Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora

Keita Tsuji

Library and Information Science Course, Graduate School of Education, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
Email: keita@nii.ac.jp

Abstract

The method to automatically extract translational Japanese-KATAKANA and English word pairs from bilingual corpora is proposed. The method applies all the existing transliteration rules to each mora unit in a KATAKANA word, and extract English word which matched or partially-matched to one of these transliteration candidates as translation. For instance, if there is a word '**グラフ**' (graph) in Japanese part of bilingual corpora, we generate such transliteration candidates as <graf>, <graph>, <gulerph>,... and extract similar words from English part of corpora. The method worked fairly well, achieving 83-100% precision at 75% recall against eight corpora we used for evaluation.

Keywords: transliteration, automatic construction, bilingual lexicon, bilingual corpora, KATAKANA.

1 Introduction

We propose a method to automatically extract translational Japanese-KATAKANA and English word pairs from bilingual corpora based on transliteration rules.

The bilingual lexicon is in demand in many fields such as cross-language information retrieval and machine translation. And the bilingual corpora become widely available reflecting the change in publishing activities. On these background, the methods to construct bilingual lexicon automatically based on bilingual corpora have been intensively studied. But many of the methods proposed so far depend heavily on word frequency in the corpora and cannot treat low-frequency words properly. The low-frequency words include newly-coined words which are especially in demand in many fields.

Reflecting the recent relationship between Japanese and English environment, Japanese is now borrowing many words from English. Thus the new words are often loan words and are usually represented by KATAKANA characters. Some methods to automatically extract KATAKANA-English word pairs from bilingual corpora have been proposed, but there is still much room for further investigation.

Against these background, we propose a method to extract translational KATAKANA-English word pairs from bilingual corpora. The method applies all the existing transliteration rules to each mora unit in a KATAKANA word, and extract English word which matched or partially-matched to one of these transliteration candidates as translation. For instance, if there is a word 'グラフ' (graph) in Japanese part of bilingual corpora, we generate such transliteration candidates as <graf>, <graph>, <gulerph>,... and extract similar words from English part of corpora.

In the next section, we will explain our extraction method in detail, together with some alternatives and related studies. Next, we will show the results of our experiments and discuss the future direction.

2 Method

Our method extracts translational KATAKANA-English word pairs from bilingual corpora based on transliteration rules. In this section, we will first explain the way to construct the transliteration rules, and then the way to extract KATAKANA-English translations from corpora.

2.1 Construction of Transliteration Rule

Based on a source list of KATAKANA-English word pairs (which can be obtained from Japanese-English dictionaries, etc.) and Hepburn transliteration rules, we construct the transliteration rules as follows:

(1) Decompose KATAKANA word in the source list into mora units. For instance, the word 'ディスパッチャー' is decomposed into four units: 'ディ', 'ス', 'パッ' and 'チャー'. [1]

(2) Based on the English counterpart word, extract transliteration rule for each unit manually. For instance, from a pair 'ディスパッチャー' and 'dispatcher', we

[1] We attached 'ー' to the preceding unit and regarded it as one unit, though its number of mora is two.

can obtain the rules: ‘ディ’= ‘di’, ‘ス’= ‘s’, ‘パッ’= ‘pat’ and ‘チャー’= ‘cher’. [2] [3]

(3) Repeat (1) and (2) for all the word pairs in the list, count the frequency of each rule (the number of times observed in the list) and rank them for each KATAKANA unit. The ranks are used for time-saving as we will mention later.

(4) Add Hepburn transliteration rules into the above rules, with each frequency 1. This is done to supplement the rules extracted from the source list. If the source list is large enough, this process might not be necessary.

Henceforth we represent the rules obtained by the above procedure as ‘TR’.

We assume the following four: (1) when Japanese borrows an English word, the word is transliterated on the basis of Japanese mora unit and their correspondences are stable, (2) these transliterations are free from context (have no relation to the preceding or following mora units), (3) the number of transliteration rules is small and (4) they do not vary drastically from time to time nor from domain to domain. Therefore, TR construction is a finite task and once it is obtained, it can be used for a long period of time and for various domains.

2.2 Extraction of Traslational Word Pairs

In this section, we will explain our method to extract KATAKANA-English translations from bilingual corpora. First of all, the basic procedure which in fact is computationally prohibitive is described. And then its countermeasure, the device for time-saving, is introduced. The combination of the two, i.e., the procedure which is modified to be fast enough is our method. Next, we will explain some alternatives and methods in related studies.

2.2.1 Basic Procedure

First of all, we define the following symbols and functions.

J: KATAKANA word in the corpus
E: English word in the corpus
L(w): The number of characters of word *w*
T(J): The transliteration candidate of word *J*
$S(w_1, w_2)$: The longest common subsequence of w_1 and w_2 (e.g. *S(*‘guraf’, ‘graph’*)* = ‘gra’)
*Dice(k,m,n)=k*2/(m+n)*

Extracting KATAKANA-English translations from bilingual corpora is achieved as follows:

(1) Pick up *J* and decompose it into units according to the same framework we used at TR construction.

(2) Using all the transliteration rules in TR, generate all the possible transliteration candidates for *J*. Henceforth we represent the *i*-th transliteration candidate of *J* as $T_i(J)$.

(3) Pick up *E* which co-occurred with *J* (occurred in the same aligned segment in the corpus) and identify the longest common subsequence with each $T_i(J)$.

(4) If the following *P(J, E)* exceeds certain threshold, extract pair *J* and *E* as translation.

$$P(J, E)=max_i\ Dice(L(S(T_i(J), E)), L(T_i(J)), L(E))$$

In the case that TR is as in Table 1 and *J* is ‘グラフ’, 3*6*4 transliteration candidates such as <graf>, <graph>, <graff>, ..., <gulerff> and <gulerfe> are obtained. Therefore, in the case that *E* is ‘graph’:

*P(*グラフ, graph*)*
= *max(* *Dice(L(S(*graf, graph*)), L(*graf*), L(*graph*)),*
*Dice(L(S(*graph, graph*)), L(*graph*), L(*graph*)),*
*Dice(L(S(*graff, graph*)), L(*graff*), L(*graph*)),*
…
*Dice(L(S(*gulerfe, graph*)), L(*gulerfe*), L(*graph*)))*
= *max(*0.67, 1.00, 0.60, …, 0.33, 0.33*)* = 1.00

It indicates that ‘グラフ’ and ‘graph’ are very likely to be translation. On the other hand, in the case that *E* is ‘library’, *P(J, E)* is not high (*P(*グラフ, library*)=max(*0.36, 0.33, ..., 0.29, 0.29*)*=0.36), which indicates that they are not likely to be translation.

Our basic idea is as follows. If we use all the transliteration rules and generate all the possible transliteration candidates, ‘correct’ transliteration is always in that set, though we do not know which is the correct one. We use bilingual corpora to resolve this, assuming that there exists a transliteration of KATAKANA word in English part of bilingual corpora. We regard the transliteration candidates which actually exist as words in English part as the correct ones. However, the transliteration rules we obtain are inclined to be insufficient. Therefore we do not use exact match. Instead, we use *Dice* and extract the actual word in the English part of corpora, not *T(J)*, whose maximum of *Dice* is high, as translation.

[2] We used all the characters of English counterpart. The characters which seem to have no corresponding Japanese part are attached to the preceding characters. For instance, from a pair ‘ウェイト’ and ‘weight’, transliteration rule ‘イ’ and ‘igh’ is extracted.

[3] We did not use English characters overlappingly. Only exception is ‘x’. When ‘x’ corresponds to ‘クス’ (or ‘キス’) , we extracted transliteration rules ‘ク’= ‘x’ and ‘ス’= ‘x’, though it contradicts to our mora-basis correspondence assumption.

グ	ラ	フ
g	ra	f
gue	la	ph
gu	l	ff
	lu	fe
	r	
	ler	

Table 1: *Transliteration rules*

2.2.2 Device for Time-saving

The previously-mentioned procedure requires much computational time when applied to the actual data, because (1) using all rules in TR often leads to the combinatory explosion of the number of transliteration candidates, (2) identifying the longest common subsequence often requires much time.

As for (1), we decided to apply less transliteration rules to longer KATAKANA word. At TR construction, we have ranked transliteration rules according to their frequencies. We applied top 12*/(the number of units in J)*+1 rules to each unit of *J*. In the case that TR is as in Table 1 and *J* is 'グラフ', the number of rules applied to each unit is 12/3+1=5. Therefore 3*5*4 transliteration candidates are examined instead of 3*6*4 candidates.

As for (2), we used 'NPT_score' in [1] for computing *L(S(T(J), E))* . It is an abbreviated version for identifying the longest common subsequences.

2.3 Other Alternatives

So far we have explained our method to extract translational KATAKANA-English word pairs from bilingual corpora. To verify the effectiveness of our method, we examined the following alternatives.

2.3.1 Transliteration Rule

The simple Hepburn transliteration rule (henceforth 'HR') was used on behalf of TR. The purpose and perspective of comparing the effectiveness of TR and HR is as follows: (1) HR is an well-known rule for transliterating Japanese to alphabet strings and is easily available. If HR alone can produce good result, we can save labor for constructing TR. (2) HR gives each KATAKANA unit a unique alphabet string. Thus, while TR usually generates many *T(J)*, HR generates only one *T(J)*. If HR alone can produce good result, we can save computational time.

2.3.2 Measure for Matching

Our method uses the combination of (1) *Dice* and (2) NPT_score. But the other alternatives might be more effective. From this view, we examined the following two:

(1) The following *Mdic* was used on behalf of *Dice*:

$$Mdic(k, m, n)=(1+\log k)*k*2/(m+n)$$

Dice is simply a ratio between the length of the matched part and those of the words. For instance, *Dice(*4,5,5*)=Dice(*8,10,10*)=Dice(*12,15,15*)*… The measures which emphasize the length of the matched part more might be more effective than *Dice*.

(2) The following *Bgrm* was used on behalf of the combination of *Dice* and NPT_score:

$$Bgrm(T(J), E)=|N_{T(J)} \cap N_E|/|N_{T(J)} \cup N_E|$$

where N_w is a set of bi-grams which compose the word *w*. For instance, N_{graph} is (_g, gr, ra, ap, ph, h_). This measure was used in [2] which we will mention later.

2.3.3 Assumption of One-to-one Correspondence

The method to extract translational word pairs from bilingual corpora based on word frequency often assumes one-to-one correspondence between words in one aligned segment ([3] [4]). We examined whether the assumption is also effective for our task.

We added the following extraction procedure to our method and examined the performance improvement:

(1) From one aligned segment, extract the pair whose P-value is the highest in the segment.
(2) Eliminate the words of that pair from the segment.
(3) If there remains KATAKANA word and English word in the segment, go to (1), otherwise stop.

The procedure is applied against all the segments in the corpus. Henceforth we will represent this procedure as '1-1P' for simplicity.

2.4 Methods in Related Studies

The methods to extract Japanese-English loan word pairs from bilingual corpora were proposed in [1][5][6][7]. But [5] did not go into details of transliteration method. [7] used only consonants for Japanese English matching. The vowels which we think have useful information are ignored. [6] extracted pairs whose estimated pronunciations matched exactly. But the method to estimate the pronunciation of English word (which seems difficult) is not clearly explained.

On the other hand, [1] did not discard vowels and showed their procedure clearly. Thus we took up parts of their method and compared the effectiveness with our method. We extracted KATAKANA-English word pairs as follows: (1) Transliterate KATAKANA word into 'NPT' in [1], (2) Extract word pairs whose '*Match*' are high.

[1] aimed at extracting proper names. Toward this goal, [1] excluded from candidates, English words whose first letters are not in upper cases. But our extraction target is not limited to proper names. Thus we did not use this and other additional rules used in [1] such as limiting the minimum length of KATAKANA word.

Apart from Japanese, [2] proposed a method to identify the original English words for Korean words based on transliteration. Our method is similar to theirs in using and combining all the transliteration rules observed in the training corpora. The main difference is that while they use transition probabilities from one bi-gram (which composes the word) to another, we do not consider such transitions. We assume that Japanese loan word and original English word correspond on mora-basis and occurrences of the corresponding units do not depend on the precedent units. [4]

3 Experiment

In this section, we will first explain the data we used and then the result of our method, other alternatives and the method in related studies.

3.1 Data

The source data for TR construction and bilingual corpora from which we extract translational pairs will be explained.

3.1.1 Source Data for TR

We extracted transliteration rules from 3,742 transliterational term pairs in dictionary of artificial intelligence [8] and added Hepburn transliteration rules to them. The total number of rules in TR is 1,236. The number of types of KATAKANA units in TR is 677. Thus, the average number of transliteration rules for one KATAKANA unit is 1.83. The KATAKANA unit which has transliteration rules most is *'ク'*, which has 13 rules (*'ク'*= 'c', 'k', 'que', 'ke',..).

3.1.2 Bilingual Corpora

To evaluate the performance of our method, eight bilingual corpora were used. They are Japanese-English parallel abstracts/titles of academic papers, which were made by the authors of each paper. The domains of these corpora are artificial intelligence (AI), forestry (FR), information processing (IP) and architecture (AC). They were extracted from the Database of Academic Conference Papers provided by NACSIS.

The Japanese and English parts of these corpora are processed by morphological analyzer ChaSen2.0b and Brill part-of-speech tagger respectively. We regarded the translational KATAKANA-English single-word noun pairs in each segment of corpora as extraction target.

The basic quantities of the corpora are shown in Table 2. In Table 2, 'KEp' indicates the number of types of the extraction target pairs. 'Jw' indicates the number of tokens of ChaSen morphemes in the Japanese part of corpora and 'Ew' indicates the number of tokens of English words. We regarded one abstract (title) as one segment. Each corpus is composed of 1,000 segments except for abstract corpus of forestry (950 abstracts).

The purpose of using abstract and title corpora is to examine the influence of size and the degree of parallelism of each segment to the extraction results. Abstracts are larger and noisier than titles.

The purpose of using four domains is to examine the influence of domain difference to the results. In particular, we constructed TR based on a dictionary of artificial intelligence. If the results of the other three domains do not significantly differ from that of artificial intelligence, it can be said that we need not be so nervous about the domain of source data for TR construction.

	Domain	Jw	Ew	KEp
Abstract	AI	166,331	126,101	493
	FR	185,734	113,789	470
	IP	156,124	106,753	728
	AC	139,420	74,776	348
Title	AI	11,435	9,625	227
	FR	15,062	14,575	161
	IP	11,244	9,153	267
	AC	20,118	17,200	169

Table 2: *Basic Quantities of Bilingual Corpora*

3.2 Results

The precision and recall of our extraction method, other alternatives and the method in related studies are shown in Figure 1-8. In these Figures, 'TR_Dice_0' represents the result of our method. 'HR_Dice_0' represents the result of using HR on behalf of TR. 'TR_Mdic_0' represents the result of using *Mdic* on behalf of *Dice*. 'TR_Bgrm_0' represents the result of using *Bgrm* on behalf of *Dice* and NPT_score. 'TR_Dice_1' represents the result of our method with 1-1P. 'NR_Npts_0' represents the result of using NPT and *Match* in [1].

From Figure 1-8, we can see the following:

(1) TR is much more effective than simple Hepburn transliteration rules. For instance, while 'TR_Dice_0' achieved 93% precision at 75% recall, the precision of 'HR_Dice_0' at 75% recall remained 4% against the abstract corpora of artificial intelligence. Therefore TR is worth constructing and using.

(3) *Dice* is more effective than *Mdic*. In bilingual corpora, there are many word pairs which are long and morphologically-related but not translational. *Mdic* between these word pairs tend to become higher than those of short translational pairs, which leads to the decrease of precision.

[4] This assumption is also practically important. Using the transition probability will lead to the problem of data sparseness.

(3) The combination of *Dice* and NPT_score is more effective than *Bgrm*. One apparent reason is that the latter does not use the information of the bi-grams order in original words.

(4) 1-1P slightly improves the precision at the expense of recall. Thus 1-1P can be a potential option to our method under some circumstances or framework where higher precision is important.

(5) Our method is more effective than using combination of NPT and *Match*. The main reasons are as follows: (1) Compared with TR, the transliteration rule in [1] tends to generate strings which contain incorrect translations. For instance, a Japanese word 'グラフ' is transliterated into 'ghurlaoffphu', which contains not only correct translation 'graph', but also incorrect one 'hop'. (2) Unlike *Dice*, *Match* depends only on the length of English words and their matched parts. It does not use information about the length of no-matched parts in generated strings. Therefore, it tends to evaluate short English words which are contained by chance in generated strings as correct translations (for instance, *Match*('ghurlaoffphu', 'hop')=3/3=1).

(6) TR which was constructed based on artificial intelligence term pairs also performed well against the other three domains. This indicates the domain-independence of TR and that we do not have to construct TR for each domain.

(7) Our method achieved 96-100% precision at 75% recall against title corpora. Though it was worse than that, our method performed well against abstract corpora (83-93% precision at 75% recall). Considering the high availability of bilingual corpora aligned at abstract level, it is worth noticing.

3.3 Error Analysis

The extraction errors observed and countermeasures against them are as follows:

・Most of the translational word pairs not extracted were those containing transliteration units which were not listed in TR. For instance, 'アーキテクチャ' and 'architecture' was not extracted because 'キ' = 'chi', 'チャ' = 'ture' were not listed in TR. We can solve this problem by enriching TR.

・A few of the translational word pairs which were not extracted were acronyms and their original forms. These are translational KATAKANA-English word pairs, but cannot be extracted based on transliterations by nature.

・Many of the non-translational word pairs wrongly extracted were morphologically related pairs. For instance, 'プログラマ'(programmer) and 'program', 'クラス'(class) and 'subclass' were extracted. Emphasizing the first and the last unit matching might be effective.

4 Conclusion

We proposed a method for extracting translational KATAKANA-English word pairs from bilingual corpora. The experiment shows our method is highly effective. By enriching TR and introducing some heuristics, the performance will become higher.

Assuming that the low-frequency translational word pairs are important extraction target and many of them are composed of KATAKANA words, we proposed the present method. Our next task is to develop a method for extracting the low-frequency non-KATAKANA and English translations and high-frequency translations. We are now investigating the integration of the present method and the statistical extraction method based on word frequency.

Bibliography

1 Collier, N., Kumano, A. and Hirakawa, H. Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using KATAKANA matching. In *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*. 1997, pp. 309 - 314.

2 Jeong, K. S., Myaeng, S. H., Lee, J. S. and Choi, K. S. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*. 1999, 35(4), pp. 523 - 540.

3 Melamed, I. D. Automatic construction of clean broad-coverage translation lexicons. In *2nd Conference of the Association for Machine Translation in the Americas*. 1996, pp. 125 - 134.

4 Hiemstra, D. Deriving a bilingual lexicon for cross-language information retrieval. In *Proceedings of the fourth Groningen International Information Technology Conference for Students*. 1997, pp. 21 - 26.

5 Ishimoto, H. and Nagao, M. Automatic construction of a bilingual dictionary of technical terms from parallel texts. In *Jouhoushorigakkai Kenkyuuhoukoku NL102-11*. 1994, pp. 81 - 88. (text in Japanese)

6 Kumano, A. Building a technical term dictionary with Katakana-English matching. In *Gengoshorigakkai dai-1-kai nenjitaikai happyouronbunshuu*. 1995, pp. 221 - 223. (text in Japanese)

7 Matsuo, Y. and Shirai, S. Using pronunciation to automatically extract bilingual word pairs. In *Jouhoushorigakkai Kenkyuuhoukoku NL116-15*. 1996, pp. 101 - 106. (text in Japanese)

8 Oosuga, S. et al. (trans.) Jinko chinou daijiten. Maruzen, Tokyo, 1991. [Shapiro, S. C. and Eckroth, D. (eds.) *Encyclopedia of Artificial Intelligence*. Wiley, New York, 1987.]

9 Knight, K. and Graehl, J. Machine transliteration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. 1997, pp. 128 - 135.

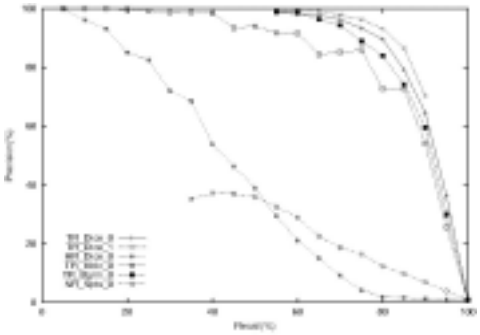

Figure 1: *Abstract of Artificial Intelligence*

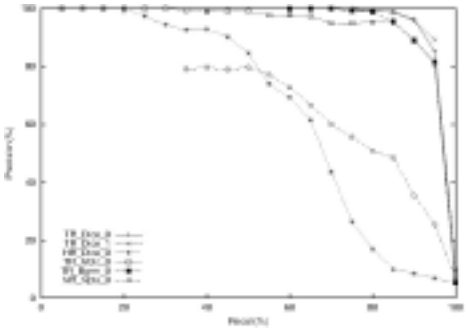

Figure 5: *Title of Artificial Intelligence*

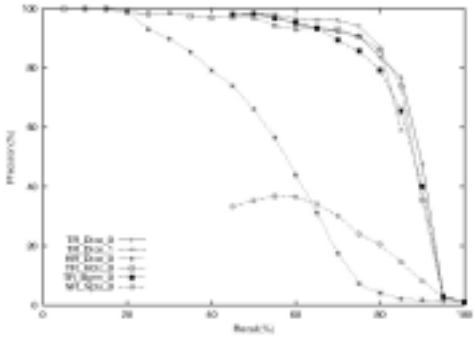

Figure 2: *Abstract of Forestry*

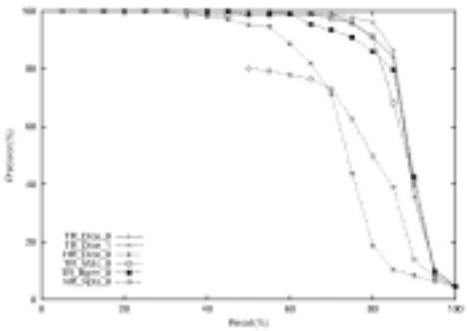

Figure 6: *Title of Forestry*

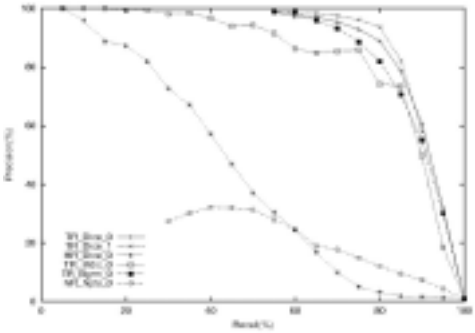

Figure 3: *Abstract of Information Processing*

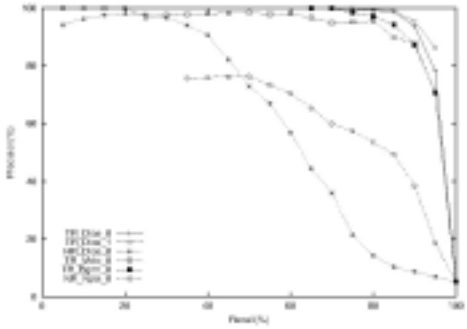

Figure 7: *Title of Information Processing*

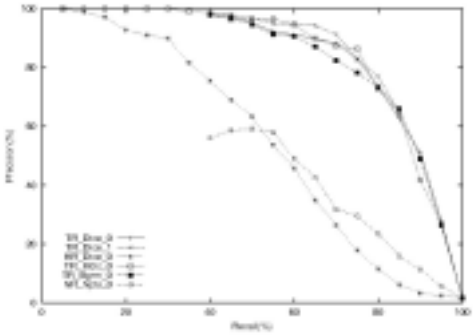

Figure 4: *Abstract of Architecture*

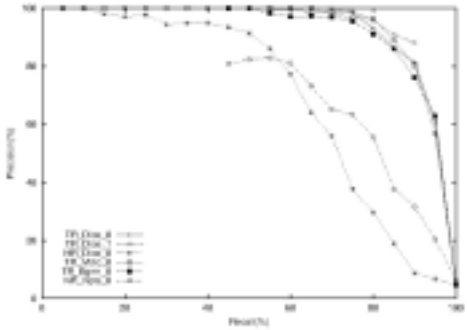

Figure 8: *Title of Architecture*