Extraction of Compound Word Traslations from Non-parallel Japanese-French Text in World Wide Web

Hiroko OMAE[*], Graduate School of Science and Technology, Keio University
Keita TSUJI, Human and Social Informatics Division, National Institute of Informatics
Kyo KAGEURA, Human and Social Informatics Division, National Institute of Informatics
Michita IMAI, Faculty of Science and Technology, Keio University
Yuichiro ANZAI, Faculty of Science and Technology, Keio University

Abstract

This paper proposes a method of extracting Japanese-French compound word translation pairs from World Wide Web. The need for enhanced Japanese-French dictionaries (or in general dictionaries of non-English language pairs) is increasing, given the poorer coverage of the existing Japanese-French dictionaries (and generally non-English language pairs) compared to that of Japanese-English dictionaries. Especially lacking is complex words, many of which have fixed forms. As for the use of Web, there are a pushing reason and a pulling reason. As a matter of fact, there are not many Japanese-French parallel corpora (note Japanese-English or French-English corpora are easy to obtain, both quantitatively and qualitatively), and the best source for the machine readable Japanese-French corresponding corpora is the Web. What is more, Web in fact includes many precious information which can never be expected from conventional corpora (even in case of the language pairs which involves English), such as names of small institutions, some events conventionally represented by some fixed names, etc.

The system proceeds as follows. First, the user inputs Japanese-French keyword pairs, which the system then submits to the search engine and obtains the web page texts. Then the system sends the texts to morphological analysers and extracts compounds based on the pre-defined part-of-speech patterns. Finally the system outputs Japanese-French compound pairs as translation if the compound pairs are similar to one of the combinations of Japanese-French single word translations in existing Japanese-French dictionaries.

Several experiments were made on the non-parallel texts which were obtained through various keywords. Small newspaper parallel text was also used for comparison. We sorted the extracted pairs based on similarity score and evaluated the top-ranked pairs. The correct translation rate of the basic method was 5.0% and that of the improved method which extract compound candidate in more refined manner was 17.9%.

The result was satisfactory enough as a first step toward the full exploitation of lexical pair extraction among non-English language pairs.

1 Introduction

In this paper we propose a method of extractiong Japanese-French compound word translation pairs from World Wide Web and describes the extraction system we implemented. Two elements, i.e. compound words and Web, are important in the definition and perspective of our research.

The reason we targeted compound words are two. Firstly, the existing French-Japanese dictionaries do not have many compound word entries or collocations. Although many compounds are independently transparent (thus given a compound we can check the meaning by the combination of the meanings of the constituents), they are not translationally unambiguous

[*] omae@ayu.ics.keio.ac.jp

(thus given, say, a Japanese compound we cannot determine what is the natural and/or standard French compound). What is more, there are a large class of compounds whose meanings are compositional but whose forms are rather fixed, ranging from so-called named entities (e.g. National Institute Informatics) through nomanclatures to technical terms, jargons, etc. The importance of these items are growing, reflecting on the growth of multilingual environment (as can be seen if we look into the anti-war web pages and mailing lists; all kinds of information from varieties of languages are translated and distributed). Though the lexical items of this ilk are not likely to be found in existing dictionaries or ordinary reference tools, they can be found if searched properly, as technical translators do.

The use of Web as a source comes from two — one negative and one positive — factors. The negative factor is that, unlike French-English or Japanese-English pairs, there are not many French-Japanese parallel corpora of news articles, technical papers, etc. especially in electronic form, so we need to turn to the Web for the basic Japanese-French language resources, without limiting ourselves to conventional types of information on the web such as newspapers online. The positive reason is that the Web is an excellent and the only source which contains the vast deposit of various types of compound expressions , such as proper names, technical terms, fixed references to events, etc.

In the long run we intend to develop a system that can explore the full varieties of these lexical expressions that exist on the Web but are not readily available, while at the same time clarifying and classifying the lexicological variations of compounds. The paper presented here is a preliminary element to this direction.

2 Related Work

Although we know of little work which addresses the type of the problem we defined above, there is much work in translation pair extraction, which of course constitutes the core part of the system we developed. The existing studies about translation pair extraction from corpora can be methodologically classified into three, i.e. those based on statistical method, those based on transliteration method and those based on existing dictionaries.

Stasitical methods rely on the cross-lingual co-occurrence of items in the alignment unit of the parallel or comparable corpora. The basic idea is simple: the more the items co-occur, the more likely the pair is a proper translation. Much research has been carried out so far on the basis of this idea [1] [3] [4] [5] [7] [9], and achieved a high performance. The statistical methods, however, are not very good at extracting low-frequency pairs and rely on the existence of parallel or at least comparable corpora.

Transliteration methods, which uses the transliteration table made automatically or manually, can take advantage of the cognates or recent trends in some languages which coin words by transliteration from mainly English [8] [10]. The methods generally show high performance, but the scope of application is limited to transliterated pairs.

Dictionary based methods resort to existing dictionaries and explore proper translational pairs of compounds, using the matching information of their constituents [2]. This method, while can be applied only to the extraction of compounds and requires a high-quality dictionary, has the advantage of making possible the extraction of pairs from non-comparable corpora. This method thus should be a natural starting point to be considered in our research agenda.

Some other researches of translation pair extraction exist. For instance, [6] extracts French-Japanese pairs by using Japanese-English/English-Japanese and French-English/English French dictionaries. Though this gives high performance, the situation in which this method is really required is rare.

3 Automatic Extraction of Compound Translation

3.1 System Overview

Fig.1 shows the overview of the system for automatically extracting compound translation.

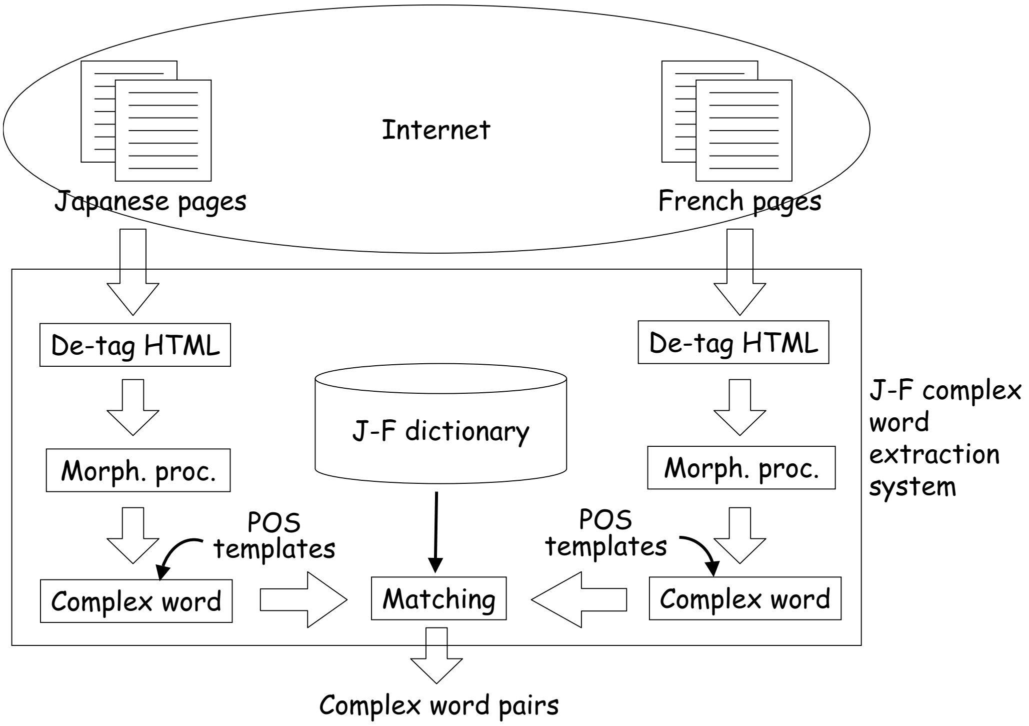

Figure 1: Overview of the system

First, this system collects the web pages both in Japanese and French, which are relevant to a specific keyword using the existing search engine. Next, it removes tag information from obtained HTML files, and generates plain texts. Morphological analysis is performed to these texts and compound words are extracted based on the part-of-speech information on the obtained words. The part-of-speech sequence pattern of a compound words is beforehand prepared as a part-of-speech template, and the word sequence which is in agreement in a text is extracted as a candidate of a compound word. The system repeats the same process for many web pages in both language and extracts the combination of translations from accumulated compound candidates.

3.2 User Interface

The front end of the inplemented system is shown in Fig.??. Using GUI enables to process from web search which is origin of a compound words to extraction of them interactively.

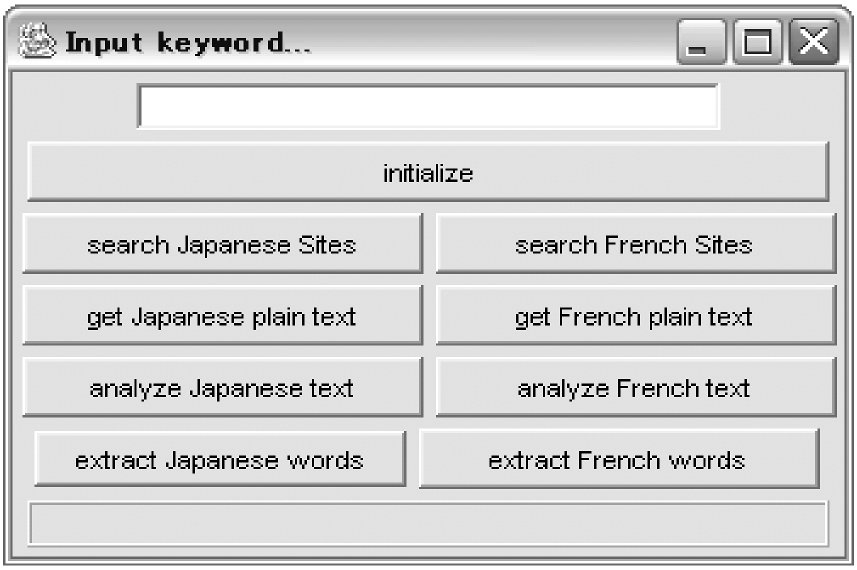

Figure 2: Frontend of the system

3.3 Text Extraction from WWW

3.3.1 Web Search

In order to search web page set that is the retrieval target of a text, we first input the keyword which serves as a topic in the text field in GUI shown in Fig.**??**. Next, we click a Japanese page reference button when a Japanese keyword is inputted and a French page reference button when a French keyword is inputted, and a keyword will be sent to a search engine.

We use Google for web page search engine, and Google Web API[14] for execution of web search. Google Web API is a Java library which Google has distributed, and can perform web search easily from the inside of a program by calling method.

The system retrieve HTML source of web page on the basis of URL of a search result. It uses wget/windows-1.5.3.1[15] for acquisition of HTML source. wget is the GNU tool for downloading web page including a picture or a link and it enables to acquire the form of a file, the depth, directory structure, a domain, etc. by specifying an option. We used the tool compiled for windows this time.

3.3.2 Text Extraction

Next, the system omits tag information and extract a plain text from HTML file. It uses HTML parser in the standard API of J2SDK.

There are several Japanese character code used on web page (e.g. Shift-JIS, JIS, EUC, etc.). By using the text converter which Java offers, it can distinguish these code automatically. However, although there are also several French character code (e.g. ISO-8859-1, windows-1252, etc.), there was no method to distinguish automatically. To resolve this problem, the system refer charset specified as CONTENT attribute within META tag of HTML. About the page in which the character code is not specified, it processed as general ISO-8859-1 as a European-languages code.

3.4 Compound Word Extraction from Text

3.4.1 Definition of Compound Word

The part-of-speech pattern which constitutes a Japanese and French compound word defined as follows, respectively.

Japanese Compound Word

Japanese compound is defined as a noun sequence. We considered an unknown word and a numeral to be also nouns this time. Moreover, when an affix was attached, it treated as a compound.

In consideration of the above thing, the regular expression of a Japanese compound part-of-speech pattern is defined as:

$$PFX?\{N|UK|NUM\}\{N|UK|NUM\}+$$

$$PFX : affix,\ N : noun,\ UK : unknown\ word,\ NUM : numeral$$

French Compound Word

A French compound is defined as a compound combination of a noun + adjective and a noun + preposition (+ article) + noun. As well as the case of Japanese, an unknown word, a numeral are regarded as a noun. When an affix is attached, it takes into consideration, too. Furthermore, in French, since a verbal past participle form and the verbal present participle may also play the role of an adjective in many cases, they are also extracted as an adjective.

As mentioned above, the regular expression of a French compound part-of-speech pattern is defined as:

$$PFX?\{N|UK|NUM\}\{ADJ|PREP\ DET?\{N|UK|NUM\}\}+$$

$$PFX : affix,\ N : noun,\ UK : unknown\ word,\ NUM : numeral,$$

$$ADJ : adjective,\ PREP : preposition,\ DET : article$$

However, not all prepositions constitute a compound word. In this research, we treated only two prepositions “à” and “de”.

3.4.2 Morphological Analyser

The system currently uses the following two morphological analyser.

- Chasen version 2.1[12]: Japanese Morphological Analyser
- WinBrill version 0.4[13]: French Morphological Analyser

Chasen is a Japanese morphological analysis tool which was developed at Matsumoto laboratory of Nara Institute of Science and Technology and which performs analysis based on statistics information.

WinBrill is the French version of English morphological analyser, Brill Tagger, which performs rule-based analysis, and is the morphological analysis tool developed at inalf research institute in France.

These tools are distributed as free software on the Internet.

3.5 Compound Translations Extraction Algorithm

The system extracts the Japanese compound word corresponding to the extracted French compound word using a bilingual dictionary.

3.5.1 Generation of Translation Candidate

The method of [2] was extended in this research. The system generate a translation candidate using a bilingual dictionary.

French compound words are taken out one by one from the compound word list generated beforehand, and translation candidate are generated for each compound word using a dictionary. A translation candidate is a total combination of the complete translation word in a bilingual dictionary enumerated for every single word.

For example, the case where there is the compound “relation culturelle”, translations of two words “relation” and “culturel” (original form of “culturelle”) are extracted from a bilingual dictionary. If there are four words like “関係”, “交際”, “知人”, “交流” as a translation of “relation” and one word “文化の” as a translation of “culturel”, the translation candidates will be 4 words * 1 word = 4words, that is,

文化の関係
文化の交際
文化の知人
文化の交流

Since there is little influence which it has on the determination of a translation, neither a preposition nor an article is taken into consideration here.

3.5.2 Method of Calculating Similarity

After translation candidates are determined, the similarities between all translation candidates and the translation pair candidates in a Japanese compound list are calculated, and the candidate whose similarity is the highest is extracted as a translation pair.

Word order is not taken into account when a translation candidate is determined, and the ambiguity of the notation might arise when treating the translation of an adjective, and adopted the similarity defined by [8]. It is computed by comparing the words per character sequence.

The similarity is computed from bi-gram. For example, when there are the two words “情報 検索” and “情報 探索”, those bi-grams will be “_情”, “情報”, “報_”, “_検”, “検索”, “索_”, and “_情”, “情報”, “報_”, “_探”, “探索”, and “索_”. It compares with the unit of every 2 characters of compound words. Using the number of corresponding bi-grams, the similarity is computed based on the following formulas.

$$sim(s,t) = \frac{|N_s \cap N_t|}{|N_s \cup N_t|} \tag{1}$$

N_s and N_t are bi-grams of word s and t. The range of the value of the similarity is set to 0-1. Since it becomes $|N_s \cup N_t| = 8$ and $|N_s \cap N_t| = 4$ in a previous example,

$$sim = \frac{4}{8} = 0.5.$$

The translation pair candidate whose value of the similarity was the highest is extracted as a translation pair.

When there are two or more translation pair candidates with the same value of the similarity, a translation pair whose number difference of Japanese-French words is the smallest is adopted. However, in French compound words, a preposition and an article are not counted to the number of words.

4 Experiments

4.1 Overview

Two types of experiments are carried out to evaluate the performance of our system, which differ in the extraction of candidate compounds (see below). The extracted translation pair candidates were evaluated according to the four categories of a “correct answer”, a “partial correct answer”, an “incorrect answer”, and a “morphological analysis error”. Each details are described below.

- correct answer
  The pair judged to be right as a translation.

- partial correct answer
  The partial correct answer was further classified into three, “the excess of Japanese”, “the excess of French”, and “others”. “the excess of Japanese” shows the pair that whole French compound and a part of Japanese compound correspond, like “**orientation de la politique**” and “宗教**政治指導**者”. Conversely, “the excess of French” is the pair that whole Japanese compound and a part of French compound correspond, like “absence de **perspective politique**” and “**政治展望**”. However, correspondence of one French word and whole Japanese word like, “islamiste **palestinien**” and “**パレスチナ人**”, or case that it is reverse are classified as “others”. In addition to it, the translation pair which partially correspond mutually is also included in “others.”

- incorrect answer
  The pair contained in neither a correct answer nor a partial correct answer nor a morphological analysis error.

- morphological analysis error
  The pair of which the morphological analysis itself has mistaken and the word that has not correspond in the part-of-speech sequence of the compound is contained in Japanese or French.

4.1.1 Experiment 1

The experiment using the method which does not allow partial duplication but takes out only the longest word sequence as a compound. For example, although the words “破壊 兵器” and “特別 委員 会” are extracted, if the word that includes two words below, like “大量 破壊 兵器 破棄 特別 委員 会” are extracted, the system omits two of the former and it extracts only the latter as a compound.

We experimented giving “United Nations”-“organisation des nations unies” as a keyword of web search and setting to follow the link under 1 class from a top page. It was presupposed that only the link of the same domain as a top page is followed.

4.1.2 Experiment 2

The experiment using the method which decompose the compound words consisted of three or more words into more than two words and extract all of them. For example, when the compound word consisted of four words “安全 保障 理事 会” is extracted it will be decomposed as follows,

安全 保障
保障 理事
理事 会
安全 保障 理事

保障 理事 会
安全 保障 理事 会

and all of them will be extracted as compound words.

4.2 Result

The result of experiment 1 is shown in table1.

Table 1: Result of Experiment 1

files		words		extracted compounds		translation pairs
Japanese	French	Japanese	French	Japanese	French	
60	26	160454	17155	8634	465	279
correct answer	accuracy (%)	partial correct answer			incorrect answer	morph analysis error
		Japanese excess	French excess	others		
14	5.0	4	11	187	60	3

The result of experiment 2 is shown in table2. The total translation obtained here was 966 pairs. However, for the comparison with experiment 1.2, we extracted 279 pairs from the higher rank (more than 0.50 of similarity), and analyzed.

Table 2: Result of Experiment 2(in top 279 pairs)

files		words		extracted compounds		translation pairs
Japanese	French	Japanese	French	Japanese	French	
60	26	160454	17155	20529	1712	279
correct answer	accuracy (%)	partial correct answer			incorrect answer	morph analysis error
		Japanese excess	French excess	others		
50	17.9	0	23	159	36	11

4.3 Discussion

We have found that the accuracy of experiment 2 is higher than that of the experiment 1. This technique was devised in order to remove the translation pairs which were the excess of Japanese and the excess of French among the numbers of partial correct answers. The rate of a correct answer improved 12.9 point from the experiment 1 so that table 2 might show. The pairs of the excess of Japanese or the excess of French which were not extracted since the similarity was low and did not reach the threshold and the pair that a part of French compound and a part of Japanese compound corresponded and which were contained in the “others” of a partial correct answer have to be extracted as a correct answer by applying the decomposing compound method. Actually, extracted translation pairs increased sharply to 966 pairs this time. By decomposing a compound, it can be said that many translation pairs with the high similarity became to be extracted.

However, as shown in table 2, the translation pairs classified into the excess of French of a partial correct answer are increasing compared with table 1. This is because the partial correct answer also marked the high similarity and has been extracted as a translation pair like the example that the pair “mission de maintien de la paix” and “平和維持” has been extracted, in spite of obtaining the correct answer “maintien de la paix” and “平和維持”, since duplication of a compound was allowed. In order to avoid such situation, when the translation pair obtained in the smaller unit already exists, it is necessary to apply the translation pair extraction method to eliminate the compound containing the word.

The common cause in two experiments among translation pairs other than a correct answer is as follows.

- Translation pair extraction with consideration to word order is not performed.
  Since the compound to which word order is reverse exactly also shows the high similarity, it has been extracted as a translation pair. We think that it can be improved by adding a weight using the relation of modification in the case of the calculation of similarity.

- The same translation is extracted repeatedly.
  The translation extracted once has been repeatedly extracted as a translation of other words. When a translation is extracted several times, it is promising to leave only what has the highest similarity.

- The two words expressed in Japanese are often one word in French that cannot be extracted.
  Although “〜的”, “〜化”, “○○人” (○○ is the name of a country), etc. were extracted as two words, it is expressed in many cases in French by one adjective, like “islamiste **palestinien**” and “**パレスチナ 人**”. Therefore, as a compound, it is not extracted and becomes the cause of a translation pair extraction error. We think that this error is avoidable to some extent by preparing huristics to remove the Japanese compound consisted from suffix such as “的” and “化” at last.

- The translation of the unknown word in French cannot be extracted.
  This time, we did not make a bi-gram and calculate the similarity of the French unknown word whose headword has not appeared in the dictionary. However, since the translation corresponds to the remaining word was found, the similarity became high, and it was extracted as a translation pair. About French unknown word, it will be necessary to consider that a capital letter is an abbreviation to leave, and, to remove the small letter from a compound list.

5 Conclusions and Outlook

We have presented the method of extracting Japanese-French complex word pairs from the Web, using the information obtained by existing dictionaries and initial “seed” keywords. Although we are at odds with the Web data which we can explore, the performance was not bad, and the results of the experiments suggested a few aspects that can lead to straightforward improvements of the proposed method as explained in 4.3. As a preliminary and core element of what we are planning to develop, the proposed method is very promising. It can be combined with the transliteration method which also works to some extent between Japanese and French [8] [10].

The overall agenda of the present research in fact has much more to it than what is reported here. In fact, the idea of using “seed” keywords came originally from actual online technical translators, who are daily obtaining and checking the necessary translations using “seed” keywords and exploring Web pages. Further consultations with technical translators show some other interesting heuristics which were not taken into account in the present system. For instance, they rely heavily on “generate (retrieve) and test” of the candidates, and use diagonal information for validation of the candidates. This means that, though the improvement and development of the present method is important, it is also important to exploit the module of *validating* the candidates. The most naive method of doing that is to input all the candidates into the Web search engine and explore the retrieved documents. Another aspect of exploring the web to enhance the compound pair extraction is to categorise and edit the Web, not topically but from different points of view. [11], for instance, develops a method of extracting definitions of input words and related words. This kind of methods will be incorporated into our system as well.

References

[1] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, Robert L. Mercer, and P. S. Roossin. A Statistical Approach To Machine Translation. *Computational Linguistics*, Vol.16, No.2, pp.79-85, June, 1990.

[2] Y. Yamamoto, and M. Sakamoto. Extraction of Technical Term Bilingual Dictionary from Bilingual Corpus. IPSJ SIGNotes NL94-12, pp.85-92, 1992.

[3] J. Kupiec. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. *ACL-31: 31th Annual Meeting of the Association for Computational Linguistics*, pp.23-30, Columbus, OH, 1993.

[4] I. D. Melamed. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. *Proceedings of 3rd Workshop on Very Large Corpora*, pp.184-198, 1995.

[5] F. Smadja, K. McKeown, and V. Hatzuvassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, Vol.22, No.1, pp.1-38, 1996.

[6] K. Tanaka-Ishii, K. Umemura, and H. Iwasaki. Construction of Bilingual Dictionary Intermediated by a Third Language. IPSJ Journal, Vol.39, No.6, pp.1915-1924, 1998.

[7] P. Fung. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *Lecture Notes in Artificial Intelligence*, Vol.1529, pp.1-17, 1998.

[8] K. S. Jeong, S. H. Myaeng, J. S. Lee, and K. S. Choi. Automatic Identification and Back-transliteration of Foreign Words for Information Retrieval. *Information Processing and Management*, Vol.35, No.4, pp.523-540, 1999.

[9] K. Yamamoto, and Y. Matsumoto. Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure. *Proceedings of the 18th International Conference on Computational Linguistics*, pp.933-939, Saarbrucken, Germany, July, 2000.

[10] K.Tsuji. Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora. *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages(ICCPOL)*, pp.245-250, 2001.

[11] Y. Sasaki, and S. Sato. Automatic Collection of Relational Terms Using Web. *Proceedings of the 9th Association for Natural Language Processing*, pp278-281, March, 2003.

[12] Chasen Homepage. http://chasen.aist-nara.ac.jp/index.html.ja

[13] WinBrill Homepage. http://www.inalf.fr/winbrill

[14] GoogleAPI Homepage. http://www.google.com/apis/index.html

[15] Wget Homepage. http://www.interlog.com/t̃charron/wgetwin.html