Web-Based Technical Term Translation Pairs Mining for Patent Document Translation

Feiliang REN
Northeastern University
Shenyang, China
renfeiliang@ise.neu.edu.cn

Jingbo ZHU
Northeastern University
Shenyang, China
zhujingbo@mail.neu.edu.cn

Huizhen WANG
Northeastern University
Shenyang, China
wanghuizhen@ise.neu.edu.cn

Abstract:

This paper proposes a simple but powerful approach for obtaining technical term translation pairs in patent domain from Web automatically. First, several technical terms are used as seed queries and submitted to search engineering. Secondly, an extraction algorithm is proposed to extract some key word translation pairs from the returned web pages. Finally, a multi-feature based evaluation method is proposed to pick up those translation pairs that are true technical term translation pairs in patent domain. With this method, we obtain about 8,890,000 key word translation pairs which can be used to translate the technical terms in patent documents. And experimental results show that the precision of these translation pairs are more than 99%, and the coverage of these translation pairs for the technical terms in patent documents are more than 84%.

Keywords:

Term translation, patent document translation, web-based, key word extraction, key word selection, machine translation

1. Introduction

Patent translation is an important translation task in natural language processing. During this translation task, it is surely to face the problem of translating a large amount of technical terms. Usually, these technical terms are the most fundamental units to express the main characteristics of a patent document, such as its domain, its novelty, its technical routine, and so on. These terms are rarely occurred in common bilingual documents. Thus it is difficult even impossible to cover them by existed bilingual dictionary or existed bilingual corpus. Besides, these technical terms can express a clear meaning only as a whole, it would introduce great ambiguity or confusion if they were segmented and translated partly. Because of these reasons, translating them is one of the greatest challenges in patent translation task.

To translate technical terms accurately, the best way is to construction a bilingual technical term dictionary. High construction cost makes it impossible to construct such bilingual dictionary by human experts. Thus how to construct bilingual technical term dictionary automatically becomes the key of technical term translation in patent translation task. In fact, automatic acquisition of bilingual technical term translation pairs has been extensively researched in the literature. Generally these automatic construction methods can be summarized as the following two categories. One is to extract bilingual term translation pairs from existed bilingual parallel or comparable corpus. The other is to acquire bilingual term translation pairs from the Web.

The state-of-art term translation pair extraction strategies tend to take Web as a big corpus. The main idea behind these strategies is that for every input term, its translation must exist somewhere on the web. Thus the term translation pair extraction problem is converted to the problem of finding these translations from the web and extraction them correctly. Cheng et al. (2004) proposed a method to translate English unknown queries with the assistance of Web resources. In their method, they used context-vector and chi-square methods to determine Chinese translations for unknown query terms via mining of top-100 returned web pages from a search engine. Zhang et al.(2009) translated OOV terms based on Web mining and a supervised learning method. In their method, they used classification and ordinal regression method to rank the extracted translation candidates for the input OOV terms. Fang et al. (2006) proposed a Chinese-English term translation method based on semantic prediction with the assistance of Web. In their method, they used and term expansion strategy and a feedback learning method to collect effective web pages. And the final translation results are extracted from the returned web pages based on multi features.

From the review above we know that a lot of related works have been done on term translation with the assistance of Web. However, these existed methods can't be used to translate technical terms in patent domain directly because of the following two reasons.

First, most of these methods are active translation methods, which mean that the term to be translated is given before translated. This kind of term translation method is not suitable for the task of translating large

978-1-4244-6899-7/10/$26.00 ©2010 IEEE

amount of terms because of the huge time cost. But in patent translation task, large number of technical terms requires more fast translation methods. Thus existed term translation methods are not suitable for the technical term translation task in patent translation. Besides, term recognition is still an open problem that hasn't been solved properly.

Secondly, existed term translation method is sensitive to the notability degree of input terms. For famous term, the translation performance is very promising. However, it is lower notability for most of technical terms in patent domain. That is to say we would not find their correct translations with existed methods for most of technical terms.

On the other hand, there is a large of number of technical papers released in all kinds of technical journals every year in China. These technical papers cover a wide range of technical domains. The most important is that there are usually bilingual key word translation pairs in these technical papers. And a large portion of these key words are technical terms that often occur in paten documents too. This phenomenon indicates us that if we could collect these technical papers and extract the key words translation pairs from them, we would construct a large bilingual technical term dictionary. This is very important for technical term translation in patent translation task. Based on this basic idea, a new translation extraction method is proposed for technical terms in patent documents. In this method, a huge amount of web pages that contain the snaps of technical papers are collected firstly. Then an extraction algorithm is used to extract those key word translation pairs from the web pages. Finally, an evaluation method based on multi features is proposed to pick up those true technical terms from the extracted key word translation pairs.

The remainder of this paper is organized as follows. In section 2, we list some related work. The overview of our mining approach is presented in section 3. In section 4, we give the detail description of the modules in our method. The experimental results are reported in section 5. And finally in section 6, we conclude our work.

2. Related Work

For term translation pairs extraction from web pages, Cao et al., (2007) and Lin et al., (2008) proposed two different methods with the parenthesis pattern. The basic idea of their methods is that authors of many mix-language web pages, especially those whose primary language is non-English (such as Chinese, Japanese and so on), usually annotate terms with their original English translations insides of a pair of parentheses. Thus they can extract some term translation pairs with parenthesis patterns. However, it is obvious that not all term translation pairs follow parenthesis pattern, especially for those technical terms in patent documents. Thus these methods can not be used to extract technical term translations for patent documents.

Apart from extracting term translations directly from mix-language web pages, more approaches have been proposed to mine term translations from snippets returned by search engines(Ren et al., 2009; Jiang et al., 2007; Zhang and Vines, 2004; Cheng et al., 2004; Huang et al., 2005). In these methods, the source language term is given and the goal is to mine the correct translations from the Web. These methods usually can achieve high translation performance. However, they are not suitable for large scale technical term translation in patent documents for the following reasons. First of all, these methods need a list of predefined source terms which is not easy to obtain. Secondly, most of these methods rely heavily on the frequency of the target translation in the returned snippets, which makes mining low-frequency technical term translations difficult.

3. Overview of the Proposed Approach

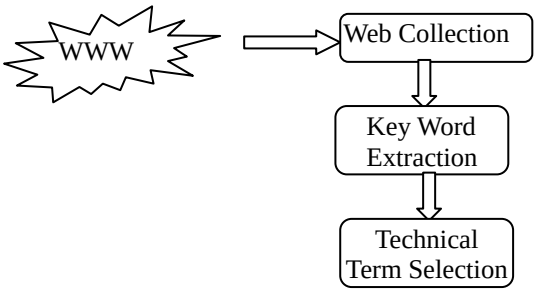

Figure 1. The Framework of Our Approach

As illustrated in Figure 1, our method consists of three main components: web collection, key word extraction and term evaluation.

In the Web collection step, we must collect those web pages that contain the snaps of technical papers, especially those web pages that contain both the Chinese key words and their translations.

In the key word extraction step, an extraction algorithm is used to extract those key word translation pairs from the technical web pages.

In the technical term selection, a multi-features based selection method is proposed to pick up those true technical term translation pairs.

4. Detail of the Proposed Approach

In this section, we will present the details about the three components in the proposed approach.

4.1. Web Collection

Generally, the technical terms in patent documents

rarely occur on the Web directly. However, in some academic websites, there are huge amount of web pages that contain some snippets of technical papers. Moreover, in these web pages, they usually contain some bilingual key word translation pairs. Figure 2 shows an example of such kind of web pages from CNKI (http://www.cnki.net/).

Figure 2. An example of technical papers from CNKI

From above Figure2 we can see that there are huge amount of available technical key word translation pairs on the Web. Most of these key words are also those technical terms that are often used in patent documents. Thus if we can collect these web pages that contain the key word translation pairs effectively, we can construct a big bilingual technical term translation dictionary which is very useful for patent document translation. The key of this web collection process is to find those appropriate web pages effectively.

Here in our approach, we design a Breadth-first search algorithm to collect those web pages that we need from CNKI. In our algorithm, we first take some technical terms as seeds. And then these seed technical terms are taken as input queries and are submitted to the search engine[2] provided by CNKI. From the returned web pages, we extract all the reference documents and download all these references.

The detail breadth-first web collection algorithm is shown in Figure 3. In this algorithm, stop criteria θ is used to control the time complexity or space complexity.

[2] http://epub.cnki.net/grid2008/index/ZKCALD.htm

Input: Seed technical term queue, $Q = \{t_1, t_2, ..., t_m\}$,URL queue $H = \Phi$, and stop criteria θ

Output: Web pages set

Algorithm:

1. For each term $t_i \in Q$, submit it to search engine provided by CNKI. Download the returned web pages.
2. Extract all the reference documents URLs from these returned web pages in previous step. And add these URLs in H.
3. WHILE $H \neq \Phi$ and θ is not satisfied DO
4. Select a URL from H, download the corresponding web page.
5. GOTO step 2.

Figure 3. Breadth-First Search Algorithm for Web Collection

4.2. Key Word Extraction

In most of time, for each obtained web page, if there were English key words in it, these English key words would be aligned to the Chinese key words in sequence one by one. This is the case shown in Figure 2. However, there are always some excerpts of web pages which don't follow this pattern. In these excerpts, the number of English key words is not equal the number of Chinese key words. This can be seen from figure 4.

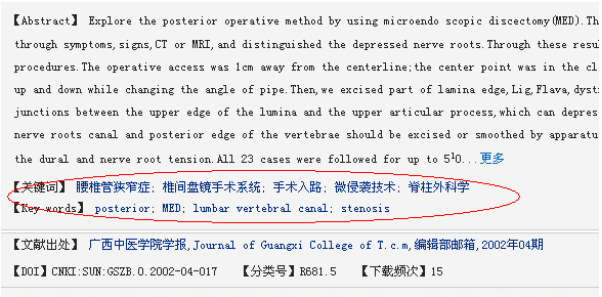

Figure 4. An Example of Web Page that has a Different Key Word Number for English and Chinese

Through observation we also find following two facts. First, both the Chinese key words and the English key words are started from some fixed strings. For Chinese key words, they are started from string "【中文关键词】|【关键词】", and for English key words, they are started from string "【英文关键词】|【Key words】|【English Key words】". Second, most of these key words can be identically formatted as $"w_1[;|,|\backslash t|\backslash s]w_2[;|,|\backslash t|\backslash s]...[;|,|\backslash t|\backslash s]w_m"$, which means that the key words are separated by a

limited number of punctuations, such as comma, semicolon, tab and so on.

Take above information into consideration; we design an extraction method to extract those key word translation pairs. The algorithm is shown in Figure 5.

Input: Web Page Set $P = \{p_1, p_2, ..., p_n\}$

Output: Key Word Translation Pair Dictionary D

Algorithm:

1. For each web page $p_i \in P$, if there is key word starting strings for both English and Chinese, extract Chinese key words sequence $CK_i = \{ck_{i1}, ck_{i2}, ..., ck_{im}\}$ and English key words sequence $EK_i = \{ek_{i1}, ek_{i2}, ..., ek_{in}\}$.
2. For each key word sequence pairs "$CK_i - EK_i$" DO
3. IF $|CK_i| = |EK_i|$, align ck_i to ek_i, and add $ck_i \leftrightarrow ek_i$ to D.
4. IF $|CK_i| \neq |EK_i|$, then for each ck_i DO
5. If part elements of ck_i and ek_j are translations of each other, align ck_i to ek_j, and add $ck_i \leftrightarrow ek_j$ to D.
6. Return D.

Figure 5. Key Word Extraction Algorithm

With above algorithm, we will extract lots of key word translation pairs. We must evaluate the confidence for every translation pair.

It is easy to understand that for a key word translation pair $ck_i \leftrightarrow ek_j$, it is unlikely to occur only in ONE technical paper. That is to say these key word translation pairs will occur in at least several technical papers. And the more of a key word translation pair occur in the technical papers, the more reliable of this key word translation pair. Based on this idea, we design our key word translation pair evaluation algorithm as in Figure 6.

Input: Key Word Translation Pair Dictionary D

Output: Refined Word Translation Pair Dictionary D'

Algorithm:

1. For each key word translation pair $ck_i \leftrightarrow ek_i$, count the documents where contain it. Denote this number as $Score(ck_i \leftrightarrow ek_i)$
2. Sort $Score(ck_i \leftrightarrow ek_i)$
3. Remove those key word translation pairs whose $Score$ value less than a given threshold α.
4. Return the refined dictionary D'.

Figure 6. Key Word Evaluation Algorithm

4.3. Technical Term Selection

4.3.1 Features Used in Our Approach

After the above key word extraction step, we will obtain a key word translation pair dictionary with a high confidence for every of its items. However, not all of these key word translation pairs are technical terms that are used in patent documents. Thus, we must pick up those true technical terms from key word translation pair dictionary D. In our method, we use a multi-features based evaluation method. And following features are used.

Inverse Domain Frequency We define the inverse domain frequency of a key word translation pair kw_i with following formula:

$$IDmF(kw_i) = \log(C / Ck) \qquad (1)$$

where C is the total domain number, and Ck is the number of domains that contain key word kw_i.

Intuitively, if a key word translation pair kw_i occurs in many different domains, it is unlikely to be a true technical term translation pair. Thus the inverse domain frequency is an important index to pick up the true technical terms from the obtained key word translation dictionary. To compute inverse domain frequency, we need two parameters C and Ck. From figure 2 we can see that we can obtain a classification ID for every downloaded web page. And these classification IDs are the labels which are used to distinguish different domains. So we can compute the inverse domain frequency for every key word translation pairs easily.

Inner-Domain Inverse Document Frequency We define the inner-domain inverse document frequency of a key word with following formula:

$$IDF(kw_i, dom_j) = \log(N_i / N_j(k)) \qquad (2)$$

where N_i is the total document number in domain dom_i, and $N_j(k)$ is the number of document that contain key word kw_i in domain dom_j.

It is the same as inverse domain frequency, if a key word translation pair kw_i occurs in many different documents, it is unlikely to be a true technical term translation pair.

Inner-Domain Average Term Frequency We

define the inner-domain average term frequency of a key word with following formula:

$$iATF(kw_i, dom_j) = TF_i / N_j(k) \quad (3)$$

where TF_i is the total term frequency of ek_i (or ck_i) in the abstract part of a technical paper in domain dom_i, and $N_j(k)$ is the number of document that contain key word kw_i in domain dom_j.

For most of technical paper, the abstract part is the most condensed description for the technical characteristic of the entire paper. Thus if a key word occurs in the abstract part, it will provide us a strong hint that the key word are true technical terms. Based on this idea, we take inner-domain average term frequency as another important index for the technical term selection.

Average Term Frequency We define the average term frequency of a key word with following formula:

$$ATF(kw_i) = TF_i / N_i \quad (4)$$

where TF_i is the total term frequency of ek_i (or ck_i) in the abstract part of a technical paper, and N_i is the number of document that contain key word kw_i.

We use the average term frequency index as global information to evaluate the reliability of a key word translation pair is a true technical term translation pair. It is also easily to understand that the more of a key word translation pair occur in the technical papers of different domains, the less likely that this key word translation pair is true technical term translation pair.

4.3.2 FeaturesCombination

Finally, we evaluate every key word translation pair so that possible technical terms get higher scores. We use a weighted sum of the above features to compute the probability of being a technical term for every key word translation pair. The formula used is shown in following formula 5.

$$\begin{aligned} &ConfidenceScore(kw_i, dom_j) \\ &= \lambda_1 IDmF(kw_i) \times ATF(kw_i) \\ &+ \lambda_2 \frac{1}{|Dom|} \sum_j IDF(kw_i, dom_j) \times iATF(kw_i, dom_j) \end{aligned} \quad (5)$$

where $|Dom|$ is the total number of domains, and $\lambda_1 + \lambda_2 = 1$.

In the above equation, $IDmF(kw_i) \times ATF(kw_i)$ can be viewed as another kind of $TF \times IDF$ value which is defined on the whole technical papers. And $IDF(kw_i, dom_j) \times iATF(kw_i, dom_j)$ can also be viewed as a kind of $TF \times IDF$ value which is defined on the technical papers in a specific domain. The final confidence score for a key word translation pair is proportional to these two factors.

5. Experimental Results

We will introduce our experiments in this section. There are three components in our experiments. Firstly we evaluate the precision of extracted key word translation pairs. Then we evaluate the coverage of the selected technical term translation pairs for the technical terms in patent documents.

5.1. Precision Evaluation

With the proposed method, we download about 4,000,000 web pages which cover the domains of IT, medical, material, chemistry, and physics and so on. From these web pages, we totally extract about 8,890,000 key word translation pairs. Some statistics about these key word translation pairs are shown in table 1. From table 1 we can see that the extracted key word translation pairs have a very wide range of domains, which is very necessary for the technical term translation in patent document translation tasks. Also the huge number of these key word translation pairs also guarantees the feasible of patent translation with them.

Domain	# (10,000)
Materials Science	13
Fiscal & Monetary	8
Philosophy and Social Sciences	63
Electronics and Telecommunications	10
Power Industry	1
Power engineering & nuclear technology	25
Law	3
Integrated Technology	18
Management Science	5
Chemistry	91
Aerospace	5
Environmental Science	2
Machinery, instruments	6
Basic Medical & Biological Engineering	37
Computer and Automation	54
Architecture and Water	2
Comprehensive Economic	4
Mining Engineering	2
Integrated Science and Engineering	92

Obstetrics and gynecology, pediatrics	8
Surgery, Dermatology and Venereology	34
Integrative Medicine	4
Comprehensive Medical Science	43
Special medical oncology	18
ENT(ears, nose and throat)	9
Pharmacy	3
Internal Medicine	28
Neurology and psychiatry	6
Mechanics	10
Comprehensive Agricultural Sciences	49
Biological Sciences	89
Mathematics, Nonlinear Science and System Science	21
Fisheries, Fisheries	2
Astronomy, earth science	23
Livestock animal industry	3
Integrated Engineering	98

Table 1. Statistics of Extracted Key Word Translation Pairs

To evaluate the precision of these extracted key word translation pairs, we randomly selected 1,000 translation pairs as test data. We take two evaluation methods here. One is human evaluation, and the other is automatic evaluation. For human evaluation, we present these test data to a human translator to judge whether there translation pairs correct. For automatic evaluation, we submit the Chinese language parts of these test data to Google Translator[3] to obtain their corresponding English translations. Then we submit the English language parts of these test data to Google Translator to obtain their corresponding Chinese translations. We think it is a correct translation pair if Google Translator returns a translation that has some overlap fragments with the translation we extracted. And the corresponding experimental results are shown in table 2.

	Human Evaluation	Automatic Evaluation
Precision	97.6%	94.3%

Table 2. Evaluation Results

From table 2 we can see that there are still lots of wrong extracted key words translation pairs. However, after thoroughly analysis we found that many of these wrong translation pairs by our evaluation methods are those key words pairs whose translations are abbreviated. For example, for the key word translation pair "光码分多址/OCDMA", the English translation part is the abbreviation of the Chinese part's translation that is "Optical Code Division Multiple Access". That is to say, the true precision of the extracted key word translation pairs should be higher than the results in table 2. To validate this conclusion, we remove all the key word translation pairs whose English parts are abbreviated. After removal, we obtain 964 key word translation pairs as a new test data, and evaluate them again. The new experimental results are shown in table 3.

	Human Evaluation	Automatic Evaluation
Precision	99.69%	96.78%

Table 3. Evaluation Results

From table 3 we can see that the precision for both human evaluation and automatic evaluation improved compared with table 2. We also noticed that there are still several wrong translation pairs (3 for human evaluation and 31 for automatic evaluation)in our extracted results. After recheck these translation pairs which have be judged as wrong by human evaluation and automatic evaluation, we found that in fact the precision from automatic evaluation is wrong. And the true errors are 3, just as human evaluation outputted. And all of these errors are introduced when $|CK_i| \neq |EK_i|$ in figure 5. Despite these errors, the obtained key word translation pairs still have a high quality.

5.2. Coverage Evaluation

To evaluate the coverage of these extracted key word translation pairs for those technical terms used in patent document, we downloaded 200 patent abstracts respectively for semiconductor optoelectronics domain and radiologic domain from State Intellectual Property Office of P.R.C[4]. From these abstracts we extract all the 947 technical terms by human. We take these technical terms as test data to evaluate the coverage of our extracted key word translation pairs. Here the coverage is defined as following formula 6.

$$Coverage = \frac{M}{N} \qquad (6)$$

where M is the number of test data that can be found in our key word translation pairs, and N is the total number of test data.

The corresponding experimental result is shown in table 4.

Domain	Coverage
semiconductor optoelectronics	83%
radiologic	86.5%
Total	84.75%

Table 4. Coverage Results

From table 4 we can see that our extracted key word

[3] http://translate.google.cn/

[4] http://www.sipo.gov.cn/sipo2008/zljs/

translation pairs have a about 85% coverage for the technical terms in patent documents. For those technical terms that can't be covered by our extracted key word translation pairs, there are following reasons. One reason is that there are a big time span between the patent document and the technical papers where the key word translation pairs extracted. For example, if the patent documents are granted in an earlier year and the technical papers we used are released in a later year, the technical terms in these patent documents are unlikely to be covered by the extracted key word translation pairs. The other reason is that the patent proposers are prone to invent new technical terms to distinguish their patents with other patents.

From the table above, we can see that our method is powerful for constructing the necessary technical term dictionary for patent document translation.

Chinese Parts	English Parts
光电子器件	optoelectronic devices
GaN 基半导体	GaN-based semiconductor
紫外辐射	ultraviolet radiation
紫外探测器	UV detectors
光纤通信	optical communication
氢等离子体清洗	hydrogen plasma cleaning
紫外写入	ultraviolet-writing
阵列波导光栅	arrayed waveguide grating
火焰水解法	flame hydrolysis deposition
掺氮类金刚石薄膜	nitride carbon film
外延生长	epitaxial growth
GaN 薄膜	GaN film
光电耦合隔离放大器	optically coupled isolation amplifier
偏振模色散	polarization mode dispersion
色散补偿器	dispersion compensator
双折射系数	birefringence coefficient
异质结有机发光器件	heterojunction organic light-emitting device
界面电荷	interfacial charge
异质结有机发光器件	heterojunction organic light-emitting device

Table 5. Some Examples of Key Word Translation Pairs in Semiconductor Optoelectronics Domain

Chinese Parts	English Parts
肾移植	Kidney transplantation
动脉狭窄	Arterial Stenosis
血管成形术	Angioplasty
血管瘤	Hemangioma
颌面部	Maxillofacial region
栓塞	Embolization
局部注射治疗	Local injection treatment
脑血管疾病	Angiography
造影	Embolization
栓塞	Cerebral vascuar disease
上消化道狭窄	Upper alimentary tract stricture
食管气管瘘	Esophagotracheal fistula
动脉狭窄	Arterial Stenosis
血管成形术	Angioplasty
叶状囊肉瘤	Cystosarcoma phyllodes

Table 6. Some Examples of Key Word Translation Pairs in Radiologic Domain

6. Conclusions

Technical term translation pair collection is very important for patent document translation. Existed term translation methods are not competent for the large scale and fast technical term translation requirement in patent document translation task. In this paper, we propose a simple but powerful approach for obtaining technical term translation pairs in patent domain from Web automatically. With our method, we extract about 8,890,000 key word translation pairs with a precision above than 99%. And our experiments also show that these extracted key word translation pairs have a coverage for the technical terms in patent documents higher than 84%.

Acknowledgements

This paper is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and also is supported by the "the Fundamental Research Funds for the Central Universities" and National Science Foundation of China (60873091).

References

[1] Gaolin Fan, Hao Yu, and Fumihito Nishino. 2006. Chinese-English Term Translation Mining Based on Semantic Prediction. Proceedings of COLING/ACL 2006. pp199-206.

[2] Yuhang Yan, Qin Lu, Tiejun Zhao. 2008. Chinese Term Extraction Using Minimal Resources. Proceedings of Coling 2008. pp1033-1040.

[3] Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. pp870-878.

[4] Yuejie Zhang, Yang Wang and Xiangyang Xue. 2009. English-Chinese Bi-Directional OOV Translation based on Web Mining and Supervised Learning. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp129-132.

[5] Feiliang Ren, Muhua Zhu, Huizhen Wang, Jingbo Zho, Chinese-English Organization Name Translation Based on Correlative Expansion. Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009. pp143-151

[6] Feiliang Ren, Jingbo Zhu, Huizhen Wang. Translate Chinese Organization Names Using Examples and Web. Proceedings of 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering. 2009. pp83-89.

[7] Fei Huang, Stephan vogel and Alex Waibel. 2004. Improving Named Entity Translation Combining Phonetic and Semantic Similarities. Proceedings of the HLT/NAACL. pp281-288.

[8] Masaaki NAGATA. 2001. Using the Web as a Bilingual Dictionary. Proceedings of ACL 2001 Workshop on Data-driven Methods in Machine Translation. pp95-102.

[9] Jian-Cheng Wu, Tracy Lin and Jason S.Chang. 2005. Learning Source-Target Surface Patterns for Web-based Terminology Translation. Proceedings for the ACL Interactive Poster and Demonstration Sessions.

[10] Iñaki Alegria, Nerea Ezeiza, and zaskun Fernandez. 2006. Named Entities Translation Based on comparable Corpora. 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multi-word expressions in a Multilingual Context, pp1-8

[11] Bonnie Glover Stalls and Kevin Knight. 1998. Translating Names and Technical Terms in Arabic Text. Processing of the COLING/ACL Workshop on Computational Approaches to Semitic Languages.

[12] Y. Al-Onaizan and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp400-408.

[13] Fei Huang. 2005. Cluster-specific Named Transliteration. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp435-442.

[14] Hany Hassan and Jeffrey Sorensen. 2005. An Integrated Approach for Arabic-English Named Entity Translation. Proceedings of ACL Workshop on Computational Approaches to Semitic Languages. pp87-93.

[15] Fei Huang, Stephan Vogel and Alex Waibel. 2003. Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization. Proceedings of the 2003 Annual Conference of the Association for Computational Linguistics, Workshop on Multilingual and Mixed-language Named Entity Recognition.

[16] Feiliang Ren, Li Zhang, Minghan, Zhang and Tianshun Yao. 2007. EBMT Based on Finite Automata State Transfer Generation. Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation. pp65-74.