[TMI 99: *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, August 1999, Chester, UK]

Extracting Bilingual Collocations from Non-Aligned Parallel Corpora

Kumiko Ohmori, Masanobu Higashida
NTT Information Sharing Platform Laboratories
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa-ken,
239-0847, JAPAN
{kumiko, higasida}@isl.ntt.co.jp

Abstract

This paper proposes a new method to find correspondences of uninterrupted collocations from Japanese-English bilingual corpora without sentence-to-sentence alignment. Uninterrupted collocations in English such as "once again", "give up", or "gross national product" handled as a single word or a compound word in Japanese, can be automatically extracted with corresponding Japanese words using word co-occurrence frequencies in both corpora.

The method consists of two stages. First, English and Japanese collocations are extracted separately from given corpora. After successive word units, which become collocation candidates, are collected by using *n*-gram statistics of each word, two kinds of entropy values, after-unit and before-unit are calculated for each unit to select word units surpassing thresholds as uninterrupted collocations. Second, a correspondent translation of each uninterrupted English collocation is extracted from the Japanese corpus by calculating correlation values between the target collocation and Japanese words or collocations which co-occur in the given corpora and using a basic English to Japanese word unit dictionary. Experiments are executed on economic articles of Asahi Newspaper as corpora. A Japanese word unit for each extracted English collocation is automatically obtained with more than 70 % precision rate, whereas the rate was about 40 % if only word-to-word correspondence is used.

1 Introduction

Machine translation, which is a significant application for Natural Language Processing (NLP), has increasingly become useful in recent years. The quality of translation is strongly influenced by the quantity and the quality of bilingual dictionaries.

While much work has already been done on the automatic alignment of parallel corpora, there are several problems not addressed by their alignment algorithms. First, many papers assumed word-to-word correspondences between two languages Gale & Church (1991), Kay & Roscheisen (1993) ,Utsuro et al. (1994), Fung (1994), Fung & McKeown (1994). However, many collocations have their own meanings in the text and it is impossible to acquire the translations for the collocations using word-to-word processing methods. This is the bottleneck for the alignment of parallel corpora. One proposal to cope with this problem is to rely on the lexical information of each word and extract the correspondences of compound nouns or adjective phrases from corpus

Kupiec (1993). Yamamoto & Sakamoto (1993), Fung (1995). This approach needs to get the lexical information of each word by morphological analysis with great accuracy.

Second, most previous alignment systems assume sentence-level alignment has been done in parallel corpora Brown et al. (1988), Kay & Röscheisen (1993), Kupiec (1993), Smadja & McKeown (1994), Wu & Xia (1994), Haruno et al. (1996). For European languages there exist several sentence-level alignment bilingual corpora, such as the Canadian Hansards (parliamentary debates). On the other hand, in corpora for Asian languages, many-to-one or one-to-many sentence translations can be often observed. This makes it difficult to get sentence-aligned parallel corpora. Some of the alignment programs for Asian languages use manually produced sentence-aligned parallel corpora. This paper introduces a new method that extracts correspondences of collocations from sentence-unaligned Japanese-English parallel corpora where collocations are picked up automatically. First, collocation extraction is attempted using *n*-gram statistics and the entropy values. Correspondences of collocations from Japanese-English parallel corpora are then extracted using the word co-occurrence frequency and a basic word-unit bilingual dictionary, instead of relying on sentence alignment.

In the following sections, we elaborates the framework of the proposed bilingual collocation extraction algorithm.

2 The Framework of Our Algorithm

The overall framework of the proposed algorithm is depicted in Figure 1. Although, the framework was implemented for Japanese and English corpora, it can be applied to other languages. The goal of the system is to extract the correspondences of collocations from parallel corpora. The system consists of two stages. In the first stage, collocations are picked up from each corpus using *n*-gram statistics of each word and the entropy values of words before and after units. The second stage estimates the Japanese translations for English collocations. Japanese sentences are pre-processed by the morphological analysis tool "Chasen" (Matsumoto 1997) into Japanese word units. Since sentence-aligned parallel corpora are rather limited for English-Japanese corpora, correlations between the co-occurrence features of each English collocation and Japanese words or word units are used, with the assistance of a basic word-to-word bilingual dictionary after the co-occurrence frequencies for each English collocation and Japanese collocation are determined. Finally, a Japanese word or word units with the highest correlation values for each English collocation is obtained as the correspondent translations.

3 Extracting collocations from the corpus

This section describes the proposed method, the extraction of collocations from a corpus based on *n*-gram statistics and the entropy values. There are many collocations, called idiomatic expressions or frozen patterns, in any corpus. Processing these collocations as single units is one way to resolve morphological-level ambiguity or syntactic-level ambiguity, both of which are important problems for NLP. Although, the techniques of applying *n*-gram statistics for each character to find units from corpus have been recently proposed by Nagao & Mori (1994), Ikehara et al. (1995). The method to extract

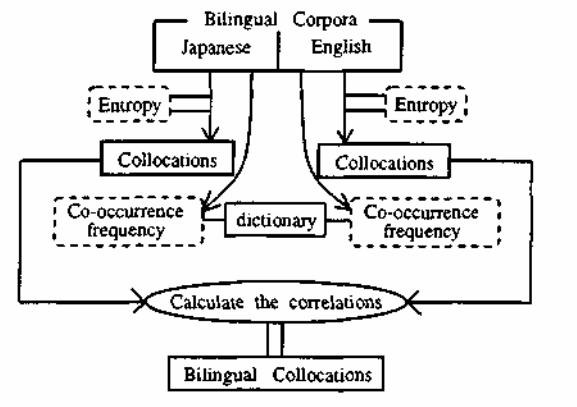

Figure 1: The framework of extraction bilingual collocations

units adopting *n*-gram regularly tries to find a sequence of characters (for Japanese texts) or words (for English texts) considering their frequencies of appearance in a given text. Two factors mentioned above, however, cannot be necessarily regarded to properly characterize features of collocations. So, we came to a thought that an entropy concept should be introduced to identify proper collocations, because entropy can represent the connective strength of neighboring two characters (Japanese) or words (English).

The more the variety of neighboring words exists, and the less bias the occurrence probability of neighboring words becomes, the higher the entropy value becomes. A unit with a high entropy value before or after it implies that a unit sequence can mostly divided at its position and consequently more likely becomes a collocation. On the contrary, when the kind of neighboring words are few, or the occurrence probability of those words is imbalanced, the entropy value tends to be lower. Thus, a unit with a low entropy value can be considered to be a part of a longer sequence of units which will form a collocation. The following procedures explains how the proposed method extracts collocations from a corpus.

(1) **Extraction of units as candidate collocations.**
After calculating *n*-gram statistics of each word in both languages based on the method by Ikehara et al. (1995), units consisting of 2 or more words whose frequency of appearance is 2 or more times are extracted.

(2) **Calculation of the entropy values.**
When the word w_b occurs after a certain unit W, the entropy between W and w_b is called "the behind entropy for W", and calculated by the following formula:

$$H_b(W) = -\sum_b P_b(W|w_b) \cdot \log P_b(W|w_b) \quad (1)$$

where $P_b(W|w_b)$ is the occurrence probability of the word w_b after the unit W. $P_b(W|w_b)$ is defined as below:

$$P_b(W|w_b) = \frac{f(W|w_b)}{f(W)} \quad (2)$$

where. $f(W|w_b)$ is the frequency of the word w_b after the unit W, and $f(W)$ is frequency of the unit W. Likewise, toward a certain unit W, "the" forward entropy for W'' is also calculated using the frequency of the word w_f before the unit W, and the frequency of the unit W.

(3) Extraction of collocations

Extract the units whose behind and forward entropy values exceed a preset thresholds. The extracted units are defined as collocations.

4 Matching Bilingual Collocations

This section describes how the proposed technique extracts the Japanese translations for the English collocations.

Since it is almost hopeless to get sentence-level alignment as stated before, we determined to use the word co-occurrence frequency and a machine readable word-unit bilingual dictionary to acquire the translations for English collocations from sentence-unaligned parallel corpora. The definition of "the word co-occurrence frequency" is the set of words appeared in those sentences within which the collocation occurs and the number of times each word or the collocation occur, after the definition of Kaji & Aizono (1996). Though Kaji proposed a method for extracting word correspondences from Japanese-English parallel corpora using the co-occurrence information for each word in both languages and a bilingual dictionary, this has a problem that words with high frequency negatively influence the results. To achieve better results, we improved Kaji's method and optimize the extraction of translations for English collocations.

The proposed processing steps of our method are as follows:

(1) For each English collocation *eu* extracted by the method in Section 3, a co-occurrence set $C(eu)$ and a co-occurrence frequency f are defined as follows:

$$C(eu) = \{(ew_i, f_i) \mid i = 1, \cdots, l\}, \quad |C(eu)| = \sum_{i=1}^{I} f_i \tag{3}$$

which shows that an English word ew_i co-occurs with the English collocation *eu* f_i times in a sentence. To avoid disadvantage of Kaji's method, a constraint for the frequency of the word ew_i in co-occurrence set $C(eu)$, is introduced: the frequency of the English words inside a co-occurrence set should be lower than the threshold decided as the upper bound. An English word whose frequency is exceeding this limitation will be excluded from the co-occurrence set.

(2) The translation for an English collocation can be a word or a collocation in Japanese. Therefore, for each Japanese collocation and each word, we define the co-occurrence set and the co-occurrence frequency as follows, the frequency of the Japanese word jw_j is lower than a preset threshold.

$$C(j) = \{(jw_j, g_j) \mid j = 1, \cdots, m\}, \quad |C(j)| = \sum_{j=1}^{m} g_j \tag{4}$$

(3) A correlation $R(eu,j)$ between an English collocation *eu* and a Japanese collocation or word j is defined as follows:

$$R(eu, j) = \frac{|C(eu) \cap C(j)|}{|C(eu)| + |C(j)| - |C(eu) \cap C(j)|} \quad (5)$$

where $|C(eu) \cap C(j)|$ is the intersection of $C(eu)$ and $C(j)$, whose elements are the pair of an English word and a Japanese word together with their frequencies. We calculate $|C(eu) \cap C(j)|$ using the approximate value of $C(eu)$. The approximation is produced using a word-unit bilingual dictionary D. Each English word inside the English collocation's co-occurrence set $C(eu)$ is translated into Japanese words by consulting dictionary D. We then define $C_p(eu)$, the approximate co-occurrence set of $C(eu)$ as below:

$$\begin{aligned} C(eu) &\simeq C_p(eu) = \{(jw'_k, g'_k) \mid k = 1, \cdots, n\} \\ g'_k &= \sum_{\{ew_i \in C(j) \,\&\, (ew_i, jw'_k) \in D\}} f_i \end{aligned} \quad (6)$$

where Japanese word jw'_k is the translation of English word ew_i in dictionary D. $C(eu) \cap C(j)$ is the approximation of the intersection of pseudo co-occurrence set $C_p(eu)$ and Japanese co-occurrence set $C(j)$.

$$\begin{aligned} &C(eu) \cap C(j) = C_p(eu) \cap C(j) \\ &= \{(jw'_k \cap jw_j, min\{f'_k, g_j\}) \mid j = 1, \cdots, m, k = 1, \cdots, n\} \end{aligned} \quad (7)$$

Finally, $|C(eu) \cap C(j)|$ is calculated as follows:

$$|C(eu) \cap C(j)| = \sum_{j,k} min\{f'_k, g_j\} \quad (8)$$

(4) For each English collocation *eu*, a Japanese collocation or a word *j* whose value of correlation *R(eu,j)* is the highest, is extracted as the Japanese translation of the *eu*.

An example of extracting the translation for the English collocation, "once again" is shown in Figure 2. This shows that "once again" in a English sentence corresponds to a word "再び(hutatabi) *Chinese character*" in a translated Japanese sentence.

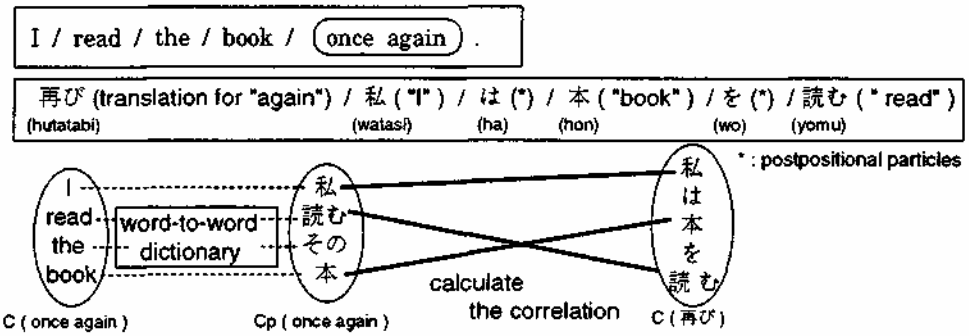

Figure 2: An example of extracting the translation for "once again"

5 Results of Extracting Collocations

This section reports the results of an experiment on extracting collocations from English and Japanese corpora.

5.1 The Results of Extracting Units

As the input parallel corpora, we used economic articles of Asahi Newspaper (8,090 sentences, 212,227 different words in total, 9.307 different vocabularies in English. 7,674 sentences, 194,708 different words, 10.728 different vocabularies in Japanese). The method described in Section 3 was used to pick up units from each corpus whose word lengths were more than 2, whose frequency of appearance was more than 2. and whose behind and forward entropy values were more than 0.3. Top 10 units are shown in Table 1. In the table, "/" is the word boundary, and "F" means the frequency of the unit. In this case, most of extracted candidates are mostly prepositions (in English) and postpositional particles (in Japanese).

Table 1: Extracted English units and Japanese units satisfying conditions

English unit	entropy	F
of / the	2.0792	119
in / the	2.0719	115
and / the	2.0569	114
of / a	2.0334	105
by / the	1.9590	91
to / the	1.9590	91
for / the	1.9085	81
from / the	1.8921	78
have / be	1.8808	76
on / the	1.8451	71

Japanese unit	entropy	F
へ (he) / の (no)	1.8261	88
に (ni) / も (mo)	1.8129	64
に (ni) / は (ha)	1.7913	167
に (ni) / よる (yoru)	1.7682	97
と (to) / いう (iu)	1.7582	121
など (nado) / の (no)	1.7102	50
と (to) / する (suru)	1.6258	129
で (de) / の (no)	1.5355	35
に (ni) / 対する (taisuru)	1.5195	54
の (no) / ようだ (yoda)	1.4936	98

5.2 A Result of Re-extracting Units

Clearly, it is hard to regard the units in Table 1 as collocations. The units with high entropy values have high frequency (Table 1), and each unit that has high entropy values consists of high frequency words, such as 'the' (17,286 times), 'of' (8,350 times), 'to' (6,225 times), 'in' (4,940 times), 'の(no) *Chinese character*' (12,817 times), 'を (wo) *Chinese character*' (7,114 times), 'は(ha) *Chinese character*' (6,848 times), 'に(ni) *Chinese character*' (6.548 times), and so on.

To exclude unnecessary words with high frequency, we adopted a restriction on the frequency of the words in the unit and disposed word units whose frequency is ranked in the top 1 % among all word units. Using Zipf's law Zipf (1949), we find that the English words in the top 1 % appear more than 228 times, and Japanese words appear more than 182 times. This constraint was utilized to re-extract "fare" collocations again. This yielded 1.282 English collocations and 720 Japanese collocations. Some of the units with high entropy values are shown Table 2.

Table 2: Some re-extracted English units and Japanese units

English unit	entropy	F
once / again	1.0735	14
still / remain	1.0140	11
billion / yen	0.9208	8
public / opinion	0.9031	8
crude / oil	0.8655	8
land / price	0.8253	10
turn / point	0.8021	7
trillion / yen	0.8021	7
atomic / power	0.8021	7
so / many	0.7782	6

Japanese unit	entropy	F
億 (oku) / 円 (en)	0.7782	6
とも (tomo) / いえる (ieru)	0.7280	6
億 (oku) / ドル (doru)	0.6990	5
貯蓄 (chochiku) / 不足 (husoku)	0.6021	4
朝鮮 (chosen) / 半島 (hanto)	0.6021	4
真剣だ / 取り組む (shinkenda) / (torikumu)	0.6021	4
三塚 (mitsuduka) / 派 (ha)	0.6021	4
軍事 (gunji) / 行動 (kodo)	0.6021	4
金融 (kinyu) / 緩和 (kanwa)	0.6021	4
短距離 (tankyori) / 核 (kaku)	0.6021	4

Since the units in Table 2 don't contain high frequency words, they can be regarded as collocations. The Japanese translations for these 1,282 English collocations are examined in the next section.

6 Results of Extracting Bilingual Collocations

In this section we will evaluate the results of our experiment using the method stated in Section 4 and Section 5.2. We used the Edict (Breen 1995) as a word-to-word public dictionary that contains about 120,000 English entry words. The corpora were the same newspaper articles used for extracting collocations in Section 3. In creating the co-occurrence sets, we decided that the upper bound of English word frequency would be 228, and Japanese word frequency 182 (we set the top 1 % in frequency rate as the threshold). The system we proposed in this paper could extract appropriate Japanese translations for the 1,282 English collocations (picked up in Section 5.2), with 70.4 % precision rate. Table 3 shows the ranking of the correctly estimated translation pairs. The translation of 70.4 % of English collocations was realized accurately by the first candidate, and 76.2 % in the top 3. Examples of collocation correspondences extracted as the first candidate, are shown in Table 4.

Table 3: Accuracy of translation estimates

first estimate	top 3 estimate
70.4 %	76.2 %

Most mistranslated English collocations were those that did not have corresponding collocations in Japanese. This means that an English collocations were divided into two or more units in Japanese. For example, the collocation "airport / construction" should have been translated as "空港(kuko) *Chinese character* / 建設 (kensetu) *Chinese*

Character", but it was divided into 空港 [translation for "airport"] and 建設 [translation for "construction"]. In this case the system often selects one of the divided words as the translation, like 空港 [translation for "airport"]. Table 5 lists some of the English collocations whose corresponding translations were selected incorrectly.

It is observed (by our experiment) that only 39.4 % of 1,282 English collocations can be given correct Japanese translations. In this case, we filtered out concatenations of translated Japanese words that didn't occur in the corpus, after producing all possible combinations of translated Japanese words, picked up from a word-to-word bilingual dictionary for each English word forming a English collocation.

Thus shows that most English collocations cannot be translated by merely concatenating Japanese words found in a dictionary.

7 Conclusion

This paper has introduced a new method for extracting the correspondences of collocations from sentence-unaligned parallel corpora. The essence of our method is to calculate correlations between English collocations and Japanese collocations or words, based on their co-occurrence frequency with assistance of a word-unit bilingual dictionary. In our experiments using newspaper articles, 1,282 fair English collocations were extracted automatically by the proposed method, and its precision for extracting the correct translations for those English collocations was 70.4 %. Also, only 39.4 % extracted English collocations can be translated by merely concatenating Japanese words found in a word-to-word bilingual dictionary.

Thus, it is clearly proved that the proposed method in this paper is sufficiently effective for the extraction of bilingual collocations from non-aligned parallel corpora.

8 Acknowledgements

We would like to thank Prof. Masakazu Nakanishi and Prof. Hiroaki Saito of Keio University for their guidance on theoretical aspects and essential remarks on experiments.

Table 4: Examples of extracted collocation correspondences

English collocation	Extracted result
once / again	再び (hutatabi)
crude / oil	原油 (genyu)
get / around	逃れる (nogareru)
give / up	あきらめる (akirameru)
chemical / weapon	化学 (kagaku) / 兵器 (heiki)
humanitarian / aid	人道 (jindo) / 援助 (enjo)
nuclear / non-proliferation	核 (kaku) / 不 (hu) / 拡散 (kakusan)
strategic / nuclear / weapon	戦略 (senryaku) / 核 (kaku)
high / school	高校 (koko)
inheritance / tax	相続税 (sozokuzei)
gross / national / product	国民総生産 (kokuminsoseisan)

Table 5: Examples of mis-extracted collocation correspondences

English collocation	Extracted Japanese word or collocation
air / pollution	汚染 (osen) [translation for "pollution"]
public / work	公共 (kokyo) [translation for "public"]
economic / growth	経済 (keizai) [translation for "economic"]
50 / years / ago	50 / 年 (gojunen) [translation for "50 / years"]
president / hussein	フセイン (husein) [translation for "hessein"]
peaceful / solution	解決 (kaiketsu) [translation for "solution"]
several / year / ago	数 (kazu) [translation for "several"]
one-party / dictatorship	一党 (itto) [translation for "one-party"]
individual / investor	個人 (kojin) [translation for "individual"]

References

Peter F. Brown et al: 1988, 'A Statistical Approach to Language Translation', in *12th International Conference on Computational Linguistics: COLING-88,* pp. 71-76.

Pascale Fung.: 1994, 'K-vec: A New Approach for Aligning Parallel Texts', in *15th International Conference on Computational Linguistics: COLING-94,* pp. 1096-1102.

Pascale Fung & Kathleen R. McKeown: 1994, 'Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping', in *Proc. the First Conference of the Association for Machine Translation in the Americas,* pp. 81-88.

Pascale Fung: 1995, 'A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora', in *Proc. the 33rd Annual Meeting of the Association for Computational Linguistics,* pp. 236-243.

James W. Breen.: 1995, 'Freeware Japanese / English dictionary',

William A. Gale & Kenneth W. Church: 1995, 'Identifying Word Correspondences in Parallel Texts', in *Proc. the 4th DARPA Speech and Natural Language Workshop,* pp. 152-157.

Ikehara, Satoru et al: 1995, 'Automatic Extraction of Uninterrupted and Interrupted Collocations from Very Large Japanese Corpora Using *n*-gram Statistics', in *Technical Report of IPSJ (Vol. 36, No. 11),* pp. 2584-2596.

Kaji, Hiroyuki & Toshiko Aizono: 1996. 'Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information', in *16th International Conference on Computational Linguistics: COLING-96.* pp. 23-28.

Martin Kay & Martin Röscheisen: 1993, 'Text-Translation Alignment', in *Computational Linguistics, Vol. 19(1),* pp. 121-142.

Kupiec M. Julian: 1993, 'An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora', in *Proc. the 31st Annual Meeting of the Association for Computational Linguistics,* pp. 17-22.

Matsumoto, Yuji et al.: 1997, 'Chasen', (http://cl.aist-nara.ac.jp/lab/nlt/chasen.html).

Nagao, Makoto & Shinsuke Mori: 1994, 'A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrase from Large Text Data of Japanese', in *15th International Conference on Computational Linguistics: COLING-94,* pp. 611-615.

Frank Smadja & Kathleen R. McKeown: 1994, 'Translating Collocations for Use in Bilingual Lexicons', in *Proc. ARPA Human Technology Workshop 94.*

Utsuro, Takehito et al.: 1994, 'Bilingual Text Matching using Bilingual Dictionary and Statistics', in *15th International Conference on Computational Linguistics: COLING-94,* pp. 1076-1082.

Dekai Wu & Xuanyin Xia: 1994, 'Learning an English-Chinese Lexicon from Parallel Corpus', in *Proc. the First Conference of the Association for Machine Translation in the Americas,* pp. 206-213.

Yamamoto, Yukio & Hitoshi Sakamoto: 1993, 'Extraction of technical term bilingual dictionary from bilingual corpus', in *IPSJ SIG Notes, 93-NL-94,* pp. 85-92.

G. K. Zipf: 1949, *Human behavior and the principle of least effort,* Massachusetts: Addison-Wesley.

Frank Smadja et al.: 1996, 'Translating Collocations for Bilingual lexicons: A Statistical Approach', in *Computational Linguistics, Vol. 21(4),* pp. 1-38.

Haruno, Masahiko et al.: 1996, ''Learning Bilingual Collocations by Word-Level Sorting', in *16th International Conference on Computational Linguistics: COLING-96,* pp. 525-530.