Extraction of Broad-Scale, High-Precision Japanese-English Parallel Translation Expressions Using Lexical Information and Rules ⋆

Qing Ma[a], Shinya Sakagami[a], and Masaki Murata[b]

[a] Department of Applied Mathematics and Informatics, Ryukoku University,
1-5 Oei-Yokoya, Seta, Otsu 520-2194, Japan
qma@math.ryukoku.ac.jp
[b] Department of Information and Electronics, Tottori University,
4-101 Koyama-Minami, Tottori 680-8552, Japan
murata@ike.tottori-u.ac.jp

Abstract. Extraction was attempted of broad-scale, high-precision Japanese-English parallel translation expressions from large aligned parallel corpora. To acquire broad-scale parallel translation expressions, a new method was used to extract single Japanese and English word n-grams, by which as many parallel translation expressions as possible could then be extracted. To achieve high extraction precision, first, hand-crafted rules were used to prune the unnecessary words often found in expressions extracted on the basis of word n-grams, and lexical information was used to refine the parallel translation expressions. Computer experiments with aligned parallel corpora consisting of about 280,000 pairs of Japanese-English parallel sentences found that more than 125,000 pairs of parallel translation expressions could be extracted with a precision of 0.96. These figures show that the proposed methods for extracting a broad range of parallel translation expressions have reached a level high enough for practical use.

Keywords: parallel translation expression, extraction, rule, lexical information, English-writing support

1 Introduction

Non-native speakers often have problems explaining ideas or presenting achievements in written English, in part because of the large amount of time needed to determine which possible translation of an expression most suits the context. Trying to develop English-writing support tools that will enable non-native speakers to produce nearly perfect English sentences for mixed English-Japanese sentences–in which expressions without know translations are simply written in Japanese–Ma et al. (2008; 2009) have developed systems that can provide support at the word and phrase levels. That is, the given Japanese parts in the mixed English-Japanese sentences can be words or phrases. For phrase-level support in those systems, the Japanese parts are extracted from the mixed sentences and segmented into words, the candidate English equivalents of the segmented Japanese words are identified by searching through a Japanese-English dictionary, and the best equivalents of the Japanese phrases are selected from the combinations of the candidate translations of the single words. This kind of support, however, has two problems. One is the variety of the phrase patterns that the system can support is limited because the English phrases are generated only from the combinations of the candidate translations of single words. The other is the processing is time-consuming because a large number of the translations of single words results in an enormous number of combinations.

⋆ This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI (C) 19500133) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

We think that an effective way to solve these problems would be to use an approach based on translation patterns: to first construct a dictionary of the Japanese-English translation patterns and use that dictionary to provide English-writing support. We therefore tried to extract broad-scale, high-precision Japanese-English parallel translation expressions from large aligned parallel corpora first.[1] To acquire broad-scale parallel translation expressions, we used a new method to extract single Japanese and English word n-grams, by which we could then extract as many parallel translation expressions as possible. To achieve high extraction precision, we first used hand-crafted rules to prune the unnecessary words often found in expressions extracted on the basis of word n-grams, and we further used lexical information from a Japanese-English dictionary to refine these parallel translation expressions. Computer experiments with aligned parallel corpora consisting of about 280,000 pairs of Japanese-English parallel sentences found that more than 125,000 pairs of parallel translation expressions could be extracted with a precision of 0.96. These figures show that the proposed methods for extracting a broad range of parallel translation expressions have reached a level high enough for practical use.

2 System Overview

The inputs to the system developed by Ma et al. (2008; 2009) are mixed English-Japanese sentences that contain Japanese words or (noun and verb) phrases. The Japanese parts are first identified and extracted. For the word-level support, stemming is performed to provide the base forms of the words. The phrase-level support, on the other hand, starts with word segmentation. Words are stemmed and segmented using the Japanese morphological analysis tool Chasen. The candidate English equivalents of the single Japanese words are then identified by searching through the Japanese-English dictionary Eijiro. Queries for word-level support are then constructed by using the equivalent candidates and their contexts, and queries for phrase-level support are constructed by using the combinations of the candidate equivalent of the single words.[2] All queries constructed in this way are searched for in high-quality corpora using an originally developed, flexibly searchable system and searched for in huge amounts of Web data using Google Web APIs to obtain their hits. Finally, the N-best English equivalent candidates with which the queries are constructed that have the largest N numbers of hits are selected as the answers for the systems.

One problem, however, is that the system can support only an extremely limited variety of the phrase patterns. The only noun phrases that can be supported are compound nouns and noun phrases composed of adjectives and nouns, and the only verb phrases that can be supported are those composed of nouns and verbs. Gerund phrases such as “走っている犬” [(1) *hashitteiru inu*, (2) *running dog*, (3) *a running dog*][3], for example, cannot be supported. Neither can translations to different patterns, such as noun phrases to infinitive phrases, be supported. Another problem is that the processing for phrase-level support is time-consuming because the number of combinations of the translations of single words is enormous.

We think that an effective way to solve these problems would be to take an approach based on translation patterns: to first construct a dictionary of the Japanese-English translation patterns and then use that dictionary to provide English-writing support. We therefore tried first to extract broad-scale, high-precision parallel translation expressions of Japanese-English from large aligned parallel corpora.

[1] Related work is discussed in Sec. 3.1.

[2] Acquiring candidate equivalents in phrase-level support is more complicated. See Ma et al. (2009) for details.

[3] The non-English examples are also rendered in (1) *alphabetized form*, (2) *English glosses*, and (3) *English translation* (or grammar explanations if there are no proper translations) complying with the PACLIC 25 Paper Submission Guidelines. The English translations, however, will be skipped if they appear in the pairs of the parallel translation expressions.

3 Extraction of Parallel Translation Expressions

3.1 Related work

The earlier studies most closely related to our work were done by Sainoo et al. (2003) and Kitamura and Matsumoto (1996; 2005). The principal techniques used in those studies were used as our baseline in this study and are described in Sec. 3.2.1. In the study of Sainoo et al., the research object was limited to a small corpus of 8,500 Japanese-English parallel sentences, which were examples from a number of bilingual dictionaries, and only a couple dozen parallel translation expressions were extracted. The first study of Kitamura and Matsumoto (1996), on the other hand, used Japanese-English parallel corpora of three distinct domains: a computer manual, a scientific journal, and business correspondence letters. Each had about 10,000 parallel sentences, and lots of stock phrases would have been used because of the natures of the domains. Pairs therefore can be expected to be easier to extract from these corpora than the copora that cover diverse fields. From each corpus, however, no more than 3,300 pairs were extracted with precisions of 0.87, 0.92, and 0.80. The extraction scale was extremely small, and the extraction precisions were low. In addition, the recall computed on the basis of the definition given in Sec. 4.1.3 was less than or equal to 0.29, a very low figure. Furthermore, single Japanese and English word n-grams were extracted and pairs of the parallel translation expressions were acquired through a repetition processing, which fed back the interim results. However, the whole processing is time-consuming and the extraction is therefore not easy to scale up. In Kimura and Matsumoto's second study (2005) they confirmed the interim results manually, used more linguistic resources than they did in their earlier study, and introduced a method for halving the extraction time by dividing a corpus into quarters. Their method is still hard to scale up, however, because the repetition processing is costly and the results can be manually confirmed only in small-scale extraction.

Among other interesting studies on the extraction of Japanese-English parallel translation expressions are those done by Yamamoto and Matsumoto (2000) and Sato and Saito (2002-1; 2002-2). The methods proposed in those studies, however, like those proposed in the earlier studies mentioned above, have low precisions and recalls and also cannot be scaled up. These methods are thus not practical for use in the extraction of large-scale parallel translation expressions. In a study by Sato and Saito (2002-1), for example, even though the test corpus used for extraction consisted of only about 3,000 parallel sentences, whereas the training corpus consisted of 30,287 parallel sentences, more than ten times the number in the test corpus, the extraction precision was still less than or equal to 0.65. Also, in the study by Yamamoto and Matsumoto (2000), the extraction recall was extremely low: only 461 pairs of parallel translation expressions were extracted with a precision of 0.9 from 13,000 pairs of parallel sentences (meaning the recall was 0.03 based on our definition).

In contrast to the approaches used in these earlier studies, our approaches are practically oriented and intended to be used for extracting broad-scale, high-precision parallel translation expressions useful in English-writing support for all kinds of phrases. Since none of methods proposed by Sato and Saito (2002-1; 2002-2) or by Yamamoto and Matsumoto (2000), the repetition processing used by Kitamura and Matsumoto (1996), or the manually confirmation conducted by Kitamura and Matsumoto (2005) are suitable for this, we instead adapted the methods proposed by Sainoo et al. (2003) and part of the methods proposed by Kitamura and Matsumoto (1996) as the baseline method. To acquire broad-scale parallel translation expressions, we improved the baseline method for extracting single Japanese and English word n-grams so that as many parallel translation expressions as possible can be extracted. To achieve high extraction precision, we first used hand-crafted rules to prune the unnecessary words often found in expressions extracted on the basis of word n-grams, and we further used the lexical information from a Japanese-English dictionary to refine the parallel translation expressions. In addition, we used very large corpora

covering a much wider range of areas and extracted a broad range of parallel translation expressions.

3.2 Methods of extraction

3.2.1 The baseline method Pairs are extracted with the baseline method of parallel translation expressions by (1) extracting the single Japanese and English word n-grams from parallel corpora, and (2) acquiring the parallel translation expressions by computing the similarity between the Japanese and English word n-grams extracted in step (1).

In step (1) each Japanese and English word sequence with an arbitrary length less than or equal to n words (referred to as "word n-gram" in this paper) that appears at least twice in both the Japanese and English sides of the parallel corpora is extracted. Because this can result in many fragmented expressions that do not make sense being extracted, we used inhibition treatment to inhibit the generation of these kinds of expressions. Suppose there are two sentences in a corpus:

s1: a b c d e f

s2: g b c d h

where a, b, $\cdots$, h are words and sequences (b c d), (b c), and (c d) are the word n-grams appearing at least twice in both the Japanese and English sides of the parallel corpora. Only (b c d) is extracted, and (b c) and (c d) are not extracted because they are the parts of (b c d).

In step (2) the similarity between Japanese and English word n-grams is calculated using the *Dice coefficient* (Kay and Röschesen, 1993) below.

$$sim(x_j, x_e) = \frac{2f(x_j, x_e)}{f(x_j) + f(x_e)}, \quad (1)$$

where x_j and x_e are respectively the Japanese and English word n-grams, $f(x_j)$ and $f(x_e)$ are respectively the appearance frequencies of x_j and x_e in the parallel corpora, and $f(x_j, x_e)$ is the frequency of x_j and x_e co-occurring in the same pairs of Japanese-English sentences. That is, the similarity between the Japanese and English word n-grams in each pair extracted in step (1) is computed, and a pair is considered a pair of parallel translation expressions if the similarity is larger than a threshold value.

3.2.2 The improved method for n-gram extraction To acquire as many parallel translation expressions as possible, we improved the method for extracting single Japanese and English word n-grams, step (1) of the baseline method described in Sec. 3.2.1.

By examining the experimental results with the baseline method, we found that a number of parallel translation expressions that could be extracted when the maximum n-gram lengths was set to a small value could not be extracted when the maximum n-gram lengths was set to the larger one. The main reasons can be considered to be as follows. In the baseline method we used inhibition treatment to remove the fragmented expressions. Thus, if the expressions that can originally be extracted when the maximum n-gram length is short, are included in the longer ones that can be extracted when the maximum n-gram length is longer, then the short ones are removed. The parallel translation expression {J: 湾岸諸国と [(1) *wagan shokoku to*, (2) *the gulf states with*], E: with the gulf states}, for example, could not be acquired when the maximum n-gram lengths was set to six because the longer English expression "E: with the gulf states is" was extracted.

To resolve this problem, we do not extract word sequences with an arbitrary length until n-gram in a lump as done in the baseline method. Instead, we extract the word sequences from 1-gram to 2-gram, from 1-gram to 3-gram, $\cdots$, from 1-gram to n-grams in independent steps. When 3-grams are extracted from the sentence "Natural language processing is a field of computer science", for example, the word sequences from 1-gram to 2-gram, i.e.,

1-grams: Nature / language / processing / is / a / field / of / computer /science

2-grams: Natural language / language processing / processing is / is a / a field / field of / of computer / computer science

and the word sequences from 1-gram to 3-gram, i.e.,

1-grams: Nature / language / processing / is / a / field / of / computer /science
2-grams: Natural language / language processing / processing is / is a / a field / field of / of computer / computer science,
3-grams: Natural language processing / language processing is / processing is a / is a field / a field of / field of computer / of computer science

are extracted in each independent step.[4] We also perform the inhibition treatment on them and acquire the parallel translation expressions in these independent steps. We finally assemble the parallel translation expressions obtained in these steps and remove the overlapping ones.

Parallel translation expressions are therefore acquired with the improved method by (1) initially setting the n-gram length to two; (2) extracting the single Japanese and English word sequences from 1-gram to the n-gram from parallel corpora, performing the inhibition treatment on them, and acquiring the parallel translation expressions by computing the similarity between the Japanese and English n-grams extracted; (3) incrementing the n-gram length with 1 and repeating steps (2) and (3) until the maximum n-gram length set beforehand is reached; and finally (4) assembling the parallel translation expressions obtained by steps (2) and (3), and removing the overlapping ones.

3.2.3 The hand-crafted rules By examining the experimental results with the baseline method, we found that ungrammatical Japanese expressions starting with words such as "こと" [(1) *koto*, (2)(3) *thing*] and "いる" [(1) *iru*, (2)(3) a progressive form] and ending with the words such as "この" [(1) *kono*, (2)(3) *this*] and "いろいろな" [(1) *iroiro-na*, (2)(3) *various*] were extracted. Similarly, ungrammatical English expressions starting with possessive forms and ending with articles and auxiliary verb, for example, were extracted. We therefore manually created the following rules and used them to suppress the extraction of ungrammatical single Japanese and English word n-grams.

Rule1
Expressions are eliminated if they end with the words in the list.
List
- Japanese
  - adnominal
    e.g., この [(1) *kono*, (2)(3) *this*], あの [(1) *ano*, (2)(3) *that*], いろいろな [(1) *iroiro-na*, (2)(3) *various*]
  - prefix
    e.g., もと [(1) *moto*, (2)(3) *the former*], アンチ [(1) *anchi*, (2)(3) *anti-*], 超 [(1) *chou*, (2)(3) *super*], 反 [(1) *han*, (2)(3) *anti-*], 新 [(1) *shin*, (2)(3) *new*]
  - adverb
    e.g., たいそう [(1) *taisou*, (2)(3) *very*], いつも [(1) *itsumo*, (2)(3) *always*], 人一倍 [(1) *hotoichibai*, (2)(3) *more than others*]
  - conjunctive noun
    e.g., 兼 [(1) *ken*, (2)(3) *and*], 対 [(1) *tai*, (2)(3) *versus*]
- English
  article (e.g., a, the), preposition (e.g., about, at, by, in, out, of), auxiliary verb (can, may, will, must, shall), to, 4W1H (what, where, who, which, how)

Rule2
Expressions are eliminated if they start with the words in the list.
List

[4] In actual n-gram extraction, the article "a" was not counted as a word in making N-grams. For details see Sec. 4.1.2.

Japanese
- non-independent verb
  - e.g., しまう [(1) *shimau*, (2)(3) *have*], 願う [(1) *negau*, (2)(3) *wish*]
- verb suffix
  - e.g., する [(1) *suru*, (2)(3) *do*], られる [(1)*rareru*, (2)(3) a passive, させる [(1) *saseru*, (2)(3) *make*], いる [(1) *iru*, (2)(3) a progressive form]
- adjective suffix
  - e.g., ったらしい [(1) *ttarashii*, (2)(3) *new*], っぽい [(1) *ppoi*, (2)(3) *-ish*]
- non-independent adjective
  - e.g., づらい [(1) *durai*, (2)(3) *difficult*], がたい [*gatai*, (2)(3) *difficult*], よい [(1) *yoi*, (2)(3) *can/may*]
- non-independent noun
  - e.g., こと [(1) *koto*, (2)(3) *thing*], ため [(1) *tame*, (2)(3) *for*]
- non-independent noun-verb
  - e.g., ごらん [(1) *goran*, (2)(3) *try*], ちょうだい [(1) *choudai*, (2)(3) *please*]
- noun suffix
  - e.g., 化 [(1) *ka*, (2)(3) *-lization*], ぎみ [(1) *gimi*, (2)(3) *tend to*]
- particle
  - e.g., は [(1) *ha*], から [(1) *kara*], を [(1) *wo*]
- auxiliary verb
  - e.g., らしい [(1) *rashii*, (2)(3) *seem*], た [(1) *ta*, (2)(3) a past form]

English
- possessive form (e.g., 's)

3.2.4 The lexical information To further improve the extraction precision, we used the lexical information to refine the parallel translation expressions extracted with the methods described in the previous sections. Each pair of parallel translation expressions is refined as follows. (1) Independent words are extracted from the Japanese side, and their English translation candidates are obtained from a Japanese-English dictionary.[5] However, the word sequences (i.e., word n-grams) with arbitrary lengths that include both independent and function words are regarded as single words, and their English translation candidates are searched for through the dictionary. (2) The translation candidates obtained in the first step and the English words of the English side are compared. For the translation candidates that match the English words, their original Japanese words and the matching English words are classified into the "matching list". Those that do not match, on the other hand, are classified into the "non-matching list". Also, the words for which translations cannot be found in the dictionary are not classified into either list. (3) The numbers of words in the two lists are compared. If the "matching list" has more words than the "non-matching list", the pair of parallel translation expressions then passes. Otherwise, the pair is eliminated. Furthermore, if the passing pairs are in many-to-many relationships, which means a Japanese expression has multiple English translations and vice versa, then the pair with the largest number of words in the "matching list" is finally selected.

We use the parallel translation expression {J: 湾岸戦争の勝利 [(1) *wagansensou no shouri*, (2) *the gulf war of victory*], E: victory in the gulf war} as an example to explain the refinement procedure described above. In step (1), we first extract the independent words "湾岸" [(1) *wangan*, (2)(3) *gulf coast*], "戦争" [(1) *sensou*, (2)(3) *war*], and "勝利" [(1) *shouri*, (2)(3) *victory*] from the

[5] The nominal verbs, which have been segmented by the morphologic analysis, are connected again as single words for dictionary searching.

Japanese side. However, although the words in pair {J: 湾岸戦争 [(1) *wangansensou*, (2)*gulf war*], E: gulf war} are in correct parallel translation, the single words "湾岸" [*gulf coast*] and "gulf" are not. This means that even a pair in perfectly correct parallel translation might be judged to be incorrect by using the lexical information if only the independent words are used. Thus, we also make n-grams with arbitrary lengths and regard them as single words for dictionary searching. As a result, we obtain the following eight expressions[6]: "湾岸" [*gulf coast*], "戦争" [*war*], "勝利" [*victory*], "湾岸戦争" [*gulf war*], "戦争の" [*of war*], "戦争の勝利" [*victory of war*], and "湾岸戦争の勝利" [*victory in the gulf war*]. We then obtain their English translation candidates by searching the dictionary. In step (2), we first check the English translation candidates of the independent words and the English expressions. As a result, "戦争" and "war" and "勝利" and "victory" are matched and classified into the "matching list", while "湾岸" and "gulf" are not matched and are classified into the "non-matching list". We then check the English translation candidates of the n-gram words and the English expression. Since the English translation candidates of "湾岸戦争" match "gulf war", we compare them with the words in the "non-matching list". Since these words partially match the words "湾岸" and "gulf", which are in the "non-matching list", they are moved to the "matching list". In this way, "matching list" has six words and and the "non-matching list" has none. Since the "matching list" has more words than the "non-matching list", the parallel translation expression passes. Furthermore, if this parallel translation expression and the other passing parallel translation expressions are in many-to-many relationships, the one with the largest number of words in the "matching list" is selected.

4 Experiments

4.1 Experimental setup

4.1.1 Data and tools The aligned Japanese-English parallel corpora that we used in the experiments consisted of the NICT corpus (Uchimoto et al., 2004), the JENNAD corpus, and the aligned Reuters corpus (Utiyama and Isahara, 2003). The NICT corpus is composed of sentence-aligned Mainichi newspaper articles (Japanese) and their translations done by professional translators and has about 40,000 pairs of parallel sentences. The JENNAD corpus is composed of automatically sentence-aligned Yomiuri newspaper articles (Japanese) and Daily Yomiuri newspaper articles (English) and has about 180,000 pairs of parallel sentences. The aligned Reuters corpus is composed of automatically sentence-aligned Reuters Japanese and English news articles and has about 70,000 pairs of parallel sentences. After the overlapping sentences were excluded, the corpora had 278,323 pairs of parallel sentences.

Eijiro was used for the Japanese-English dictionary, and Chasen and TreeTagger were used for Japanese and English morphologic analysis.

4.1.2 Extraction requirement On the basis of a preliminary examination, we found that few pairs are in correct parallel translation if either the Japanese side or English side has an n-gram longer than six or if the difference between the number of words in Japanese and English sides is larger than three. On the basis of these observations, the maximum n-gram length for extraction was set to six (however, to see the extraction results with shorter values, the experiment with a maximum n-gram of three was also performed for the baseline method.) Also, the pairs in which the difference between the number of words in Japanese and English sides is larger than three were removed from the parallel expression candidates before their similarities were computed. Single Japanese and English n-grams were extracted if they appeared at least twice on both the Japanese and English sides of the corpora. The parallel translation expression candidates were acquired on the condition that the Japanese and English n-grams co-occurred at least twice in the same pairs

[6] To reduce the processing time for making n-grams, we used the rules described in Sec. 3.2.3 to exclude expressions such as "の勝利" [(1) *no shouri*, (2) *of victory*, (3) *victory of*].

of Japanese-English sentences and their similarities were greater than or equal to 0.5. In handling English expressions, we did not distinguish between person, tense, or singular or plural forms. We also did not distiguish between capital and lower-case letters. Furthermore, we did not count the articles “a”, “an”, and “the” as single words when making n-gram, i.e., we regarded the word sequence “the structure reform of the labor market” as a 5-gram, not a 7-gram.

4.1.3 Evaluation methods Extraction precisions, recalls, and F-measures were used for evaluation. The precisions were obtained by manually evaluating 500 pairs selected at regular intervals from all extracted parallel translation expressions. The recalls were used to estimate how many correct expressions have been extracted relative to the size of the corpora used for extraction and were defined as follows.

$$Recall = \frac{N_{ex} \times Precision}{N_{all}},$$

where N_{ex} is the number of the extracted pairs of parallel translation expressions and N_{all} is the total number of the pairs of parallel sentences in the corpora used for extraction. Since we could not know the exact number of correct parallel translation expressions existing in the parallel corpora, we thus used the size of the corpora (i.e., the figure N_{all}), instead.[7]

4.2 Results

The extraction results are listed in Table 1, where the *Baseline (3-gram)* and *Baseline (6-gram)* are the baseline methods with the maximum n-gram lengths of three and six, respectively, the *Improved ex. (6-gram)* is the improved method with the maximum n-gram length of 6 for n-gram extraction, the *Rule* is the method applying the hand-crafted rules to the expressions when extracting n-gram using the *Improved ex. (6-gram)*, and the *Lexicon* is the method using lexical information to further refine the expressions that have already been extracted using the *Rule*.

Table 1: Number of extracted pairs of parallel translation expressions, precisions, recalls, and F-measures.

Methods	Number	Precision	Recall	F-measure
Baseline (3-gram)	258,994	0.09	0.08	0.08
Baseline (6-gram)	190,895	0.14	0.09	0.11
Improved ex. (6-gram)	808,556	0.18	0.52	0.27
Rule	481.396	0.35	0.61	0.44
Lexicon	**125,165**	**0.96**	**0.43**	**0.59**

From the table we first see that both the precision and recall of the baseline method were extremely low and the different values for the maximum n-gram length did not affect the performance much. We then see that the *Improved ex. (6-gram)* dramatically increased the number of extracted expressions, while achieving higher precision than the baseline methods. As a result, the recall was tremendously improved, and the F-measure was also largely improved. By examining the results obtained by the baseline methods and the *Improved ex. (6-gram)*, on the other hand, we found lots of parallel translation expressions, as shown in the below examples, that could be extracted using the *Improved ex. (6-gram)* but not the baseline methods. That is, the *Improved ex. (6-gram)* extracted more numerous phrases with a larger variety of patterns such as past-participle phrases with the form “V された N” [(1) *V sareta N*, (2)(3) *Ved N*] and infinitive phrases with the form “N を V するため” [(1) *N wo Vsuru tame*, (2) *N to V*, (3) *to V N*], than the baseline methods.

Examples extracted by *Improved ex. (6-gram)*

J1: 暗殺されたラビン首相

[(1) *ansatsu sareta rabin shushou*, (2) *assassinated rabin prime minister*]

E1: prime minister rabin who was assassinated

[7] As a matter of course, the recall defined in this way can be a value larger than 1 in theory, and the F-measures cannot be therefore used in a strict sense.

J2: 不均衡を縮小するため
[(1) *fukinkou wo shukushou suru tame*, (2) *imbalance reduce to*]
E2: to reduce imbalance
J3: 拉致疑惑問題
[(1) *rachigiwaku mondai*, (2)*kidnapping issue*]
E3: the kidnapping issue
J4: 揶揄された
[(1) *yayu sareta*, (2) *ridiculed*]
E4: have been ridiculed

From the table we further see that the *Rule* almost doubled the extraction precision of the *Improved ex. (6-gram)*. Thus, although the number of extracted expressions largely reduced, the recall and F-measure were still largely better than those of the *Improved ex. (6-gram)*. Furthermore, although the number of extracted expressions largely reduced, by examining the results we could also find some parallel translation expressions that could only be extracted by using the *Rule*. The parallel translation expression {J: 寄与する必要がある [(1) *kiyo suru hitsuyou ga aru*, (2) *contribute need*], E: need to contribute}, for example, has different numbers of Japanese words (five words) and English words (three words). To extract such an expression, we first need to extract all the English word sequences between 1-gram and 5-gram, as well as all those of the Japanese, and then compute their similarities. However, since the 4-gram "need to contribute to" existed in the corpora, the 3-gram "need to contribute" was removed by the inhibition treatment in the baseline methods and the *Improved ex. (6-gram)*. This parallel expression therefore could not be extracted in these methods. In the *Rule*, however, since the 4-gram "need to contribute to" ended with the word "to" and was removed by the rules, the word sequence "need to contribute" was therefore not inhibited. Thus, this parallel translation expression could be extracted by the *Rule*.

Finally, from Table 1 we see that the *Lexicon* dramatically improved the extraction precision and had a precision of 0.96, thus reaching a level high enough for practical use. Although its recall was lower than that of the *Rule*, its F-measure is much higher than that of the *Rule* and is the highest among the all methods. One reason for the lower recall with the *Lexicon* can be considered to be as follows. As mentioned in Sec. 4.1.1, the NICT corpus, a part of the parallel corpora we used, for example, is composed of Mainichi newspaper articles (Japanese) and their translations done by professional translators. This means that a number of parallel sentences of the corpora are in free translation relationships. Thus, the English words used are usually different to the translations of the Japanese words obtained from the Japanese-English dictionary, and the expressions in correct parallel translations might be judged as incorrect and removed from the final extraction results. Specifically, the parallel translation expression {J: 邱政雄財政部長 [(1) *chiu masao zaiseibuchou*, (2) *chiu paul finance minister*], E: finance minister paul chiu}, for example, could be extracted with the *Rule* and should be regarded as a correct one. Since the *Lexicon*, however, has no translations of "邱" [(1) *chiu*], "政雄" [(1) *masao*], or "部長" [(1) *buchou*] that matched "chiu", "paul", or "minister", these were classified into the "non-matching list". Also, since there were no translations for the n-grams such as "邱政雄" [(1) *chiu masao*, (2) *chiu paul*, (3) *paul chiu*] and "財政部長" [(1) *zaiseibuchou*, (2)(3) *finance minister*], the words in the "non-matching list" could not be moved to the "matching list". Thus, the parallel translation expression was judged as incorrect and removed from the final extraction results.

5 Conclusion

We presented practically oriented approaches for extracting a broad range of parallel translation expressions for use in English-writing support. Our study differs from the earlier related studies in two ways. One is that we used very large corpora covering a much wider range of areas so

that we could extract a broader range of parallel translation expressions. The other is that we used a new n-gram extraction method for acquiring as many parallel translation expressions as possible, and adopted hand-crafted rules and lexical information for refining the extracted expressions. Computer experiments with aligned parallel corpora consisting of about 280,000 pairs of Japanese-English parallel sentences found that more than 125,000 pairs of parallel translation expressions could be extracted with a precision of 0.96. We think that these figures show that the proposed methods for extracting a broad range of parallel translation expressions have reached a level high enough for practical use, and to our knowledge, ours are the first practically oriented approaches for extracting a wide range of Japanese-English parallel translation expressions.

We plan to construct a dictionary of Japanese-English translation patterns using the parallel translation expressions acquired in this study and build it into an English-writing support system.

References

Kay, M. and M. Röschesen. 1993. *Text-translation alignment*. Computational Linguistics, 19(1), pp. 121–142.

Kitamura, M. and Y. Matsumoto. 1996. *Automatic extraction of word sequence correspondences in parallel corpora*. Proceedings of the 4th Workshop on Very Large Corpora, pp. 79–87.

Kitamura, M. and Y. Matsumoto. 2005. *Practical translation pattern acquisition from combined language resources*. The First International Joint Conference on Natural Language Processing, Revised Selected Papers, Lecture Notes in Artificial Intelligence 3248, pp. 244–253.

Ma, Q., K. Nakao, M. Murata and H. Isahara. 2008. *Selection of Japanese-English equivalents by integrating high-quality corpora and huge amounts of web data*. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008).

Ma, Q., R. Mori and M. Murata. 2009. *Development of English-writting support systems*. Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (Pacling2009), pp. 171–176.

Sainoo, D., J. Murakami, M. Tokuhisa and S. Ikehara. 2003. *Toward automatic extraction of Japanese/English expression pattern-pairs*. IPSJ SIG Technical Reports, 2003-NL-153, pp. 113–118 (in Japanese).

Sato, K. and H. Saito. 2002-1. *Extracting word sequence correspondences with support vector machines*. Proceedings of the 20th International Conference on Computational Linguistics (COLING2002), pp. 870–876.

Sato, K. and H. Saito. 2002-2. *Extracting bilingual word pairs with maximum entropy modeling*. Proceedings of Sixth World Multiconference on Systemics, Cybernetics and Informatics (SCI2002). Vol. 3, pp. 412–417.

Uchimoto, K., Y. Zhang, K. Sudo, M. Murata, S. Sekine and H. Isahara. 2004. *Multilingual aligned parallel treebank corpus reflecting contextual information and its applications*. Proceedings of the Workshop on Multilingual Lingustic Resources (MLR2004), pp. 63–70.

Utiyama, M. and H. Isahara. 2003. *Reliable measures for aligning Japanese-English news articles and sentences*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003), pp. 72–79.

Yamamoto, K. and Y. Matsumoto. 2000. *Acquisition of phrase-level bilingual correspondence using dependency Structure*. Proceedings of the 18th International Conference on Computational Linguistics (COLING2000), pp. 933–939.