Improving Web-Based OOV Translation Mining for Query Translation

Yun Dong Ge, Yu Hong, Jian Min Yao, and Qiao Ming Zhu

Provincial Key Laboratory of Computer Information Processing Technology,
Soochow University, Suzhou, China, 215006
geyundong@gmail.com, {hongy,jyao,qmzhu}@suda.edu.cn

Abstract. Query translation is the most widely used approach for cross-language information retrieval (CLIR). The major challenge of query translation is translating Out-Of-Vocabulary (OOV) terms. This paper proposes three methods to improve OOV translation mining for query translation. Firstly, Co-occurrence information is utilized to extract topic words and to expand the source language query with the translations of topic words for collecting relevant bilingual snippets. Secondly, an improved frequency change measurement method which combines context dependency is utilized to extract valid OOV translation candidates from noisy, small-sized bilingual snippets. Thirdly, for choosing the proper translation, a combination model considering frequency-distance, surface patterns matching and phonetic features is proposed to pick out the appropriate translation(s). Experimental results show that this OOV translation mining approach for query translation has substantial CLIR performance improvement.

Keywords: Out Of Vocabulary, OOV, Translation Mining, Query Translation.

1 Introduction

CLIR enables people to retrieve documents written in one language by another language query. Although CLIR has been advancing rapidly, a major bottleneck remains for translating OOV in queries. Because OOVs are translated incorrectly, irrelevant documents may rank highly, which will cause worse performance (precision, recall etc.) of CLIR.

Conventional CLIR uses a bilingual dictionary to translate queries. Pirkola & al. [12] analyzed the shortcomings of dictionary-based approaches and proposed the possible resolutions. As real queries are usually short, dynamic and diverse, even using the state-of-the-art dictionary cannot avoid the OOV problem.

As increasing number of bilingual resource is available on Internet, exploiting the web to translate OOV is feasible and reliable. For bilingual resource collecting, if directly send the source language OOV to search engine, the returned snippets (contain titles and summaries) are usually monolingual, which cannot be used to extract the target language translation of OOV. That is to say, OOV without expansion usually cannot extract effective snippets. Besides, mostly existing query expansion methods for

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 576–587, 2010.

collecting snippets need to segment the OOV. This segment will introduce extra errors and lower the quality of snippets.

Extracting MLUs (Multi-Lexical Unit) as candidates from the gathered snippets is the second main problem. If the candidate of OOV is not extracted correctly, the substantial translation of OOV cannot be mined. The most common method for candidate extraction is taking continuous characters after removing stop words as candidates. However, most of these continuous strings are invalid lexical units.

How to select the most appropriate translation from the large candidate set is the third challenge. The intuitive frequency feature of candidate is mostly adopted in existing methods. More effective features (distance, surface patterns and phonetic etc.) can be used for OOV translation selection.

For the above problems, this paper proposes a novel solution to mine high quality translation of OOV for query translation. Briefly, OOV translation mining method consists of three main parts:

1. Bilingual snippets collection. Gather the bilingual snippets containing the OOV in English and its translation in Chinese from a search engine.
2. Candidate extraction. Extract MLUs from snippets as OOV translation candidates.
3. Appropriate translation selection. Rank candidates generated in 2 for picking out the correct OOV translation.

The Co-occurrence information is used to extract the topic words of the OOV. The OOV expanded with translations of the topic words are sent to search engine to collect relevant bilingual web snippets. The translation of topic words based cross-language expansion method can get more relevant bilingual snippets. To enhance the quality of candidates, a variation of frequency change measurement term extraction method is adopted. Features such as frequency, distance, surface patterns and transliteration are exploited for better translation selection.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 proposes the solution to bilingual resource collection. In section 4, we present the OOV translation candidates extraction method. Candidates ranking for choosing the OOV translation and experiments are presented in section 5 and section 6 respectively. Finally, conclude the paper in the last section.

2 Related Work

Nagata & al. [11] firstly attempted to use the web for translating Japanese OOV. They downloaded the top 100 full web pages as resource. Lu & al. [9] extracted translations of terms through mining of web anchor texts and link structures. For the dependence on full web pages, these two methods need big network bandwidth, large storage capacity and long computing time.

Cheng & al. [3] and Lu & al. [8] utilized the top 100 snippets to mine the OOV translation. Thus the complexity of collecting bilingual resource by these two approaches is sharply reduced. The main shortcoming, however, is the snippets seldom contain corresponding translations since the source query is not expanded with cross-language words; and most of the snippets returned are usually monolingual, which cannot be used to extract the target language translations.

Fang & al. [4] segmented the source OOV, and then expanded the terms with cross-language words to collect web resource. Sun & al. [15] used forward-backward maximum matching method to segment the source term and looked up the target language translations of segmented units to expand the source term for collecting bilingual snippets. These approaches all exploited cross-language query expansion to collect more relevant bilingual snippets and needed to segment the source term, unfortunately, the meaning of a source term is not always the simple combination of the component words meanings. Thus segmenting source term will introduce extra noise in query expansion and further lead to translation mining errors.

Zhang & al. [18] used all substrings of twenty Chinese characters immediately before OOV and twenty Chinese characters immediately after OOV as candidates. Huang & al. [6] and Sun & al. [15] exploited punctuations to segment the snippets and continuous English strings were taken as candidates. Taking the continuous string as candidate or taking substrings of strings before and after OOV as candidates has a drawback because this kind of candidate set often contains lots of invalid lexical units. If the valid candidate of OOV can not be extracted, then correct translation of OOV can hardly be mined. Chien & al. [2] proposed a variant method of mutual information called Significance Estimation (SE). Silva & al. [13] extracted MLUs from large corpora with Local Maxima algorithm, the formula used in the algorithm is symmetric conditional probability (SCP). Cheng & al. [3] introduced Context Dependency (CD) to improve SCP for Chinese MLUs extraction. SE, SCP and SCPCD achieved good performance in large corpora situation.

3 Bilingual Snippets Collection

Unfortunately, not all the snippets gathered from a search engine with the source OOV contain both the OOV and the translation. Take Kim Dae-Jung as an example(Kim Dae-Jung is translated as “金大中” in Chinese), only 1 snippet in the top 10 returned snippets contains target language information since most of the snippets are in English which cannot be used to extract Chinese translations. Ballesteros & al. [1] proposed that query expansion lays a substantial basis for translation extraction. The most expansion methods in previous researches must segment the source OOV, but the meaning of a source term usually is not the combination of individual word meaning in the source term. For example in “风凉话”, which means sarcastic remarks, the meanings of the component characters are “wind”, “cool” and “talk” respectively. The component characters meanings are irrelevant to the meaning of “风凉话”. So an expansion method based on co-occurrence information is utilized in our study which doesn’t need segmentation. We first take the OOV as a whole unit with quotation marks and submit it to a search engine. And then extract the topic words from the returned snippets in the source language. Topic words are that with high relevant relation to the source OOV in the same topic or domain. After that, we send the OOV together with the translations of the topic words in the target language respectively to a search engine to collect bilingual resource.

We filter out the non-noun words and English stop words from the bilingual resource, and then we can get an English noun word list. TF*IDF metric is used to extract

topic words from the noun English word list. Then we select the top 5 of the list as the topic words.

In the previous example of "Kim Dae-Jung", the topic words extracted from top 20 snippets are "Korea", "president", "winner", "peace" and "prize". Their corresponding translations (if they have several meanings, we choose the first meaning in order to simplify the procedure) are then used as the cross-language expansion words. In this example, the cross-language expansion words are "韩国(Korea)", "总统(president)", "胜利者(winner)", "和平(peace)", and "奖品(prize)". Then we send the source OOV "Kim Dae-Jung" together with translations of topic words respectively to retrieve bilingual snippets. The quality of these bilingual snippets is greatly improved than using only the source OOV.

4 Candidates Extraction

The translation of the OOV may be either a MLU or a single word. As the scale and domain of the dictionary, conventional dictionary-based segmentation approaches are not able to identify the OOV in the snippets, thus the translation of the OOV can hardly be obtained by these approaches.

The approaches for extracting MLUs from large corpus (SE, SCP, SCPCD) are not satisfactory in search engine based candidates extraction, as the size of snippets is quite small. One snippet usually just contains 2 or 3 sentences. Moreover, the snippets are usually fragments of sentences.

We use Frequency Change Measurement together with Context Dependency (FCMCD) to extract MLUs from the bilingual collection. It combines the Frequency Change Measurement (FCM) [8] and Context Dependency (CD). FCM is based on two observations as follows: the first observation is that the component characters of the term have similar frequencies in a collection returned from a search engine. Such as "金(Kim)", "大(Dae)"and "中(Jung)" have similar frequency in the mixed language snippets gathered by "Kim Dae-Jung" with cross-language expansion; the second is that when a valid MLU is extended with an extra character; the frequency of extended term drops apparently. The frequency of "金大中几"is quite lower than that of "金大中" in the snippets. In the FCM method, the following equation is used to evaluate the probability of string S being a MLU.

$$R(S) = \frac{f(S)}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1}$$

where S is a Chinese string; f(S) is the frequency of S; x_i is the frequency of each character in S and $\bar{x}$ is the average frequency of all the characters in S.

The candidates extracted by FCM still contain some invalid fragments which are sub-sequences of the valid MLUs. Thus some correct candidates will not be extracted by FCM only. In order to deal with this problem, we combine the Context Dependency (CD) to improve the quality of candidates since we discovered that valid candidates usually have diverse adjacent characters while their sub-terms have relative less

and fixed adjacent characters. The CD reflects the degree of a string stands alone as a word. In FCMCD method, Equation (1) is modified as follows.

$$R'(S) = \frac{LN(S) \times f(S) \times RN(S)}{1 + \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2}$$

where LN(S) is the number of unique left neighboring characters of S; RN(S) is the number of unique right neighboring characters of S. Virtual sentence start mark "B" and sentence end mark "E" were inserted into each sentence at the sentence boundary. If S occurs at the start of the sentence, LN(S) is one. If S occurs at the end of the sentence, RN(S) is one.

We do not simply remove the Chinese stop words from a snippet directly. That will lead to the possibility that the left neighboring character of the stop word and the right neighboring character are extracted as MLU and thus introduce extra noise. Take the sentence "前总统慰问金大中*的*妻子以及他们*的*三个儿子" (viz. Former president condoled to the wife of Kim Dae-Jung and their three sons.) as an example. If we directly remove the stop word "的", "中妻" may be extracted as MLU, but "中妻" is an invalid lexical unit in Chinese.

5 Translation Selection

A method that combines frequency-distance, surface patterns matching and phonetic features of the Chinese candidates is used in choosing the correct translation(s).

5.1 Frequency Distance Model

We consider the intuitive features: the frequency of the candidate, the distance between the candidate and the OOV. Intuitively, the authentic translation of the OOV usually co-occurs with the OOV frequently in the snippets. The nearer the candidate to the OOV, the more probable it is the final translation. The measurement between one candidate and the OOV is calculated as:

$$FD(s,t) = \frac{\sum_{J} \sum_{k} \frac{1}{d_k(s,t)}}{\max_{fre-dis}} \tag{3}$$

where s is the OOV, t is one candidate. $d_k(s,t)$ is the k-th distance between s and t in one snippet, for s and t might co-occur more than once in a snippet. J is the number of the snippets and K is the number of co-occurrence between s and t. The denominator is the max reciprocal of the distance among all the candidates. We adopt the count of words between s and t as the distance other than the byte distance, for the snippets may contain several kinds of symbols such as Chinese characters, English characters and punctuation marks etc., which are encoded differently. Thus it reflects more linguistic information to take word count as the distance measure. If there are no character between s and t, the distance is one and so forth.

5.2 Surface Patterns Matching Model

Some Asian languages users usually annotate terms with their translations in English inside a pair of parentheses. These punctuation marks can be used to enhance the precision of the final OOV translation(s). In our solution, some English-Chinese pairs are submitted to a search engine to learn the surface patterns automatically [16], [7]. Some surface patterns obtained by this procedure are listed in table 1. E is one English term; C is the Chinese translation of E.

Table 1. Surface patterns example

No.	Surface patterns
1	C(E, E(C, C (E, E (C
2	C[E, E[C, C [E, E [C
3	C.E, E.C, C,E E,C
4	C>>(E, C|E, C-E

If one candidate matches most of surface patterns in a bilingual snippet set, its probability of being the right translation will increase greatly. The cost of the surface patterns matching is formulated as:

$$SP(s,t) = \frac{N_{matching}}{\max_{num}} \tag{4}$$

where s is the OOV, t is one candidate. The numerator is the number of times that s and t matches the surface patterns. The denominator is the maximum number of matching patterns among all the candidates.

5.3 Transliteration Model

Many OOVs are translated based on phonetic pronunciations, which we called transliteration. Firstly, our transliteration model resolves a sort of matching problem, computing the phonetic similarity between the English OOV and Chinese candidates. We already have the Chinese candidates and thus we don't need to generate the Chinese transliterations. Secondly, to avoid the double errors from English phonetic representation to pin-yin and from pin-yin to Chinese characters, we use a method proposed by [14], [10] to segment an English name into a sequence of syllables, computing the probability between an English syllable and a Chinese character to estimate the possibility. The aim is to compute the phonetic similarity for selecting the right translation. First of all, we segment the English OOV into a sequence of syllables based on heuristic rules and then compute the transliteration cost using the following equation.

$$Trl(s,t) = P(s,t) \Big/ D(s,t) \tag{5}$$

where P(s,t) is the co-occurrence probability of s and t which is defined as:

$$P(s,t) \approx \prod_{i=1}^{\min(m,n)} (1-\gamma_1) prob(e_i, c_i) \quad (6)$$

where $\gamma_1$is the smoothing weight. prob(e_i,c_i)is the probability between an English syllable e_i and a Chinese character c_i and is computed based dynamic programming from the training corpus contains 37,665 proper name pairs. D(s,t) is the number of syllable difference between an English OOV s and a Chinese candidate t, which is defined as:

$$D(s,t) = \varepsilon + |m-n| \quad (7)$$

here ε is a decaying parameter, m is the total number of English term syllables and n is the total number of Chinese characters.

In order to improve incorrect transliteration mapping between English syllables and Chinese characters, we combine the forward and backward mapping. The final transliteration cost is defined as the average of forward value and backward value.

5.4 Model Combination

We use the frequency-distance model as the baseline model, and re-rank the results by the surface pattern matching model. Not all the OOVs are transliteration terms, directly combing with the transliteration model value decreases the whole performance. If one source term is transliteration word, the transliteration model value of it is much higher than values of those are not transliteration words. The threshold of the transliteration model is computed. If the transliteration value of one candidate is greater than the threshold, then we re-rank the candidate by the transliteration value. Otherwise the rank of the candidate stays the same.

6 Experiments and Analysis

TDT4 corpora were used in our CLIR experiment. TDT4 contains topics, documents and relevance judgments. This document set contains the complete set of English, Arabic and Chinese news text used in the 2002 and 2003 Topic Detection and Tracking technology evaluations. TDT4 have 80 topics. Each topic has four parts: topic number and Title, Seminal Event, Topic Explication Rule of Interpretation. There are 27,142 Chinese documents in TDT4 corpora. We use the title of the topic as our English query, that is to say we have 80 source language queries. There are 4 queries in all the source queries have no OOV; the other 76 queries have 82 OOVs.

6.1 Snippets Collection Experiment

We sent the 82 English OOV terms (without cross-language expansion) to search engine to gather snippets and also used our expansion method to gather snippets. We used top 50 snippets returned by search engine for each OOV. Each method gathered 4,100 snippets. If one snippet contains the target language information of the OOV, we consider this snippet is effective.

Without expansion method collected 545 effective snippets. With our expansion method collected 2,636 effective snippets. Our proposed snippets collection method can get more 2,091 effective snippets than the no expansion method, the result shows our method is valid in resource collection stage. High quality of snippets is a key fundamental resource for MLUs extraction and translation selection.

6.2 Candidates Extraction Experiment

50 OOV terms were randomly selected from OOV set. Our expansion method was used to collect snippets for each selected OOV. MLUs were extracted from the 2,500 of 4,200 snippets by SE, SCP, SCPCD, FCM and FCMCD respectively. If one candidate is extracted with valid lexical boundary, this candidate is regarded as correct. Precision is used to evaluate candidate extraction. The precision is defined as the percentage of correct candidates in all candidates. As we do not know the exact number of valid candidates in the specific snippets in advance, so the number of candidates extracted by each methods is used to measure how close to the exact number. That is to say the number of extracted candidates works as an indirect measure of recall. Table 2 contains the results of MLUs extraction.

Table 2. The quality of MLUs extraction

Methods	Precision	Total
SE	55.30%	38,573
SCP	61.32%	47,904
SCPCD	68.54%	48,397
FCM	84.28%	37,355
FCMCD	91.84%	36,006

SE, SCP and SCPCD do not work well on this small size snippets resource. These methods extracted more MLUs, but the precisions are very low. FCM alone achieves better performance than SCPCD, the precision increases 15.74%. FCM together with CD achieves the best performance 91.84% which increases 7.56% precision compared to FCM, but the number of MLUs is lower than that of FCM. Using CD can filter more invalid candidates, although it reduces the recall. As for query translation, precision is more important than recall, the high quality of candidate set is the primary element.

6.3 OOV Translation Selection Experiments

The top *n* inclusion rate [3] is used as evaluation metric for translation selection, which is defined as the percentage of terms whose translations are included in the top *n* returned translations and we implemented the Chi-square and Context Vector (Chi+CV) translation selection method which was proposed in [3]. Each OOV was expanded by our method and then 100 snippets were gathered for candidate extraction. Different translation selection models were combined in turn for evaluating the power of each model.

Table 3. Translation mining result

	TOP1	TOP3	TOP5	TOP10
Chi+CV	30.50%	45.12%	56.10%	60.98%
FD	54.88%	71.95%	79.27%	87.80%
FD+SP	60.98%	72.17%	82.93%	92.68%
FD+SP+Trl	65.85%	75.61%	84.15%	93.90%

FD is frequency-distance model. SP is surface patterns matching model. Trl is transliteration model. FD achieves better performance than that of Chi+CV. Chi+CV doesn't use the cross-language expansion and the CV sometimes misguides the selection when the distribution of the incorrect translation is confirmed with the distribution of the OOV. FD together with SP get further improvement. Top 1 inclusion rate increased 6.1% compared to FD. FD+SP+Trl get the best performance, improving the Top 1 inclusion rate with 4.87% increase compared to FD+SP. This indicates three models are complementary for each other.

In order to examine the influence of the number of snippets, we use 50 snippets, 100 snippets and 150 snippets to mine translation with FD+SP+Trl for each OOV respectively. The mining result is shown in table 4.

Table 4. OOV translation mining results with different numbers of snippets

	TOP1	TOP3	TOP5	TOP10
50	52.44%	63.41%	69.51%	85.37%
100	65.85%	75.61%	84.15%	93.90%
150	67.07%	80.49%	85.37%	95.12%

Using 100 snippets improves the Top 1 inclusion rate by 13.41% compared to using 50 snippets. Using 150 snippets slightly improves Top 1 inclusion rate by 1.22% compared to using 100 snippets. The performance improves when the number of snippets increases. The number of relevant snippets increases when use more snippets; the correct OOV translation co-occurs more often and matches more surface patterns with the OOV when using more snippets. However, the main drawback of using more snippets is needs more bandwidth and time for extracting candidates and selecting translation.

6.4 CLIR Experiments

TDT4 English titles are used to retrieve Chinese documents. The retrieval system is constructed with Lucene. Documents were indexed using 2-gram based inverted file index. Mean Average Precision (MAP) values were used to evaluate the performance of retrieval system. Five runs are compared to investigate the performance of different query translation methods. Query disambiguation and relevance feedback [5] were not

applied in retrieval because our main aim is to evaluate the improvement of our web based OOV translation mining for query translation.

RUN 1: Monolingual retrieval. English titles were translated into Chinese by professional translators and the translated titles are used to retrieve the Chinese documents. This run provides a comparison of "ideal" retrieval case.

RUN 2: English Queries were translated using a dictionary (containing 286,932 single word pairs) ignoring the OOV.

RUN 3: Translation equivalents were extracted from parallel corpus. This parallel corpus was constructed using the method proposed in [17]. There are 760,000 sentences in this corpus. English queries were translated using the same dictionary in RUN2, and then the translation of OOV was looked up from the translation equivalents.

RUN 4: English queries were translated using the same dictionary in RUN2, and then the translation of OOV was mined using our proposed web based method.

RUN 5: English queries were translated with dictionary, translation equivalents and the translations of OOV which mined by our web based method.

Table 5. MAP values obtained by different query translation methods

	MAP	Per. of RUN 1（%）
RUN 1	0.4945	-
RUN 2	0.0975	19.72
RUN 3	0.1496	30.25
RUN 4	0.3850	77.86
RUN 5	0.4070	82.31

Two main reasons account for the poor performance of RUN 2 .First, our dictionary only contains common words, such compound words and proper names in the queries cannot be found in this dictionary. Second, some terms have several translations and we just select the first entity as their translation which introduces extra error. RUN 3 gets the performance of 30.25% percentage of RUN 1. The improvement is not remarkable because most of the translation equivalents are common words, thus most OOV cannot be translated. RUN 4 improves the performance greatly with 77.86% percentage of RUN 1. RUN 5 achieves the best performance. That is because the high quality translation equivalents of common words obtained from parallel corpus and the power of OOV translation mining using web are all benefit to the query translation.

Our web based OOV translation mining method returns 10 translations for every OOV. We found that some returned translations are full names or abbreviations of the correct translations; some are the different transliterations using different Chinese characters of the correct translations; although some results are not the correct translations, they have high relation with the source OOV. All these kinds of OOV translations can be used as natural query expansion for OOV. Whether using more translations of OOV improves the performance of CLIR or not? We investigate this problem using the method of RUN4. The result of web method is shown in table 6 using top 1, top3, top5 and top 10 translations of OOV respectively.

Table 6. MAP values obtained by RUN 4 with different number of translations

	TOP1	TOP3	TOP5	TOP10
MAP	0.3850	0.3901	0.3921	0.3853
Per. of RUN 1（%）	77.86	78.89	79.30	77.91

Using top 5 translations achieves the top performance. However, while using top 10 translations the performance drops slightly. The reason is the last 5 translations are usually incorrect and decrease the whole performance.

7 Conclusion

This paper proposes a web-based OOV translation mining for query translation. The topic words based method was used to expand the OOV for collecting relevant bilingual snippets. Then an improved Frequency Change Measurement method which combines Context Dependency (FCMCD) is used to extract valid MLUs from noisy, small bilingual snippets. A method using frequency-distance, surface patterns and transliteration modeling is proposed to select the correct translation. Experimental results show that this method has impressive improvement in English-Chinese CLIR on TDT4 test set.

For further work, we will combine our candidate extraction method with POS tagging to extract more reliable candidates. Semantic relation between OOV and candidates will be used to improve Top 1 inclusion rate of translation selection. Our experiment will be conducted on other CLIR test set (CLEF, NTCIR etc.). For better CLIR performance, query disambiguation and relevance feedback will be integrated into our CLIR architecture.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grant No 60970057.

References

1. Ballesteros, L. and Croft, W. B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: Proc. of 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 84-91. Philadelphia, USA (1997)
2. Chien, L.-F.: PAT-tree-based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, Philadelphia (1997)
3. Cheng, P.–J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., Chien, L.-F.: Translating Unknown Queries with Web Corpora for Cross-language Information Retrieval. In: The Proceedings of 27th ACM SIGIR, pp. 146–153. ACM Press, New York (2004)

4. Fang, G., Yu, H., Nishino, F.: Chinese-English Term Translation Mining Based on Semantic Prediction. In: Proceedings of the 21th International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Main Conference Poster Session, pp. 199–206 (2006)
5. He, D., Wu, D.: Enhancing query translation with relevance feedback in translingual information retrieval. Information Processing and Management, In Press, Corrected Proof, Available online (2009)
6. Huang, F., Zhang, Y., Vogel, S.: Mining Key Phrase Translation from Web Corpora. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 483–490 (2005)
7. Lin, D., Zhao, S., Derme, B.V., Pasca, M.: Mining Parenthetical Translations from the Web by Word Alignment. In: Proceedings of 46th Meeting of the Association for Computational Linguistics: Human Language Technology, pp. 994–1002 (2008)
8. Lu, C., Xu, Y., Geva, S.: Web-Based Query Translation for English-Chinese CLIR. Computational Linguistics and Chinese Language Processing 13(1), 61–90 (2008)
9. Lu, W.-H., Chien, L.-F., Lee, H.-J.: Translation of Web Queries Using Anchor Text Mining. Asian Language Information Processing 1(2), 159–172 (2002)
10. Lu, W.-H., Lin, J.-H., Chang, Y.-S.: Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names. Computational Linguistics and Chinese Language Processing 13(1), 91–120 (2008)
11. Nagata, M., Saito, T., Suzuki, K.: Using The Web as a Bilingual Dictionary. In: 39th Meeting of the Association for Computational Linguistics 2001 Workshop Data-Driven Methods in Machine Translation, pp. 95–102 (2001)
12. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based Cross-language Information Retrieval: Problems, Methods, and Research Findings. Information Retrieval 4(3/4), 209–230 (2001)
13. Silva, J.F.d., Dias, G., Guilloré, S., Pereira, J.G.: Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: Progress in Artificial Intelligence: 9th Portuguese Conference on Artificial Intelligence, pp. 113–132 (1999)
14. Wan, S., Verspoor, C.M.: Automatic English-Chinese Name Transliteration for Development of Multilingual Resources. In: Proceedings of 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 1352–1357 (1998)
15. Sun, J., Yao, J.-M., Zhang, J., Zhu, Q.-M.: Web Mining of OOV Translations. Journal of Information & Computational Science 5(1), 1–6 (2008)
16. Wu, J.-C., Lin, T., Chang, J.S.: Learning Source-Target Surface Patterns for Web-based Terminology Translation. In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics Interactive Poster and Demonstration Sessions, Ann Arbor, pp. 37–40 (2005)
17. Yan, Z.-X., Feng, Y.-H., Hong, Y., Yao, J.-M.: Parallel Resources Mining From Bilingual Web Pages. In: Conference of China Information Retrieval (CCIR), pp. 513–524 (2009)
18. Zhang, Y., Vines, P.: Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In: Proceedings of SIGIR Conference, pp. 162–169 (2004)