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

Extraction of Translation Equivalents from Non-Parallel Corpora

Takaaki TANAKA and Yoshihiro MATSUO
NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0237, JAPAN
{takaaki,yosihiro}@cslab.kecl.ntt.co.jp

Abstract

This paper presents a widely applicable method for extracting bilingual expressions from non-parallel corpora. The algorithm first collects word sequences as candidates for translation equivalents that match given patterns of word sequences from each corpus. Then, translation equivalents are selected from these candidates by aligning component words from within word sequences. We show the results of acquiring Japanese and English compound nouns from unrelated financial newspapers. We also demonstrate that the method can collect pairs that do not appear in terminological dictionaries.

1 Introduction

There is a lot of research on the acquisition of translation knowledge from bilingual corpora. Most approaches (Kaji et al. 1992; Dagan & Church 1994; Smadja et al. 1996) use sentence alignments in parallel corpora and effectively extract bilingual expressions or translation templates. However, there are not always enough well-aligned parallel corpora to extract pairs of expressions in two languages, and adapting these methods to corpora that cannot be well aligned is not simple. Fung proposed a method of using pattern matching to overcome this constraint (Fung 1995). Her algorithm can cope with parallel corpora that are difficult to align at the sentence level. Nevertheless, this method is still limited to parallel corpora, because it uses position information closely concerned with each corpus.

Even non-parallel corpora, which are not translations, include some phrases and compounds[1] that have the same meanings and functions. Some algorithms for word sense disambiguation resolve lexical ambiguities by syntactic relations in a monolingual corpus (Dagan & Itai 1994; Yarowsky 1995). Such research suggests that information in other language corpora, which are not necessarily related to the source language corpus, is useful for disambiguation. The problem of acquiring a translation of a compound is close to that of disambiguation when determining the translation of its component (e.g., selecting *personal* or *individual* as a translation of *kojin* in the Japanese compound *kojin toushika* meaning "individual investor"). However, for the extraction of translation pairs, more correspondence candidates of a word are required, because some correspondence components in two language are not typical translations from a dictionary. Therefore, we use a thesaurus to increase the number of candidates.

[1] In this paper, we treat complex nominate (e.g., *travel agent)* as compounds.

Rapp (1995) and Fung & Yee (1998) identified translations of words using non-parallel corpora. Their methods are based on the assumption that the patterns of word co-occurrences in a language are similar to those in another language. In addition, some expressions and their translations have a syntactic pattern. For example, Japanese *noun-noun* compounds are often translated into English *noun-noun, adjective-noun,* etc.

Fung (1997) also proposed a method that finds terminology translations from non-parallel corpora. She estimated the correspondence between words by matching their "Word Relation Matrices". The elements of "Word Relation Matrices" denote the mutual information between a word and seed words that are given translations. The method is significant in that it can find translations even if non-parallel corpora in two languages are provided. However, it is difficult to match the expressions that do not frequently appear because their matrices lack information. Moreover, the manner of defining terms is not positively represented, and in extracting translations determination of an extracted unit is an inevitable problem.

We propose a method that collects word sequences from each corpus by using syntactic patterns of translations as a clue to translation equivalents; the method then searches for translation equivalents by matching collected expressions. Accordingly, we can exploit monolingual corpora in different languages instead of parallel corpora to acquire bilingual expressions.

2 Translation Patterns

We know that some language translation equivalents have syntactic features. Table 1 shows patterns in translations between Japanese and English[2]. In each row, a word in one language corresponds to a word that has the same suffix in the other language. Note that corresponding words in two languages can have different parts-of-speech. In the example (α2) in Table 1, the noun *kankyou* "environment" corresponds to the adjective *environmental.*

The translation patterns (αl)-(α3) are examples of noun phrases. Many languages can construct compound nouns with two or more nouns, and their translations are often compound nouns as well. For example, both *trade friction* and its Japanese translation *boueki masatsu* are compound nouns consisting of two nouns. Here, these expressions have the same part-of-speech (POS) sequences, *noun-noun,* and each component corresponds to that of the other language.

In sentence (2), which is not related to sentence (1), a translation *(current profits)* of the Japanese compound *keijou rieki* in (1) appears. If we use the translation pattern (α2) in Table 1, the following word sequences are collected: "*keijou*/**N** *rieki*/**N**" from (1), and "*these*/**J** *investments*/**N**" and "*current*/**J** *profits*/**N***"* from (2). Then, a pair *(keijou rieki* and *current profits)* is selected, since their components correspond to each other. Therefore, we consider translation patterns as a clue to extracting expressions.

[2] Note: In this paper, we use bold characters as symbols for parts-of-speech: **N**(noun), **V**(verb), **J**(adjective), **P**(preposition), and **X**(affix). ***NP*** represents a noun phrase.

(1) *Chuukanki/***N**-*no keijou/***N** *rieki/***N**-*wa hobo zennennami/***N**-*o*
interim ordinary profits nearly the level of last year
*kakuho-suru/***V** *mitoushi/***N**-*da.*
hold prospect
'Interim current profits are expected to hold to nearly the level of last year'

(2) These/**J** investments/**N** have/**V** had/**V** a negative/**N** effect/**N**
on/**P** current/**J** profits/**N**.

We can also see other patterns in the following sentences.

(3) *Beikoku-wa nippon-to hikakushi-te kabunushi-no chikara-ga tsuyoi.*
U.S. Japan compare stockholders' power strong
'U.S. stockholders have more power than Japanese ones.'

(4) The absolute amount of investment isn't large - particularly in comparison with Japanese overseas investment.

The second sentence (4) is not a translation of the first sentence (3), but sentence (4) includes the expression, *in comparison with,* which has a similar meaning and function as the Japanese phrase *to hikakushi-te.* The expression ***NP*** *to hikakushi-te,* which consists of ***NP****+particle+verb+particle,* makes a dependent clause. This clause works as an adverbial clause like the English prepositional phrase *in comparison with* ***NP,*** although the internal syntax of these expressions is quite different.

We can use these translation patterns to collect candidates of translation equivalents, because there are many pairs that have the same patterns. It is necessary to select translation pairs from word sequences collected by these patterns. In many translation equivalents, their components (words) are related to the components of their counterparts. Thus, we estimate the correspondence of the expressions by matching their components and choose pairs that have high values. For this estimation, three types of word correspondence are considered; lexical, similarity and co-occurrence correspondence. The details of these types will be described in 3.2.

	Japanese	English	Examples (J/E)
(α1)	N_1 N_2	N_1 N_2	*boueki masatsu* / trade friction
(α2)	N_1 N_2	J_1 N_2	*kankyou hogo* / environmental protection
(α3)	N_1 N_2	N_2 of N_1	*boueki shuushi* / balance of trade
(β1)	[***NP***]-*ni* V_1-*te*	in N_1 to [***NP***]	[*NP*]-*ni kotae-te* / in response to [*NP*]
(β2)	[***NP***]-*o* V_1-*te*	in N_1 of [***NP***]	[*NP*]-*o sagashi-te* / in search of [*NP*]
(β3)	[***NP***]-*to* V_1-*te*	in N_1 with [***NP***]	[*NP*]-*to hikakushi-te* / in comparison with [*NP*]

Table 1: Translation Patterns

3 Extraction of Translations

As described in section 2, many translation pairs have patterns. We collect word sequences from each corpus by first using translation patterns to acquire candidates for bilingual expressions. Next, we search for pairs of words that satisfy the correspondences of the sequences.

In short, our method is composed of two phases.

- Collection Phase: Collecting word sequences from each corpus.
- Correspondence Phase: Searching collected expressions for valid translation pairs.

3.1 Collecting Word Sequences

First, we collect the n-grams of POSs appearing in a translation pattern (e.g., NN, JN, etc.) from each corpus. These corpora only have to be tagged with POSs, and the corpora do not have to be related. As the method simply extracts word sequences according to POS tags, it also collects noisy sequences. However, most meaningless sequences will be eliminated, since they cannot match other sequences in the correspondence phase.

3.2 Matching Word Sequences

The method searches for translation equivalents by matching collected word sequences.

3.2.1 Matching Words

We define correspondences between words to compute the similarity of two expressions. We consider the following three types of word correspondences.

(a) Lexical Correspondence This is a basic relation between a pair of words. When one word appears as a translation of another in an ordinary bilingual dictionary, we say these words have a lexical correspondence. In the following example, the Japanese word *shouken* means "securities", so the two words have a lexical correspondence.

$$(\underline{shouken})_{lex1}\ gaisha\ -\ (\underline{securities})_{lex1}\ company$$

Since a component word of one expression is sometimes translated into a word that has a different POS from the original word (e.g., noun into adjective), we allow a word to correspond to a derivative of its translation. For instance, the Japanese noun *keizai* means "economy", and therefore, *keizai* can be linked to the adjective *economic* as well as the noun *economy*.

We use the dictionary from the Japanese-to-English machine translation system ALT-J/E (Ikehara et al. 1991), which contains about 300,000 entries.

(b) Similarity Correspondence When the meaning of one word is similar to another, we say these words have a similarity correspondence. A component of one expression is sometimes not directly translated. In the following example, *keiei* means management in a business or a similar organization. *Keiei* and *business* do not have the same meaning, but both words relate to activities in a company.

$(\underline{keiei})_{sim1}$ *keikaku* - $(\underline{business})_{sim1}$ *plan*

We adopt the semantic categories in *Nihongo Goitaikei - A Japanese Lexicon* (Ikehara et al. 1997) as a measure of word similarity. *Goitaikei* has approximately 300,000 Japanese words with semantic attributes that are classified into about 3,000 categories. To classify English words into the same semantic categories as Japanese, the English words are once translated into Japanese words, and the categories of Japanese are taken as the original English. English words are often given several different categories, because English words can be translated into different Japanese words in context, and each Japanese word may have one or more categories. In this paper, a English word is considered to have all such categories. For instance, "*bank"* is ambiguous because it can be translated into *ginkou* or *teibou* in Japanese, so "*bank"* is assigned three categories as follows.

bank { *ginkou* [enterprise/company/industry]; [work place]; *teibou* [man-made embankment] }

If Japanese word *A* and English word *B* have the same semantic categories, *A* and *B* are linked.

(c) Co-occurrence Correspondence When translations of expressions containing word *A* often include another word B, we say these words have a co-occurrence correspondence. In the pair *keijou rieki* and *current profits, keijou* means status that is kept at a constant level, and does not mean "current". However, the compound noun "*keijou* **N**" is often translated into "*current* **N**"; therefore, we link *keijou* with *current.*

$(\underline{keijou})_{co1}$ $(rieki)_{lex1}$ - $(\underline{current})_{co1}$ $(profits)_{lex1}$

$\underline{keijou}$ $(torihiki)_{lex2}$ $\underline{current}$ $(transaction)_{lex2}$

$\underline{keijou}$ $(kakaku)_{lex3}$ $\underline{current}$ $(price)_{lex3}$

To examine this correspondence, we make a co-occurrence dictionary from pairs of expressions. After we match each word of one expression with its counterpart, we add the remaining words, which cannot be aligned to other words, to the co-occurrence dictionary. For example, when *keijou rieki* and *current profits* are given, *rieki* is aligned to *profits,* because *rieki* has the same meaning as *profits.* Then, we add the pair of remaining words *keijou* and *current* to the co-occurrence dictionary.

3.2.2 Estimating Correspondence of Word Sequences

We evaluate the correspondence of two word sequences by adapting the word correspondences to each combination of words. A correspondence measure $cor(J_x,E_y)$ is defined as (1), where J_x and E_y are a Japanese word sequence $(w_{J1},...,w_{Jm})$ and an English word sequence $(w_{E1},...,w_{En})$, respectively. A function $links_{JE}(J_x,E_y)$ (resp. $links_{EJ}(J_x,E_y)$) represents the correspondence between word sequences J_x and E_y in linking Japanese to English (resp. English to Japanese).

The value of function $link(w_J, w_E)$ is determined according to the relation between w_J and w_E (2). The value of the lexical correspondence is 1, and the values of the

other two correspondences are lower. When either w_J or w_E is the head and the other is a modifier (e.g., *kenkyuu* "research" in *kenkyuu dantai* "research association" and *research* in *basic research),* the value of $link(w_J,w_E)$ is reduced by weight $wg(\,)$ (3).

$$cor(J_x, E_y) = \frac{links_{JE}(J_x, E_y) + links_{EJ}(J_x, E_y)}{n(J_x) + n(E_y)} \quad (1)$$

where

$n()$: the number of words in a word sequence

$$links_{JE}(J_x, E_y) = \sum_i max_j(link(w_{Ji}, w_{Ej})wg(w_{Ji}, w_{Ej}))$$

$$links_{EJ}(J_x, E_y) = \sum_j max_i(link(w_{Ji}, w_{Ej})wg(w_{Ji}, w_{Ej}))$$

$$link(w_a, w_b) = \begin{cases} 1: & \text{lexical correspondence} \\ \lambda_s(0 < \lambda_s < 1): & \text{similarity correspondence} \\ \lambda_c(0 < \lambda_c < 1): & \text{co-occurrence correspondence} \\ 0: & \text{no relation} \end{cases} \quad (2)$$

$$wg(w_a, w_b) = \begin{cases} 1: & \text{both } w_a \text{ and } w_b \text{ are heads or modifiers} \\ \alpha(0 < \alpha < 1): & \text{otherwise} \end{cases} \quad (3)$$

4 Evaluation

4.1 Ability to collect and filter expressions

To evaluate our method, we choose compound nouns including terminology as extracted expressions. We use a bilingual terminological dictionary containing about 105,000 economic and other terms to estimate the ability to collect translation equivalents in this section. Table 2 show POS patterns of Japanese and English terms in the terminological dictionary. Several syntactic patterns of terms are common. 30% of the Japanese terms consist of **NN**, and 37% of the English terms consist of **NN** or **JN**.

We use for the extracting source Japanese and English newspapers that are not related in contents: Nihon Keizai Shimbun issued in 1994 (Nikkei corpus CD-ROM 1994, 200MB) and Wall Street Journal in 1988 (WSJ corpus, 300MB). The articles of both newspapers are generally related to economics or politics, but the main topics and publication dates are different. Each corpus is automatically tagged in advance. We use ALT-J/E's morphological analyzer, ALTJAWS for Japanese and Brill's tagger (Brill 1992) for English.

About a quarter of the Japanese (**NN**) and English (**NN** and **JN**) terms in the terminological dictionary appear in each newspaper covering a period of one year, as shown in Table 3. Moreover, 1,778 translation pairs **NN-NN** (1,555 Japanese expressions) and 1,337 pairs **NN-JN** (1,185 Japanese expressions) are used in these corpora. These pairs are sets including Japanese NN-type compound nouns and their English translations that can be collected from the corpora (existing as entries in the dictionary). We use these translation pairs as answer sets in this section.

N	34644	(32.8%)	**N**	33040	(31.3%)
NN	30912	(29.3%)	**NN**	20285	(19.2%)
NX	9903	(9.4%)	**JN**	18316	(17.3%)
NNN	6714	(6.3%)	**NPN**	4202	(4.0%)
NNX	4667	(4.4%)	**NNN**	2853	(2.7%)
Others	18770	(17.8%)	Others	26914	(25.5%)
Total	105610	(100%)	Total	105610	(100%)

(a) Japanese (b) English

Table 2: POS Patterns of Terms in a Terminological Dictionary

Japanese **NN**	8126	
English **NN**	4357	
English **JN**	4082	
Both Japanese **NN** and English **NN** appear	1778	pairs
Both Japanese **NN** and English **JN** appear	1337	pairs

Table 3: Terminology Appearing in Corpora

We experiment with the translation patterns (α1) Japanese **NN** – English **NN** and (α2) Japanese **NN** – English **JN** in Table 1. The parameters of a function *cor*() in the last section are determined to make the similarity correspondence the weakest of the three types. This is because semantic similarity allows for many more linked words than the other two correspondences. The values we used are: $\lambda_s = 0.8$, $\lambda_c = 0.9$, and $\alpha = 0.7$. Since in English and Japanese **NN**s the rightmost **N** is normally the head of the compound, we use this as the head. The pairs that satisfy $cor(J,E) \geq 0.9$ are selected.

882,250 Japanese **NN** sequences, 419,928 English **NN** and 295,484 **JN** sequences are extracted in the collection phase from two corpora. Table 4 shows the result of picking valid **NN-NN** and **NN-JN** pairs from these sequences in the correspondence phase.

Correspondence	(a)	(a)+(b)	(a)+(c)	(a)+(b)+(c)
NN-NN pairs				
Correct † (Japanese)	610 (601)	876 (820)	990 (925)	1117 (1,032)
Extracted (Japanese)	1,218 (842)	12,873 (1,312)	4,183 (1,186)	15,206 (1,404)
Recall	34.3%	49.3%	55.7%	62.8%
Precision †	50.0%	6.8%	23.7%	7.3%
NN-JN pairs				
Correct †(Japanese)	216 (214)	301 (290)	722 (669)	757 (700)
Extracted (Japanese)	428 (303)	3,660 (574)	2,783 (824)	5,838 (932)
Recall	15.7%	21.9%	52.4%	55.0%
Precision†	50.5%	8.2%	25.9%	13.0%

†Here we only consider pairs appearing in the dictionary as correct.

Table 4: Comparison with the correspondence types

Japanese	English	Frequency (English)	Included in Dictionary
setsubi toushi	equipment investment	32	yes
	facility investment	2	no
akaji sakugen	deficit reduction	148	yes
	deficit cut	28	no
	loss cut	4	no

Table 5: English Translation Candidates of Japanese Compounds

The first column of Table 4 gives the results when our algorithm searches all the English sequences for the counterpart of each Japanese in the answer sets using only (a) lexical correspondence. The low recall indicates that the entries in an ordinary translation dictionary are insufficient for linking the components of compounds even though the dictionary is large. The precision is not high either, because many Japanese words match more than two English words, as shown in Table 5. In this example, the most frequent expressions, "equipment investment" and "deficit reduction", are adequate as the counterparts to the Japanese words, so it is possible to choose a probable candidate by sorting them in order of frequency. The expressions "facility investment" and "deficit cut" are not in the dictionary, but both are acceptable translations. This shows that the method can collect various translations that are not in a dictionary.

Next, we combine (b) similarity correspondence or (c) co-occurrence correspondence with (a) lexical correspondence. We search for translations of the 1,555 Japanese (**NN-NN**) and 1,185 Japanese (**NN-JN**) in the answer sets by (a)+(b) and (a)+(c). For use in (c), a co-occurrence dictionary is compiled from pairs of Japanese and English compounds in the terminological dictionary (excluding pairs in the answer sets).

In the **NN-NN** in Table 4, (a)+(b) exhibits a 15% higher recall than (a) because a similarity correspondence expands the translation candidates. In contrast, its precision is the worst of all on account of the many candidates. The number of candidates must be statistically narrowed down. The recall of correspondence (a)+(c) is higher than (a) and (a)+(b), and the precision is much better than (a)+(b). Note here that the correct pairs are counted only when they appear in the terminological dictionary. We then choose the 476 pairs whose Japanese terms appear most frequently in a corpus from those extracted by (a)+(c), and we manually examine their correctness. As a result, we find that 313 pairs in 476 are correct and the substantial precision is 65.6%. That is, 191 pairs can be new entries in the dictionary.

Finally, we extract translations using (a)+(b)+(c), as shown in the last column of Table 4. Although we must eliminate more inadequate pairs, we can improve the ability of collecting expressions by combining the three types of correspondences.

These results show that extracted pairs contain incorrect translations and these expressions should be filtered for dictionary compilation. To do this we order extracted candidates by using the function *cor*() and frequency of the expressions.

The likelihood of correspondence between two word sequences is indicated by the value of function *cor*(). To differentiate the translation candidates that have the same value of *cor*(), we rank the expressions in descending order of frequency. Figure 1 shows the variation of recall and precision when the top n candidates are adopted. The

Adequate Japanese Compounds	Correct English	17	(23%)	
	Incorrect English	34	(46%)	
	Eliminated	23	(32%)	
	Subtotal	74	(100%)	(18%)
Noise	Eliminated (filtered out)	310	(90%)	
	Incorrect English	35	(10%)	
	Subtotal	345	(100%)	(72%)
Total		419		(100%)

Table 6: Filtering Noisy Word Sequences

precision tends to decline relatively steeply as more candidates are allowed. In contrast, the increase in recall is gradual. These results suggest that the present ranking method is reasonable and the method can produce good results in selecting upper candidates.

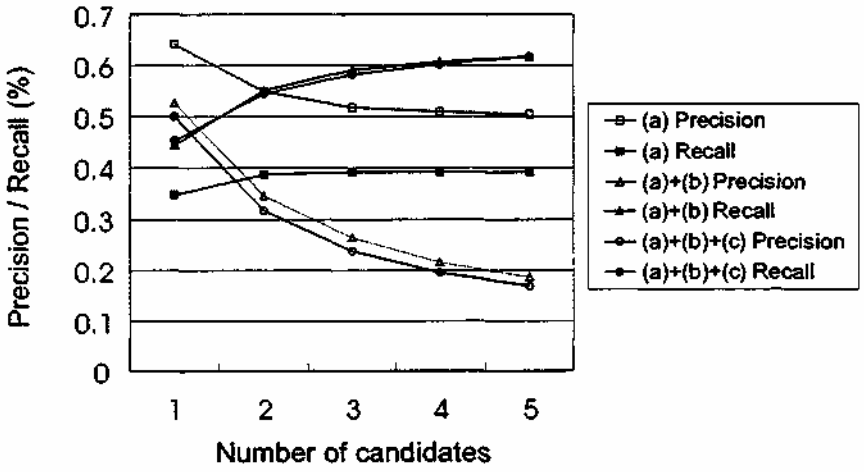


Figure 1: Recall and Precision of top n candidates

Until now, target expressions have been limited to those in the answer sets. To evaluate the capability of filtering out noisy word sequences, we randomly pick 419 word sequences that include a lot of noise from Japanese **NN**, and search for their translations by our method. Table 6 shows that 90% of 345 noisy sequences are filtered out because their counterparts are not found in the correspondence phase. Most of the remaining noise comes from fragments of longer expressions (e.g., **NNN**). Some of these could be removed by using mutual information before the correspondence phase.

Not all translations of Japanese, including expressions peculiar to Japanese, exist in the English corpus, so a 23% rate of finding correct English compounds is a useful result. In addition, 16 out of the 17 correct pairs acquired appear neither in an economic (105,000 entries), an electronics (5,600 entries) or a medical dictionary (83,000 entries); hence, these expressions are worth collecting.

Table 7 shows examples of extracted bilingual expressions. We find that the method can extract many translation equivalents, including pairs that do not appear in the bilingual terminological dictionary, from non-parallel corpora.

Japanese	English	Freq.(Eng.)	Corres.	in Dict.
ryuutsuu kikou	distribution system	87	(a)(b)	yes
kihon keikaku	master plan	13	(c)(a)	yes
shuushin koyou	life employment	1	(c)(a)	yes
tenkan kakaku	conversion value	3	(c)(a)	yes
	conversion price	37	(c)(a)	no
eigyou sonshitsu	operating loss	126	(c)(a)	yes
kiki kaihatsu	hardware development	4	(b)(a)	no
seiken soudatsusen	administration contest	2	(a)(a)	no
houshuu kakusa	pay differences	2	(a)(a)	no

Table 7: Extracted Translations

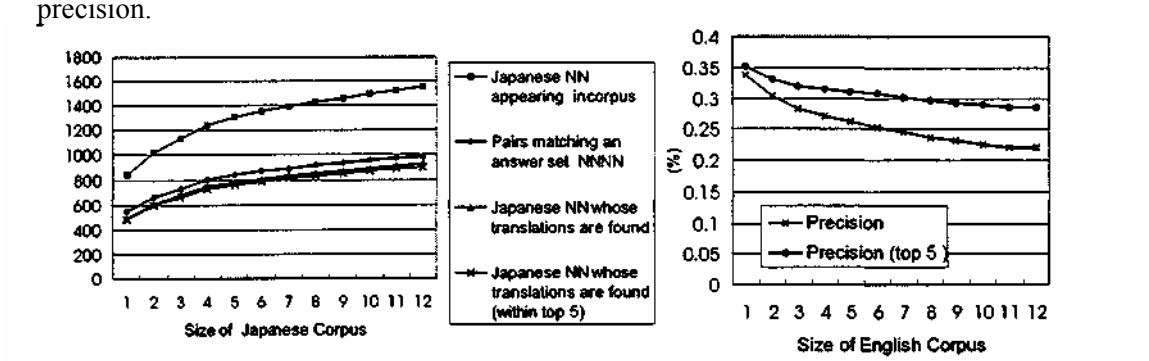
4.2 Influence of size of corpora

The size of corpora directly affects the quantities of the learning knowledge. We divide each corpus into 12 parts and change the number of parts.

Figure 2 (a) shows the number of extracted expressions in the answer set **NN-NN**, when the size of the English corpus is fixed to a full capacity and that of the Japanese corpus changes. The horizontal axis represents the number of corpus parts. As the corpus increases in size, the expressions appearing in the corpus and the extracted translation pairs increase at the same rate when the size is more than 5. They do not become saturated.

On the other hand, when the size of the Japanese corpus is fixed, the precision declines as the size of English corpus increases in Figure 2(b). However, the decrease rate of the precision is weak when the top 5 candidates are selected.

Therefore, if the size of the corpus is enlarged and adopted pairs are limited to upper candidates, we can obtain more translation equivalents while keeping the reasonable precision.



(a) Collected expressions (b) Precision

Figure 2: Variation result according to corpus size

5 Conclusion

We have shown that the proposed algorithm successfully derives Japanese compound nouns and their English translations from non-parallel corpora that have no relation-

ship to one another. Since the method simply uses POS tags, any corpora can be used. We found that introducing two types of correspondences, semantic similarity and co-occurrence, is effective for collecting more translation equivalents than by only using an ordinary dictionary. By ordering collected translation candidates in their likelihood of correspondence and frequency, the method can select more favorable pairs. Furthermore, the method can acquire many expressions worth collecting, which do not appear in bilingual dictionaries.

In the future, we will improve the precision by effectively filtering statistic and linguistic information. We will also investigate the application of the method to other translation patterns.

References

Brill, Eric: 1992, A Simple Rule-Based Part of Speech Tagger, in *Third Conference on Applied Natural Language Processing: ANLP-92,* Trento, pp. 152-155.

Dagan, Ido & Ken Church: 1994, Termight: Identifying and Translating Technical Terminology, in *Fourth Conference on Applied Natural Language Processing: ANLP-94,* Stuttgart, pp. 34-40.

Dagan, Ido & Alon Itai: 1994, Word Sense Disambiguation Using a Second Language Monolingual Corpus, *Computational Linguistics,* 20(4), pp. 563-596.

Fung, Pascale: 1995, A Pattern Matching Method for Finding Noun and Proper Noun Translation from Noisy Parallel Corpora, in *33rd Annual Meeting of the ACL,* Boston, Massachusetts, pp. 236-243.

Fung, Pascale: 1997, Finding Terminology Translations from Non-parallel Corpora, in *5th Workshop on Very Large Corpora,* Hong Kong, pp. 192-202.

Fung, Pascale & Lo Yuen Yee: 1998, An IR Approach for Translating New Words from Non-parallel, Comparable Texts, in *36th Annual Meeting of the COLING-ACL,* Montreal, pp. 414-420.

Gale, William A. & Kenneth W. Church: 1994, A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics,* 19(1), pp. 75-102.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, Toward an MT System without pre-editing – Effects of new methods in ALT-J/E –, in *3rd Machine Translation Summit: MT Summit III,* Washington D.C., pp. 101-106.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama & Yoshihiko Hayashi: 1997, *Nihongo Goitaikei – A Japanese Lexicon,* Iwanami Shoten.

Kaji, Hiroyuki, Yuuko Kida & Yasutsugu Morimoto: 1992, Learning Translation Templates From Bilingual Text, in *14th International Conference on Computational Linguistics: COLING-92,* Nantes, pp. 672-678.

Rapp, Reinhard: 1995, Identifying Word Translations in Non-Parallel Texts, in *33rd Annual Meeting of the ACL,* Boston, Massachusetts, pp. 320-322.

Smadja, Frank, Kathleen R. McKeown & Vasileios Hatzivassiloglou: 1996, Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics,* 22(1), pp. 1-38.

Yarowsky, David: 1995, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *33rd Annual Meeting of the ACL,* Boston, Massachusetts, pp. 189-196.