Mining Key Phrase Translations from Web Corpora

Fei Huang **Ying Zhang** **Stephan Vogel**

School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
{fhuang, joy, vogel}@cs.cmu.edu

Abstract

Key phrases are usually among the most information-bearing linguistic structures. Translating them correctly will improve many natural language processing applications. We propose a new framework to mine key phrase translations from web corpora. We submit a source phrase to a search engine as a query, then expand queries by adding the translations of topic-relevant hint words from the returned snippets. We retrieve mixed-language web pages based on the expanded queries. Finally, we extract the key phrase translation from the second-round returned web page snippets with phonetic, semantic and frequency-distance features. We achieve 46% phrase translation accuracy when using top 10 returned snippets, and 80% accuracy with 165 snippets. Both results are significantly better than several existing methods.

1 Introduction

Key phrases such as named entities (person, location and organization names), book and movie titles, science, medical or military terms and others [1], are usually among the most information-bearing linguistic structures. Translating them correctly will improve the performance of cross-lingual information retrieval, question answering and machine translation systems. However, these key phrases are often domain-specific, and people constantly create new key phrases which are not covered by existing bilingual dictionaries or parallel corpora, therefore standard data-driven or knowledge-based machine translation systems cannot translate them correctly.

As an increasing amount of web information becomes available, exploiting such a huge information resource is becoming more attractive. (Resnik 1999) searched the web for parallel corpora while (Lu et al. 2002) extracted translation pairs from anchor texts pointing to the same webpage. However, parallel webpages or anchor texts are quite limited, and these approaches greatly suffer from the lack of data.

However, there are many web pages containing useful bilingual information where key phrases and their translations both occur. See the example in Figure 1. This example demonstrates web page snippets[2] containing both a Chinese key phrase "浮士德" and its translation**,** "Faust".

We thus can transform the translation problem into a data mining problem by retrieving these mixed-language web pages and extracting their translations. We propose a new framework to mine key phrase translations from web corpora. Given a source key phrase (here a Chinese phrase), we first retrieve web page snippets containing this phrase using the Google search engine. We then expand queries by adding the translations of topic-relevant hint words from the returned snippets. We submit the source key phrase and expanded queries again to Google to retrieve mixed-language web page snippets. Finally, we extract the key phrase translation from the second-round returned snippets with phonetic, semantic and frequency-distance features.

[1] Some name and terminology is a single word, which could be regarded as a one-word phrase.

[2] A snippet is a sentence or paragraph containing the key phrase, returned with the web page URLs.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 483–490, Vancouver, October 2005.

Figure 1. Returned mixed-language web page snippets using source query

Figure 2. Returned mixed-language web page snippets using cross-lingual query expansion

We achieve 46% phrase translation accuracy when using 10 returned snippets, and 80% accuracy with 165 snippets. Both results are significantly better than several existing methods.

The reminder of this paper is organized as follows: cross-lingual query expansion is discussed in section 2; key phrase translation extraction is addressed in section 3. In section 4 we present experimental results, which is followed by relevant works and conclusions.

2 Retrieving Web Page Snippets through Cross-lingual Query Expansion

For a Chinese key phrase *f*, we want to find its translation *e* from the web, more specifically, from the mixed-language web pages or web page snippets containing both *f* and *e*. As we do not know *e*, we are unable to directly retrieve such mixed-language web page using (*f*,*e*) as the query.

However, we observed that when the author of a web page lists both *f* and *e* in a page, it is very likely that *f'* and *e'* are listed in the same page, where *f'* is a Chinese hint word topically relevant to *f*, and *e'* is *f*'s translation. Therefore if we know a Chinese hint word *f'*, and we know its reliable translation, *e'*, we can send (*f*, *e'*) as a query to retrieve mixed language web pages containing (*f*, *e*).

For example, to find web pages which contain translations of "浮士德"(Faust), we expand the query to "浮士德+goethe" since "歌德" (Goethe) is the author of "浮士德" (Faust). Figure 2 illustrates retrieved web page snippets with expanded queries. We find that newly returned snippets contain more correct translations with higher ranks.

To propose a "good" English hint *e'* for *f*, first we need to find a Chinese hint word *f'* that is relevant to *f*. Because *f* is often an OOV word, it is unlikely that such information can be obtained from existing Chinese monolingual corpora. Instead, we

query Google for web pages containing *f*. From the returned snippets we select Chinese words *f'* based on the following criteria:

1. *f'* should be relevant to *f* based on the co-occurrence frequency. On average, 300 Chinese words are returned for each query *f*. We only consider those words that occur at least twice to be relevant.
2. *f'* can be reliably translated given the current bilingual resources (e.g. the LDC Chinese-English lexicon [3] with 81,945 translation entries).
3. The meaning of *f'* should not be too ambiguous. Words with many translations are not used.
4. *f'* should be translated into noun or noun phrases. Given the fact that most OOV words are noun or noun phrases, we ignore those source words which are translated into other part-of-speech words. The British National Corpus[4] is used to generate the English noun lists.

For each *f*, the top Chinese words *f'* with the highest frequency are selected. Their corresponding translations are then used as the cross-lingual hint words for *f*. For example, for OOV word *f* = 浮士德 (Faust), the top candidate *f'*s are "歌德 (Goethe)", "简介 (introduction)", "文学 (literature)" and "悲剧 (tragedy)". We expand the original query "浮士德" to "浮士德 + goethe", "浮士德 + introduction", "浮士德 + literature", "浮士德 + tragic", and then query Google again for web page snippets containing the correct translation "Faust".

3 Extracting Key Phrase Translation

When the Chinese key phrase and its English hint words are sent to Google as the query, returned web page snippets contain the source query and possibly its translation. We preprocess the snippets to remove irrelevant information. The preprocessing steps are:

1. Filter out HTML tags;
2. Convert HTML special characters (e.g., "&lt") to corresponding ASCII code (">");
3. Segment Chinese words based on a maximum string matching algorithm, which is used to calculate the translation probability between a Chinese key phrase and an English translation candidate.
4. Replace punctuation marks with phrase separator '|';
5. Replace non-query Chinese words with placeholder mark '+', as they indicate the distance between an English phrase and the Chinese key phrase.

For example, the snippet

> 《 廊 桥 遗 梦 》 (the bridges of madison county)[review]. 发布者：anjing | 发布时间：2004-01-25 星期日 02:13 | 最新更新时间

is converted into

> | 廊 桥 遗 梦 | the_bridges_of_Madison_county | review | ++ + | anjing | ++ ++ | 2004-01-25 +++ 02 13 | + + ++ ++,

where "" and "" mark the start and end positions of the Chinese key phrase. The candidate English phrases, "the bridges of madison county", "review" and "anjing", will be aligned to the source key phrase according to a combined feature set using a transliteration model which captures the pronunciation similarity, a translation model which captures the semantic similarity and a frequency-distance model reflecting their relevancy. These models are described below.

3.1 Transliteration Model

The transliteration model captures the phonetic similarity between a Chinese phrase and an English translation candidate via string alignment. Many key phrases are person and location names, which are phonetically translated and whose written forms resemble their pronunciations. Therefore it is possible to discover these translation pairs through their surface strings. Surface string transliteration does not need a pronunciation lexicon to map words into phoneme sequences; thus it is especially appealing for OOV word translation. For non-Latin languages like Chinese, a romanization

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L27
[4] http://www.natcorp.ox.ac.uk/

script called "pinyin" maps each Chinese character into Latin letter strings. This normalization makes the string alignment possible.

We adopt the transliteration model proposed in (Huang, et al. 2003). This model calculates the probabilistic Levinstein distance between a romanized source string and a target string. Unlike the traditional Levinstein distance calculation, the character alignment cost is not binary (0/1); rather it is the logarithm of character alignment probability, which ensures that characters with similar pronunciations (e.g. `p` and `b`) have higher alignment probabilities and lower cost. These probabilities are automatically learned from bilingual name lists using EM.

Assume the Chinese phrase f has J Chinese characters, $f_1, f_2, ... f_J$, and the English candidate phrase e has L English words, $e_1, e_2, ..., e_L$. The transliteration cost between a Chinese query f and an English translation candidate e is calculated as:

$$C_{trl}(e, f) \approx \sum_j \log p(e_{a_j} \mid y_j) = \sum_j \sum_i \log p(e_{a_{(i,j)}} \mid y_{i,j}).$$

where y_j is the pinyin of Chinese character f_j, $y_{j,i}$ is the i th letter in y_j, and e_{a_j} and $e_{a_{(j,i)}}$ are their aligned English letters, respectively. $p(e_{(i,j)} \mid y_{i,j})$ is the letter transliteration probability. The transliteration costs between a Chinese phrase and an English phrase is approximated by the sum of their letter transliteration cost along the optimal alignment path, which is identified based on dynamic programming.

3.2 Translation Model

The translation model measures the semantic equivalence between a Chinese phrase and an English candidate. One widely used model is the IBM model (Brown et al. 1993). The phrase translation probability is computed using the IBM model-1 as:

$$P_{trans}(f \mid e) = \frac{1}{L^J} \prod_{j=1}^{J} \sum_{l=1}^{L} p(f_j \mid e_l)$$

where $p(f_j \mid e_l)$ is the lexical translation probabilities, which can be calculated according to the IBM models. This alignment model is asymmetric, as one source word can only be aligned to one target word, while one target word can be aligned to multiple source words. We estimate both $P_{trans}(f \mid e)$ and $P_{trans}(e \mid f)$, and define the NE translation cost as:

$$C_{trans}(e, f) = \log P_{trans}(e \mid f) + \log P_{trans}(f \mid e).$$

3.3 Frequency-Distance Model

The more often a bilingual phrase pair co-occurs, or the closer a bilingual phrase pair is within a snippet, the more likely they are translations of each other. The frequency-distance model measures this correlation.

Suppose S is the set of returned snippets for query f, and a single returned snippet is $s_i \in S$. The source phrase occurs in s_i as $f_{i,j}$ ($j \geq 1$ since f may occur several times in a snippet). The frequency-distance weight of an English candidate e is

$$w(e) = \sum_{s_i} \sum_{f_{i,j}} \frac{1}{d(f_{i,j}, e)}$$

where $d(f, e)$ is the distance between phrase $f_{i,j}$ and e, i.e., how many words are there between the two phrases (the separator `|` is not counted).

3.4 Feature Combination

Define the confidence measure for the transliteration model as:

$$\phi_{trl}(e \mid f) = \frac{\exp[C_{trl}(e, f)] w(e)^m}{\sum_{e'} \exp[C_{trl}(e', f)] w(e')^m},$$

where e and e' are English candidate phrases, and m is the weight of the distance model. We empirically choose m=2 in our experiments. This measure indicates how good the English phrase e is compared with other candidates based on transliteration model. Similarly the translation model confidence measure is defined as:

$$\phi_{trans}(e \mid f) = \frac{\exp[C_{trans}(e, f)] w(e)^m}{\sum_{e'} \exp[C_{trans}(e', f)] w(e')^m}.$$

The overall feature cost is the linear combination of transliteration cost and translation cost, which are weighted by their confidence scores respectively:

$$C(e,f) = \lambda\phi_{trl}(e \mid f)\exp[C_{trl}(e,f)] + (1-\lambda)\phi_{trans}(e \mid f)\exp[C_{trans}(e,f)]$$

where the linear combination weight λ is chosen empirically. While ϕ_{trl} and ϕ_{trans} represent the relative rank of the current candidate among all compared candidates, C_{trl} and C_{trans} indicate its absolute likelihood, which is useful to reject the top 1 incorrect candidate if the true translation does not occur in any returned snippets.

4 Experiments

We evaluated our approach by translating a set of key phrases from different domains. We selected 310 Chinese key phrases from 12 domains as the test set, which were almost equally distributed within these domains. We also manually translated them as the reference translations. Table 1 shows some typical phrases and their translations, where one may find that correct key phrase translations need both phonetic transliterations and semantic translations. We evaluated *inclusion rate*, defined as the percentage of correct key phrase translations which can be retrieved in the returned snippets; *alignment accuracy*, defined as the percentage of key phrase translations which can be correctly aligned given that these translations are included in the snippets; and *overall translation accuracy*, defined as the percentage of key phrases which can be translated correctly. We compared our approach with the LiveTrans[5] (Cheng et.al. 2004) system, an unknown word translator using web corpora, and we observed better translation performance using our approach.

4.1 Query Translation Inclusion Rate

In the first round query search, for each Chinese key phrase *f*, on average 13 unique snippets were returned to identify relevant Chinese hint words *f'*, and the top 5 *f*'s were selected to generate hint words *e'*s. In the second round *f* and *e'*s were sent to Google again to retrieve mixed language snippets, which were used to extract *e*, the correct translation of *f*.

Figure 3 shows the inclusion rate vs. the number of snippets used for three mixed-language web page searching strategies:

Movie Title	廊桥遗梦 the Bridges of Madison-County 阿甘正传 Forrest Gump
Book Title	红楼梦 Dream of the Red Mansion 茶花女 La Dame aux camellias
Organization Name	圣母大学 University of Notre Dame 大卫与露西派克德基金会 David and Lucile Packard Foundation
Person Name	贝多芬 Ludwig Van Beethoven 奥黛丽赫本 Audrey Hepburn
Location Name	勘察加半岛 Kamchatka 塔克拉玛干沙漠 Taklamakan desert
Company / Brand	汉莎航空 Lufthansa German Airlines 雅诗兰黛 Estee Lauder
Sci&Tech Terms	遗传算法 genetic algorithm 语音识别 speech recognition
Specie Term	秃鹫 Aegypius monachus 穿山甲 Manispentadactyla
Military Term	宙斯盾 Aegis 费尔康 Phalcon
Medical Term	非典型性肺炎 SARS 动脉硬化 Arteriosclerosis
Music Term	空山鸟语 Bird-call in the Mountain 巴松管 Bassoon
Sports Term	休斯敦火箭队 Houston Rockets 环法自行车赛 Tour de France

Table 1. Test set key phrases

- Search any web pages containing *f* (Zhang and Vines 2004);
- Only search *English* web pages[6] containing *f* (Cheng et al. 2004);
- Search any web pages containing *f* and hint words *e'*, as proposed in this paper.

The first search strategy resulted in a relatively low inclusion rate; the second achieved a much higher inclusion rate. However, because such English pages were limited, and on average only 45 unique snippets could be found for each *f*, which resulted in a maximum inclusion rate of 85.8%. In the case of the cross-lingual query expansion, the search space was much larger but more focused and we achieved a high inclusion rate of 89.7% using 32 mixed-language snippets and 95.2% using 165 snippets, both from the second round retrieval.

[5] http://livetrans.iis.sinica.edu.tw/lt.html

[6] These web pages are labeled by Google as "English" web pages, though they may contain non-English characters.

Features	No Hints (Inc = 44.19%) (avg. snippets = 10)	With Hints (Inc = 95.16%) (avg. snippets=130)
Trl	51.45	17.97
Trans	51.45	40.68
Fq-dis	53.62	73.22
Trl+Trans	63.04	51.36
Trl+Trans+ Fq-dis	65.94	86.73

Table 2. Alignment accuracies using different features

These search strategies are further discussed in the section 5.

4.2 Translation Alignment Accuracy

We evaluated our key phrase extraction model by testing queries whose correct translations were included in the returned snippets. We used different feature combinations on differently sized snippets to compare their alignment accuracies. Table 2 shows the result. Here "Trl" means using the transliteration model, "Trans" means using the translation model, and "Fq-dis" means using Frequency-Distance model. The frequency-distance model seemed to be the strongest single model in both cases (with and without hint words), while incorporating phonetic and semantic features provided additional strength to the overall performance. Combining all three features together yielded the best accuracy. Note that when more candidate translations were available through query expansion, the alignment accuracy improved by 30% relative due to the frequency-distance model. However, using transliteration and/or translation models alone decreased performance because of more incorrect translation candidates from returned snippets. After incorporating the frequency-distance model, correct translations have the maximum frequency-distance weights and are more likely to be selected as the top hypothesis. Therefore the combined model obtained the highest translation accuracy.

4.3 Overall Translation Quality

The overall translation qualities are listed in Table 3, where we showed the translation accuracies of the top 5 hypotheses using different number of snippets. A hypothesized translation was considered to be correct when it matched one of the reference translations. Using more snippets always increased the overall translation accuracy, and with all the 165 snippets (on average per query), our approach achieved 80% top-1 translation accuracy, and 90% top-5 accuracy.

Snippets Used	Accuracy of the Top-N Hyp. (%)				
	Top1	Top2	Top3	Top4	Top5
10	46.1	55.2	59.0	61.3	62.3
20	57.4	64.2	69.7	72.6	72.9
50	63.2	74.5	77.7	79.7	80.6
100	75.2	84.5	85.8	87.4	87.4
165	**80.0**	**86.5**	**89.0**	**90.0**	**90.0**
Babel-Fish[7] MT	31.3	N/A	N/A	N/A	N/A
CMU-SMT	21.9	N/A	N/A	N/A	N/A
LiveTrans (Fast)	20.6	30.0	36.8	41.9	45.2
LiveTrans (Smart)	30.0	41.9	48.7	51.0	52.9

Table 3. Overall translation accuracy

We compared the translations from a research statistical machine translation system (CMU-SMT, Vogel et al. 2003) and a web-based MT engine (BabelFish). Due to the lack of topic-relevant contexts and many OOV words occurring in the source key phrases, their results were not satisfactory. We also compare our system with LiveTrans, which only searched within English web pages, thus with limited search space and more noises (incorrect English candidates). Therefore it was more difficult to select the correct translation. Table 4 lists some example key phrase translations mined from web corpora, as well as translations from the BabelFish.

5 Relevant Work

Both (Cheng et al. 2004) and (Zhang and Vines 2004) exploited web corpora for translating OOV terms and queries. Compared with their work, our proposed method differs in both webpage search

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

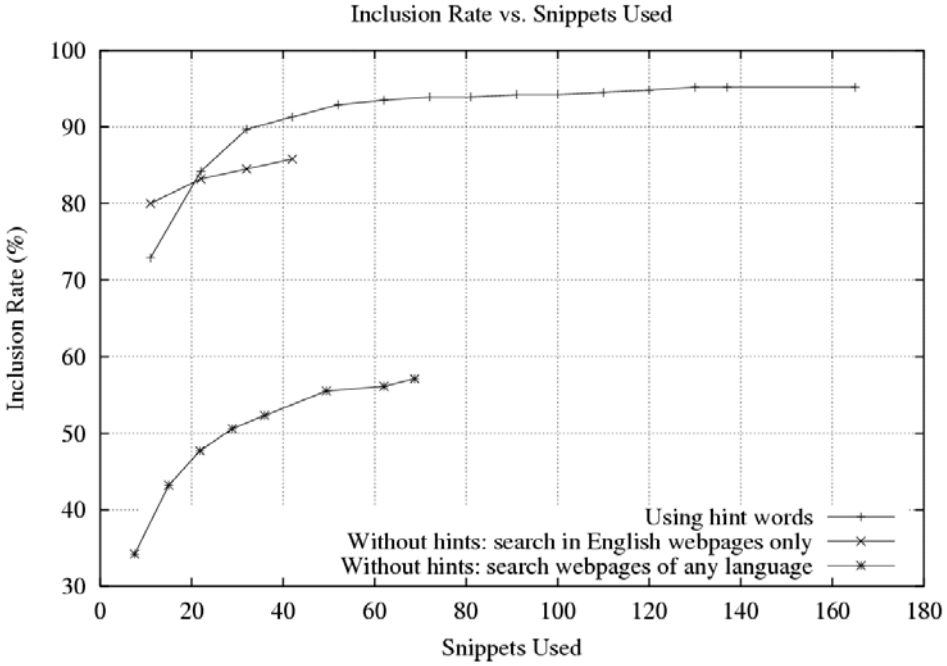

Figure 3. Inclusion rate vs. the number of snippets used

Category	Examples		
	Chinese Key Phrase	**Web-Mining Translation**	**BabelFish™ Result**
Movie Title	廊桥遗梦	the Bridges of Madison County	*Love has gone and only good memory has left in the dream
Book Title	理智与情感	Sense and Sensibility	*Reason and emotion
Organization Name	伍德威尔逊全国联谊基金会	Woodrow Wilson National Fellowship Foundation	*Wood the Wilson nation gets together the foundation
Person Name	小泽征尔	Seiji Ozawa	*Young Ze drafts you
Location Name	柴达木盆地	Tsaidam Basin	Qaidam Basin
Company / Brand	倩碧	Clinique	*Attractive blue
Sci&Tech Terms	贝叶斯网络	Bayesian network	*Shell Ye Si network
Specie Term	海象	walrus	walrus
Military Term	同温层堡垒	stratofortress	stratofortress
Medical Term	青光眼	glaucoma	glaucoma
Music Term	巴松管	bassoon	bassoon
Sports Term	环法自行车赛	Km Tour de France	*Link law bicycle match

*: Incorrect translations

Table 4. Key phrase translation from web mining and a MT engine

space and translation extraction features. Figure 4 illustrates three different search strategies. Suppose we want to translate the Chinese query "浮士德". (Cheng et al. 2004) only searched 188 English web pages which contained the source query, and 53% of them (100 pages) had the correct translations. (Zhang and Vines 2004) searched the whole 55,100 web pages, 10% of them (5490 pages) had the correct translation. Our approach used query expansion to search any web pages containing "浮士德" and English hint words, which was a larger search space than (Cheng et al. 2004) and more focused compared with (Zhang and Vines 2004), as illustrated with the shaded region in Figure 4. For translation extraction features, we took advantage of machine transliteration and machine translation models, and combined them with frequency and distance information.

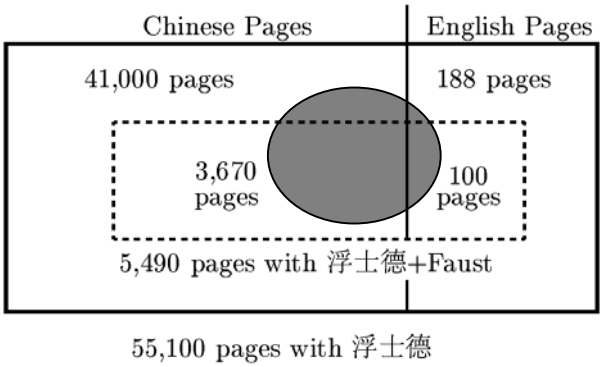

Figure 4. Web search space strategy comparison

6 Discussion and Future Work

In this paper we demonstrated the feasibility of the proposed approach by searching for the English translation for a given Chinese key phrase, where we use punctuations and Chinese words as the boundary of candidate English translations. In the future we plan to try more flexible translation candidate selection methods, and apply them to other language pairs. We also would like to test our approach on more standard test sets, and compare the performance with other systems.

Our approach works on short snippets for query expansion and translation extraction, and the computation time is short. Therefore the search engine's response time is the major factor of computational efficiency.

7 Conclusion

We proposed a novel approach to mine key phrase translations from web corpora. We used cross-lingual query expansion to retrieve more relevant web pages snippets, and extracted target translations combining transliteration, translation and frequency-distance models. We achieved significantly better results compared to the existing methods.

8 References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R.L. Mercer. The Mathematics of Machine Translation: Parameter Estimation. *In Computational Linguistics,* vol 19, number 2. pp.263-311, June, 1993.

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 the Proceedings of 27th ACM SIGIR, pp146-153. ACM Press, 2004.

F. Huang, S.Vogel and A. Waibel. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In the Proceedings of the 41st ACL. Workshop on Multilingual and Mixed-language Named Entity Recognition, pp124-129, Sapporo, Japan, July 2003.

W.-H. Lu, L.-F. Chien, H.-J. Lee. Translation of web queries using anchor text mining. ACM Trans. Asian Language Information Processing (TALIP) 1(2): 159-172 (2002)

P. Resnik and N. A. Smith, The Web as a Parallel Corpus, Computational Linguistics 29(3), pp. 349-380, September 2003

S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao and A. Waibel. The CMU statistical machine translation system. *In Proceedings of the MT Summit IX Conference* New Orlean, LA, September, 2003.

Y. Zhang and P. Vines. Detection and Translation of OOV Terms Prior to Query Time, In the Proceedings of 27th ACM SIGIR. pp524-525, Sheffield, England, 2004.

Y. Zhang and P. Vines 2004, 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.

Y. Zhang, F. Huang and S. Vogel, 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.