2016 International Conference on Intelligent Transportation, Big Data & Smart City

Knowledge Acquisition Method for Large-Scale Bilingual Corpus Based on Web Mining

Wei Yin
Jincheng College of Sichuan University Chengdu, Sichuan 611731
yinwei1983@qq.com

***Abstract*—This paper describes a method to acquire multi-word translational equivalences from English-Chinese parallel corpora based on Web mining. To solve the correspondence problem of multiple word, N-gram model is adopted to extract candidate translate units. Then the co-occurrence information is used to acquire subject words related to resource proper noun from search engine. The subject terms translation is adopted to perform language-crossed extension, and the extended query will obtain bilingual abstract resources with high quality from the search engine. We also extract the candidate translate units such as compound words and phrases, based on frequency change information and adjacency information, and make final selection of proper nouns integrated transliteration features, statistical features and template features. The experiments show that the translation mining method proposed in this paper has good performance.**

Keywords-knowledge acquisition; corpus; OOV; translation

I. Introduction

Recently, popularization and rapid development of internet offer massive and abundant electronic information. Due to globalization needs, more and more websites have become bilingual websites and more and more online information is published in form of multi-languages so this offers bilingualism lots of resources. Internet is an inexhaustible and increasing information source so it is a potential and huge multi-language corpus. It is no doubt an effective way to construct bilingual corpus and acquire translating knowledge such as studying effective methods, automatically mining these massive and genuine bilingual document, applying the mined Chinese-English corpus for extracting equivalences in Chinese-English translation and applying search engine for translation version mining.

From the studies of acquiring translation equivalences in bilingual corpus, this paper discusses how to apply Chinese-English parallel corpus to extract Chinese-English translation equivalences. At first, this paper respectively performs part of speech tagging and segmentation, then utilizes N-gram model to acquire candidate translating units, calculates translating probability of candidate translating units according to statistical co-occurrence function and finally adopts iteration strategy to realize extraction of translation equivalences. Search engine-based translation acquisition studies firstly apply query expansion method of key words translation to acquire effective bilingual abstract resources from search engines. Then, it utilizes frequency change information and adjacency information to extract multi-words candidate translation units such as compound word and phrase from relatively smaller-scaled abstract resources with noise. Finally, it selects proper nouns based on transliteration feature, statistical feature and module feature. The tests verify the feasibility of our scheme.

II. Web-based Bilingual Corpus Acquisition

The existing form of bilingual parallel corpus can divide Web resource into two categories. That is, Chinese-English parallel texts respectively exist two Chinese-English parallel web pages and the same one page. We respectively call them parallel resource between web pages and parallel resource inside web pages.

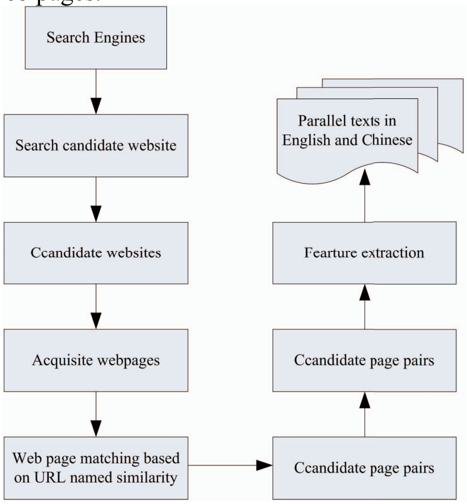

Figure 1. Basic flow chart of parallel texts

As is shown by figure 1, the detailed procedures steps are described as follows:

Step 1: According to anchor text information in web pages, through search engine to search for possibly obtaining websites of bilingual comparison web pages, results are called “bilingual candidate websites”.

Step 2: All web pages are collected in bilingual candidate websites.

Step 3: Possible noise in all web pages needs to be performed a series of pretreatment to remove noise and web page parsing in each bilingual candidate website. According to the feature that bilingual parallel web pages usually have similarity during URL naming, bilingual web pages pair of possibly interactive translation will be acquired.

Step 4: After feature extraction, that is, character length probability feature and web page structure feature, bilingual candidate web pages filter pseudo-parallel bilingual webpage pair to acquire the genuine parallel bilingual webpage pair.

978-1-5090-6061-0/17 $31.00 © 2017 IEEE
DOI 10.1109/ICITBS.2016.76

Step 5: On the basis of step 4, bilingual equivalence of mutual translation are obtained from bilingual parallel web pages. This is called “bilingual parallel equivalence”. Thus, sentence-scaled bilingual parallel corpus is acquired.

III. Knowledge Acquisition Scheme

A. *Principle Idea*

Web mining-based query translation method has become the main method in CLIR query translation method.

Source language query firstly adopts bilingual dictionary translation. If dictionary does not contain this query translation, this query is OOV so search engine is used to mine its translation. The center of this method is to apply search engine to mine OOV translation and there are usually three main stages:

- Abstract Obtaining Stage: bilingual abstract resources containing source query and its translation versions are acquired from search engine. The key problem is how to obtain effective abstract resource which is related to source query.
- Candidate Translation Unit Extraction: candidate translation unit containing legal vocabulary boundary information is extracted from bilingual abstract resources in last step. Abstract resources from search engine only contain 2-3 sentences or sentence segment. Compared to traditional bilingual corpus, its scale is very small and then the acquired abstract usually contains OOV. Therefore, it needs deep study on extracting candidate unit.
- Translation version selection: candidate translation unit set from abstract selection is usually very large so the best translation of OOV needs to be selected from this set.

Its system framework is shown as figure 2.

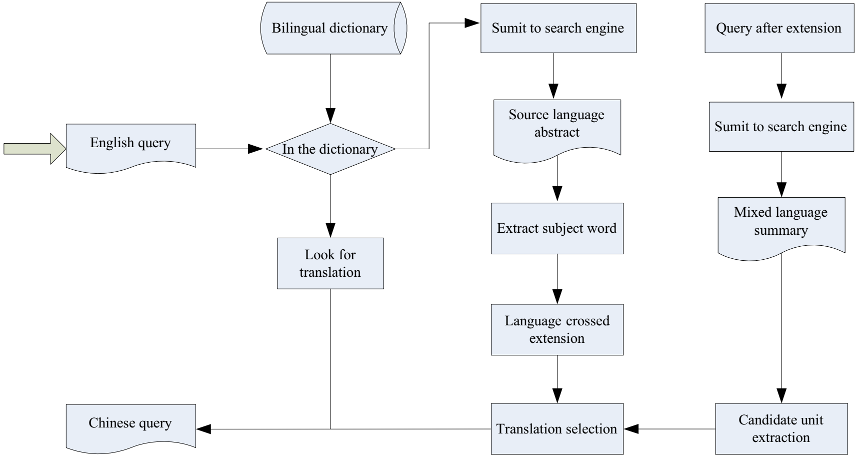

Figure 2. System structure of query translation method based on search engine

B. *Automatic Acquisition of Bilingual Abstract Resources*

In order to improve effectiveness and correlation of bilingual abstract resources, this paper adopts co-occurrence information-based key words translation query expansion for extracting abstract. Co-occurrence information-based key words translation query expansion is based on such discoveries. It is set up on the basis of this basic hypothesis. If it appears in large-scaled corpus, these two words usually co-occur in the same window unit of document. Therefore, these two words are mutually correlated in meaning. The higher the co-occurrence probability, the closer the mutual correlation.

Based on words co-occurrence figure in this paper, we use co-occurrence strength between words as basis of graph partition to divide graph into non-connecting clusters so that there is not connection between clusters but there is connection inside clusters. Then, one cluster is the author’s connecting sub-graph of his basic idea so as to form author’s certain specific ideas. At the same time, author’s global theme is formed by multi-themes in semantics and different positions of text. Therefore, we also depend on connecting words between clusters to describe author’s connecting feature in different positions. Thus, we also depend on conjunctions between clusters to describe author’s connecting features between different themes. This paper applies this idea to automatically extract key words in text. This method will be discussed in detail:

Step 1: The generating words co-occur node set Dhf in graph G. To input one text D which will extract key words, if it is Chinese text, words will be segmented. If it is an English text, it will be performed its root. Then, according to stop-word table to eliminate stop-words in text, text D is further converted into word set D_{terms}. Meanwhile, this word set D_{terms} will be sorted according to words frequency. Finally, previous n high-frequency words from D_{terms} in order will be taken as node set D of words co-occurrence G, that is, the full black nodes as shown in figure 3.

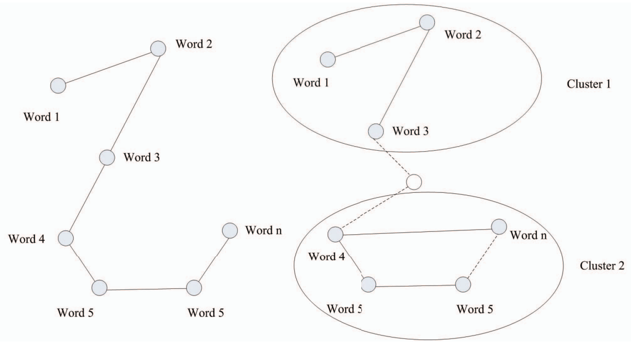

Figure 3. Singly connected G and multiply connected G

Step 2. To set up boundary between nodes in words co-occurrence graph G. Boundary establishment is based on co-occurrence degree between words in words set to be for boundary connection. If the co-occurrence degree between node corresponding words is large enough, their boundary will be connected. Meanwhile, it only gets node n– 1 connecting boundaries in graph G because n–1 is the minimal value which forms connecting graph. If words co-occurrence graph G is a single connection graph, this indicates that the theme in this text is single without multi-themes. If graph G is not a single connecting graph, it means this graph G will be segmented into multi-connected part to form cluster. Each cluster corresponds one theme but there is conjunctions between clusters. This word has an important function in relatively low word frequency but can connect different clusters. Furthermore, this word can be used to describe connecting features between different themes.

C. Candidate Translation Unit Extraction

Many English-Chinese vocabulary is one-to-many or many-to-many. Meanwhile, since there is no natural interval between words in Chinese text, in order to process words in Chinese texts, words must be separated. Words separation will inevitably appear errors especially some technical terms. Since the covering area of words separation dictionary is not enough, most of these words cannot be correctly separated. In order to eliminate effect of multi-words correspondence and words segmentation errors, N-gram model is used to extract candidate translation unit. At first, Chinese text and English text are respectively performed POS tagging. Stop-words are used to segment sentence into chunks which are performed N-gram extraction. That is, approximate N words combination of each word and its chunk can be taken as candidate translation units. Here, stop-words refer to those words with weak word-building ability and high-frequency.

Through the extracting method of N-gram candidate translation unit, there are many candidate translation units which cannot form reasonable syntax mode. Because of this, we develop a filter. This filter uses a syntax mode table to check whether sequence of POS in multi-words structure makes up a reasonable syntax mode. If it cannot make up reasonable syntax mode, it will be deleted. If it makes up a reasonable mode, it can be taken as a multi-words combination structure. Figure 4 depicts a syntax mode of filter. The left part in figure is Chinese syntax mode while the right one is English syntax mode.

“a+n”	“NN+NN”
“b+n”	“NN+NNS”
“n+n”	……
…	“NN+IN<of>”
“MWU+n”	“JJ+NN”
“n+MWU”	...
“MWU + MWU”	“MWU+ MWU”

Figure 4. Syntactic pattern in filter

D. Translation selection

1) Transliteration model

In order to avoid double errors of conversion from English to phoneme, from phoneme to pronunciation, and from pronunciation to Chinese characters, English is directly converted into syllables and then similarity between each syllable and each Chinese character is calculated. According to heuristic principle, after English OOV is separated into syllable, below formula is used to calculate similarity between OOV and one candidate unit.

$$Trl(s,t) = P(s,t) / D(s,t) \qquad (1)$$

s denotes certain OOV and t denotes any candidate unit. The numerator is co-occurrence probability of s and t, and the denominator is the number of different syllables in s and t. The definition of $P(s,t)$ is :

$$P(s,t) \approx \prod_{i=1}^{\min(m,n)} (1-\gamma) prov(e_i, c_i) \qquad (2)$$

γ is smoothing factor; $prov(e_i, c_i)$ is matching probability of English syllable e_i and Chinese character c_i, which is acquired form the training corpus containing all the English-Chinese transliteration pairs.

The definition of $D(s,t)$ is:

$$D(s,t) = \varepsilon + | m - n | \qquad (3)$$

ε is attenuation parameter; m is the total number of English syllable; n is the total number of one candidate unit.

To avoid the error of syllable separation, OOV is separated by positive and reverse direction. The average value under two separations are taken as the similarity of candidate unit.

2) Statistical model

In given bilingual abstract sets, the genuine translation version usually appears with source language words and they have similar frequency. Next, if the distance between one candidate unit and source language words is nearer, its possibility to become genuine translation version is larger. Similarity between source language words and one candidate unit adopts below formula for calculation:

$$F\_Q(s,t) = \frac{\sum_j \sum_k \frac{1}{d_k(s,t)}}{\max_{fre-dis}} \qquad (4)$$

j is the total number of abstracts; k the times of co-occurrence of s,t, since they may occur multiples times in one abstract; $d_k(s,t)$ is the distance of k_{th} co-occurrence s,t in one abstract.

3) *Module match model*

It is extremely important for signal information between OV and its corresponding translation version to extract translating version and this can apply these information to improve the extracted translation version quality. First some of English-Chinese word pairs are submitted to automatic learning surface template of search engine. If one candidate translation unit and OOV match most templates, it has large probability to be taken as correct translation. The matching contribution value of template is:

$$SP(s,t) = N_{matching} / Max_{num} \qquad (5)$$

IV. Performance Evaluation

The experimental test set is acquired in NTCIR4 and NTCIR5 cross-language information search task and there are totally 110 queries. From these query English title fields, there are totally 129 words which cannot be searched in bilingual dictionary. That is, there are totally 128 OOV and each OOV corresponding Chinese part is taken as correct translation reference. NITCIR query is similar to practical users and this is used to test translation with method in this paper for practical query.

To verify co-occurrence information-based effectiveness of key words translation version query expansion, this paper randomly selects 50 OOVs from OOV set and adopts expansion method in this paper to collect the previous 50 page abstracts from search engines towards each OOV under condition without expansion. Each method respectively collect 2500 page abstracts. If one abstract contains its corresponding OOV target language translation information, this abstract is effective. Otherwise, if this abstract only contains source language information or does not contain target language translation of OOV, this abstract is considered to be invalid. Statistical situation of effective abstracts from two abstract sets under these two conditions is shown as table 1.

Table 1.The statistics of efficient abstract in different methods

Method	The number of efficient abstracts
Without extension	331
With extension	996

From table 2, we can see that abstract quality of expansion method in this paper is higher than abstract quality of direct obtaining without expansion and it obtains 635 more effective abstracts. Expansion method in this paper can collect more abstract resources relating to the mining translating OOV in obtaining stage and this fully proves the effectiveness of expansion method in this paper. Bilingual abstract resources in high quality lay a solid foundation for extracting multi-words candidate units.

We adopt 50, 100, and 150 webpage abstract to extract the translations. Inclusion rate of T1 increases from 51 .71 % to 60 .24 and 62 .06%; Inclusion rate of T2 increases from 84 .05 % to 91 .7 %和 95 .34 %. When using 150 webpages the highest performance can be acquired, while the defects of more page abstracts are more network width and extraction time of more candidate units.

We also extract through local bilingual corpus which is set up by WEB mining with method in this paper. Figure 5 depicts the results in the form of graph. The result is clear that search engine-based obtaining translation is superior to corpus-based acquisition translation pair.

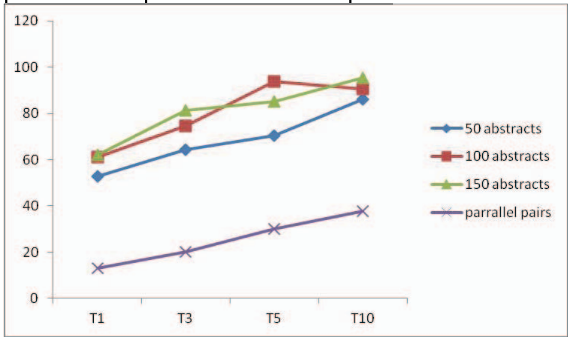

Figure 5. Acquisition results of translation pairs

V. Conclusions

It is an effective way to construct bilingual corpus and obtain translating knowledge on how to automatically construct a large-scaled bilingual parallel corpus, utilize the mining Chinese-English parallel corpus for extracting Chinese-English translation equivalences and apply search engine for translation version mining. This paper adopts co-occurrence information-based key words expansion method to collect effective bilingual abstract resources for cross-language expansion OOV. Then, it adopts the improved frequency change information and adjacent information-based candidate translating unit extraction to extract high-quality multi-words candidate units. Finally, OOV translation version is selected with multi-models. Experiment proves that various methods integrity is effective to check translation. Local resources can be taken as search engine method-based effective supplement and can also acquire higher search performance in comparison with the similar method.

References

[1] Lu YaJuan, Li Sheng, Zhao TieJun. Automatically acquiring Chinese parsing knowledge based on a bilingual language model. Chinese Journal of Computers, 2003, 26: 32-38

[2] Liu PengYuan, Zhao TieJun, Li Sheng. Resolving error accumulation of automatically acquiring bilingual lexical knowledge by semantic similarity. Journal of Harbin Engineering University, 2006, 27:575-579

[3] Zhang Jie. A Narrative Research of the Effectiveness of Teaching Strategies in Autonomous Learning of English Majors, Journal of Nanchang College of Education, 2016, 6: 59-61.

[4] Hashemi Homa B., Shakery Azadeh. Mining a Persian-English comparable corpus for cross-language information retrieval. Information Processing and Management, 2014, 50:384-398