Study of Query Translation Dictionary Automatic Construction in Cross-Language Information Retrieval

Su-Mei Xi[1] and Young-Im Cho[2]

[1] College of Information,
Shandong Polytechnic University
3501 University-ro, Changqing-gu, Jinan 250-353, China
xsm@ suwon.ac.kr
[2] College of Information Technology,
University of Suwon
San 2-2, Bongdam-eup, Hwaseong-si, 445-743, Korea
ycho@suwon.ac.kr

Abstract. The bilingual machine-readable dictionary is the commonly-used resources for query and translation based on the cross-language information retrieval; however, the traditional method of constructing the bilingual dictionary manually wastes time and energy. This paper uses the method of statistics to automatically obtain the translation dictionary from the English-Chinese parallel corpus for query and translation.

Keywords: cross-language, information retrieval translation dictionary.

1 Introduction

Cross-language information retrieval (CLIR) tries to identify relevant documents in a language different from that of the query. Its main problem is matching between query and documents of different languages. At present the main approach is to add language conversion mechanism (query translation or document translation) on the basis of monolingual information retrieval system [1].

2 Principle of Automatic Construction of Query Translation Dictionary

Based on the existing ambiguity problem-solving approach, we follow the following principles in constructing a query translation dictionary.

- Part of speech information term marked. There are many words which have more than one part of speech in natural language, and different part of speech generally means different meaning [2]. Polysemy problem can be solved to some degree by combining part of speech information to translate query.

S. Lee et al. (Eds.): Intelligent Autonomous Systems 12, AISC 194, pp. 585–592.
springerlink.com

- Provide phrase-level translation. Average precision can be increased 25% when translating query using phrase unit compared to the word unit, but the quality of phrase translation will largely affect the retrieval results
- Provide translation of named entities as detailed as possible [3].
- Provide the using information of words [4].

3 Query Translation Dictionaries Automatically Construction Based on Statistics

The method of translation dictionary automatically construction based on sentence aligned parallel corpus can be divided into five steps, specific process shown in Figure 1. The core issue what will be solved during the translation dictionary construction is acquisition of the candidate translation unit (including words unit and phrases) and generation of the translation equivalent pairs.

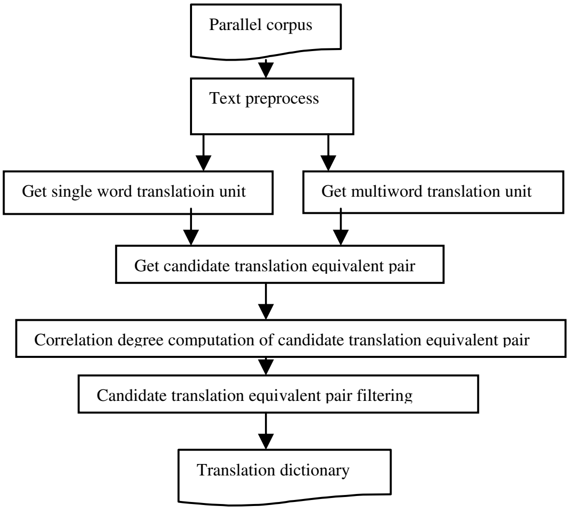

Fig. 1. Flowchart of translation dictionary automatically construction

3.1 Preprocess

The main task of preprocess is to process sentence aligned English-Chinese parallel corpora, including word segmentation and part of speech tagging of Chinese corpora and part of speech tagging of English corpora. The tools of word segmentation and part of speech tagging both are open-source toolkit developed by Stanford University [5]. Here is a sentence pair after the above processing.

English: Making _VBG sure_RB the_DT column_NN is_VBZ used_VBN for_IN unique_JJ identification_NN ._.
Chinese: 请_VV 确保_VV 唯一_JJ 标识_NN 列_NN 。_PU

Among them the identification after "_" is part of speech, here we use the Penn Treebank Tag Set.

3.2 Obtain the Candidate Translation Unit

Considering the roles of part of speech and phrase identification in word sense disambiguation and affecting query translation, nouns and verbs will be obtained separately when obtaining the candidate translation units and noun phrases will be identified.

1) Single word. Firstly, Single word translation unit of noun and verb can be obtained using the results of word segmentation and part of speech tagging. The reason which selects these two parts of speech as candidate translation unit is that these two parts of speech ratio are high among the queries and they also determine the main meaning of queries. Secondly, we can filter stop words to generated verb candidate translation unit and delete those unmeaning translation units such as ”可以”、”应该”、”能够”、”be”、” have、has、had、s、re、ve” and so on.

2) Noun phrase. Recognizing the noun phrase by using part of speech pattern constraint method, generating candidate noun phrase, calculating the correlation degree of all of the adjacent word pairs of candidate noun phrase by combining statistic information, if the adjacent words’ correlation degree is lower than a given threshold value in this noun phrase, this phrase will be deleted and the final noun phrase translation unit will be obtained.

Firstly, defining some noun phrase part of speech patterns using linguistic knowledge, combining result of text part of speech tagging, the candidate noun phrase will be extracted. The part of speech patterns in this paper includes the followings:

AN, NN, AAN, ANN, NNN, NAN, ANNN, AANN, AAAN, NNNN.

Where A is adjective, N is noun, the longest length of noun phrase is 4 and the shortest is 2.

We can use statistic information to filter the candidate noun phrase. The detailed process is as follows: assuming one candidate binary noun phrase includes two words N_1,N_2, counting the frequency of phrases N_1N_2 and single word N_1 and N_2 appearing in corpora, and then calculate the correlation degree of N_1,N_2 using formula Log Likelihood Ratio(LLR). The reason of selecting LLR coefficient is that this formula can process the correlation strength of low-frequency pair, the correlation degree calculation formula is as follows:

$$LLR(N_1,N_2) = 2[\log L(p_1,a,a+b) + \log L(p_2,c,c+d) - \log L(p,a,a+b) - \log L(p,c,c+d)] \quad (1)$$

Where a= freq(N_1, N_2), denotes the frequency of binary noun phrase N_1N_2 in the corpora, b= freq(N_1)-freq(N_1, N_2),denotes the number of sentence of N_1 appearing but N_2 disappearing, c=freq(N_2)-freq(N_1, N_2), denotes the number of sentence of N_2 appearing but N_1 disappearing, d=N-a-b-c, denotes the number of sentence of N_1 and N_2 both disappearing, and N denotes is total number of sentence of corpora. $\log L(p,k,n) = k\log(p)+ (n-k)\log(1-p)$, $p_1 = a/(a+b)$, $p_2= c/(c+d)$, $p= (a+c)/(a+b+c+d)$, $\log(0)=0$.

According to the calculated LLR value, for each noun phrase $N_1,\ldots,N_k$ ($2<= k<=4$), if the LLR value of all binary phrases $N_{i-1}N_i$ included in this noun phrase are greater than threshold value α, this phrase will be regarded as a multi-word units, otherwise, if existing $N_{i-1}N_i$, whose LLR value is lower than threshold value α, this phrase will be deleted from candidate phrase list. And then stemming for all obtained English translation units by using Porter Stemmer, the final candidate translation unit will be generated. Table 1 is a candidate translation unit from two example sentence of 3.1.

Table 1. case of candidate translation unit extraction

	English candidate translation unit	*Chinese candidate translation unit*
Noun	Column, ident Unique ident	标识列，唯一标识 唯一标识列
verb	Make us	确保

3.3 Generate Translation Equivalent Pairs

In this section we obtain candidate translation equivalent pairs according to English-Chinese translation unit, calculate the correlation degree of translation equivalent pairs, filter the expect value and English-Chinese Chinese-English dictionaries, and generate the final translation dictionary.

1) Obtain Candidate Translation Equivalent Pairs

We omit the length of candidate translation unit when generating noun candidate translation equivalent pairs. Only if pair of noun or noun and noun phrase appears in a pair of bilingual sentence they will be regarded as candidate translation equivalent pairs. The process of verb candidate translation equivalent pairs uses same approach.

2) Correlation Degree Calculation

Firstly count appearing frequency of all candidate translation equivalent pairs and each candidate translation unit, delete the candidate translation equivalent pairs whose co-occurrence frequency is less than 5, and then calculate the correlation degree of each translation equivalent pair.

There are four common formulas about calculating translation equivalent pair correlation degree, including LLR, Dice coefficient mentioned above, also including MI andΦ^2 coefficient. We use these four methods to calculate candidate translation equivalent pair correlation degree in order to comparing them in this paper. The detailed formulas are as follows:

$$MI= \log[a/(a+b)\ (a+c)] \tag{2}$$

$$Dice(cp,ep) = 2a\ /(2a+b+c) \tag{3}$$

$$\Phi^2 = (ad-bc)^2/\ [(a+b)\ (a+c)\ (b+d)\ (c+d)] \tag{4}$$

Where a= freq(cp, ep), denotes the number of sentence pairs which including Chinese candidate translation unit cp and English candidate translation unit ep, b=freq(cp)-freq(cp,ep),denotes the number of sentence pairs which only including cp but not including ep, c= freq(ep)-freq(cp, ep), denotes the number of sentence pairs which only including ep but not including cp,, d= N-a-b-c, denotes the number of sentence pairs which not including cp and ep, and N denotes is total number of sentence pairs of corpora.

3) Filtering

We descending sort all English translation items of each Chinese candidate translation unit according to its correlation degree of Chinese translation unit, generating Chinese-English bilingual dictionary and English-Chinese bilingual dictionary. Final translation dictionary can be obtained through expect value filtering, Chinese-English English-Chinese dictionary combining filtering, deleting some translation equivalent pairs.

4 Experimental Result and Analysis

The corpora of experiment is English-Chinese parallel corpora of computer field, which come from web sites and has been processed by sentence aligned, and it includes 300000 sentence pairs. The total number of bytes is 87 612 118.

According to above method, translation equivalent pairs can be extracted automatically from corpora and then two translation dictionaries about English-Chinese dictionary and Chinese-English dictionary will be generated. The detailed information about dictionaries is shown as table 2.

It can be seen from table 2 that generally not only total tokens but also number of equivalent pairs are the least, which generated by LLR coefficient, but they are the most which generated by MI coefficient. Furthermore, the number of noun tokens and equivalent pairs are similar whatever it generated by each of four coefficients. However, the case of verb is very different.

Table 2. Statistic result of generated dictionary

	coefficient	*Total noun tokens*	*Noun phrase tokens*	*Total noun equivalent pairs*	*Verb tokens*	*Verb equivalent pairs*
Chinese-English dictionary	LLR	6015	2831	8195	1413	1906
	Φ^2	6431	3015	8844	2148	3231
	Dice	7501	3328	13345	3251	10380
	MI	8424	4338	12417	4314	11113
English -Chinese dictionary	LLR	5734	3032	8195	1161	1853
	Φ^2	6215	3414	8844	1707	3130
	Dice	7070	3828	13345	2129	10129
	MI	7576	4484	12417	2480	10878

150 tokens have been selected randomly from the dictionary, including 50 noun phrase tokens and 50 verb tokens. The translation equivalent pairs that come from four statistic formulas based on co-occurrence have been evaluated separately. For calculation of precision of translation dictionary, the following method is used in this paper.

Precision=(number of equivalent pairs of correct translation in dictionary+ 0.5×number of equivalent pairs of partly correct translation in dictionary)/the total number of translation equivalent pairs.

According to above method, we can obtain noun word unit, verb and noun phrase evaluation results, shown as table 3, table 4 and table 5.

Through analyzing table 3, table 4 and table 5, the following conclusions can be drawn.

1) Generally speaking under the corpora environment this paper used, the statistical property of LLR coefficient is better than the other three coefficients, and Φ^2 is better than LLR only for English-Chinese noun phrase equivalent pairs. The reason is that the correct number (43) of English-Chinese noun phrase equivalent pairs is less than Φ^2 (45), but the wrong number (2) is greater than Φ^2 (1).

Table 3. Evaluation result of noun word unit equivalent pair

		Correct equivalent pairs	*Partly correct equivalent pairs*	*Total equivalent pairs*	*precision*
Chinese-English	LLR	143	61	206	0.8792
	Φ^2	144	71	215	0.8758
	Dice	154	92	284	0.7042
	MI	142	100	258	0.8235
English -Chinese	LLR	149	41	206	0.8228
	Φ^2	150	40	214	0.7943
	Dice	165	64	261	0.7548
	MI	155	70	270	0.7037

Table 4. Evaluation result of verb equivalent pair

		Correct equivalent pairs	*Total equivalent pairs*	*precision*
Chinese-English	LLR	50	68	0.7353
	Φ^2	57	79	0.7216
	Dice	83	147	0.5646
	MI	87	147	0.5918
English -Chinese	LLR	61	77	0.7922
	Φ^2	76	107	0.7103
	Dice	116	222	0.5225
	MI	118	231	0.5108

2) For noun dictionary, the main factor which affects dictionary precision is indirectly related problems, i.e. some idiomatic and phrases may make some bilingual words that not responds directly having high co-occurrence frequency. For example, in generated noun dictionary, “系统“ has two translations, one is “system” and the other is “oper”, among which “oper” means ”Operating”. Because ” Operating System” is a fixed phrase in computer field ”系统” and ” Oper” will be extracted as translation equivalent pair.

Table 5. Evaluation result of np equivalent pair

	coefficient	***Correct equivalent pairs***	***Partly correct equivalen t pairs***	***Wrong equivalent pairs***	***Precision of multiword unit***
Chinese-English	LLR	45	35	0	0.7813
	Φ^2	45	37	0	0.7743
	Dice	49	46	1	0.7500
	MI	48	43	2	0.7473
English -Chinese	LLR	43	26	2	0.7887
	Φ^2	45	27	1	0.8013
	Dice	48	34	3	0.7647
	MI	49	36	3	0.7614

Although it will be regarded as wrong translation equivalent pair when evaluating dictionary, during CLIR process if “System” and ”Oper” are submitted to the system as translation of ”系统”, the final retrieval effect may be high for it equivalent to the query expansion, increasing retrieval contextual information.

3) All equivalent pairs that partly correct translation in noun dictionary can play a query expansion role in CLIR.

4) The main reason about the number of verb token equivalent pairs of MI and Dice is much larger than LLR andΦ^2 is that the filtering is more loose of MI and Dice coefficient, which reserving lots of wrong translation items. The detailed analysis is as follows:

In the sample set of Chinese-English verb dictionary generated by MI and LLR coefficient, correct translation equivalent pairs are 87 and 50, ratio is 87 /50 =1.74. Wrong translation equivalent pairs are 60 and 18, ratio is 60/18= 3.33. The total equivalent pairs are 147 and 68, ratio is 147/68= 2. 16. 3.33> 2.16> 1.74, i.e. the ratio of wrong translation equivalent pairs > the ratio of total equivalent pairs > the ratio of correct translation equivalent pairs. The result is same about English-Chinese verb dictionary.

5 Conclusions

This paper mainly explores how to construct translation dictionary which suit for cross-language information retrieval query translation, summarizing the characteristics of translation dictionary which suit for CLIR query translation, automatically constructing a query translation according these characteristics, comparing their performance of four common statistic models based on co-occurrence. By analyzing experimental result, we find that the ambiguity problem which CLIR faced can be solved in some degree by editing query translation based on corpora.

References

1. Chen, A., Kishidak, Jiang, H.: Automatic construction of a Japanese-English lexicon and its application in cross-language information retrieval (2008), http://www.clis.umd.edu/conferences/midas/papers/chen.ps.
2. Davism, D.: A trec evaluat ion of query translation methods for multi-lingual text retrieval (September 15, 2008), http://trec.nist.gov/pubs/trec4/papers/nmsu.psgz
3. Lin, C.-H., Chen, H.: An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) Documents (December 15, 2008), http://dlist.sir.arizona.edu/526/01/lin.pdf
4. Davism: New experiments in cross-language text retrieval at NM SU's computing research lab. In: Proceedings of TREC-5. Gaithersburg, pp.447–453 (1997)
5. Bdksterosl, C.W.B.: Dictionary-based methods for cross-lingual in formation retrieval. In: Proceedings the 7th International DEXA Conference on Database and Expert Systems Applications, Zurich, Switzerland, pp. 791–801 (1996)
6. Ballesterosl, Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Develpment in Information Retrieval (SIGIR 1997), USA, Philadelphia, pp. 84–91 (1997)
7. Demner-Fushmand, O.: The effect of bilingual term list size on dictionary-based cross-language information retrieval (May 30, 2007), http://ieeexplore.ieee.org/iel5/8360/26341/01174250.pd?ftp=&isnumber=&arnumber=117425
8. http://nlp.stanford.edu/software/index.shtml