Analysis of Japanese Compound Nouns by Direct Text Scanning

Toru Hisamitsu and Yoshihiko Nitta
Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-03, JAPAN
{hisamitu, nitta}@harl.hitachi.co.jp

Abstract

This paper aims to analyze word dependency structure in compound nouns appearing in Japanese newspaper articles. The analysis is a difficult problem because such compound nouns can be quite long, have no word boundaries between contained nouns, and often contain unregistered words such as abbreviations. The non-segmentation property and unregistered words cause initial segmentation errors which result in erroneous analysis. This paper presents a corpus-based approach which scans a corpus with a set of pattern matchers and gathers co-occurrence examples to analyze compound nouns. It employs boot-strapping search to cope with unregistered words: if an unregistered word is found in the process of searching the examples, it is recorded and invokes additional searches to gather the examples containing it. This makes it possible to correct initial *over-segmentation* errors, and leads to higher accuracy. The accuracy of the method is evaluated using the compound nouns of length 5, 6, 7, and 8. A baseline is also introduced and compared.

1. Background

1.1 Compound Nouns in Japanese Newspaper Articles

This paper analyzes the word dependency structure in compound nouns appearing in Japanese newspaper articles. Assume that you are given a large number of articles and a compound noun such as "改正大店法施行". This noun actually consists of three nouns "改正" (revision), "大店法" and "施行" (application), where "大店法" is the abbreviation of "大規模小売店舗法"(大規模: large, 小売店舗: retail shop, 法: law). However, it is highly unlikely that such a word can be found in an ordinary dictionary. Newspaper articles are full of this kind of difficult compound nouns which can be infinitely generated, and such compound nouns often convey substantial information through which the articles can be summarized .

In Japanese newspapers, compound nouns are especially useful because they convey a lot of information in a compact expression (even a single *kanji*, or Chinese character, can represent complex meaning). The number of nouns forming a compound noun often exceeds three, and may reach as much as ten. This means that a compound noun can contain up to twenty *kanji* characters or more. Therefore, an analysis of noun compounds has to deal with both segmentational and structural ambiguities.

As for the example above, an initial *morphological analysis* (segmentation + tagging) causes an over-segmentation error such as "改正 sn/大 adj/店 n/法 n/施行 sn" because "大"(large), "店"(shop) and "法"(law) are all meaningful expressions by themselves.

1.2 Existing Methods and Problems

Compound noun analysis has been researched for many years because it is important for understanding natural language. A concise review of this research area can be found in, for instance, Lauer (1995), which dates back to Finin (1980). When applying the existing methods to Japanese compound nouns in newspaper articles, however, a problem arises:

(1) All the methods are difficult to apply because they use training schemes such as (partial) parsing of the whole corpus and counting word occurrence in word windows.

As Lauer (1995) pointed out, using (partial) parsing of the text is too costly. Thus, the word co-occurrence approach seems to be more appropriate. However, counting the frequency of a given word is not an easy task in a non-segmented Japanese text. Ordinary pattern matching algorithms cannot count the number of occurrences of a word in non-segmented Japanese text because of the ambiguity in how sentences should be segmented. Thus, whatever method one chooses, he is first confronted with the high cost of Japanese morphological analysis and its inaccuracy caused by unregistered words.

Thus, researchers of Japanese compound noun analysis have been obliged to employ manually written syntactic rules for compound nouns (Miyazaki, 1984) or the conceptual dependency model (Kobayashi *et al.*, 1994) which employs a thesaurus and a limited co-occurrence data, for example, a collection of four *kanji* sequences (Tanaka, 1992) extracted from a corpus.

The problems in existing methods are:

(2) It is costly to manually prepare the rules for the analysis of compound nouns.

(3) Methods employing a conceptual dependency model are brittle when unregistered words occur often. One has to properly allocate an unregistered word in the thesaurus,

which is another tough problem.

For these reasons, the existing methods are not effective for compound noun analysis in newspaper articles. A scheme for collecting collocational information
(1) must be practical for large amounts of Japanese raw text, and also collect *reliable* data.
(2) should cope with unregistered words.

1.3 Direct Text Scanning Method

To satisfy the requirements mentioned above, we used a direct text scanning method which collects *external evidence* (McDonald, 1993) of a modifier-modifee relationship between two words using a set of simple pattern matchers.

In this method, a Japanese morphological analyzer (JMA) first determines the most plausible segmentation for a given compound noun by using an ordinary dictionary. At this initial stage, the segmentation often contains an *over-segmentation* error. That is, when the analyzer encounters an unregistered word, it is likely to segment the word into a sequence of registered words of short length (we empirically confirmed that *word boundary crossing type* errors make up less than 5% of all errors caused by unregistered words). Our method corrects many of *over-segmentation* errors automatically.

Every word in the initial output of the JMA is used as a key in pattern matching. Twenty-three pattern matchers gather various types of word co-occurrence, and many unregistered words can be detected in the process of pattern matching.

For example, in the searches for L={"改正" "大" "店" "法" "施行"}, a pattern matcher finds evidence that "大店法" appears as a single word. Then, "大店法" is registered, added into L, and invokes a search of word co-occurrence around "大店法" itself. This bootstrapping search makes it possible to correct initial *over-segmentation* errors and to obtain the correct solution of morphological analysis.

A comparison of possible dependency structures is conducted by using mutual information and syntactic constraints. Lauer (1995) compared a *dependency model* with *adjacency models*, and found that the dependency model is better. We used the dependency model as well.

We did not use a conceptual dependency model. This is because:
(1) it is difficult to assign a proper position in a thesaurus to an unregistered word.
(2) we aimed to evaluate the performance of the genuine direct scanning approach, since no one has reported whether or not it works, or if it works, how large the corpus should be.

Finally we also introduce a baseline that has yet not been introduced in the literature of Japanese compound noun analysis. The baseline works fairly well, and the text scanning method will turn out to be much better than the baseline.

Section 2 describes the algorithm of text scanning method in detail, section 3 shows the results of our experiments and introduces the baseline. Section 4 discusses problems for future research.

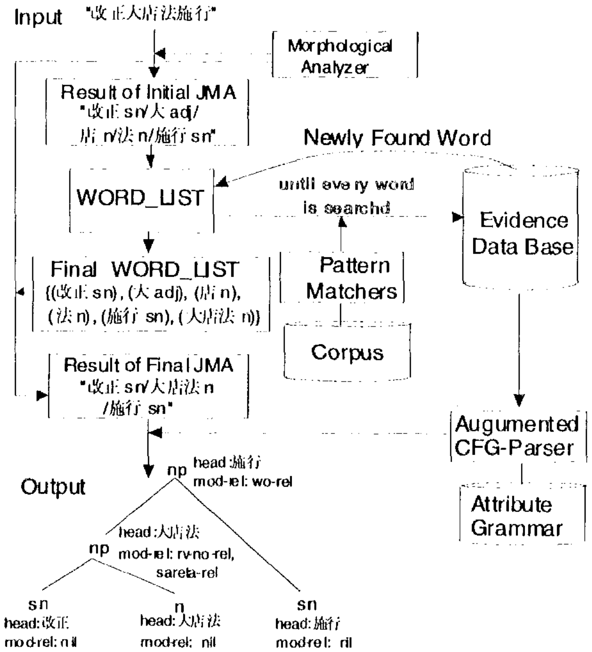
2. Text Scanning Approach

2.1 Overview

Figure 1 illustrates the processing flow. An input compound noun is first analyzed by JMA and segmented into a sequence of registered words. The output is stored as an initial value in a list called WORD_LIST (WL).

For every word in WL, a search for its collocational pattern is conducted, and the results are stored in the evidence data base (EDB). It is important that there is a feedback loop from EDB to WL through which newly found words can be added to WL. The search is continued until every word in WL is used as a key. This feedback enables the *bootstrapping* acquisition of evidence.

Figure 1
Architecture of Direct Scanning Method


Input "改正大店法施行"
Morphological Analyzer
Result of Initial JMA "改正 sn/大 adj/店 n/法 n/施行 sn"
Newly Found Word
WORD_LIST
until every word is searchd
Evidence Data Base
Final WORD_LIST {(改正 sn), (大 adj), (店 n), (法 n), (施行 sn), (大店法 n)}
Pattern Matchers
Corpus
Result of Final JMA "改正 sn/大店法 n /施行 sn"
Augumented CFG-Parser
Attribute Grammar
Output
np head:施行 mod-rel: wo-rel
np head:大店法 mod-rel: rv-no-rel, sareta-rel
sn head:改正 mod-rel: nil
n head:大店法 mod-rel: nil
sn head:施行 mod-rel: nil

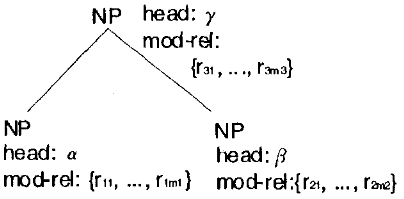
After the searches, the input is re-analyzed using newly found words. The final result of JMA is then passed to a CFG-parser which calculates the cost of possible structures and the attribute-values attached to each node in a solution. In the case that there is ambiguity in the final morphological analysis of a given compound noun, the morphological analyzer picks up the solution with the least number of segmentations.

The procedure of the cost calculation of a dependency structure is basically the same in Kobayashi *et al.* (1994). The cost of the dependency between two nodes is given by using mutual information between the lexical heads of the nodes (fig. 2).

Here two kind of attributes are used; *head*, which records the head of a node as a value, and *mod-rel*, which records the kind of relationship found between two heads of children.

In Japanese, if the two children are both content words, the value of the *head* attribute of the parent node is usually identical to the value of the *head* attribute of the right daughter.

Figure 2
Dependency Representation Using
Attribute-Value Pairs


NP head: γ
mod-rel:
$\{r_{31}, \ldots, r_{3m3}\}$
NP
head: α
mod-rel: $\{r_{11}, \ldots, r_{1m1}\}$
NP
head: β
mod-rel:$\{r_{21}, \ldots, r_{2m2}\}$

2.2 Basic CFG Rules

The category which the morphological analyzer assigns to a word is one of the following: sn (stem of a sino-verb), n (noun), pn (proper noun), num (number), adj (stem of an adjective or an adjectival verb), prfx (nominal prefix), sfix (nominal suffix), num-prfx (numerical prefix), and num-sfix (numerical suffix). CFG rules for compound noun construction use these categories as non-terminals. The following two rules are the most basic: [np → np np] and [np → n]. These rules construct the basic framework of the dependency-structure of a compound noun. We assume that the structure of a compound noun can be represented in the framework of binary-tree grammar by using attribute-value pairs.

2.3 Co-occurrence Data Collection by Direct Text Scanning

This subsection describes the most important part of our method: the pattern matchers and heuristics on unregistered word treatment.

Table 1 shows the main part of the pattern matchers. We will describe the procedure for collecting evidence by using the example mentioned previously, "改正大店法施行".

The initial segmentation of the compound noun is "改正 sn/大 adj/店 n/法 n/施行 sn". Thus the WL initially contains these five words. The words are used as keys for the search. As mentioned in the previous section, this solution contains an *over-segmentation* error, which is the most likely error in the situation when unregistered words appear. Therefore this example captures the typical problem faced in our task.

In Table 1, 'A' stands for a given key, 'B' stands for a sequence of *kanji* characters (we only treat *kanji*-compound nouns in this paper), and 'D' stands for an "extended" delimiter: D is identical to a space, a symbol, a katakana or a hiragana except "の" (*no*; *of*). After preliminary experiments, we decided to eliminate "の" from the delimiters because if it is used, a pattern such as "AのBのC"(roughly C *of* B *of* A) could be picked up, and it may add erroneous evidence because of its ambiguity in dependency structure.

Table 1
Part of Pattern Matchers

1.1	D・AB・D	D・AするB・D	1.5
	D・BA・D	D・AしたB・D	
		D・AされるB・D	
1.2	D・AのB・D	D・AされたB・D	
	D・BのA・D	D・BするA・D	
1.3	D・AいB・D	D・BしたA・D	
	D・AなB・D	D・BされるA・D	
	D・BいA・D	D・BされたA・D	
	D・BなA・D		
		D・AおよびB・D	1.6
1.4	D・AがBする・D	D・AとB・D	
	D・AをBする・D	D・BおよびA・D	
	D・AにBする・D	D・BとA・D	
	D・BがAする・D	D・AについてのB・D	1.7
	D・BをAする・D	D・Aに関するB・D	
	D・BにAする・D	D・BについてのA・D	
		D・Bに関するA・D	

Patterns in 1.1 collect evidence of inner-word collocation of A and B. If the length of A is more than or equal to 2, The length of B is limited to less than or equal to 3. If the length of A is 1, the length of B is limited to less than or equal to 2. Additional explanation will be given later in this subsection.

Patterns in 1.2 collect the evidence of particle-combined collocation of A and B. A and B are combined by a particle "の" which is similar to "*of*" in English. Note that no part of a phrase such as "AのBのC" is picked up so that erroneous evidence can be to avoided. The length of B is limited to less than or equal to 3 (in 1.3, ..., 1.7, the same condition on B is used).

Patterns in 1.3 collect the evidence of an *adjectival modifier-modifiee* relationship between an adjective (or an adjectival noun) and a noun.

Patterns in 1.4 collect the evidence of a *predicate-argument* relation between a sino-verb and a noun. Particles "が" (*ga*), "を"(*wo*) and "に"(*ni*) roughly indicate AGENT, OBJECT and GOAL, respectively.

Patterns in 1.5 collect the evidence of a modifier-modifiee relationship between a sino-verb and a noun, the sino-verb which appears at the tail of a noun modifier phrase and the noun which is modified by the phrase.

Patterns in 1.6 collect the evidence of a coordination relationship between two words.

Patterns in 1.7 collect phrases such as "A about B" and "B about A".

Here we omit the others. One can add any pattern as long as it supplies reliable evidence.

In the following part of this subsection, we will illustrate the search procedure using the initial value of WL {(改正 sn), (大 adj), (店 n), (法 n), (施行 sn)}.

From the first item "改正", evidence shown in 3.1 of figure 3 is collected, and the result is stored in the form shown in 3.1'. Note that the number of occurrences and the observed relationships are recorded. At this stage, the unregistered word "大店法" is already captured by using a pattern matcher in 1.5.

As for the second word, however, one has to be careful because a word with length 1 is very likely to appear through an *over-segmentation* error. The pattern matchers gather evidence such as "大きな変化" (大きな:big; 変化: change), "大学" (university), "大型" (large), "大店法" (large retail-shop law) etc. as given in 3.2. This evidence contains not only correct examples (such as "大きな変化") but also registered words (such as "大学", "大型") and unregistered words (such as "大店法").

To classify the evidence, we developed the following rules:

R-(a)

If (1) the length of A is 1, and the length of B is 1, and (2) there is no entry for the concatenated string AB (BA) in the dictionary used by JMA,
then recognize the concatenated string as an unregistered word, and apply R-(c).

R-(b)

If (1) the length of A is 1, and the length of B is 2, (2) there is no entry for the concatenated string AB (BA) in the dictionary, (3) the category of B is not 'sn' (the condition for AB), and (4) the concatenated string AB (BA) cannot be segmented as a sequence of two registered words A'B'(B'A'), where A'≠A,
then recognize the concatenated string as an unregistered word and apply R-(c).

R-(c)

If (1) the character string consisting of B is identical to the concatenated string of the first or the first two words following A in the initial solution (the condition for AB), or (2) the character string consisting of B is identical to the concatenated string of the first previous or the first two previous words preceding A in the initial solution (the condition for BA), then record AB in WL as an unregistered word, which will invoke pattern matching using AB as a key.

R-(d)

If (1) the length of A is larger than or equal to 2, and
(2) the concatenated string AB (BA) cannot be segmented as a sequence of two registered words A'B'(B'A'), where A'≠A, then, record an evidence of *inner-word* co-occurrence of A and B.

We admit that the definition of a word might be controversial. However, we do not mention the arguments here because of the lack of space. We only say that the standpoint we chose is simple and machine-tractable, and works well for our purpose.

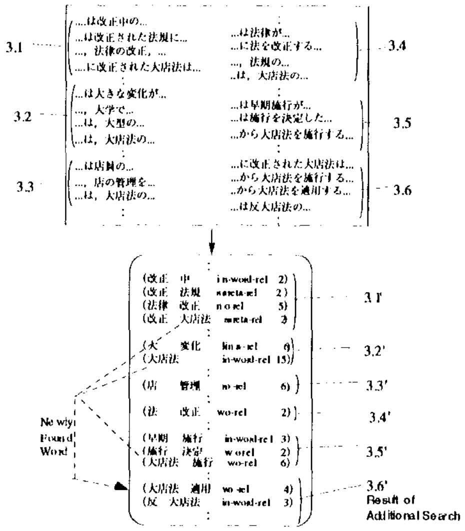
"大きな変化" is recorded as evidence of a straightforward adjectival modifier-modifiee relationship between "大" and "変化".

According to R-(a), "大学" and "大型" are neglected. According to R-(b) and R-(c)-(1), "大店法" is recorded as an unregistered word and stored in WD, which invokes a search of the patterns around it.

Having worked through all the elements in WD, the evidence given in 3.1', 3.2', 3.3', 3.4', 3.5' and finally 3.6' is obtained.

At this stage, JMA re-analyzes the input compound noun by using newly found words. Thus the correct segmentation "改正 sn / 大店法 n / 施行 sn" is obtained, and passed to the CFG-parser.

Figure 3
Example of Evidence Collection



2.4 Selection of Proper Analysis

2.4.1 Cost Calculation and Mutual Information

The rest of the procedure is straightforward. An augmented bottom-up CFG parser chooses the minimum cost tree for the given word sequence. Let NP_3 be the parent of NP_1 and NP_2 in a subtree. Each node has three kinds of attributes: *head*, *mod-rel* and *accum-cost*. *head* has the lexical head of the subtree under NP_i as its value. *mod-rel* keeps the observed relationships captured by the pattern matchers between the two lexical heads of child nodes (this value is not actually used in the following experiments). *accum-cost* c_i records the accumulated cost of the subtree which has NP_i as its root. c_3 is calculated as follows:

$$c_3 = c_1 + c_2 - log_2\left(\frac{N(head_1, head_2)}{N(head_1)N(head_2)}\right)$$

where $N(head_i)$ stands for the number of patterns containing $head_i$, $N(head_1, head_2)$ stands for the number of the patterns containing both $head_1$ and $head_2$. The value of *accum-cost* of each leaf node is set to 0.

2.4.2 Preference to Analysis Containing Observed Evidence

The corpus based approach inevitably encounters the *sparseness problem*. Our approach also encounters this problem, although it turned out to be *not serious*, as will be explained in section 3.3. This subsection describes the heuristic that is employed when the evidence cannot cover any of the entire trees.

Figure 4 shows two possible dependency structures in a three-word compound noun. For simplicity, the values of the *head* attribute are indicated instead of the non-terminal symbols. For three noun words, the following rule is applied:

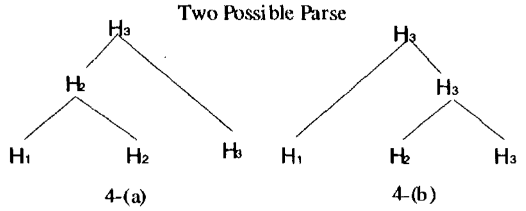
If only the dependency between H_1 and H_2 was observed, then 4-(a) is chosen, else if only the dependency between H_1 and H_3 was observed, then 4-(b) is chosen, else if only the dependency between H_2 and H_3 was observed, then 4-(b) is chosen.

In general, priority is given to the solution containing more subtrees which directly reflect the observed evidence.

In our experiments, the analysis which has multiple minimum cost solutions was considered to have failed.

Figure 4

Two Possible Parse



3. Results

3.1 Test Data

We used the articles contained in "*Nikkei Shinbun*" for January and February in 1992 as the corpus for the experiments. The number of the articles is about 27,000, which contain about 7 million characters.

Experiments were carried out using 400 compound nouns: 100 for 5-*kanji* words, 100 for 6-*kanji* words, 100 for 7-*kanji* words and 100 for 8-*kanji* words. The frequency of these word lengths is about the same in the corpus. After randomly selecting the test samples, we confirmed that they were all compound nouns.

Numerical expressions appeared in 10% of the test samples, and such expressions were pre-processed as follows:

"約百八十業者" → "約 pr-num/ 百八十 num/ 業者 n"
(約: about; 百: hundred; 八: eight; 十: ten; 業者: dealer)

3.2 Baseline

Baselines have rarely been introduced in research on Japanese noun compounds. This paper introduces a baseline to facilitate our evaluation of the effectiveness of our method.

The baseline we used is *leftmost derivation*. This is an extension of *left branch preference* in Lauer (1995). The baseline is also a well-known heuristic method to analyze Japanese noun phrases combined with "の" (such as "AのBのC"). As shown below, this heuristic method works well especially when the length of a compound noun is relatively short. Note that the baseline correctly analyzes "改正大店法施行" if "大店法" is registered. However, the baseline actually fails because it cannot capture the unregistered word.

3.3 Results and Comparison

Table 2 shows the results of the proposed method. The first line indicates the number of samples for which the correct dependency structure was given as the single minimum cost solution. The second line indicates the accumulated number of samples for which the correct dependency structure was given as one of the minimum cost solutions. Table 3 shows the results of the baseline, and indicates the number of samples for which the correct dependency structure was given.

Table 2

word length	5	6	7	8
1	89	70	58	58
~1	92	81	76	83

The result of Direct Scanning

Table 3

word length	5	6	7	8
1	83	63	41	39

The result of baseline

Comparing the two tables reveals that the proposed method is more accurate than the baseline. For longer word length, the difference is greater.

Our result cannot be compared accurately with the existing result (Kobayashi *et al.*, 1995) because we used a different test corpus, and only the results on 4-, 5- and 6-*kanji* compound nouns were reported. However, the accuracy of their results on 6-*kanji* compound nouns is 53%, unless they combine their conceptual dependency model with a heuristic using the distance of modifier and modifee. After combining the model and the heuristic, accuracy improves to 70%, which is the same as ours.

An 8-*kanji* compound noun usually contains four nouns. The performance of our method (accuracy of 58%) is encouraging, since most of the errors were caused by proper nouns. This problem can be solved using a pre-processor (explained below).

3.4 Causes of Errors

Forty-two percent of the error was caused by proper nouns, 16% by time expressions, and 15% by monetary expressions. This means that proper nouns are a major cause of the errors, as pointed out in previous research. There are several reasons for this:

(1) an identical proper noun normally does not appear

many times in the corpus.

(2) proper nouns sometimes cause *cross-boundary* errors at the initial morphological analysis.

We can be optimistic about eliminating these three types of errors. If we use a preprocessor (for proper nouns, see Kitani *et al.*, 1994), most of them can be eliminated.

4. Future Directions

This paper discussed performance of the direct text scanning method. There remain several interesting problems:

(1) We did not employ the conceptual dependency model. A method for combining a conceptual dependency model with the proposed approach should be investigated and the results analyzed.

(2) A proper noun pre-processing module should be combined with the proposed method.

(3) The effect of varying the corpus size should be investigated.

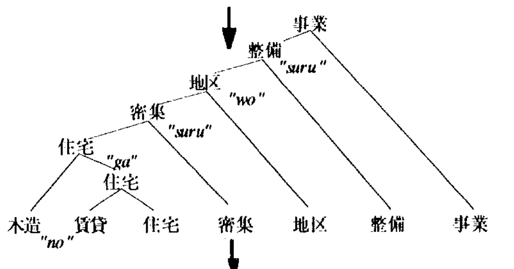
(4) The distance between a compound noun and its evidence should be reflected in the cost calculation in comparing solutions.

(5) Parallel search should be employed to speed up the process.

(6) How to obtain an expanded expression from a given compound noun should be investigated. At the moment, the value of the *mod-rel* attribute is not used. Some compound nouns can be rephrased with an ordinary Japanese sentence. Figure 5 shows an example of expansion.

Figure 5
Analysis of Long Word and
Expansion to Ordinary Japanese

"木造賃貸住宅密集地区整備事業"



((((木造"no"(賃貸 住宅))"ga"密集)"suru"地区)"wo"整備)"suru" 事業)

"木造の賃貸住宅が密集する地区を整備する事業"
"an enterprise which aims at improving the area where many wooden apartments for rent stand close together"

木造: mokuzo; wooden	賃貸: chintai; rental
住宅: juutaku; apartment	密集: misshuu; crowd
地区: chiku; area	整備: seibi; improve, maintain
事業: jigyou; enterpise	

5. Conclusion

A corpus-based approach for analyzing Japanese compound nouns was proposed. This method scans a corpus with a set of pattern matchers and gathers external evidence to analyze compound nouns. It employs a bootstrapping procedure to cope with unregistered words: if an unregistered word is found in the process of searching the co-occurrence examples, the newly found word is recorded and invokes additional searches, which enable necessary evidence to be gathered for the given compound noun. This also makes it possible to correct *over-segmentation* errors in the initial segmentation, and leads to higher accuracy. The method is also very portable because it depends little on a dictionary of a morphological analyzer and treats registered words and unregistered words in the same manner. The accuracy of the method was evaluated using the compound nouns of length 5, 6, 7, and 8. A baseline, which takes leftmost derivation strategy, was also investigated for comparison with our method. The proposed method is much more accurate than the baseline in the experiments for words of four different lengths.

Acknowledgement

We would like to express our gratitude to Professor Yorick Wilks (Sheffield) and Dr. Shojiro Asai (Hitachi, Ltd.), who gave the first author the opportunity to do this research at the University of Sheffield as a visiting researcher (from January to December, 1995).

References

Finin, Tim. 1980. *The Semantic Interpretation of Compound Nominals*, PhD Thesis, Co-ordinated Science Laboratory, University of Illinois, Urbana, IL

Lauer, Mark. 1995. *Corpus Statistics Meet the Noun Compound: Some Empirical Results*, in Proc. of ACL, pp.47-54

McDonald, David B. 1993. *Internal and External Evidence in the Identification and Semantic Categorization of Proper Names*, in Proc. of SIGLEX workshop on *Acquisition of Lexical Knowledge from Text*, pp. 32-43, Ohio, USA

Miyazaki, Masahiro. 1984. *Automatic Segmentation Method for Compound Words Using Semantic Dependent Relationships between Words*, in Trans. of IPSJ, Vol. 25, No. 6, pp.970-979

Kitani, T. and Mitamura, T. 1994. *An Accurate Morphological Analysis and Proper Name Identification for Japanese Text Processing*, in Trans. of IPSJ, Vol. 35, No. 3, pp.404-413

Kobayashi, Y., Tokunaga, T. and Tanaka, H. 1994. *Analysis of Japanese Compound Noun using Collocational Information*, in Proc. of COLING, pp. 865-869

Tanaka, Yasuhito. 1992. *Acquisition of knowledge for natural language; the four kanji character sequence* (in Japanese), in National Conference of Information Processing Society of Japan