Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2, August 2003, pp. 77-92

Preparatory Work on Automatic Extraction of Bilingual Multi-Word Units from Parallel Corpora

Boxing Chen*, Limin Du*

Abstract

Automatic extraction of bilingual Multi-Word Units is an important subject of research in the automatic bilingual corpus alignment field. There are many cases of single source words corresponding to target multi-word units. This paper presents an algorithm for the automatic alignment of single source words and target multi-word units from a sentence-aligned parallel spoken language corpus. On the other hand, the output can be also used to extract bilingual multi-word units. The problem with previous approaches is that the retrieval results mainly depend on the identification of suitable Bi-grams to initiate the iterative process. To extract multi-word units, this algorithm utilizes the normalized association score difference of multi target words corresponding to the same single source word, and then utilizes the average association score to align the single source words and target multi-word units. The algorithm is based on the Local Bests algorithm supplemented by two heuristic strategies: excluding words in a stop-list and preferring longer multi-word units.

Key words: bilingual alignment; multiword unit; translation lexicon; average association score; normalized association score difference;

1. Introduction

1.1 The Background of Automatic Extraction of Bilingual Multi-Word Units

In the natural language processing field, which includes machine translation, machine assistant translation, bilingual lexicon compilation, terminology, information retrieval, natural language generation, second language teaching etc., the automatic extraction of bilingual multi-word units (steady collocations, multi-word phrases, multi-word terms etc.) is an

* Center for Speech Interaction Technology Research, Institute of Acoustics, Chinese Academy of Sciences
Address: 17 Zhongguancun Rd. Beijing 100080, China
E-mail: {chenbx , dulm}@iis.ac.cn

important aspect of the automatic alignment of bilingual corpus technology. Since the 1980's, the technique of automatic alignment of a bilingual corpus has undergone great improvement; and during the mid- and late-1990's, many researchers began to research the automatic construction of a bilingual translation lexicon [Fung 1995; Wu *et al.* 1995; Hiemstra 1996; Melamed 1996 etc.] Their works have focused on the alignment of single words. At the same time, the extraction of multi-word units in singular languages has been also studied. Church utilized mutual information to evaluate the degree of association between two words [Church 1990]; hence, mutual information has played an important role in multi-word unit extraction research, and it is used most often with this technology by means of a statistical method. Many researchers [Smadja 1993; Nagao *et al.* 1994; Kita *et al.* 1994; Zhou *et al.* 1995; Shimohata *et al.* 1997; Yamamoto *et al.* 1998] have utilized mutual information (or the transformation of mutual information) as an important parameter to extract multi-word units. The shortcoming of these methods is that low frequency multi-word units are easy to eliminate, and the output of extraction mainly depends on the verification of suitable Bi-grams when the iterative algorithm initiates.

Automatic extraction of bilingual multi-word units is based on the automatic extraction of bilingual word and multi-word units in singular languages. Research in this field has also proceeded [Smadja *et al.* 1996; Haruno *et al.* 1996; Melamed 1997 etc], but the problem with this approach is that it relies on statistical methods more than the characteristics of the language per se and is mainly limited to the extraction of noun phrases.

Because of the above problems and the fact that Chinese-English corpuses are commonly small, we provide an algorithm that uses the average association score and normalized association score difference. We also apply the Local Bests algorithm, stopword filtration and longer unit preference methods to extract Chinese or English multi-word units.

1.2 The Object of Our Research

In research on the results produced by single-English-word to single-Chinese-word alignment, we have found an interesting phenomenon: During the phase of Chinese word segmentation, if the translation of an English word ("*A*") comprises of several Chinese words ("*BCD*"), the mutual information and the t-score for each "*B-A, C-A, D-A*" mapping are both very high and close to each other. Thus, we can use the average association score and the normalized association score difference to extract the translation equivalent pairs of single-English-word to multiple-Chinese-word mappings.

For example, when names and professional terms are translated, "*Patterson*" is translated as "佩特逊," which includes three entries in a Chinese dictionary ("佩," "特," and "逊"); "*Internet*" is translated as "因特网," which includes three entries in a Chinese dictionary

(“因,” “特,” and “网”). Furthermore, the same situation occurs with some non-professional terms. For example, “*my*” is translated as “我的.” Also, the same rule applies to Chinese-English translation. For example, “不三不四” is translated as “*get funny*,” and “放肆” as “*get fresh.*”

Therefore, the research presented in this paper is focused on single-source-word to multi-target-word-unit alignment. The alignment of bilingual multi-word units will be the focus of our future research.

2. Algorithm

The method we use to align single source words with target multi-word units from a parallel corpus can be divided into the following steps (we use the mutual information and t-score as the association score):

(1) Word segmentation:

We do word segmentation first because Chinese has no word delimiters.

(2) Calculating the co-occurrence frequency:

If a word pair appears once in an aligned bilingual sentence pair, one co-occurrence is counted.

(3) Computing the association score of single word pairs:

We calculate the mutual information and t-score of the source words and their co-occurrence target words.

(4) Calculating the average association score and normalized association score:

We calculate the average mutual information and normalized mutual information difference, and the average t-score and normalized t-score difference of every source word and its co-occurrence target words’ N-gram (N: 2-7, since most phrases have of 2-6 words).

(5) The Local Bests algorithm:

We utilize the Local Bests algorithm to eliminate non-local best target multi-word units.

(6) Stop-word list filtration:

Some words cannot be used as the first or the last word of a multi-word unit, so we use the stop-word list to filter these multi-word units.

(7) Bigger association score preference:

After the above filtration, from among the remaining multi-word units, we choose N items with the maximal average mutual information and average t-score as the

candidate target translation.

(8) Longer unit preference:

We extract multi-word units but not words, so if the longer word string C_1 entirely contains another shorter word string C_2, then string C_1 is taken as the translation of the source word.

(9) Lexicon classification:

According to the above four parameters, we classify the lexicons into four levels of translation lexicons.

We will use "*Glasgow*: 格拉斯哥," which appears in the corpus as shown in Figure 1, as an example to explain the whole process.

(1.a) I'd like to fly to Glasgow on the fifth of May.
(1.b) 我想 5 月 5 日飞往格拉斯哥。
(2.a) Can I take this train to Glasgow?
(2.b) 我可以乘这次列车去格拉斯哥吗?

Figure 1. Sentence Example.

The reasons why we choose "*Glasgow*" are: (1) the occurrence frequency of "*Glasgow*" is quite low, only two times, which is easily ignored by the previous algorithm; (2) the Chinese translation of "*Glasgow*" is unique, so the correct extraction of this lemma can prove the accuracy of our algorithm; (3) "*Glasgow*" contains four single-character words, and it will be found later that our algorithm is more effective with multi-word units made up of two words, so here we use "*Glasgow*" to prove that our algorithm is also effective with multi-word units made up of more than two words.

2.1 Chinese Word Segmentation

We used the "maximum probability word segmentation method" [Chen 1999] and *The Grammatical Knowledge-base of Contemporary Chinese* published by Peking University [Yu 1998]. The idea behind this method is: first find out all the possible words in the input Chinese string on a vocabulary basis and then find out all the possible segmentation paths, from which we can find the best path (with the maximal probability) as the output. We randomly sampled 1000 sentences to check: if we did not take "un-listed words that are divided" as an error, then the precision rate was 98.88%; but if it was being taken as an error, the precision rate was 88.74%. The unlisted words in DECC1.0 (Daily English-Chinese Corpus) were mainly the Chinese translations of foreign personal names and place names. The main focus of our research here was the aggregation of single Chinese characters that are produced through

segmentation. The results of word segmentation are shown in Figure 2:

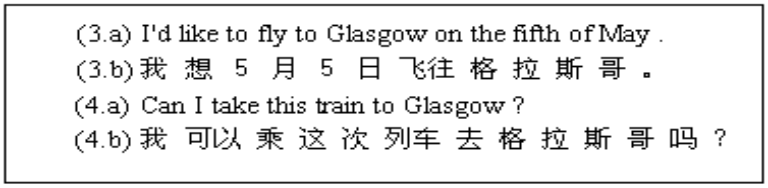

Figure 2. Word Segmentation Results.

2.2 Calculate the Co-occurrence Frequency

There were many translation sentence pairs in the corpus. For each possible word pair in these translation sentence pairs, the higher the probability of appearance it had, the higher the probability it had of being the correct translation word pair. We built a co-occurrence model to count the number of appearances: it was counted as a co-occurrence each time the word pair appears in a sentence pair. The reasons are as follows: First, the length of a sentence in spoken language is usually shorter than that in a written language; for example, in the corpus DECC1.0, the average length of English sentences is 7.07 words, and the average length of Chinese sentences is 6.87 words and expressions. Secondly, the corresponding sense units of English-Chinese sentence pairs in spoken language are not always aligned in terms of position, as shown in Figure 3.

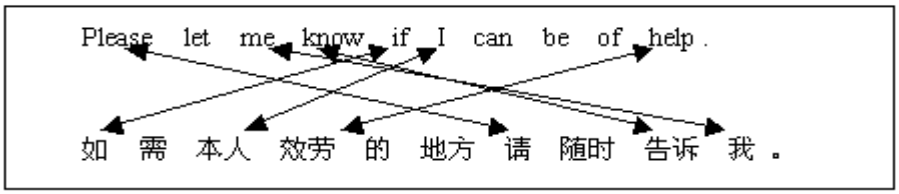

Figure 3. Example of Word Alignment.

2.3 Calculate the Mutual Information and T-Score

Having calculated the word pair's co-occurrence frequency and the frequency of every word, we use formulas (1) and (2) to calculate the mutual information $MI(S,T)$ and t-score $t(S,T)$ of any source word and its single target word. As for the association verifying score [Fung 1995], the higher the t-score, the higher the degree of association between S and T:

$$MI(S,T) = \log \frac{\Pr(S,T)}{\Pr(S)\Pr(T)}, \quad (1)$$

$$t(S,T) \approx \frac{\Pr(S,T) - \Pr(S)\Pr(T)}{\sqrt{\frac{1}{N}\Pr(S,T)}}. \quad (2)$$

Here, N is the total number of sentence pairs in the corpus, S is the source word, T is the target word, and *Pr(.)* is the probability of the source word or target word. For the "*Glasgow*" example, the outcome of Formula (1) is shown in Figure 4, and the outcome of Formula (2) is shown in Figure 5.

Glasgow:
哥：8.004633
格：6.723699
拉：6.669632
飞往：6.087710
列车：5.455188
乘：5.008900
日：4.793789
月：4.686817
斯：4.637337
次：3.518246
去：2.772455
可以：2.451673
想：2.194690
这：1.460433
吗：1.339204
我：0.794849

Figure 4. Mutual Information Score

Glasgow:
哥：1.413741
格：1.412514
拉：1.412419
斯：1.400519
飞往：0.997729
列车：0.995726
乘：0.993322
日：0.991719
月：0.990784
次：0.970349
去：0.937492
可以：0.913851
想：0.888607
我：0.775485
这：0.767864
吗：0.737946

Figure 5. T-Score.

2.4 Calculate the Average Association Score and its Normalized Difference

The Average Association Score (AAS) is the average association score of the source word and every word in the target language N-gram. It can measure the association degree between the source language and target language. The Normalized Difference (ND) is the normalized difference for the association score of the source word and every word in the target language N-gram. It can measure the internal association of the target multiword units. Therefore, we use the AAS and ND to build the association model of the single source word and target multiword units. We compute the average mutual information, normalized mutual information difference, average t-score, and normalized t-score difference of the consecutive Chinese word string N-gram (N: 2-7), which co-occurs with "*Glasgow.*" Vintar's research indicated that the

length of 95% of English phrases and Slavic phrases is between 2-6 words [Vintar *et al.* 2001], and from our experience, we can conclude that Chinese multiword units of more than 6 words are also very rare. To reduce the complexity of calculation, we only consider multiword units with 6 words or less. Suppose a Chinese word string C (chunk) is expressed by the following symbols:

$$C = W_1 W_2 ... W_i ... W_n . \quad (3)$$

Then the formulae of AMI (Average Mutual Information), MID (Mutual Information Difference), AT (Average T-score) and TD (T-score Difference) are as follows:

$$AMI\ (C,T) = \frac{1}{n}\sum_{i=1}^{n} MI\ (W_i,T), \quad (4)$$

$$MID\ (C,T) = \frac{1}{n \times AMI\ (C,T)}\sum_{i=1}^{n} |\ MI\ (W_i,T) - AMI\ (C,T)\ |, \quad (5)$$

$$AT\ (C,T) = \frac{1}{n}\sum_{i=1}^{n} t(W_i,T), \quad (6)$$

$$TD\ (C,T) = \frac{1}{n \times AT\ (C,T)}\sum_{i=1}^{n} |\ t(W_i,T) - AT\ (C,T)\ |. \quad (7)$$

Here, t(.) is the t-score, MI(.) is the mutual information, T is the target word. The results obtained using formulae (4)-(7) are shown in Table 1. (There were 108 outputs from each parameter; we chose only 16 that were connected with the correct answer "*Glasgow*" and could be used to explain the algorithm.)

2.5 Local Bests Algorithm

Currently, the algorithms for extracting multiword units are mainly based on setting a global threshold for some association score (mutual information, entropy, mutual expectation etc.), and if only the association score of the checked word string is bigger or smaller than that threshold, then the word string is considered to be a multiword unit. However, the threshold method has many limitations because the threshold will change with the type of language, the size of the corpus, and the difference of the selected association score, and because of the threshold cannot be easily chosen.

The Local Bests algorithm [Silva *et al.* 1999] is a more robust, flexible and finely tuned approach to the extraction of multiword units, which is based on the local context, rather than on the use of global threshold methods. If a word string (n-gram) is a multiword unit, there

should be stronger internal association, and the association score will be high. Also, as a local structure, a multiword unit can show the best association in a local context. Thus, when we find the association score of a word string that is high in a local context, we may consider it as a phrase. For example, there is a strong internal association within the Bi-gram <*ice*, *cream*>, i.e., between the words *ice* and *cream*. On the other hand, one cannot say that there is a strong internal association within the Bi-gram <*the*, *in*>. Therefore, let us suppose that there is a function S(.) that can measure the internal association of each n-gram.

Let Ω_{n-1} be the set of all the (n-1)-grams contained in the N-gram word string C (Chunk), and let Ω_{n+1} be the set of all the (n+1)-grams containing this N-gram word string C. Suppose the bigger the association score S(.), the better the result. The Local Bests algorithm can be described as follows:

Algorithm 1. Local Bests Algorithm

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$ if

(length(C) = 2 and S(C) > S(y)) or

(length(C) > 2 and S(x) $\leq$ S(C) and S(C) > S(y))

then word string C is a multiword unit.

Here, S(.) is the internal association score of the Multi-Word Units, and length (C) is the number of words included in C.

In our algorithm, it is better if AMI and AT are bigger, and if MID and TD are smaller; every n-gram of the local best co-occurring with "*Glasgow*" is shown in boldface in Table 1. As we can see in the table, the normalized mutual Information difference of "格拉斯哥" is not a global best score, but it is a local best score, so we may exclude this Multi-Word Unit if we use the global threshold but not the local best algorithm.

Table 1. AMI, MID, AT and TD of Chinese N-gram (N=2~7) co-occurring with "Glasgow."

	AMI	MID	AT	TD
飞往格	6.405704	0.049642	1.205121	0.172093
飞往格拉	6.493680	0.041679	1.274221	0.144659
飞往格拉斯	6.029595	0.115452	1.305795	0.117961
飞往格拉斯哥	6.424602	0.132251	1.327384	0.099340
格拉	**6.696666**	**0.004037**	**1.412466**	**0.000034**
格拉斯	6.010223	0.152283	1.408484	0.003770
格拉斯哥	**6.508825**	**0.143765**	**1.409798**	**0.003291**
格拉斯哥吗	5.474901	0.363350	1.275428	0.168565
拉斯	5.653485	0.179738	1.408469	0.004230
拉斯哥	6.437201	0.186402	1.408893	0.003962
拉斯哥吗	5.162702	0.421181	1.241156	0.202718
去格	4.748077	0.416089	1.175003	0.202136
去格拉	5.388595	0.323664	1.254142	0.168322
去格拉斯	5.200781	0.287627	1.290736	0.136838
去格拉斯哥	5.761551	0.285570	1.315337	0.114904
去格拉斯哥吗	5.024493	0.546907	1.219105	0.208561

There are still two main problems with using the Local Bests algorithm to extract multiword units: (1) A fraction of the extracted multiword units are not correct, such as "的传球" and "没法把," with improper words at the beginning or the end of a multiword unit; the same is true with English multiword units, such as "*and, or*" appearing at the beginning of a multiword unit, and "*the, may, if*" at the end of a multiword unit. (2) For a source word, several multiword units are extracted, but not all of them are correct translations.

We utilize a stop-word list to solve the first problem, and the methods based on the association score best and longer unit preference are used to solve the second.

2.6 Stop-word List Filtration

A stop-word is a word that cannot be used at the beginning or the end of a multiword unit. By analyzing the parts of speech and the characteristics of specific words arrangements, we manually create four types of stop-word lists: non-beginning and non-ending Chinese words, and non-beginning and non-ending English words. Samples of lists are shown in Table 2.

Table 2. Stopword List.

Stop-word List	Content
Non-beginning Chinese words	quantifier (个), auxiliary word (的), modal word (吗) etc. 267 words
Non-ending Chinese words	conjunction (和, 或者), part preposition (从) etc. 189 words
Non-beginning English words	part adverb (*not*), part conjunction (*and, or*) etc. 23 words
Non-ending English words	article (*the*), conjunction (*when*), aux verb (*ought to*), part pronoun (*my*) etc. 78 words

Using the stop-word lists to filter multiword units, we can the first problem mentioned above.

2.7 Association Score Best Filtration

The association score (mutual information and t-score) is a measure used to judge whether the source word and the target multiword unit are translations of each other, so if a source word corresponds to several target multiword units, then the target multiword unit with a higher association score is more likely to be a translation of this source word. Then we can choose from among the remaining multiword units after two filtrations and take N items with the maximal average mutual information and average t-score as the candidate target translations. According to the results of sample tests, after local bests filtration, the association score of the correct target translation is usually among the best three scores, so we assume that N equals 3.

2.8 Longer Units Preference

A short unit is more likely to be a word [Tanapong *et al.* 2000], but for the following reasons, we apply the Longer Units Preference: (1) Our algorithm determines that the multiword units of two words, especially the two words of the maximal association score with the source word, have the higher average association score and the lower association score difference. For example we can see that "格拉" is better than "格拉斯哥" based on four parameters. (2) We extract multiword units but not words, and if a longer word string has the local best result, then this word string is a comparatively steady structure. Therefore, if a longer words string C_1 entirely contains another shorter word string C_2, then string C_1 is taken as the translation of the source word. This method might choose Multi-Word Units that are longer than necessary, a situation we call "translation units expansion," but it is useful for the extraction of bilingual Multi-Word Units, and it is can be used in the phase of bilingual Multi-Word Unit extraction.

2.9 Lexicon Classification

Thus, the work of extracting a multiword unit translation of every source word is basically accomplished. There are four parameters used in the algorithm. The Average Association Score can measure the association degree between the source language and target language. The Normalized Difference can measure the internal association of the target multiword units. If a pair of bilingual word strings can match more parameters after Local Best and N-bests association score filtering, then it must have higher probability of being correct. Based on the four parameters, four bilingual lexicons are constructed, and they can be subjected to the merge application or intersection application according to different application requirements. We calculate four outcome tables using Formulae (4), (5), (6) and (7), each of them based on a certain measure. Then we pick translation word pairs from those four tables to form five lexicons. The 1^{st} level lexicon composed of word pairs which has appeared only once in the tables; the 2^{nd} level lexicon composed of word pairs which has appeared twice in the tables; and the same rule applies to the 3^{rd} and 4^{th} level lexicons. The higher level one word pair belongs to, the more precision it has. The 0^{th} lexicon is a union of the other four lexicons; that is, any word pairs that have appeared in the tables go into the 0^{th} lexicon. If a source word has several target entries, we calculate the co-occurrence frequency of every entry with the source word in the corpus and then normalize the probability of every entry.

3. Results and Analysis

3.1 Bilingual Corpus

The bilingual corpus we used was DECC1.0, which consists mostly of daily life dialogues, including 14,974 aligned bilingual sentence pairs and a total of 1,039,183 bytes. In this corpus,

there are 7,491 English word types and 7,344 Chinese word types.

3.2 Lexicon Evaluation

Taking English as the source language and Chinese as the target language, we provide an example of the 4th level lexicon and the 0th level lexicon in Figures 6 and Figure 7.

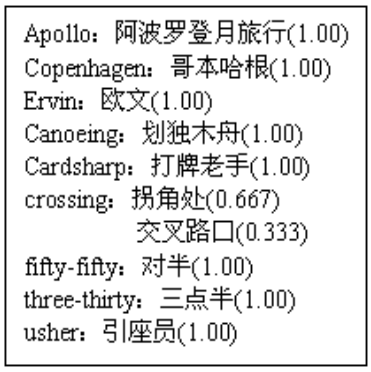

Figure 6. 4th level lexicon.

Figure 7. 0th level lexicon.

There is no uniform method for calculating the precision of translation lexicons, so we take the following approach: the corpus is the measure – if and only if the lexicon entry has an exact match in the corpus, it is taken as correct. For example, the meaning of "*fifty-fifty*" in the English-Chinese dictionary is "平分为二的, 对半地, 平分为二分地," and in the corpus the corresponding translation of "*fifty-fifty*" is "对半," so we consider that the translation "*fifty-fifty*: 对半" in Figure 6 is correct, but in Figure 7, "*Adam*: 亚当和夏娃" is considered to be incorrect because in the corpus, the pair is "*Adam*: 亚当." The recall rate is the number of English words in each lexicon divided by the number of all the English words in the whole corpus.

The F-measure is an important parameter for balancing precision and recall [Langlais *et al.* 1998]. Table 3 shows the precision, recall and F-measure results of the English-Chinese, Chinese-English 0~4 level lexicons. For lexicons that had more than 200 entries, we randomly chose 200 entries from each of them; for those that had less than 200 entries, we used all the entries for calculation:

$$F = 2\frac{recall \times precision}{recall + precision}. \quad (8)$$

Table 3. Precision and recall results of all levels of lexicons.

i^{th} level lexicons	precision（%）	Recall（%）	F-measure
0th E-C	41.394	98.63	0.583
1st E-C	23.535	84.22	0.368
2nd E-C	52.388	31.56	0.394
3rd E-C	78.323	5.18	0.097
4th E-C	94.900	1.36	0.027
0th C-E	38.266	96.94	0.549
1st C-E	18.943	82.58	0.308
2nd C-E	47.564	29.92	0.367
3rd C-E	75.092	7.54	0.137
4th C-E	88.293	2.83	0.055

"E-C" lexicons take the single-English-word as the source language and the multi-Chinese-word unit as the target language, and vice versa.

3.3 Analysis of the Result

By analyzing the precision and recall results, and the lemmas of all levels of lexicons, we reached the following conclusions:

(1) There are many lemmas satisfying one qualification (viz. the 1st level lexicon). Almost every English word and Chinese word and expression has at least one target word string satisfying the local best and other qualifications, but the precision of the 1st level lexicon is very low. This shows that (1) depending on a single qualification is not sufficient to construct a bilingual lexicon with high precision, and that (2) not every source word has a corresponding target phrase.

(2) Compared with the 1st level lexicon, the precision of the 2nd level lexicon is greatly increased. According to the sketchy statistics, the two qualifications satisfied by most of the correct portion of the 2nd level lexicon are mutual information and t-score, which shows that for a certain parameter (mutual information or t-score), simultaneously using the difference and average value can improve the results greatly.

(3) Compared with the 2nd level lexicon, the precision of the 3rd level lexicon is also greatly increased and recall is decreased, which shows that after one parameter has been satisfied, if a qualification of another parameter can be also satisfied, then the translation is very likely to be correct. In similar works, many other researchers needed to consider multiple parameters, and the selection of parameters was very important. From early works on word alignment and our current work on phrase extraction, we find that a combination of mutual information and t-score provides a reliable measure.

(4) Only a little manual collation work is needed to make the 4th level lexicon practical. The English-Chinese 4th level lexicon has only 98 lemmas, which, except for some common phrases with high appearance frequency, are mainly personal names, place names and specialized terms; and all of these terms have low appearance frequency, many occurring only once. This shows that for the extraction of low frequency phrases, our algorithm also is good.

(5) The higher the lexicon's level, the lower its recall rate. This shows that the cases of single source words corresponding to a target word string are comparatively few. On the other hand, it shows that our corpus is too small. If the corpus could be increased, the result would be better.

(6) There are cases of "translation unit expansion" in all levels of lexicons; for example, in the 4th level lexicon for "*Apollo:*阿波罗登月旅行," "*Apollo*" corresponds to "阿波罗," but there is only one sentence pair in which "*Apollo*" appears in the whole corpus (Figure 8). In addition, "阿波罗登月旅行" exists as a sense unit, so according to the longer units preference method, our algorithm selected "阿波罗登月旅行." It should be made clear that, although "*Apollo*: 阿波罗登月旅行" is an incorrect lemma, it provides a basis for constructing a translation lexicon in which the source language and the target language are both multi-word phrases. Especially in the 0th level lexicon, we can see that the two translations of "*moon*" are "阿波罗登月" and "登月旅行," from which, using a certain algorithm, we can extract the correct phrase "*Apollo's trip to the moon:* 阿波罗登月旅行," and this will be the focus of our future research.

E: Is it possible to obtain any information about Apollo 's trip to the moon ?
C: 有 可能 了解 些 有关 阿 波 罗 登 月 旅行 的 消息 吗 ?

Figure 8 Sentence pair in a corpus with "Apollo."

(7) Another fact that affects the precision is that the corpus we used contains 171 bilingual proverbs, and such sentence pairs can rarely be translated word for word, as demonstrated by the example shown in Figure 9.

E: You must reap what you have sown
C: 种瓜得瓜，种豆得豆。

Figure 9. Bilingual proverb.

4. Conclusion and Future Research

4.1 Conclusion

Because there are many cases of single source words corresponding to target multi-word units, for example, English personal names and place names, we have provided an algorithm for the automatic alignment of single source words and target multi-word units from a sentence-aligned parallel spoken language corpus, which makes a translation lexicon more practical. It will be of great help for machine translation, especially Chinese-English translation. On the other hand, the outputs can also be used to extract bilingual multi-word units. Compared with other similar researches, this algorithm differs in the following ways:

(1) It utilizes the normalized association score difference as the criterion for extracting phrases.

(2) It simultaneously uses the Local Bests algorithm, stop-word filtration, and the longer units preference method to extract phrases.

(3) Classify lexicon. Different levels of lexicons can be applied to obtain practical translation lexicons or can be used as the basis for further research.

Mutual information has been used in many other similar researches, but these processes are mainly based on algorithms of iterating the Bi-gram calculation, and the retrieval results mostly depend on the identification of suitable Bi-grams for the initiation of the iterative process. Errors can accumulate during the iteration process, thus greatly affecting the precision of multi-word phrase extraction [Dias *et al.* 2000]. Our algorithm solves this problem by calculating the normalized association score difference of the target words corresponding to the same source word. The use of t-score increases the precision of the phrase translation lexicon, and the classification of the lexicon reduces the number of the incorrect entries in the high level lexicon effectively, which makes the translation lexicon more practical.

4.2 Future Research Plan

Currently, “translation unit expansion” is a common problem, and we shall utilize the outcome to extract bilingual multi-word units in our future research.

Reference:

Chen, X.H. “Automatic Analysis of Contemporary Chinese Using Visual C++,” Beijing: Beijing Language and Culture University Press, (It's published in Chinese) 1999, pp.97-103.

Church, K.W. and P. Hanks. "Word Association Norms, Mutual Information & Lexicography." *Computational Linguistics,* 16(1) 1990, pp.22-29.

Dias, G., Guilloré,S. and Pereira L.J.G. "Normalization of Association Measures for Multiword Lexical Unit Extraction." *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA'2000).* Monastir, Tunisia, 2000, pp. 207-216.

Fung P. "A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora." *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Boston, USA. 1995, pp. 236-243.

Haruno M., Ikehara S. and Yamazaki T. "Learning Bilingual Collocations by Word-level Sorting. *COLING96.* 1996, pp. 525~530.

Hiemstra, D. "Using Statistical Methods to Create a Bilingual Dictionary." *Master's Thesis*, University of Twente. 1996.

Kita, K., Kato, Y., Omoto T. and Yano Y. "A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria." *Journal of Natural Language Processing*, 1 (1), 1994, pp. 21-33.

Langlais P., Simard M. and Véronis J. "Methods and Practical Issues in Evaluating Alignment Techniques." *Proceedings of COLING-ACL*, 1998, Montréal, Canada, pp. 711-717.

Melamed I. D. "Automatic Construction of Clean Broad-Coverage Translation Lexicons." *Conference of the Association for Machine Translation in Americas*, Montreal, Canada. 1996, pp. 125-134.

Melamed I. D. "Automatic Discovery of Non-Compositional Compounds in Parallel Data." *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing.* Providence, RI. USA. 1997, pp. 97-108.

Nagao, M. and Mori, S. "A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese." *Proceedings of the 15th International Conference on Computational Linguistics.* 1994. pp.611-615.

Sayori Shimohata, Toshiyuki Sugio and Junji Nagata "Retrieving Collocations by Co-occurrences and Word Order Constraints." *35th Conference of the Association for Computational Linguistics (ACL'97)*, Madrid, 1997, pp. 476-481.

Silva J.F., Dias G., Guillor S. and Lopes J.G.P. "Using Localmaxs Algorithm for Extraction of Contiguous and Non-contiguous Multiword Lexical Units." *9th Portuguese Conference in Artificial Intelligence, Lecture Notes, Springer-Verlag*, Universidade de Evora, Evora, Portugal, 1999, pp. 113-132.

Smadja, F. "Retrieving Collocations from Text: Xtract." *Computational Linguistics*, 1993. Vol.19, No.1. pp. 143-177.

Smadja F., McKeown K.R. and Hatzivassiloglou V. "Translation Collocations for Bilingual Lexicons: a Statistical Approach." *Computational Linguistics* 1996, 22(1), pp. 1~38.

Tanapong Potipiti, Virach Sornlertlamvanich and Thatsanee Charoenporn. "Towards Building a Corpus-based Dictionary for Non-word-boundary Language." *Workshop on Terminology Resources and Computation, Workshop Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece. 2000, pp. 82-86.

Vintar, Spela. "Using Parallel Corpora for Translation-Oriented Term Extraction." *Babel Journal*, John Benjamins Publishing. 2001.

Wu, D. and Xia, X. "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon." *Machine Translation* (4). 1995, pp. 285-313.

Yamamoto, M. and Church, K.W. "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus." *Proceedings of the 6th Workshop on Very Large Corpora*, Montreal, Canada, 1998, pp.28-37.

Yu, S.W. The Grammatical Knowledge-base of Contemporary Chinese, Beijing: Tsinghua University Press, (It's published in Chinese) 1998.

Zhou, J. and Dapkus, P. "Automatic Suggestion of Significant Terms for a Predefined Topic." *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge1995, pp.131-147.