NTCIR-6 Experiments using Pattern Matched Translation Extraction

Dong Zhou
School of Computer Science and IT
University of Nottingham
Nottingham, NG8 1BB, UK
dxz@cs.nott.ac.uk

Mark Truran
School of Computing
Univeristy of Teesside
Teesside, TS1, 3BA, UK
M.A.Truran@tees.ac.uk

Tim Brailsford
School of Computer Science and IT
University of Nottingham
Nottingham, NG8 1BB, UK
Tim.brailsford@nottingham.ac.uk

Helen Ashman
School of Computer Science and IT
University of Nottingham
Nottingham, NG8 1BB, UK
hla@cs.nott.ac.uk

Abstract

This paper describes our experiment methods and results in the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies. We introduce a Pattern Matched Translation Extraction (PMTE) approach to the analysis of mixed-languages web pages, which makes use of pattern matching to automatically extract the translation pairs. The experiment results demonstrated the proposed method is effective when translating Out-of-Vocabulary (OOV) terms, a well-known problem in fields of cross-language information retrieval (CLIR), question-answering (QA), machine translation (MT) and knowledge discovery (KD). We also report the experiment results of single-language information retrieval (SLIR) and illustrate the performance through different collections in STAGE 2 of NTCIR-6.

Keywords: *CLIR, SLIR, Experiment, PMTE, OOV terms.*

1 Introduction

There are two ways to translate terms [1] from one language to another. One uses *parallel corpora* [1, 2, 8] and the other employs a *dictionary based translation* [6, 9]. Both approaches face the *coverage problem*. Coverage refers to the linguistic limitations of parallel corpora and dictionaries. Certain types of words are not commonly found in this type of resource and are likely to cause translation problems. These include compound words; proper names, such as the identifiers attached to persons or organizations; technical terms, including newly invented words and phrases from specialist disciplines such as science. Approaches using parallel corpora cannot really address this problem for a number of reasons, including: the insufficiency of parallel corpora in enough different languages; the fact that most corpora belonging to a specific domain, and the obvious flaw that most of translation probabilities induced from parallel corpora are typically based on single-word mappings. Neither can a dictionary based approach realistically remedy this situation. Dictionaries are in most cases too general to translate specialized vocabulary.

Recent attempts to solve the unknown terms problem have concentrated on purely statistical methods [7, 3, 11, 12]. In this paper we introduce web-based Pattern Matched Translation Extraction (PMTE) method which generates translation candidates for unknown terms using linguistic and symbolic features of the extracted text. Our method is based on an observation relating to the frequency of mixed language pages on the World Wide Web. Using a web search engine we can target web pages using a mix of human languages and analyse the content of the page according to the specific features of the relevant language. We then generate translation candidates using a theoretically simple algorithm. This method merges relevant findings from the fields of CLIR, natural language processing, MT and

[1] Here *terms* refer to single word or phrases expressing a meaningful chunk of information

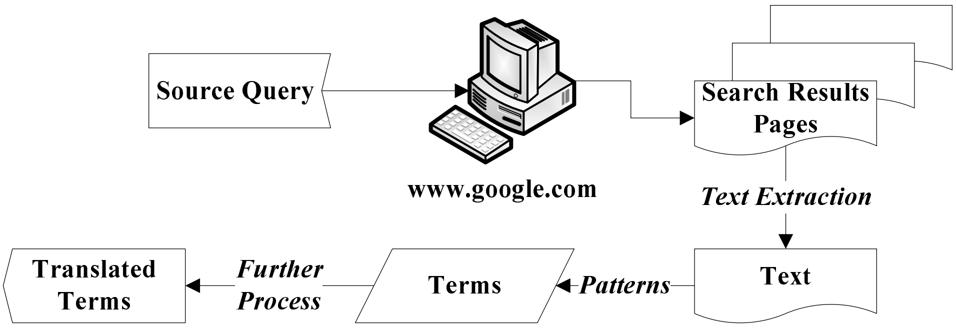

Figure 1. PMTE Translation System Overview

knowledge discovery areas.

The paper is organized as follows: section 2 describes the proposed PMTE method and algorithm, section 3 reports on SLIR(Chinese-Chinese), CLIR(English-Chinese) and baseline experiments performed and results obtained using NTICR-5 data collection; section 4 reports results for NTCIR-6 and cross-collection experiments, section 5 concludes and speculates on further work.

2 PMTE Approach

Our research exploits the *information redundancy* of the World Wide Web, that is, mixed-language information which is contained on many websites and pages. The main problem is how to efficiently extract the translation terms. The translation extraction process described here is inspired by several studies that used statistical techniques to extract translation terms [3, 7, 10, 11, 12]. However, the PMTE technique relies heavily on pattern matching driven by linguistic knowledge, as follows.

First the unknown term is submitted to several search engines to obtain a large number of result pages. Thereafter, the text content of the result pages is parsed for analysis purposes, and several different patterns are used to extract the candidate terms. Finally a frequency based term-weighting techniques is applied to extract the final translation term. This system is illustrated in Figure 1.

2.1 Querying the Web

We submit the unknown term to well-known search engines such as Google and cache the first 500 results. The search option is set to retrieve documents written in languages different from the unknown term. For example, if we want to translate an unknown English term into Chinese, the search option will be set to retrieve Chinese web pages. This can be done through the advanced search options provided by most search engines. The work done by several other researchers only used the first 100 documents [3, 11, 12]. We fetch the first 500 results to increase the accuracy of our results. Titles and summaries are removed from the results pages. This is done by removing the HTML tags and other non-content symbols.

2.2 Candidate Terms Extraction

We now have the raw text of the results page and we need to extract possible translation terms. In a purely statistical fashion this could be done by calculating the co-occurrence frequency of the unknown term with accompanying foreign language terms, but this is prohibitive expensive and error prone. Instead we apply our own technique, which relies on symbolic and linguistic patterns. Figure 2 is all the patterns (written in simple format expressions) used in our implementation followed by illustrative sentence fragments (the bold font type means the matched patterns), which will be introduced afterwards.

There are two different types of symbolic patterns. One is symmetric (see pattern 1 and 2 in figure 2). Some authors tend to include the translation of unknown terms in brackets. This type of pattern includes all the different types of brackets, quotation marks, and other forms of parenthetic symbols.

Our next pattern is the punctuation pattern (see pattern 3 in figure 2). Aside from the symmetric symbols introduced above, we tend to find the extraction boundaries using punctuation symbols. Punctuation symbols usually mark the boundaries between the concepts. The closer the position between the text and the unknown term, the greater the chance that the former will provide a translation of the latter.

Patterns With Examples	***Term Definition***
1. ***SOURCE {SP} NP * {SP}*** ➢ …starbucks (星巴克)… **or** …Nanotechnology [奈米科技]…	● **SOURCE**–*Source term, we assume source term is in English* ● **NP**–*Noun phrases in other languages (e.g. Chinese)* ● **SP**–*Symmetric patterns which must be appear in correspondent pairs* ● **PP**–*Punctuation patterns* ● **ET**–*Eliminator terms (words/phrases)* ● **DT**–*Discriminator phrases* ● *****–*Any noun phrases*
2. ***{PP}{ET}{PP}NP {SP} SOURCE {SP}*** ➢ …, …星巴克 (starbucks) **or** …, …奈米科技 [Nanotechnology]…	
3. ***{PP} NP SOURCE NP {PP}*** ➢ …, 星巴克 starbucks **or** Nanotechnology奈米科技, …	
4. ***{PP}{ET} NP SOURCE NP {ET}{PP}*** ➢ …, **还是**星巴克 starbucks **or** Nanotechnology奈米科技**是一间**…	
5. ***{DT} SOURCE {DT} NP {ET} {PP}*** ➢**The translation of** starbucks **is** 星巴克 … **or** Nanotechnology **的中文翻译是**奈米科技	

Figure 2. Patterns with illustrative examples

We also analyse the linguistic features of the target text. There are two types of linguistic patterns used for this purpose: eliminator terms and discriminator phrases.

The eliminator terms [2] (words/phrases, see pattern 4 in figure 2) are somewhat like the stop words used in corpus linguistics, but common stop word lists are not suitable for our term extraction purpose because some stop words maybe important components in the translation. If we include the eliminator terms in the translation candidate, a problematic translation may be generated. If we use eliminator terms we can successfully extract the correct translation term. We first appoint some words by examination of the web pages extracted and further analyse the terms situated around the source query to extract the most frequent terms as eliminator terms. The left side chunk of the source term will be processed from the words adjacent to the source term to the first eliminator term detected. The right side chunk will be extracted until the first right eliminator term is detected.

Discriminator phrases (see pattern 5 in figure 2) can be regarded as a special syntactic rule which indicates the relations between different terms/concepts. This is very important to our process as we are trying to detect a relation between the unknown term and candidate translations. This type of phrases often includes the explicit clue that the noun phrases after the source term are translation of it. We include all the common phrases people used in writing mixed-language pages. Examples can be viewed from the example sentences in pattern 5 in figure 2.

2.3 Selection of Candidates

Further process in the system flow chart means that several terms might extracted by different patterns, we need to determine which candidate terms are the correct translation of the query. To do this we calculate the frequency (*Hit Counts*) of the translation candidates in our database. From the statistical point of view, if the source query and the candidate term co-occur frequently there is a greater chance the former is a correct translation.

Furthermore, the patterns themselves will have different impact on the importance of extracted translation candidates, hence different weightings will be allocated to the terms derived by individual patterns. Some symbolic patterns tend to be more important than others (for example, people tend to include translation terms in brackets rather than quotation marks). Therefore, the candidates extracted by this type of patterns should be given more weight in the final selection process. The weighting assigned to each term candidate depends on the presence or absence of different patterns. Generally speaking, the symmetric patterns and the discriminator terms will be given double weighting score than other patterns, for example, if the terms extracted by punctuation are assigned weighting score 1, the terms extracted by brackets for same source term will be assigned score 2. In our approach, when a term was calculated once for hit count score, it will be multiplied by the weighting score and add up to final hit count score . We select

[2] Please note this type of terms also include the digit from 1 to 9 or the digit in Foreign language format. They are used as an extraction boundary because the unknown term does not contain numbers.

the term with highest hit count score as final translation term.

3 Experiment Results Using NTCIR-5 Data

We first report test results using NTCIR-5 data to evaluate our PMTE method in this section because we conduct a baseline experiment to evaluate our method using NTCIR-5 data only. We will report results using NTCIR-6 data and discuss the cross-collection results in the next section.

3.1 Experiment Setup

Details of NTCIR-5 document collection, topics and formats can be referred to [4], we will not give details here. The document pre-processing is done through the following way: Since the word boundary in Chinese text is less distinguishable, we first use a segmentation tool downloaded from www.mandarintools.com to do the text segmentation. No further sophisticated procedures for text processing was applied in our experiment, such as stop words remove or phrase identification.

The NTCIR-5 queries consist of 50 topics and we used all of them. Each topic in NTCIR-5 is composed of four parts: *Title, Description, Narrative, and Concepts*. We conduct title runs (experiment uses the Title field) and description runs (experiment uses the Description field). The narrative and concepts fields are not used due to the search words or phrases in title and description fields are short and more close to the real-world user queries. The English queries used in our experiments are all stemmed and stop words are removed. The Chinese queries are processed same as text processing.

An English-Chinese dictionary machine readable dictionary was used in our experiment: ldc_ec_dict 2.0 from the Linguistic Data Consortium [3]. This dictionary contains 110,843 entries of the English words and Chinese counter-part. It is generated from a large bilingual corpus. The dictionary is compiled from a set of diverse resources, partly LDC-internal but mainly from the Internet. The machine translation system we used was the BabelFish [4] web translation interface.

Three types of results were produced - single language information retrieval in Chinese (SLIR), cross language information retrieval in English-Chinese (CLIR), and baseline CLIR in English-Chinese. The SLIR result used Chinese topics prepared by the NTCIR-5 workshop committee. The CLIR and baseline tasks both jointly used methods of dictionary-based translation and machine translation (Babelfish). The result is represented and analyzed below.

To evaluate the effectiveness of the proposed method, we conducted six runs named *C-C-T, C-C-D, E-C-T, E-C-D, E-C-T-baseline, E-C-D-baseline*, in which *C* means Chinese, *E* means English, *T* means retrieving using Title field as queries and D means retrieving using Description field as queries. For example, *C-C-T* means retrieving Chinese documents using Chinese queries in title fields.

We generated SLIR results in order to measure our CLIR results. The baseline measure is done through simple translation through the MT system without unknown terms translation. We used the Lemur Toolkit [5] and the Okapi [6] BM25 retrieval functions without feedback in our retrieval process. The relevance judgments provided by NTCIR are at two levels: strictly relevant documents known as rigid relevance, and likely relevant documents, known as relax relevance. In this paper, we used both relevance results to report our results.

The source terms not found in the dictionary and not recognized by Babelfish were sent to the PMTE translation system to locate the possible correct translation. After the translation terms obtained, they will be sent to Lemur for information retrieval process in monolingual environment. In total we sent 32 unknown terms in 50 topics to our translation system for processing.

3.2 Results and Discussion

Results Analysis. Table 1 lists the mean average precision (MAP) for all six runs. As indicated in Table 1 the proposed method PMTE is able to outperform the baseline model in both *title* and *description* runs for relax and rigid results. For example, PMTE method can improve the retrieval performance by 47.6% in the title filed run and 19.6% in the description field run through rigid relevance assessment comparing to the baseline. Based on these results, we can confirm that the PMTE method performs substantially better than simple translation CLIR but additional works required for improving the whole system performance. This is a quite reasonable result because our goal is to show the effectiveness of our method in translating the unknown terms.

[3] http://www.ldc.upenn.edu/

[4] http://babelfish.altavista.com/

[5] http://www.lemurproject.org/

[6] "Okapi" in the City University of London IR system originally stood for "Online Keyword Access to Public Information." (Stephen Robertson, personal communication). The OKAPI information retrieval system can be found at http://www.soi.city.ac.uk/ andym/OKAPI-PACK/

Table 1. Comparison of PMTE method (E-C-T, D), monolingual run (C-C-T, D), and baseline run (E-C-T, D Baseline) in title and description fields. Results are measured in mean average precision and precision at 10 documents with relax and rigid assessment. %change denotes the percent change in performance comparing to the baseline. Bold figures indicate statistically significant differences in performance between PMTE run, monolingual run and the baseline with a 95% confidence according to the Wilcoxon test. *E-English, C-Chinese, T-Title, D-Description*

Run	MAP		%Change		Precision@10	
	Relax	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*
E-C-T	**0.313**	**0.2681**	53.51%	47.55%	0.416	0.33
C-C-T	**0.4294**	**0.3692**	110.59%	103.19%	0.534	0.416
E-C-T-Baseline	0.2039	0.1817	0.00%	0.00%	0.296	0.226
E-C-D	**0.3262**	**0.2696**	20.99%	19.56%	0.462	0.35
C-C-D	**0.4243**	**0.3619**	57.38%	60.49%	0.54	0.412
E-C-D-Baseline	0.2696	0.2255	0.00%	0.00%	0.392	0.298

Table 2. Results of NTCIR-6 CLIR task (E-C-T, D) and SLIR task(C-C-T, D) in *title* **and** *description* **fields. Results are measured in mean average precision, R-Precision and precision at 10 documents with relax and rigid assessment.** *Percentage* **denotes the percent change in CLIR performance comparing to the SLIR. Average Performance of all NTCIR-6 participants is also included.** *E-English, C-Chinese, T-Title, D-Description*

Run	MAP		Percentage		R-Precision		Precision@10		Average	
	Relax	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*
WTG-E-C-T-01	0.1647	0.1237	64.36%	67.05%	0.1984	0.1505	0.312	0.206	0.2071	0.1521
WTG-E-C-D-02	0.1562	0.1207	60.40%	64.93%	0.1894	0.1501	0.33	0.232	0.1971	0.1424
WTG-C-C-T-03	0.2559	0.1845	N/A	N/A	0.3051	0.2282	0.442	0.286	0.3214	0.232
WTG-C-C-D-04	0.2586	0.1859	N/A	N/A	0.3046	0.229	0.474	0.288	0.3339	0.2379

Table 3. Results of NTCIR-6 cross-collection (STAGE 2) CLIR task (E-C-T, D) and SLIR task(C-C-T, D) in *title* **and** *description* **fields. Results are measured in mean average precision, R-Precision and precision at 10 documents with relax and rigid assessment.** *Percentage* **denotes the percent change in CLIR performance comparing to the SLIR. Average Performance of all NTCIR-6 participants is also included.** *E-English, C-Chinese, T-Title, D-Description*

Run	MAP		Percentage		R-Precision		Precision@10		Average	
	Relax	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*	*Relax*	*Rigid*
WTG-E-C-T-01-N3	0.0741	0.0588	65.98%	61.76%	0.1189	0.0918	0.1952	0.1238	0.1237	0.1035
WTG-E-C-D-02-N3	0.0663	0.0514	60.16%	54.56%	0.1159	0.088	0.1738	0.1119	0.1293	0.1047
WTG-C-C-T-03-N3	0.1123	0.0952	N/A	N/A	0.1651	0.1365	0.3167	0.2286	0.2847	0.2232
WTG-C-C-D-04-N3	0.1102	0.0942	N/A	N/A	0.1699	0.1342	0.319	0.2214	0.2822	0.2291
WTG-E-C-T-01-N4	0.0776	0.0606	71.00%	66.45%	0.1089	0.0836	0.1322	0.0847	0.1138	0.0985
WTG-E-C-D-02-N4	0.0729	0.0535	70.03%	65.48%	0.112	0.0737	0.1356	0.0932	0.1089	0.0894
WTG-C-C-T-03-N4	0.1093	0.0912	N/A	N/A	0.1559	0.1284	0.2102	0.1441	0.2269	0.1841
WTG-C-C-D-04-N4	0.1041	0.0817	N/A	N/A	0.1511	0.1211	0.2034	0.1339	0.2212	0.1746
WTG-E-C-T-01-N5	0.1719	0.1374	53.17%	49.66%	0.1831	0.1524	0.264	0.186	0.1848	0.1544
WTG-E-C-D-02-N5	0.1802	0.1405	61.02%	58.93%	0.1972	0.163	0.314	0.222	0.1966	0.1659
WTG-C-C-T-03-N5	0.3233	0.2767	N/A	N/A	0.3347	0.2853	0.486	0.37	0.4013	0.3483
WTG-C-C-D-04-N5	0.2953	0.2384	N/A	N/A	0.307	0.2518	0.46	0.336	0.3851	0.3233

A further examination of Table 1 gives rise to the following observations:

1 The improvement of performance in description runs appears to be worse than that for title runs. This result confirms the unknown terms translation has a large contribution the CLIR performance in short queries. Particularly when the key search words are unknown terms that will severely reduce the retrieval performance.

2 The results using PMTE method do not represent important monolingual retrieval effectiveness. The reason is we do not perform complex text processing techniques and future works required for yielding better performance. For example, we just choose first entry in dictionary for our initial translation process.

Effectiveness of Translation Process. There are 32 unknown terms in 50 topics in NTCIR5 data set. Only 6 terms are not successfully translated. The top-one translations of all other 26 terms returned by PMTE are correct translations and that shows about 81.25% translation rate. This is a significant improvement and demonstrates the effectiveness of employing linguistics patterns alongside statistical analysis because the translation candidates acquired by statistical methods often need further disambiguation process [12, 3]. Although the translation rate is high, as mentioned above, we did not successfully obtain all the translation terms. The main reason for that is some terms are out-of-date and receive less interest from the present web society. Since our method collects all information from the contemporaneous web, it is correspondingly harder to find historical terms. This may also cause another problem: when searching for certain terms after some time, it is possible that the correct translation will have switched from one meaning to another different one. We save that for future work.

4 NTCIR-6 Experiment Results

Results Analysis. The NTCIR-6 data set description can be located at [5]. The pre-processing process and whole system are identical to those as described in last section. The only difference is that we used simple TFIDF retrieval function(vector space model) instead of Okapi BM25 in our experiments. The results for STAGE 1 and STAGE 2 are presented in Table 2 and Table 3.

Table 2 lists the results for the formal runs in STAGE 1 and table3 lists cross-collection results for runs in STAGE 2. The *percentage* denotes the overall performance of each run against the monolingual (SLIR) runs. We also include average MAP of all systems in NTCIR-6 workshop. From the results we can tell that the TFIDF system performs worse than Okapi BM25 retrieval function.

By examining the results we can notice that the PMTE method performed well on NTCIR-5 data collection while it performed worse on NTCIR-3 and NTCIR-4 data collections. This indicates that our previous finding is correct: it is hard to find correct OOV terms translations in contemporary web for out-of-date terms. This also indicates that the importance of translation of OOV terms in CLIR process. There is no difference between the *percentage* field in CLIR against SLIR in all runs over different data collections shows the robust of our method.

An interesting phenomenon is that the performance using the *Descriptive* field is consistently lower than the performance using only the *Title* field. The experiments using the *Descriptive* field should provide more context information and should give better results. This is probably because our pre-processing, translation and retrieving processes do not consider the dependency of the terms. It also shows our pre-processing and translation processes have a huge space to improve.

Comparison with other systems. The most well-known systems using mixed language web pages to translating unknown terms are described in [3] and [12]. As previous mentioned, the advantage of using our system to obtain the translations normally does not need further disambiguation process. This means using patterns are more accurate than pure statistical analysis. The improvement of performance is nearly same as those in reported systems and we do not apply complex text and query processing. The performance can easily increase by using more powerful initial translation system rather than dictionary and using more complicated text processing techniques.

5 Conclusions and Future Work

In this paper we have introduced a new technique to extract translation terms using pattern matching and reported the experiment results using various NTCIR data. Patterns are not only used in the extraction process but also in the term weighting scheme. This technique has been shown to improve retrieval effectiveness and has three main virtues: *Simplicity, Extensible and Effectiveness*.

Future work may include identifying more accurate patterns and selection rules. The Out-of-Date terms problem is another interesting facet that should be examined carefully

References

[1] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, Melbourne, Australia, 1998. ACM Press.

[2] J. Chen and J.-Y. Nie. Parallel web text mining for cross-language ir. In *Proceedings of RIAO-2000: content-based Multimedia Information Access*, pages 188–192, CollCge de France, Paris, France, 2000.

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

[4] K. Kishida, K.-h. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of clir task at the fifth ntcir workshop. In *the fifth NTCIR workshop meeting*, Tokyo, Japan, 2005. NTCIR.

[5] K. Kishida, K.-h. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of clir task at the sixth ntcir workshop. In *the sixth NTCIR workshop meeting*, Tokyo, Japan, 2007. NTCIR.

[6] Y. Liu, R. Jin, and J. Y. Chai. 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*, pages 536–543, Salvador, Brazil, 2005. ACM Press.

[7] W.-H. Lu, L.-F. Chien, and H.-J. Lee. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(2):159–172, 2002.

[8] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81, Berkeley, California, United States, 1999. ACM Press.

[9] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63, Melbourne, Australia, 1998. ACM Press.

[10] J.-H. Wang, J.-W. Teng, P.-J. Cheng, W.-H. Lu, and L.-F. Chien. Translating unknown cross-lingual queries in digital libraries using a web-based approach. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 108–116, Tuscon, AZ, USA, 2004. ACM Press.

[11] Y. Zhang and P. Vines. Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, Sheffield, United Kingdom, 2004. ACM Press.

[12] Y. Zhang, P. Vines, and J. Zobel. Chinese oov translation and post-translation query expansion in chinese–english cross-lingual information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):57–77, 2005.