Extracting of Translation Unit from Chinese-English Parallel Corpora

CHANG Baobao
Institute of Computational Linguistics
Peking University, P.R.China

Pernilla DANIELSSON and
Wolfgang TEUBERT
Centre for Corpus Linguistics
Birmingham University, United Kingdom

1. Background

The field of machine translation has changed remarkably little since its earliest days in the fifties. So far, useful machine translation could only obtained in very restricted domain. We believe one of the problems of traditional machine translation lies in how it deals with unit of translation. Normally Rule-Based Machine Translation system takes word as basic translation unit. However, word is normally polysemous and therefore ambiguous, which causes many difficulties in selecting proper target equivalent words in machine translation, especially in translation between unrelated language pairs, such as Chinese and English. On the other hand, human translation is rarely word-based. Human translators always translate group of words as a whole, which means human do not view words as the basic translation units, and it seems they view language expressions that can transfer meaning unambiguously as basic translation units instead. Following this observation, we believe translation unit shall be unambiguous words and words groups (Multi-Word Unit) and a collection of bilingual translation unit will be certainly a very useful resource to machine translation.

Manual compilation of such a database of translation unit is certainly labor intensive. But following the recent progress in Corpus Linguistics, especially in parallel corpus research[2][3][7]. Automatic identification of translation unit and its target equivalents from existed authentic translation might be a feasible solution; at least it can be used to produce a candidate list of bilingual translation unit.

As a first step towards building a database of bilingual translation units, we selected the Hong Kong Legal Documents Corpus (HKLDC) as the parallel corpus for the feasibility study. This paper elaborates the methods we adopted. We will first give our model of (semi-) automatic acquisition of bilingual translation unit based on parallel corpora in section 2. Then we will show how the corpus could be preprocessed in section 3. In section 4, several statistic measurements will be introduced which will serve as a basis for late steps in extracting of bilingual translation units. Section 5 will focuses on identification of multi-word units. Section 6 will describe how translation equivalents could be extracted. In section 7, we give some evaluation regarding to the performance in extracting the translation equivalent pairs.

2. Framework of automatic acquisition of bilingual translation unit

The whole process of identification of bilingual translation unit could be further divided into three major steps as depicted in Figure 1.

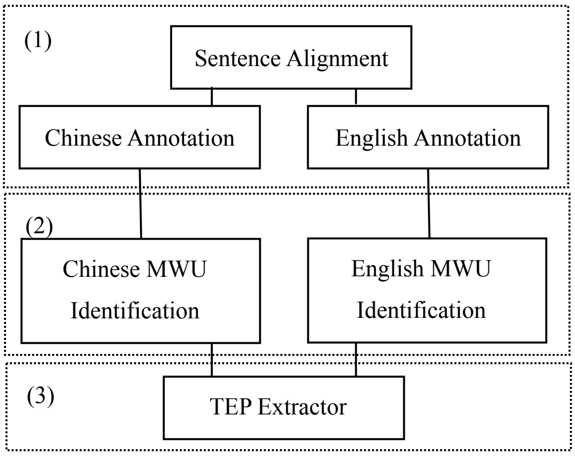

Figure 1. Framework of translation unit acquisition

(1) Preprocessing of the parallel corpora
For the purpose of extracting bilingual translation unit, some prior processing of the corpus is necessary. These include alignment of the bilingual texts at sentence level and unilingual annotation of the Chinese and English texts respectively.

(2) Identification of multi-word unit in the aligned the texts
As we mentioned before, translation unit shall not be only single words, but also multi-word units. In this step, Both the Chinese and English multi-word units are identified separately from the corpus.

(3) Extraction of the bilingual translation units
After identification of the multi-word units for texts of both languages, this step tries to set the correspondence between Chinese and English translation units. The result of this step will be a list of bilingual Translation Equivalent Pairs and every TEP is composed of a Chinese Translation unit and an English Translation unit.

3. Preprocessing the corpus

The Hong Kong legal documents were collected from Internet. The corpus is composed of laws and amendments issued by the Hong Kong Special Administration Region (HKSAR) during. All the texts in it are in both Chinese and English. We selected about 6 million words of both Chinese texts and English words (6,833,762 Chinese words and 6,391,919 English words).

All the Chinese texts in the corpus are encoded with Big-5 code. Since all our Chinese tools can only deal with Chinese GB code. We firstly converted all the Chinese texts from Big-5 code into GB code. Then the Corpus was aligned with a length-based sentence aligner. For the legal documents have been already well arranged with section by section, which makes the sentence alignment much easier and the precision is high. The Chinese texts were then segmented and pos-tagged with a program developed by the institute of Computational linguistics, Peking University. And all the English Texts were tokenized, lemmatized, and pos-tagged with a freely available tree-based tagger. Two tag sets were used for Chinese and English respectively, ICL/PKU tag set for Chinese texts and UPENN tag set for English texts. Figure 2. shows a sample of the corpus after preprocessing.

Chinese texts		English Texts		
…		…		
<s id=5>		<s_id=5>		
本	r	This	DT	this
条例	n	Ordinance	NN	ordinance
可	d	may	MD	may
…		...		
通则	n	General	JJ	general
条例	n	Clauses	NNS	clause
》	w	Ordinance	NN	ordinance
。	w	.	.	.
<s id=6>		<s_id=6>		
附注	n	Remarks	NNS	remark
：	w	:	:	:
…				

Figure 2. Samples of the corpus after preprocessing

In Figure 2., both corpus was arranged one token per line. The XML-like tag <s> marks the start of the sentence. The single-letter tags right to the Chinese tokens are their part of speech tags. The two columns right to the English tokens are part of speech tags and lemmas.

4. Statistical measurement used

Four statistical measurements were used in identification of unilingual multi-word units and the correspondences of the bilingual translation units. All four statistical formulas measures the degree of association of two random events.

Given two random events, *X* and *Y*, they might be two Chinese words appears in the Chinese texts and two translation units appears in an aligned region of the corpus. The distribution of the two events could be depicted by a 2 by 2 contingency table.

	Y	$\neg Y$
X	a	b
$\neg X$	c	d

Figure 3. A 2 by 2 contingency table

The numbers in the four cells of the table has the following meanings:

a : all counts of the cases the two events X and Y co-occur.
b : all counts of the cases that X occurs but Y does not
c : all counts of the cases that X does not occur but Y does
d : all counts of the cases that both X and Y do not occur

Based on the above-mentioned contingency table, different kinds of measurements could be used. We have tried four of them, namely, point-wise mutual information, DICE coefficient, χ^2 score and Log-likelihood score. One other measurement used by Gale[2] is ϕ^2 score, which is equivalent to the χ^2 score. All the four measurements could be easily calculated using the following formula.

(1) Point-wise mutual information

$$MI(st,tt) = \log_2 \frac{n \times a}{(a+b)\times(a+c)}$$

(2) DICE coefficient

$$DICE(st,tt) = \frac{2a}{(a+b)\times(a+c)}$$

(3) χ^2 score

$$\chi^2(st,tt) = \frac{n \times (a \times d - b \times c)}{(a+b)\times(a+c)\times(b+d)\times(c+d)}$$

(4) Log-Likelihood score

$$LL(st,tt) = 2 \times (a \times \log \frac{a \times n}{(a+b)\times(a+c)} + b \times \log \frac{b \times n}{(a+b)\times(b+d)} \\ + c \times \log \frac{c \times n}{(c+d)\times(a+c)} + d \times \log \frac{d \times n}{(c+d)\times(b+d)})$$

5. Identification of multi-word units

What might constitute multi-word units is probably a question critical to identification of them. It seems rational to assume Multi-word units are something between phrases and words, which might have the following properties:

1) The component words of a multi-word unit should tend to co-occur frequently. In the significance of statistics, multi-word unit should be word group that co-occur more frequently than expectation.
2) Multi-words units are not arbitrary combinations of arbitrary words; they shall form valid syntactic structure in the meaning of linguistics.

Based on the above-mentioned observations, we used an iterative algorithm using both statistical and linguistics means. The algorithm runs as follows: firstly the algorithm tries to find all word pairs that show strong coherence. This could be done using the measurements listed in section 4. After this step, all the word pairs in both of Chinese texts and English Texts whose association value is greater than a predefined threshold are marked. But this can only list of word groups of length of 2. Word groups of length more than 3 words could not be found by only one run of the algorithm. But apparently they could be found by a series of runs until there are no word groups having greater association value than the threshold anymore. The algorithm is designed as recursive structure, it marks longer word groups by viewing the shorter word group marked in the previous run as one word.

It is no doubt that pure statistics cannot perform very reliable. Some word groups found by the algorithm are awkward to be accepted as multi-word unit. The result of the algorithm shall be viewed as a candidate list of multi-words units. Some kind of refinement of the results might be required. For thinking that multi-word unit shall form valid syntactic pattern, we use a filter module which check all the word groups found and see if they fall into a set of predefined syntactic patterns.

"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 patterns

Figure 4. shows some patterns used by the filter. Patterns in the left side are for Chinese while the right side for English.

6. Extracting of the bilingual translation units

We adopt the same hypothesis-testing approach to set the correspondence between the Chinese-English translation units. It follows the observations that words are translation of each other are more likely to appear in aligned regions[2][3]. But we also take the multi-word units into consideration.

The whole procedure could be divided logically into two phases. The first phase could be called a generative phase, which lists all possible translation equivalent pairs from the aligned corpus. And the second phase can be viewed as a testing operation, which selects the Translation Equivalent Correspondences show an association measure higher than expected under the independence assumption as translation equivalence pairs. Again we use DICE coefficient, point-wise mutual information, LL score and χ^2 score to measure the degree of association.

One of problems of above-mentioned approach is its inefficiency in processing large corpus. Because in the generative phase, the above-mentioned approach will list all translation equivalent pairs and can lead to huge search space. To make the approach more efficient, we adopted the following assumption: Source translation units tend to be translated into translation units of the same syntactic categories. For example, English nouns tend to be translated into Chinese nouns, and English pattern "JJ+NN" tend to be translated into Chinese pattern "a+n" or "b+n". Apparently, this assumption is not always true for translation of Chinese into English and vice versa. But it really makes the algorithm much more efficient while the precision does not fall severely.

7. Experiments and Results

We have performed some preliminary experiments to test the performance of different statistic measurements, performance change when the categorial hypothesis is used. We get the following results:

For the experiments, we used a very small portion of the corpus of 500 sentence pairs. And we count how many correct and partially correct correspondences there are in the first hundred of translation equivalent pairs produced by the algorithm.

	MI	DICE	LL	χ^2
Correct	39	5	70	75
Partially correct	5	1	10	6
Accuracy	44%	6%	80%	81%

Figure 5. Performance variations of different statistical measurements

Figure 5. shows LL score and χ^2 score achieves better accuracy over mutual information and DICE coefficient. The reason might be that LL score and χ^2 score take the cell *d* of the contingency table into its consideration while point-wise mutual information and DICE coefficient do not.

Experiments also show the categorial hypothesis might lead to fall in accuracy, we did tests on the above-mention 500 sentence pair corpus using the hypothesis, the precision fall by 4% but the efficiency improved by more than 200%.

Figure 6.shows a sample of extracted translation equivalent pair from the test corpus. Some of them are wrong(see no 2), but most of them are correct translation equivalent pairs.

```
1.  见  see  /* CHI2 score=496.471 */
2.  追溯力_的   see  /* CHI2 score=496.471 */
3.  款   subsection   /* CHI2 score=496.237 */
4.  废除     repeal     /* CHI2 score=495.814 */
5.  命令     order      /* CHI2 score=493.195 */
7.  豁免     exemption      /* CHI2 score=490.829 */
25.  附属_法例 subsidiary_legislation   /* CHI2 score=477.173 */
26.  公共_机构 public_body   /* CHI2 score=475.711 */
28.  财政司_司长    Financial_Secretary      /* CHI2 score=475.711 */
31.  条例   ordinance      /* CHI2 score=470.081 */
34.  基本_文书 primary_instrument      /* CHI2 score=468.068 */
41.  卫生_主任 health_officer /* CHI2 score=468.068 */
42.  裁判官      magistrate      /* CHI2 score=468.068 */
43.  担当   discharge/* CHI2 score=468.068 */
45.  合约   contract  /* CHI2 score=468.068 */
46.  终审_法院_首席_法官   Chief_Justice_of_Final Appeal   /*   CHI2 score=468.068 */
53.  香港_特别_行政区  Hong_Kong_Special_Administrative_region   /* CHI2 score=448.576 */
63.  审裁处      tribunal   /* CHI2 score=420.579 */
64.  宣布   declare   /* CHI2 score=420.579 */
```

Figure 6. sample of results extracted from the corpus

8. Conclusion

As we see in the last section, the approach used in this paper does really list many real translation equivalent pairs from the corpus. It seems not all the results could be taken as translation units, but it really offers a candidate list from which useful translation unit could be selected by means of human validation.

Acknowledgment

We would like to give our thanks to Professor Dan Tufis. His help and suggestions are very important for our work.

References

[1] Teubert,W.(1997). Translation and the corpus, proceedings of the second TELRI seminar, 147-164.

[2] Gale,W., Identifying words correspondences in parallel Texts, DARPA speech and Natural language workshop. Asilomar, CA. ,1991

[3] Tufis,D., Computational bilingual lexicography: automatic extraction of translation dictionaries, In Journal of Information Science and Technology, Romanian Academy, Vol. 4, No. 3, 2001

[4] Maynard, D., Term Recognition using Combined Knowledge Sources, PH. D. thesis, Manchester University, United Kingdom.

[5] Yu Shiwen, Specification of Chinese text segmentation and POS tagging, see: http://www.icl.pku.edu.cn/research/corpus/coprus-annotation.htm

[6] Manual of Upenn Tree bank tag set, see: http://www.cis.upenn.edu/~treebank/

[7] Wu, D., Xia, X., Leaning an English-Chinese Lexicon from a Parallel Corpus, in AMTA-94, Association for MT in the Americas, Columbia, MD:Oct,94, pp206-213