Automatic Acquisition of Basic Katakana Lexicon from a Given Corpus

Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi

University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
{nakazawa, kawahara, kuro}@kc.t.u-tokyo.ac.jp

Abstract. Katakana, Japanese phonogram mainly used for loan words, is a trou-blemaker in Japanese word segmentation. Since Katakana words are heavily domain-dependent and there are many Katakana neologisms, it is almost impossible to construct and maintain Katakana word dictionary by hand. This paper proposes an automatic segmentation method of Japanese Katakana compounds, which makes it possible to construct precise and concise Katakana word dictionary automati-cally, given only a medium or large size of Japanese corpus of some domain.

1 Introduction

Handling words properly is very important for Natural Language Processing. Words are basic unit to assign syntactic/semantic information manually, basic unit to acquire knowledge based on frequencies and co-occurrences, and basic unit to access texts in Information Retrieval.

Languages with explicit word boundaries, like white spaces in English, do not suffer from this issue so severely, though it is a bit troublesome to handle compounds and hyphenation appropriately. On the other hand, languages without explicit boundaries such as Japanese always suffer from this issue.

Japanese character set and their usage. Here, we briefly explain Japanese character set and their usage. Japanese uses about 6,000 ideogram, Kanji characters, 83 phonogram, Hiragana, and another 86 phonogram, Katakana.

Kanji is used for Japanese time-honored nouns (including words imported from China ancient times) and stems of verbs and adjectives; Hiragana is used for function words such as postpositions and auxiliary verbs, and endings of verbs and adjectives; Katakana is used for loan words, mostly from the West, as transliterations.

Japanese is very active to naturalize loan words. Neologisms in special/technical domains are often transliterated into Katakana words without translations, or even if there are translations, Katakana transliterations are more commonly used in many cases. For example, コンピュータ, transliteration of "computer" is more commonly used than the translation, 計算機(*keisanki*).

Even for some time-honored Japanese nouns, both Japanese nouns and translitera-tions of their English translations are used together these days, and the use of

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 682 – 693, 2005.

translit-erations is increasing, such as デスクワーク, transliteration of “desk work” vs. 机仕事(*tsukue shigoto*). Furthermore, some Japanese nouns, typically the names of animals, plants, and food, which can be written in Kanji or Hiragana, are also written in Katakana sometimes [4, 6].

Word segmentation and Katakana words. Let us go back to the word segmentation issue. Japanese word segmentation is performed like this: Japanese words are registered into the dictionary; given an input sentence, all possible words embedded in the sentence and their connections are checked by looking up the dictionary and some connectivity grammar; then the most plausible word sequence is selected. The criteria of selecting the best word sequence were simple heuristic rules preferring longer words in earlier times, and some cost calculation based on manual rules or using some training data, these days.

Such a segmentation process is in practice not so difficult for Kanji-Hiragana string. First of all, since Kanji words and Hiragana words are fairly stable excepting proper nouns, they are most perfectly registered in the dictionary. Then, the orthogonal usage of Kanji and Hiragana mentioned above makes the segmentation rather simple, as follows:

彼	は	大学	に	通う
(Kare	wa	daigaku	ni	kayou)
he	postp.	Univ.	postp.	go

Kanji compound words can cause a segmentation problem. However, since large num-ber of Kanji characters lead fairly sparse space of Kanji words, most Kanji compounds can be segmented unambiguously.

A real troublemaker is Katakana words, which are sometimes very long compounds such as エクストラバージンオリーブオイル “extra vergin olive oil” and ジャパンカップサイクルロードレース “Japan cup cycle road race”. As mentioned above, many neologisms are written in Katakana, it is almost impossible to register all or most Katakana words into a dictionary by hand. To handle such an insufficiency of a dictionary, conventional Japanese word segmentation incorporates a fall-safe method, which considers a whole continuous Katakana string as a word, when it is neither a registered-word, nor a combination of registered-words. And, Japanese word segmentation basically prefers longer registered words. These mechanism leads that, for example, the Katakana string トマトソース “tomato sauce” is properly segmented to トマト “tomato” and ソース “sauce”, only when トマト and ソース are in the dictionary and トマトソース is not. When ソース alone is in the dictionary (means an imperfect dictionary) or トマトソース is in the dictionary (means a redundant dictionary), トマトソース is regarded as one word.

Considering the importance of words as a basic unit of NLP, it is quite problematic to handle トマトソース as a single word. We cannot use information that トマトソース is a kind of ソース, which is very important for deeper/semantic processing of texts; a text including トマトソース cannot be retrieved with the word トマトor ソース. Note that a rough treatment using partial string matching causes a tragedy that リソース(*risōsu*) “resource” matches ソース “sauce” and スライス(*suraisu*) “slice” matches ライス(*raisu*)“rice” and イス(*isu*)“chair”!

To solve this severe problem, this paper proposes a method of constructing precise and concise Japanese Katakana word dictionary, by automatically judging a given

Katakana string is a single-word or compound, and registering only single-words to the dictionary. We suppose only a medium or large size of Japanese corpus is given, and Katakana strings and their frequencies in the corpus are extracted as follows. We call this data as a *word-occurrence data* hereafter.

ラーメン(*rāmen*):28727 “noodle”
スープ(*sūpu*):20808 “soup”
レシピ(*resipi*):16436 “recipe
カレー(*karē*):15151 “curry”
メニュー(*meny ū*):14766 “menu”
エスニック(*esunikku*):14190 “ethnic”
サラダ(*sarada*):13632 “salad”
トップ(*toppu*):11642 “top”
トマトソース(*tomatosōsu*):11641 “tomato sauce”
...
トマト(*tomato*):7887 “tomato”
...
ソース(*sōsu*):7570 “sauce”
...

Our proposed method consists of the following three methods, which utilize only a word-occurrence data and publicly available resources:[1]

- A method using a Japanese-English dictionary.
- A method using a huge English corpus and a Japanese-English dictionary.
- A method using relation in a word-occurrence data.

Since most Katakana words are transliterations of English words, we exploit Japanese-English translation information as much as possible, using a Japanese-English dictio-nary and a huge English corpus. Since these methods, however, cannot achieve high-recall, the third method uses a word-occurrence data itself: a Katakana word is regarded as a compound if it is a combination of other, frequent Katakana words in the word-occurrence data. These three methods vary from high-precision to high-recall, and their appropriate combination leads to high-precision, high-recall analysis.

We explain these three methods in detail, and then report the experimental results and discussion.

2 A Method Using a Japanese-English Dictionary

The first method utilizes a Japanese-English dictionary, judging some Katakana words as compounds and others as single-words. Words that are judged here will not be pro-cessed by the next two methods.

The basic idea using a dictionary is as follows. Suppose the input word is トマトソース and the dictionary provides the following information:

[1] There are some Katakana words that are not loan words, such as the names of animals, plants and food. We deal with these words as single-words exceptionally, if they are registered in a Japanese dictionary.

トマトソース= tomato sauce
トマト= tomato
ソース= sauce

If the translation of the input word consists of multi-words and those words correspond to Katakana substrings just enough based on the dictionary information, the input word is considered as a compound. In the case of the above example, トマトソース is divided into トマト+ ソース by these criteria.

On the other hand, if the translation of the input word is one word in the dictionary, it is considered as a single-word (that is, the other two methods are not applied to the input word any more), like the following example:

サンドウィッチ(*sandowicchi*) = sandwich

The Japanese-English dictionary can be used in such a straightforward way. In prac-tice, however, we handle some exceptional cases more carefully as follows:

– When the dictionary provides multi-word translation for an input, and all of them are capitalized, the input is regarded as a proper noun and treated as a single-word.

  ブエノスアイレス(*Buenosuairesu*) = Buenos Aires
  ミルキーウェイ(*Mirukīwei*) = Milky Way
  サザンクロス(*Sazankurosu*) = Southern Cross

– When the dictionary provides multi-word translation for an input, but the alignment of translation words and Katakana substrings fails, still if the final translation word corresponds to the Katakana suffix-string, the input is regarded as a compound, as follows:

  モルネソース(*Morunesōsu*) = Mornay sauce
  ソース= sauce

  スモークハム(*sumōkuhamu*) = smoked ham
  ハム(*hamu*) = ham

– The judgment of being a single-word is invalidated, when the translation corre-sponds to only a partial Katakana string by another dictionary entry as follows:

  シフォンケーキ(*shifonkēki*) = chiffon
  シフォン(*shifon*) = chiffon

  キャッチボール(*kyacchibōru*) = catch
  キャッチ(*kyacchi*) = catch

  シフォンケーキand キャッチボールare not disposed in this method and transfered to the next methods.

3 A Method Using a Huge English Corpus and a Japanese-English Dictionary

A dictionary contains only basic compounds, but there are many more Katakana compounds in real texts. That is, the direct use of dictionary is not enough to handle real Katakana compounds.

Therefore, we have developed a method which utilizes a Japanese-English dictionary to get a basic translation relation, and judges whether a Katakana string is a compound or not by referring to a huge English corpus.

Given an input Katakana string, all possible segmentations to Katakana words registered in the Japanese-English dictionary are detected, and those words are translated into English words. Then, the frequencies of those possible English translations are checked by referring to a huge English corpus, and the most frequent translation is selected as a resultant segmentation. As an English corpus, we use the web, and the hit number of a search engine is used as the frequency.

Forexample, パセリソース(*paserisōsu*) can be segmented in two ways, and the first segmentation can have two different translations, totaling to the three possible translation as follows:

パセリ(*paseri*)+ ソース(*sōsu*) parsley source:554
パセリ(*paseri*)+ ソース(*sōsu*) parsley sauce:20600
パセ(*pase*)+ リソース(*risōsu*) pase resource:3

The web search shows that the second translation, “parsley sauce” is by far the most frequent, supporting パセリソースis a compound パセリ+ ソース.

The important issue is how much we believe the frequency of the web. Some web pages are very messy, and even inappropriate segmentation and its mad translation has some frequency in the web, as follows:

デミ(*demi*)+ グラス(*gurasu*) demi glass:207
バン(*ban*)+ バンジー(*banjī*) van bungee:159

In order to exclude such inappropriate segmentations, we need to set up some threshold to accept the segmentation. Considering that the longer the Katakana word is, the more probable it is a compound, we set the following threshold:

$$C/N^{L},$$

where L denotes the length of the Katakana word, and C and N are constant, optimized using some development data set.

4 A Method Using Relation in a Word-Occurrence Data

Though the method using an English corpus is reliable and accurate, it can be applied only when the constituent words are in the dictionary, and the compound is a natural term in English. However, some neologisms and some words that are not usually written in Katakana are not registered in the dictionary. Furthermore, there are many

Japanese-made English-like compounds like “gasoline stand” (means “service station”), which are rarely found in native English corpus.

To handle such cases robustly, we try to find compounds only based on the infor-mation in a word-occurrence data. For example, if トマト and ソース are sufficiently frequent in the word-occurrence data, we consider トマトソース as a compound, トマト+ ソース.

Again, we have to carefully design the threshold to accept the segmentation. Since the word-occurrence data contains very many varieties of Katakana strings, most single-words can be somehow divided into two or more Katakana strings. For example, even イタリアン(*itarian*) “Italian” can be divided into イタ *ita*)+ リアン(*rian*).

Then, we established the basic criteria as follows: if the geometric mean of fre-quencies of possible constituent words (F_g) is larger than the frequency of the original Katakana word (F_o), then we accept the segmentation. Similar to the method using an English corpus, considering that the longer the Katakana word is, the more probable it is a compound, we modified the condition as follows:

$$F_o < F'_g, \qquad F'_g = F_g / (C/N^l + \alpha)$$

where l denotes the average length of constituent words (equal to the length of the Katakana word divided by the number of constituent words), C, N and α are constant, optimized using some development data set. α is a term to provide the upper bound of F'_g when l becomes large.

When there are segmentations into different number of words, the coarse segmenta-tion, that is, the segmentation into a small number of words is selected. When there are two or more possible segmentations into the same number of words, that of the largest F_g is selected.

Here are some examples in the cooking corpus (the details of this corpus are de-scribed in Section 6.1):

イタリアンレストラン(*itarianresutoran*):207
↔イタリアン(*itarian*):1421 + レストラン(*resutoran*):7922 (F_g = 3355)

スパイスライス(*supaisuraisu*):3
↔スパイ(*supai*):9 + スライス(*suraisu*):2000 (F_g = 134)
↔スパイス(*supaisu*):2203 + ライス(*raisu*):980 (F_g = 1896)

イタリアン(*itarian*):421
↔イタ(*ita*):91 + リアン(*rian*):11 (F_g = 31)
↔イタリ(*itari*):7 + アン(*an*):301 (F_g = 45)

イタリアンレストラン“Italian restaurant” and スパイスライス “spice rice” are not segmented by the English corpus method, because イタリアン is not registered in the Japanese-English dictionary, and “spice rice” does not occur frequently (though “spicy rice” is frequent). However, they are properly segmented by this method. On the other hand, イタリアン is not segmented, since neither of two possible segmentations イタ+ リアンor イタリ+ アン have large F_g.

5 Registration to Katakana Word Dictionary

Given a word-occurrence data, the three methods are applied to exclude compounds, and the remaining single-words are registered to the dictionary of Japanese segmenta-tion program.

In order to handle the ambiguity of compound segmentation, the word is registered with the cost, $C - \log f$, where f is its frequency in the word-occurrence data. Since the Japanese segmentation program JUMAN[4] selects the segmentation with the minimum cost, this cost assignment is consistent with the segmentation selected by the method using relation in the word-occurrence data. For example, the cost of segmenting スパイスライスis calculated as follows:

スパイ+スライス:

$$(C - \log 9) + (C - \log 2000) = 2C - \log (9 \times 2000)$$

スパイス+ライス:

$$(C - \log 2203) + (C - \log 980) = 2C - \log (2203 \times 980)$$

As a result, スパイス+ライス, whose cost is smaller than that of スパイ+スライス, is selected.

This cost calculation is not necessarily consistent with the segmentation supported by the English corpus method. To handle this, Katakana words are once registered into the dictionary with these costs, and then Katakana compounds handled by the English corpus method are fed to the segmentation program. Then, if the segmentation is incorrect, the compound word is registered into the compound word dictionary with its correct segmentation position.[2] All of these treatments can be done automatically based on the results of our compound detection methods.

Note that how much frequent words should be registered into the dictionary depends on the policy of the dictionary maintenance, and the system capability of handling unknown words. These issues are out of the scope of this paper.

6 Evaluation and Discussion

6.1 Experimental Results

We prepared two data sets for experiments: 87K Katakana words appearing more than once in 12-year volume of newspaper articles (5.8M sentences), and 43K Katakana words appearing more than once in web pages of cooking domain (2.8M sentences).

For both data sets, we randomly selected 500 Katakana words, and assigned correct segmentation positions to those words by hand. Then, these manual segmentation posi-tions were compared with automatic segmentation positions, calculating precision and recall scores. Note that the unit of evaluation is not words, but segmentation

[2] Japanese segmentation system has a compound dictionary to deal with exceptional (hard-to-segment) compound words, which are not limited to Katakana words. It is one possible way to reigster all Katakana compounds to the compound dictionary, but it is not reasonable from the view point of the dictionary maintenance.

positions. The average number of segmentation positions of 500 words in news domain was 1.39; that in cooking domain was 1.62.

As explained so far, our proposed methods consist of the following three methods:

- A method using a Japanese-English dictionary (D).
- A method using a huge English corpus and a Japanese-English dictionary (C).
- A method using relation in a word-occurrence data (R).

To see the effectiveness of each method, we tested four types of their combination: D, D+C, D+R, D+C+R. In all types, the D method is applied first. Then both C and R method are applied to the words which are not dealt with in D method. Results of C method are prior to those of R method. The parameters were set to $400{,}000/2^{L}$ for the second method and $F'_g = F_g/(2{,}500/4^{l} + 0.7)$ for the third method. As a Japanese-English dictionary, we used two free-to-use dictionary: Eijiro (931K all entries and 137K Katakana entries) and Edict (140K all entries and 14K Katakana entries). Table 1 shows the results, indicating that the combination of D+C+R achieved both highprecision and high-recall.

Table 1. Experimental results

News domain

	D	D+C	D+R	D+C+R
Precision/Recall	1.0/0.822	0.996/0.909	0.986/0.945	0.985/0.949
F-measure	0.902	0.950	0.965	0.966

Cooking domain

	D	D+C	D+R	D+C+R
Precision/Recall	1.0/0.717	1.0/0.836	0.990/0.948	0.991/0.956
F-measure	0.835	0.910	0.968	0.973

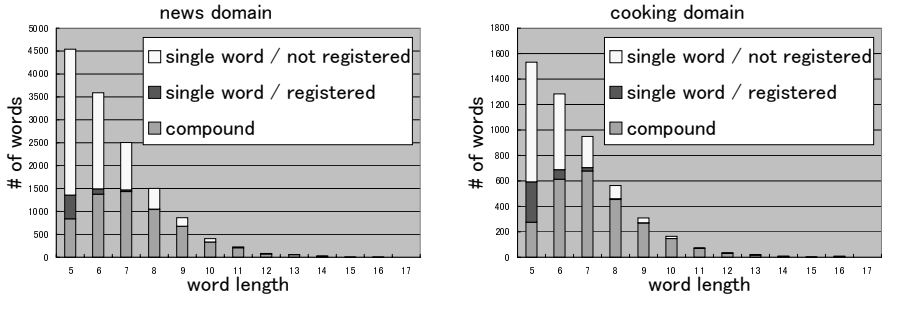

Fig. 1. Statistics of compounds and single-words

Figure 1 shows, among 10 or more frequent words of each length in news domain and cooking domain, the number of compounds, the number of single words registered in the dictionary of the segmentation program JUMAN, and the number of single words not registered in JUMAN. In total, 6K words were judged as compounds out of 13.8K words in news domain; 2.5K words out of 4.9K words in cooking domain.

6.2 Discussion

As shown in Table 1, the method using the dictionary is precise, but the recall is not high enough. Combining it with the methods of using the English corpus and the relation in the word-occurrence data leads to both high-precision and high-recall.

The causes of the incorrect results can be analyzed as follows. When a word is incorrectly segmented, the Japanese-English dictionary overlooks the word as a single word. Then, it is passed to the next methods, and segmented incorrectly. The overlook of the dictionary took place in the following cases:

- **Neologisms or words rarely written in Katakana**

  セル(*seru*)+ ライト(*raito*) cell light:15100 >12500
  (セルライト(*seruraito*) is “cellulite”)

  シュレッドチーズ(*syureddochīzu*):24 “shred cheese”
  ↔シュ(*syu*):41 + レッド(*reddo*):112 + チーズ(*chīzu*):7199 (F'_g **= 143)**

- **Not original forms**
  Transliterations of words in not original forms are often used in Katakana com-pounds, but they are not usually listed in the Japanese-English dictionary.

  セーフ(*sēfu*) + ティー(*thī*)safe tea:16500>6250
  (セーフティー(*sēfuthī*)is “safety”)

  リストラクチャリング(*risutorakucyaringu*):150 “restructuring”
  ↔リストラ(*risutora*):5081 + クチャ(*kucya*):3 + リング(*ringu*):743 (F'_g = 238)

- **Spelling variation problem**
  Though representative Katakana spellings are in the dictionary, their spelling variations are not. Handling of spelling variation is a target of our future work.

  レイン(*rein*)+ ボー(*bō*)rain bow:22100 >12500
  (The representative spelling is レインボウ(*reinbou*) “rainbow”)

  プラスティック(*purasuthikku*):48 “plastic”
  ↔プラ(*pura*):67 + スティック(*suthikku*):224 (F'_g = 143)
  (The representative spelling is プラスチック(*purasuchikku*))

- **Proper nouns**
  Proper nouns are not well covered in the dictionary. We are planning to reexamine this problem with the help of an NE detection method.

  パス(*pasu*)+ ツール(*tūru*)path tool:13700 >12500
  (パスツール(*pasutūru*) is “Pasteur”)

  コネティカット(*konethikatto*):108 “Connecticut”
  ↔コネ(*kone*):177 + ティ(*thi*):166 + カット(*katto*):4144 (F'_g = 108)

  On the other hand, the reason of lowering recall, that is, the overlook of compounds, can be summarized as follows:

– Especially for shorter words, it is actually very hard to set up clear criteria for compounds. In constructing the test sets, we regarded a word as a compound when the head (the last constituent) has an independent meaning and an is-a relation with the original word. However, whether an English translation is one word or not is not necessarily consistent with these criteria.

  バイオサイエンス(*baiosaiensu*) = bioscience
  フレックスタイム(*furekkusutaimu*) = flextime
  プールサイド(*pūrusaido*) = poolside

– Similar to the precision problem, when the constituent word is not in the dictionary, the compound could not be handled by the English corpus method, and the third method overlooked it sometimes.

  ベイエリア(*beieria*):163 “bay area”
    ↔ベイ(*bei*):116+ エリア(*eria*):1377 (F'_g=127)
    (ベイ is not in the dictionary)

  シュガーローフ(*syug¯ofu*):19 “sugar loaf”
    ↔シュガー(*syuga¯*):40 + ローフ(*rōfu*):6 (F'_g = 18)
    (ローフis not in the dictionary)

– Sometimes segmentation score cannot pass the threshold.

  ペパー(*pepa¯*)+ミント(*minto*) pepper mint:5400 < 6250
  ペパーミント(*pepāminto*):41
    ↔ペパー:8+ミント:56 (F'_g = 16)

  ヘア(*hea*)+ケア(*kea*)+チェック(*chekku*)
      hair care check:397 < 1562
  ヘアケアチェック(*heakeachekku*):458
    ↔ヘアケア(*heakea*):32+チェック:1350 (F'_g = 281)

Some Katakana strings are ambiguous and their segmentation depends on the context, such as タコス(*takosu*)+ライス(*raisu*) “tacos rice” and タコ(*tako*)+スライス(*suraisu*) “octopus slice”. However, there were few such cases in our experiments.

7 Related Work

To our knowledge, there has been no work so far handling the automatic segmentation of phonogram compounds in such a real large-scale. German compound nouns have a similar problem, like Lebensversicherungsgesellschaftsangestellter (“life insurance company employee” in English), and can be a target of our method.

There are several related work which can contribute the modification and extension of our methods. When using a Japanese-English dictionary, if we understand the translation is transliteration, we can utilize the information more effectively, handling inflections. In this sense, work by Knight and Graehl can be incorporated into our method [2].

In order to handle spelling variation problems, there have been many methods proposed [3], and we can utilize recently proposed robust treatment of Japanese Katakana spelling variation by Masuyama et al. [5].

Our second method using Japanese-English dictionary and the English corpus can be considered as a translation acquisition method. It is interesting to compare these results with other web-based methods, such as Utsuro et al. [8, 1].

There have been many studies that extract compound nouns. Nakagawa et al. focused on the tendency that most of technical terms are compound nouns, and proposed a method of extracting technical terms by using frequency and variety of its neiboring words [10, 7].

In view of information retrieval, Yamada et al. aimed at imporving information retrieval using matching of compounds [9]. It is similar to our study in handling compounds.

8 Conclusion

This paper proposed an automatic segmentation method of Japanese Katakana compounds, which makes it possible to construct precise and concise Katakana word dictionary automatically, given only a medium or large size of corpus of some domain. Since Katakana is often used for English transliteration, our method exploited a Japanese-English dictionary and a huge English corpus. Combining translation-based high-precision method with more robust, monolingual, frequency-based method, we could achieve both high-precision and high-recall compound segmentation method.

The results of this method were already successfully used to enhance a Japanese word segmentation program. We are planning to handle Katakana spelling variation and to incorporate our method with an NE detection method.

References

1. Mitsuhiro Kida, Takehito Utsuro, Kohei Hino, and Satoshi Sato. Estimating bilingual term correspondences from japanese and english documents. In *Information Processing Society of JAPAN*, pages 65–70, 2004.
2. Kevin Knight and Jonathan Graehl.Machine transliteration. *Computational Linguistics*, 24(4):599–612, 1998.
3. Junichi Kubota, Yukie Shoda, Masahiro Kawai, Hirofumi Tamagawa, and Ryoichi Sugimura. A method of detecting KATAKANA variants in a document. *Information Processing Society of JAPAN*, 35(12):2745–2751, 1994.
4. Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28, 1994.
5. Takeshi Masuyama, Satoshi Sekine, and Hiroshi Nakagawa.Automatic construction of Japanese katakana variant list form large corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1214–1219, 2004.

6. Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. *Morphological Analysis System ChaSen version 2.3.3 Users Manual*, 2003.
7. Hirokazu Ohata and Hiroshi Nakagawa. Automatic term recognition by the relation between compound nouns and basic nouns. In *Information Processing Society of JAPAN*, pages 119–126, 2000.
8. Takehito Utsuro, Kohei Hino, Mitsuhiro Kida, Seiichi Nakagawa, and Satoshi Sato.Inte-grating cross-lingually relevant news articles and monolingual web documents in bilingual lexicon acquisition. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1036–1042, 2004.
9. Koichi Yamada, Tatsunori Mori, and Hiroshi Nakagawa.Information retrieval based on combination of japanese compound words matching and co-occurrence based retrieval. *Information Processing Society of JAPAN*, 39(8):2431–2439, 1998.
10. Hiroaki Yumoto, Tatsunori Mori, and Hiroshi Nakagawa.Term extraction based on occurrence and concatenation frequency. In *Information Processing Society of JAPAN*, pages 111–118, 2001.