Extracting French-Japanese Word Pairs from Bilingual Corpora based on Transliteration Rules

Keita Tsuji*, Beatrice Daille†, Kyo Kageura*

*National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{keita, kyo}@nii.ac.jp

†University of Nantes
IRIN, 2, rue de la Houssinire BP 92208, 44322 Nantes cedex 3, France
daille@irin.univ-nantes.fr

Abstract

It has been shown so far that using transliteration rules to extract Japanese Katakana and English word pairs is highly useful and promising. But for Japanese-French pairs, the method is not guaranteed to work, because only a very few Japanese Katakana words are borrowed directly from French. In this paper we will show the possibility of extracting Japanese Katakana and French word pairs based on transliteration from loosely aligned Japanese French bilingual corpora. The method applies all the existing transliteration rules to each mora unit in a Katakana word, and extracts the French word which matches or partially-matches one of these transliteration candidates as translation. For instance, if we have 'グラフ' in the Japanese part of a bilingual corpora, we generate such transliteration candidates as <graf>, <graphe>, <gulerph>,... and identify similar words from French part of the corpora. The method performed reasonably well, achieving 80% precision at 20% recall. We had also observed that Japanese-English transliteration rules worked well for extracting Katakana-French word pairs.

1. Introduction

In this paper we intend to discuss the possibility of extracting Japanese Katakana and French word pairs from loosely aligned Japanese French bilingual corpora. It was shown that using transliteration rules to extract Japanese Katakana and English word pairs is highly useful and promising, as it can extract low frequency pairs with high precision (because most Japanese Katakana words originated from English) and even contributes to the improvement of the statistical methods.

For Japanese-French pairs, the method is not guaranteed to work, because only a very few Japanese Katakana words are borrowed directly from French. Nevertheless, this is worth pursuing for the following reasons. (a) Using transliteration rules is expected to work at least to some extent, due to the similarity between English and French words. (b) In addition, if, for instance, proper names are tagged in French, the extraction of Japanese Katakana equivalent will work even better, due to the constancy of the spellings of proper names. (c) Given the situation that there is a dearth of Japanese-French parallel or comparable corpora, all the doors should be knocked to promote the improvement of the availability of parallel or comparable corpora between non-English language pairs. If the result of Japanese Katakana and French word pair extraction is reasonable, especially in terms of precision, we can use them as anchoring points for matching corresponding articles, thus enriching the parallel resources.

2. Method

Our method is basically similar to (Tsuji, 2001), which is invented for Japanese Katakana-English word pair extraction. The method first construct transliteration rules manually and then extract translational Katakana-French (or English) word pairs from bilingual corpora based on those rules. The difference between our method and that of (Tsuji, 2001) is that while he used Hepburn (Hebon) transliteration rules as supplement, we did not use it.

2.1. Construction of Transliteration Rule

The procedure is as follows (henceforth we represent the obtained rule as 'TR'):

(1) Decompose Katakana word in some source list into mora units.
(2) Based on the French counterpart word, extract transliteration rule for each unit manually.
(3) Repeat (1) and (2) for all the word pairs in the list and rank the rules based on their frequency.

2.2. Extraction of Traslational Word Pairs

First we define the following symbols and functions:

J: Katakana word
F: French word
$L(w)$: The number of characters of word w
$T(J)$: The transliteration candidate of word J
$S(w1, w2)$: The longest common subsequence of $w1$ and $w2$
$Dice(k, m, n) = k * 2/(m + n)$

Extracting Katakana-French 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 $Ti(J)$.
(3) Pick up F which co-occurred with J and identify the longest common subsequence with each $Ti(J)$.
(4) If the following $P(J,F)$ exceeds certain threshold, extract pair J and F as translation: $P(J,F) = max_i Dice(L(S(Ti(J),F)), L(Ti(J)), L(F))$.

In the case that J is 'グラフ' and the transliteration rule is as in Table 1, transliteration candidates such as <graf>, <graphe>, <graff>, ..., <gulerff> and <gulerfe> are obtained. Therefore, P(グラフ, graphe) becomes as follows:

P(グラフ, graphe)
= max(Dice(L(S(graf,graphe)),L(graf),L(graphe)),
Dice(L(S(graphe,graphe)),L(graph),L(graphe)),
Dice(L(S(graff,graphe)),L(graff),L(graphe)),
...
Dice(L(S(gulerff,graphe)),L(gulerff),L(graphe)),
Dice(L(S(gulerfe,graphe)),L(gulerfe),L(graphe)))
= max(0.60, 1.00, 0.55, ..., 0.31, 0.31)
= 1.00

It indicates that 'グラフ' and 'graphe' are likely to be translation.

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

Table 1: Transliteration rule for 'グ', 'ラ' and 'フ'

Our basic idea is as follows. If we use all the transliteration rules and generate all the possible transliteration candidates, the '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 French part of bilingual corpora. We regard the transliteration candidates which actually exist as words in the French part as the correct ones. However, the transliteration rules we can obtain are always insufficient. Therefore we do not use exact matching. Instead, we use $Dice$ and extract the actual word in the French part of corpora, not $T(J)$, whose maximum of $Dice$ is high, as translation.

2.3. Device for Time-saving

The previously-mentioned procedure requires much computational time when applied to 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 words. At TR construction, we have ranked transliteration rules according to their frequencies. We applied the top $12/(the\ number\ of\ units\ in\ J)+1$ rules to each unit of J. [1] 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 $3\times5\times4$ transliteration candidates are examined instead of $3\times6\times4$ candidates.

As for (2), we used 'NPT_score' in (Collier et al., 1997) for computing $L(S(T(J),E))$. It is an abbreviated version for identifying the longest common subsequences.

3. Data

We will explain the bilingual corpus we used for extraction experiment and the source list for constructing transliteration rules.

3.1. Bilingual Corpus

We extracted French-Japanese translations from bilingual corpus of news articles (Le monde diplomatique: 21 articles). The Japanese and French parts of these corpora are processed by morphological analyzer ChaSen2.0b and Brill part-of-speech tagger respectively. We regarded one article as one segment from which we extract translation pairs. Our extraction target was the translational Katakana-French single-word noun pairs in each segment of corpus. The number of these Katakana-French translation pairs was 1,202.

The basic quantities of the corpora are shown in Table 2. In Table 2, 'Token' indicates the number of tokens of ChaSen morphemes in the Japanese part of corpus and the number of French words in the French part of corpus respectively. 'Type' indicates the number of types of these morphemes or words. 'SWN' indicates the number of types of single word nouns and 'SWNK' indicates the number of types of Katakana single word nouns.

	Token	Type	SWN	KSWN
Japanese	82,062	8,501	4,904	1,098
French	60,855	9,751	4,356	–

Table 2: Basic Quantities of Bilingual Corpus

[1] If $12/(the\ number\ of\ units\ in\ J)+1$ is not an integer, we adopted the maximum integer which does not exceed it. In the case that the number of units in J is 5, we used 3 rules for each unit.

3.2. Source List for TR

We used two kinds of transliteration rules, one of which was extracted from Katakana-French word pairs and the other of which was extracted from Katakana-English word pairs. The reason we used the latter is that Katakana-English word pairs are easy to obtain and we assume that there is only small difference between French and English words.

We extracted 1,000 Katakana-French and Katakana-English word pairs from *Concorde Japanese-French Dictionary* and *EDICT* respectively. The basic quantities are shown in the Table 3. In Table 3, 'Tot' indicates the total number of Katakana-French and Katakana-English word pairs. 'Crr' indicates the pairs whose phonetic units completely correspond and, therefore, from which we can extract the transliteration rules. Note that 90% of Katakana-English word pairs belong to this type and more than half of Katakana-French word pairs belong to this type. 'Abb' indicates the number of the pairs one of which is the abbreviation of the latter. For instance, (アパート, appartement) and (ガム, chewing-gum) belong to this type. 'Inv' indicates the number of the pairs whose order of constituent unit is different. For instance, (サテライトスタジオ, studio-satellite) and (アルペンスキー, ski alpin) belong to this type. Katakana-French pairs contains this type of pairs more than Katakana-English pairs do, but the number of pairs was not so large as we expected. 'Rlt' indicates the pairs which do not fall on to the above types but part of which correspond. And 'Noc' indicates the pairs which have no corresponding part. Note that about 63% of Katakana-French pairs have some correspondences from which we can extract transliteration rules.

We used 512 Katakana-French pairs in 'Crr' as source for TR construction. Besides, we randomly selected 512 Katakana-English pairs from 900 pairs in 'Crr' and used them for the other TR construction. The former TR is the Japanese-French TR and the latter is the Japanese-English TR. We extracted Katakana-French pairs from bilingual corpus based on the former TR and the latter TR.

	Tot	Crr	Abb	Inv	Rlt	Noc
J-French	1,000	512	32	13	76	368
J-English	1,000	900	37	1	9	53

Table 3: Basic Quantities of Katakana-French and Katakana-English Pairs

4. Related Studies

The methods to extract Japanese Katakana word and French word pairs from bilingual corpora based on transliteration has not been proposed so far. The methods to extract Japanese-English loan word pairs from bilingual corpora were proposed in (Ishimoto and Nagao, 1994), (Kumano, 1995), (Matsuo and Shirai, 1996), (Collier et al., 1997) and (Tsuji, 2001). But (Ishimoto and Nagao, 1994) did not go into details of transliteration method. (Matsuo and Shirai, 1996) used only consonants for Japanese English matching. The vowels which we think have useful information were ignored. (Kumano, 1995) extracted pairs whose estimated pronunciations exactly matched. But the method to estimate the pronunciation of English word (which seems difficult) was not clearly explained. On the other hand, (Collier et al., 1997) did not discard vowels and showed their procedure clearly. (Tsuji, 2001) took up parts of the method of (Collier et al., 1997) and compared the effectiveness with his own method. And his method was found more effective than that of (Collier et al., 1997). Therefore we adopted the method of (Tsuji, 2001).

Apart from Japanese, (Jeong et al., 1999) 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 the Japanese loan word and the French word correspond on mora-basis and occurence of the corresponding units does not depend on the previous units. This assumption is also practically important. Using the transition probability will lead to the problem of data sparseness.

5. Extraction Result

We used precision and recall to measure the result. The precision is the ratio of correct pairs extracted against pairs extracted. And the recall is the ratio of correct pairs extracted against correct pairs in the whole corpus.

The precision and recall of our method are shown in Figures 1. In this figure, 'J-French' represents the result of our method based on Katakana-French TR. 'J-English' represents the result of our method based on Katakana-English TR.

From this figure, we can say that the method based on Katakana-French TR performed reasonably well, achieving 80% precision at 20% recall. Besides it is observed that the effectiveness of using Katakana-French TR and using Katakana-English TR do not differ significantly. We examined the actual pairs but could not find any tendency which might have produced the difference.

6. Conclusions

It can be said that our method based on Katakana-French TR performed well considering that not a few Katakana-French pairs do not correspond in view of transliteration. We can use frequency-based method to extract these non-corresponding pairs. The integration of transliteration-based method and frequency-based method is promising.

We obtained interesting result that Katakana-English TR performed well against Katakana-French pair extraction. It is worth examining whether the Katakana-English TR works also well against Japanese and the other language such as German and Spanish.

The weak point of this study is that the size of the corpus we used is quite small and the number of pairs from which we constructed TR is also small. We would like to enlarge the size of them and confirm the result.

7. References

J. Breen. Edict. In http://www.csse.monash.edu.au/ jwb/edict.html.

E. Brill, G. Kacmarcik, and C. Brockett. 2001. Automatically harvesting katakana-english term pairs from search engine query logs. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pages 393–399.

N. Collier, A. Kumano, and H. Hirakawa. 1997. 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, pages 309–314.

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

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

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

A. Kumano. 1995. Building a technical term dictionary with katakana-english matching. In Gengoshorigakkai dai-1-kai nenjitaikai happyouronbunshuu, pages 221–223 (text in Japanese).

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

Y. Takatsuka, A. Ogata, T. Yamagata, Y. Yajima, M. Suzuki, J. Shorei, Y. Soga, and T. Nakai, editors. 1990. Concorde Japanese-French Dictionary. Hakusui-sha.

K. Tsuji. 2001. Automatic extraction of translational japanese-katakana and english word pairs from bilingual corpora. In Proceedings of the 19th International Conference on Computer Processing of Oriental Langages (ICCPOL2001), pages 245–250.

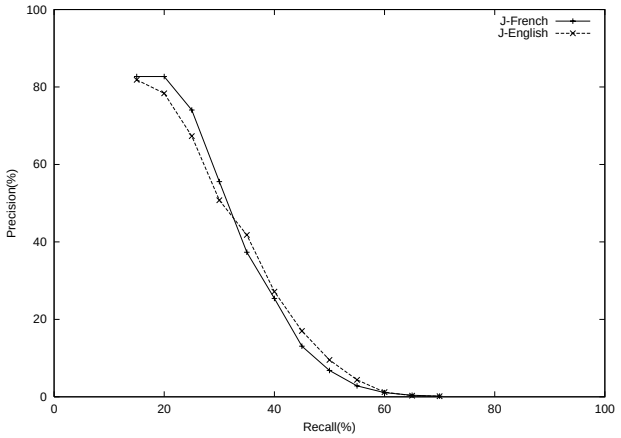

Figure 1: Extraction Result based on Katakana-French and Katakana-English TR