arbor, June 2005.

Zhang, Y. and Vines, P., “Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval.” *In Proceedings of 27th ACM SIGIR*, pp.162-169, Sheffield, United Kingdom, 2004.

Zhang, Y., Huang, F., and Vogel, S., “Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion.” *In the Proceedings of the 28th ACM SIGIR*, Salvador, Brazil, August 2005.