Detecting Japanese Term Variation in Textual Corpus

Fuyuki Yoshikane, Keita Tsuji
(Graduate School of Education, University of Tokyo, Japan),
Kyo Kageura
(National Center for Science Information Systems (NACSIS), Japan),
and Christian Jacquemin
(Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur (LIMSI-CNRS), France)

Abstract

In this paper, we describe a rule-based mechanism that detects Japanese term variations from textual corpora. The system operates on the basis of meta-rules that map syntactic and morpho-syntactic variations of terms to the original forms of terms. The framework used here has been successfully applied in such languages as English and French, and we show here that this also works well in detecting Japanese term variants, once we properly take into account specific characteristics of Japanese language. We also discuss the potential of this work for IR related applications.

1 Introduction

Recently the role of technical terminology in communication has become more and more important, not only in limited technical communication but also in more general communication. Among scientific terminology, complex or multi-word terms consisting of two or more elements or single words constitute the majority (Ishii, 1987)[1] . The ratio of multi-word terms in technical terminology in general is growing, so multi-word terms are expected to play an important role in information retrieval (IR).

However, multi-word terms do not always appear in text in their canonical forms. Terms tend to be syntactically and/or morphologically transformed into their variants when appearing in an actual context, while keeping the identity of concepts they represent. By detecting those variants properly, therefore, we can expect an improvement of recall in retrieval, without affecting precision.

Fastr (Jacquemin, 1994) is a system which was developed for the recognition of term variants in textual corpora and has been successfully applied to such languages as French and English (Jacquemin, 1996; Jacquemin, Klavans, & Tzoukermann, 1997). In this paper, we report the preliminary results of applying the same framework to Japanese term variant detection, i.e. the use of Fastr for Japanese term variant detection.

[1] Here we use multi-word terms and complex terms as synonyms, although they are not exactly the same.

We briefly summarise the general characteristics of the Fastr system in section 2. In section 3, we describe the Japanese rules for detecting term variants. In section 4, the results of a preliminary experiment in detecting term variants using Fastr is reported.

2 General Framework for Term Variant Detection using Fastr

In this section, we describe the framework for term variant detection using Fastr, a unification-based partial parser aimed at detecting term variants in texts (Jacquemin, 1994), which we used for the Japanese term variant detection reported in this paper. In the Fastr framework, one has to describe the linguistic knowledge and rules for detecting term variants using three types of rules, i.e. (1) meta-rules which describe the skeleton of variation patterns, (2) term rules which list and describe the structure of original terms whose variations are to be detected, and (3) single-word rules which define morphemes or single-words used in original terms. For instance, a term rule of “language processing” and single-word rules of constituent words of the term, i.e. “language” and “processing”, are described as follows.

Rule ($N_1 \rightarrow N_2\ N_3$)
< N_2 lemma> = ‘language’
< N_3 lemma> = ‘processing’

Word ‘language’ :
<cat> = N.
Word ‘processing’ :
<cat> = N.

By applying term rules such as this to meta rules, Fastr generates new rules which detect variants of the term.

Meta rules, which take the form of context-free skeletons, define the syntactic structures of original terms and their potential variants. The following is an example of a meta-rule:

Meta-rule Coordination ($X_1 \rightarrow X_2\ N_3$)
= ($X_1 \rightarrow X_2\ N_5\ C_4\ N_3$)

The left side of the equation represents the structure of original terms, and the right side represents the associated variations. Here N represents a noun, C represents a coordinating conjunction, and X represents any syntactic category. The same subscript on the left and right side of the equation shows that these elements are the same. So this meta rule indicates that a complex term which consists of an element of an arbitrary category followed by a noun may have a variation whose structure is the first element of the original term followed by the coordinating conjunction of a noun and the noun element in the original term.

Assume that we have defined a term rule of “language processing” as described above. Then, by applying the term rule to the meta rule ‘Coordination’, Fastr generates a new rule:

Rule ($X_1 \rightarrow$ language processing)
= ($X_1 \rightarrow$ language $N_5\ C_4$ processing)

By this rule, Fastr can detect terms such as “language understanding or processing” as as a variant of “language processing”.

Fastr can recognise not only syntactic transformations but also morphological transformations. This is done by adding constraint equations, which link morphological derivatives to meta-rules. For instance, the following meta-rule generates a variant including Adjective (A_4) which is a derivative of Noun (N_2).

Meta-rule Noun-Adjective($X_1 \rightarrow N_2\ N_3$)
$= (X_1 \rightarrow A_4\ N_3)$
$< A_4$ root$> = < N_2$ root$>$

By this meta-rule, Fastr detects, for instance, "categorial grammar" as a variant of "category grammar" because "categorial" and "category" share the same root.

3 Japanese Meta-Rules

3.1 Basic Syntactic Categories

The syntactic and morphological structures of Japanese are different from those of languages such as English or French. For instance, there is no delimiting symbol between words, and it is not possible to formally distinguish derivation from compounding (cf. Koyama, Yoshioka & Kageura, 1998). As our aim is to develop a practically applicable system, we decided to follow the convention of word-boundary delimitation and broad-level syntactic categories used in the most widely used morphological analyser for Japanese, i.e. JUMAN (Matsumoto, et. al, 1998).

Word categories must be defined accurately in meta-rules as well as in term and single-word rules. For this purpose, we have also used some detailed-level categories used by JUMAN in some cases. All in all, 20 syntactic categories were taken from the JUMAN tagset.

The basic syntactic categories used in describing Japanese rules for term variant detection are given in Table 1. Categories of functional words which are not used in describing Fastr rules are not listed in Table 1.

Some categories in Table 1 need explanation. Firstly, VS, the s-inflexional or sahen verb *suru* ('do'), is distinguished from common verbs (V) because the sahen-verb functions as both a common and special auxiliary verb[2] . Sahen-nouns (NS), which can be turned into verbs by adding the sahen-verb to the end are distinguished from other nouns (N) as their syntactic behaviour is different.

S, 'postpositional particle connecting coordinate words', corresponds to English coordinating conjunctions such as "and" and "or". We distinguished them from the other postpositional particles because this is necessary in order to describe an important variation pattern of Japanese terms.

Suffixes which are relevant to describing term variations are divided into three. $TPNN$ is the suffix that produces the noun or adjectival stem from the noun, and the majority are one-Chinese character suffixes (Nomura, 1978; Kageura, 1994). $TPAN$ is the suffix that produces the adjective from the noun. Other suffixes are categorised into one group. As we will discuss shortly, this distinction presupposes the decomposition of JUMAN, which is practically useful but not necessarily ideal from the theoretical point of view. Note that suffixes are regarded here as independent units, just like other categories for independent words. This comes from the characteristic use of Chinese characters in Japanese, and will make the distinction of syntactic and morpho-syntactic variations as defined in the Fastr framework less clear in defining meta-rules.

[2] The sahen verb or s-inflexional verb is referred to as such because the inflexional paradigm is bound to such patterns as 'sa', 'si', 'su', 'se', etc.

Symbol	Category
L	Delimiter
V	Verb (except VS)
VS	Sahen-verb *suru*
A	Adjective
N	Noun (except NS)
NS	Sahen-noun
RB	Adverb
S	Postpositional particle (except SC)
SC	Postpositional particle connecting coordinate words
TP	Suffix (except $TPNN$ and $TPAN$)
$TPNN$	Suffix deriving noun from noun
$TPAN$	Suffix deriving adjective from noun

Table 1: Major categories in the part-of-speech scheme

3.2 Patterns of Japanese term variations

In order to define the meta-rules (or the linguistic rules) that can describe term variants, we observed actual variations of terms that occur in texts. Firstly, we prepared terminological and textual data. The terminological data were extracted from the terminological part of the EDR database, which lists information processing terms (EDR, 1996). Among the approximately 120,000 EDR terminological data, we sampled about 40,000 terms. For textual data, we used 1827 titles and abstracts of conference presentations in the field of artificial intelligence, a subdomain of the field of information processing, extracted from the NACSIS Academic Conference Database (NACSIS, 1998).

Both terminological and textual data were processed by JUMAN 3.5. For each multi-word term in the terminological data, we extracted corresponding sentences from the textual data, i.e. sentences which include all the constituent elements of the term. At this stage, only verbs, adjectives, nouns and adverbs were checked. After examining the data, we defined the following patterns.

A. Modification :

Insertion of a modifier makes variants, which have narrower or more informative senses than the original terms. Nouns, adjectives and verbs are used for modification of nouns. However, only nouns can be inserted between nouns as a modifier without a postpositional particle.

For example,

zenbun kensaku → *zenbun kiiwaado kensaku* (full text search → full text key word search)

B. Decompounding/Compounding :

A compound noun can be decompounded by inserting a postpositional particle. Such a decompounded term is a variant of the original compound noun. In some decompounding variations, the appropriate verb is inserted together with the case-marking postpositional particle such as 'wo'.

For convenience, the variation in which a delimiter is inserted between nouns is classified into this pattern[3] .

In the compounding variation, on the other hand, the postpositional particle or the delimiter between nouns is omitted.

For example,
jouhou kensaku ↔ *jouhou no kensaku* (information retrieval ↔ retrieval of information)
fureemu waaku ↔ *fureemu - waaku* (frame work)
kumiawase mondai → *kumiawase wo motomeru mondai* (combination problem → combination seeking problem)

C. Coordination :

Terms which consist of two coordinated elements with a common head word or a common argument are considered to be variants of the original terms which consist of one of the coordinated elements with the head or the argument. The variation combining two terms with a common sahen-verb *suru* is also classified into this pattern.

For example,
bunkai genri → *bunkai -ketsugou genri* (resolution theory → resolution and combination theory)
chishiki no kakutoku → *chishiki no seisei ya kakutoku* (knowledge acquisition → knowledge generation and acquisition)
hikaku suru → *hikaku - kentou suru* (compare → compare and examine)

D. Sahen-Noun-Verb variations :

The Sahen-verb *suru* can be connected with sahen-nouns and constitute various verbs derived from the sahen-nouns. We call the variations relating to such types of derivation sahen-noun-verb variations.

For example,
gainen gakushuu ↔ *gainen wo gakushuu suru* (concept learning ↔ learn concept)

E. Noun-Noun variations :

Some nouns can be syntactically transformed into their noun derivatives or adjectival stems by adding special suffixes, e.g. *teki*, *ka*, etc. Let us conveniently call this type of variation the noun-noun variation.

For example,
kika moderu ↔ *kika teki moderu* (geometry model ↔ geometrical model)

F. *Na*-Adjective-Noun variations :

By eliminating the inflexion of a *na*-adjective, we can morphologically transform the adjective into a noun. Likewise, by attaching the inflexion to the noun which is the root of the *na*-adjective, we can morphologically transform the noun into an adjective. The inflexion *na* is not an independent morpheme but a part of *na*-adjective. Therefore, *na*-adjective-noun variations are morpho-syntactic ones.

For example,
aimai jouhou ↔ *aimaina jouhou* (ambiguous information)

G. Noun-Adjective variations :

Some nouns are syntactically transformed into adjectives by adding suffixes which can derive adjectives from nouns. We call the variations relating to such types of derivation noun-adjective variations[4] .

[3] In Japanese, loan terms may or may not be divided into words by a delimiter.

[4] In fact, this can be defined as the combination of meta-rule types D and G. We established this type because for now we

For example,

kika moderu ↔ *kika tekina moderu* (geometry model ↔ geometrical model)

Many of the patterns of Japanese term variations above are analogous to those in French or English (cf. Jacquemin, Klavans, & Tzoukermann, 1997). In such languages as English or French, it is relatively easy and straightforward to distinguish derivation and compounding. In Japanese, however, this is not the case. Linguistically speaking, it may be more appropriate to regard the variations E, F and G as variations based on morphological operations.

However, in implementing these meta-rules in Fastr (Jacquemin, 1994), only F is categorised as a pure morpho-syntactic variation in contrast to syntactic variations. In morpho-syntactic variants, at least one of the content words W of the original term is transformed into another word W' in the variant such that W and W' have the same root form. However, as there are ambiguities between syntactic and morpho-syntactic variations especially in the case of E, which is caused by the ambivalent status of the Japanese suffix, G and F are defined in the same manner as syntactic variations here.

On the other hand, that A, B, C and D are syntactic variations is reasonably clear. We do not go into this further in this paper as this is more closely related to Japanese morphological analysis than to term variant detection. What should be emphasised here is that the borderline between syntactic and morpho-syntactic variations is less clear than in the case of English or French.

In some cases, variants are produced by combinations of these elementary patterns. For instance, the variant "*dokuritsuna hensuu* → *dokuritsu sita hensuu*" is the composition of the *na*-adjective-noun variant "*dokuritsuna hensuu* → *dokuritsu hensuu*" and the sahen-noun-verb variant "*dokuritsu hensuu* → *dokuritsu sita hensuu*".

Table 2 shows sample meta-rules for each type of variation. These meta-rules consist of context-free skeletons and constraint equations. The symbols '*' and '+' in context-free skeletons are regular expressions, expressing their well-established meaning, i.e., '*' means 'zero or more times' and '+' means 'one or more times'. The tags used in constraint equations are based on the form of Fastr.

4 Experiments and Evaluations

After having established these meta-rules for Japanese term variants, we carried out a preliminary experiment to check the validity of the meta-rules. As mentioned previously, Fastr has to be equipped with meta-rules, single-word rules and term rules. We compiled about 87,000 term rules from EDR terminological data in information processing[5] , and about 17,000 single-word rules derived from their constituent elements.

In order to evaluate the meta-rules described in the previous section, we applied them to a corpus consisting of the titles and abstracts of 914 papers, half the data belonging to the field of artificial intelligence. The number of sentences in the corpus was 5,322. By JUMAN 3.5, those sentences were decomposed into 150,406 words.

cannot control the application order of meta-rules.

[5] From 120,000 original terms, we excluded extremely long elements, which cannot be processed by Fastr, as well as borrowed elements without delimiters when their variants with delimiters are also listed.

Variation	Sample meta-rule, and example of transformation
A. Modification	$X_2\ X_3 \rightarrow X_2\ (N \mid NS)^+\ X_3$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_3\ \text{cat}> = N \mid NS$ *zenbun[N] kensaku[NS]* → *zenbun[N] kiiwaado[N] kensaku[NS]* (full text search → full text key word search)
B. Decomposition	$X_2\ X_3 \rightarrow X_2\ S_4\ A^?\ (N \mid NS)^*\ X_3$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_3\ \text{cat}> = N \mid NS, \quad < S_4\ \text{lem}> =$ '*no*' *shisutemu[N] tokuchou[N]* → *shisutemu[N] no[S] omona[A] tokuchou[N]* (system feature → main feature of system)
B. Composition	$X_2\ S_3\ X_4 \rightarrow X_2\ (N \mid NS)^*\ X_4$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_4\ \text{cat}> = N \mid NS, \quad < S_3\ \text{lem}> =$ '*no*' *jouhou[N] no[S] kensaku[NS]* → *jouhou[N] kensaku[NS]* (retrieval of information → information retrieval)
C. Coordination	$X_2\ X_3 \rightarrow X_2\ S_4\ A^?\ (N \mid NS)^+\ (SC \mid L)\ X_3$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_3\ \text{cat}> = N \mid NS, \quad < S_4\ \text{lem}> =$ '*no*' *chishiki[N] kakutoku[NS]* → *chishiki[N] no[S] seisei[NS] ya[SC] kakutoku[NS]* (knowledge acquisition → knowledge generation and acquisition)
D. Sahen-noun-verb	$X_2\ NS_3 \rightarrow X_2\ S_4\ (A \mid RB)^*\ NS3\ VS_5$ $< X_2\ \text{cat}> = N \mid NS, \quad < S_4\ \text{lem}> =$ '*wo*' *gainen[N] gakushuu[NS]* → *gainen[N] wo[S] gakushuu[NS] suru[VS]* (concept learning → learn concept)
E. Noun-Noun	$X_2\ X_3 \rightarrow X_2\ S_4\ A^?\ X_3\ TPNN_5$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_3\ \text{cat}> = N \mid NS, \quad < S_4\ \text{lem}> =$ '*no*' *gainen[N] kaisou[N]* → *gainen[N] no[S] kaisou[N] ka[TPNN]* (concept hierarchy → hierarchisation of concept)
F. *Na*-adjective-noun	$A_2\ X_3 \rightarrow X_4\ (N \mid NS)^*\ X_3$ $< X_3\ \text{cat}> = N \mid NS, \quad < X_4\ \text{cat}> = N \mid NS, \quad < A_2\ \text{root}> = < X_4\ \text{root}>$ *aimaina[A] jouhou[N]* → *aimai[N] jouhou[N]* (ambiguous information)
G. Noun-Adjective	$X_2\ X_3 \rightarrow X_2\ TPNN_4\ A^?\ (N \mid NS)^*\ X_3$ $< X_2\ \text{cat}> = N \mid NS, \quad < X_3\ \text{cat}> = N \mid NS$ *kika[N] moderu[N]* → *kika[N] tekina[TPAN] moderu[N]* (geometry model → geometrical model)

Table 2: Japanese meta-rules

4.1 Basic Results of Preliminary Experiments

After applying Fastr based on the above-mentioned meta-rules, term-rules and single-word rules, we manually judged the result. If the content-bearing elements of an original term appear in its variant with the relation, the concept represented by the original term will be kept in the variant even if the surface form is changed. We judged such a variant as correct.

The number of original terms extracted from the corpus was 5,164. The number of extracted variants are shown in Table 3. Table 3 shows the numbers of correct and incorrect variants and the precision, i.e. the ratio of correct variants to extracted variants.

The total precision of rules from A to G exceeded 90%. As for individual rules, the precision of C was 65%, but other than that, the precision of all the rules were well over 90%. As we will see shortly, further improvement is expected to be possible with respect to C.

In addition to the 'proper' meta-rules defined in the previous section, we also tried some rough meta-rules, to observe possible variations not covered by rules A–G, and also to observe the effect of the granularity of rule descriptions.

X. Others :

X1. General insertion or omission of postpositional particle. Categories of both the left and right words are not constrained.

$X_2\ X_3 \rightarrow X_2\ S_4\ X_3$

$X_2\ S_3\ X_4 \rightarrow X_2\ X_4$

hikaku[NS] suru[VS] ↔ *hikaku[NS] wo[S] suru[VS]* (compare)

X2. General insertion or omission of suffix. Categories of both the left and right words are not constrained.

$X_2\ X_3 \rightarrow X_2\ X_4\ X_3$

$< X_4$ cat$> = TP \mid TPNN \mid TPAN$

$X_2\ X_3\ X_4 \rightarrow X_2\ X_4$

$< X_3$ cat$> = TP \mid TPNN \mid TPAN$

gengo[N] tekida[TPAN] ↔ *gengo[N] gaku[TPNN] tekida[TPAN]* (linguistic)

X3. Permutation of nouns.

$X_2\ X_3 \rightarrow X_3\ (N \mid NS)^*\ X_2$

$< X_2$ cat$> = N \mid NS, \quad < X_3$ cat$> = N \mid NS$

memori[N] kyoyu[NS] → *kyoyu[NS] waakingu[N] memori[NS]* (memory sharing → shared working memory)

The precision of the rule X is very low, i.e. less than 60%. If we consider X together with proper meta-rules, the overall precision is less than 85%.

4.2 Usefulness in Indexing

In order to assess the important role of term variation in automatic indexing, we have also evaluated the result of regarding the experiment as automatic indexing over a controlled vocabulary. Our purpose was to evaluate to which extent the exploitation of variant conflation increases the coverage of indexing.

For the same 914 documents with 87,000 terms, we compared the number of indexed documents with and without taking into consideration variant forms. The statistics are shown in Table 4. If we use variants of terms in indexing documents, we can find, on average, about two additional papers per term in the corpus[6] .

RAP in Table 4 indicates the ratio of the papers indexed by variants but not by the original terms to

[6] Here, we only consider the terms whose variants appear in the corpus at least once.

Variation	Correct	Incorrect	Sum	Precision
A. Modification	366	27	393	93.13 %
B. Decompounding / Compounding	755	51	806	93.67 %
C. Coordination	51	29	80	63.75 %
D. Sahen-noun-verb	180	16	196	91.84 %
E. Noun-noun	39	3	42	92.86 %
F. *Na*-adjective-noun	8	0	8	100.00 %
G. Noun-adjective	37	1	38	97.37 %
Sum(A-G)	1436	127	1563	91.87 %
X. others	260	178	438	59.36 %
Total sum(A-X)	1696	305	2001	84.76 %

Table 3: Extracted variations

the papers indexed by the original terms. As for the terms which appear in the corpus in both their original forms and variants, the number of additional papers indexed by variants is, on average, approximately equal to the number of papers indexed by the original term.

4.3 Diagnosis

In the above preliminary experiment, we noted a few points related to meta-rules, which we briefly discuss here. There are two notable errors related to the meta-rules A to G.

Firstly, insertion of words sometimes changes the relationship between constituent elements of the terms. This accounts for many incorrect variants extracted through proper meta-rules. For example, *gakushuu shien shisutemu* (learning support system) is an incorrect variant of *gakushuu shisutemu* (learning system), as 'learner' is the 'system' in the case of the original term, while 'learner' is people not the 'system' in the variant. This type of error typically occurs when the type A (modification) meta-rule is applied, but the same type of errors are often observed in variants extracted through types B-G, because these types incorporate the same operation as modification (see Table 2). This type of error might be avoided by refining the system of syntactic categories. Another possibility for improving the performance is to take into account the external context surrounding the elements described by meta-rules.

Secondly, the performance of type C (coordination) was proved to be very low compared to other 'proper' meta-rules. Most of the incorrect variants detected by rule C were caused by incorrect extraction of the modification structure. For example, [*kioku no*] *kousei oyobi taiwa no seigyo* (organisation [of memory] and control of dialog) may not be a correct variant of *kousei seigyo* (organisation control) because *kousei* in the variant does not modify *seigyo*. This is related to the limitation of the formal mechanism of Fastr, i.e. it cannot take into account the full structure of phrases or sentences, but still some of these errors may be avoided by the refinement of meta-rules and syntactic categories. For instance, we can expect that

	DT	AP	AP/DT	RAP
Type 1 terms	213	475	2.23	0.94
Type 2 terms	416	784	1.88	–
Total	629	1259	2.00	–

Type 1 terms: both the original form and its variants appear in the corpus.

Type 2 terms: only the variants appear in the corpus.

DT: number of different terms.

AP: total number of additional papers retrieved by variants, where only variants appear.

RAP: average of the ratio of the number of additional papers retrieved
by variants to the number of papers retrieved by original terms.

Table 4: Additional papers retrieved by variants

the performance of C will be improved by subdividing the category L (delimiter). Some delimiters are mainly used for punctuation between words, and others are mainly used for punctuation between clauses. Therefore, we had better use only the former to define the meta-rule of type C shown in Table 2.

Let us now shift our observation to the additional meta-rule X. The performance of X as a whole was low (about 60% precision). However, among the three types of X, the precision of X1 was 56.79%, that of X2 was 91.41%, and that of X3 was 28.00%. As far as simple performance is concerned, the precision of X2 was close to that of proper meta-rules.

Most variations detected by the meta-rule X1 belong to the following two types.

Y1. Insertion of the case-making postpositional particle '*ga*' between nouns.
In this case, the second noun is a sahen-noun, which is followed by the passive form of the sahen-verb 'suru' (though the sahen-verb and the suffix are not extracted).

$$X_2\ NS_3\ \rightarrow\ X_2\ S_4\ NS_3\ [VS_5\ TP_6]$$
$$< X_2\ \text{cat}>= N \mid NS, \quad < S_4\ \text{lem}>= \text{`}ga\text{'}$$
$$[< TP_6\ \text{lem}>= \text{`}reru\text{'}\ \mid\ \text{`}rareru\text{'}]$$

Y2. Insertion of the case-making postpositional particle '*wo*' between a sahen-noun and the sahen-verb '*suru*'.

$$NS_2\ VS_3\ \rightarrow\ NS_2\ S_4\ VS_3,\ < S_4\ \text{lem}>= \text{`}wo\text{'}$$

When there is no passive suffix, insertion of the case-making postpositional particle '*ga*' often changes the role of $X_2[S_4]$ against $NS_3[VS_5]$. For instance, the correct variant of *shisutemu teian* (system proposal) is not *shisutemu ga teian* [*suru*] (in which *shisutemu* is a subject of *teian*) but is *shisutemu wo teian suru* (in which *shisutemu* is an object of *teian*). This affects the precision negatively.

As we previously mentioned, X2 produces high performance. The correct variants extracted through X2 were proved to be strictly defined by the following meta-rules.

Y3. Insertion or omission of $TPNN$ between a noun and suffix.

$X_2\ X_3 \ \rightarrow\ X_2\ TPNN_4\ X_3,\quad <X_2\ \text{cat}>= N \mid NS,\quad <X_3\ \text{cat}>= TPNN \mid TPAN$

$X_2\ TPNN_3\ X_4 \ \rightarrow\ X_2\ X_4,\quad <X_2\ \text{cat}>= N \mid NS,\quad <X_4\ \text{cat}>= TPNN \mid TPAN$

This is a type of D (noun-noun variation), but the original meta-rules that belong to D assumed that the suffix be inserted between nouns and it could not extract this type of variation.

X3 (permutation of nouns) gave the lowest precision. Most of the errors were caused by changes of the relationship between words which were resulted from the change of the word order. For instance, while *shisutemu shindan* (system diagnosis) implies something which diagnoses a system, *shindan shisutemu* (diagnosing system) implies a system which diagnoses something. However, when one of the nouns is a sahen-noun, the relationships are kept identical in most cases, though the head becomes the modifier and vice versa.

Y4. Permutation of nouns.

$X_2\ NS_3 \rightarrow NS_3\ X_2,\quad <X_2\ \text{cat}>= N \mid NS$

$NS_2\ X_3 \rightarrow X_3\ NS_2,\quad <X_3\ \text{cat}>= N \mid NS$

For instance, while *memori kyoyu* (memory sharing) is a process and *kyoyu memori* (shared memory) is a result, the relation of the constituent elements is the same, so in some cases the former may be regarded as a correct variant of the latter.

The rule Y1, Y2 and Y3 might be described more properly by type F (sahen-noun-verb), type B (compounding/decompounding) and type D (noun-noun variation), respectively. As for Y4 (permutation of nouns), further examination of other samples will be needed to determine whether we should add it to the proper rules or not.

5 Discussion

In this paper, we have shown that the framework of Fastr can be successfully applied to Japanese term variant detection, by which we can expect an improvement in recall without the sacrifice of precision in IR, if the framework is properly applied. However, different meta-rules give different performances. So, we must consider the selection of meta-rules which satisfy the demands of applications.

In the Japanese language, as we previously mentioned, there is no delimiting symbol which explicitly divides words. Thus, we cannot clearly distinguish between syntactic variations and morpho-syntactic ones. In this paper, by basically following the criteria of the existing morphological analyser JUMAN 3.5, we avoided the theoretical problem related to this. Even practically, however, JUMAN 3.5 has problems caused by inconsistencies. For instance, derivational suffixes are sometimes treated as independent words or morphemes[7] . By adding a post-processing filter to the results of JUMAN, we will be able to improve the consistency of meta-rules.

There is another point which should be improved concerning preparation of meta-rules. When Fastr extracts one variant, only one meta-rule is applied. In other words, we cannot extract variants by the combination of meta-rules. In order to construct a practical system, therefore, we have to describe a large number of meta-rules which are made up of the combination of the elemental variations explained in this paper.

[7] For example, *jiyuudo* (degree of freedom) is divided into two morphemes, *jiyuu [N]* and *do [TPNN]*. However, *aimaido* (degree of ambiguity) is recognised as one noun. In the latter case, linking its derivatives is not possible.

We are currently considering the possibility of building a meta-rule precompiler which can automatically compile a set of possible combinations of given meta-rules.

In addition to these, it is expected that it will be very useful to exploit the semantic links of synonyms (cf. Jacquemin, 1999). In fact, as a descriptive framework, Fastr has the facility to treat semantic links as well. By improving these aspects, it will be possible to construct a practically applicable robust term variant recogniser for Japanese.

References

EDR, (1996) *EDR electronic dictionary version 1.5 technical guide.* Tokyo: EDR.

Jacquemin, C. (1994) "Fastr: a unification-based front-end to automatic indexing." In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, p. 34–47.

Jacquemin, C. (1996) "A symbolic and surgical acquisition of terms through variation." In: Wermter, S. Riloff, E. & Scheler, G. (eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.* Heidelberg: Springer. p. 425–438.

Jacquemin, C. (1999) "Syntagmatic and paradigmatic representations of term variation." In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341–348.

Jacquemin, C., Klavans, J. L. & Tzoukermann, E. (1997) "Expansion of multi-word terms for indexing and retrieval using morphology and syntax." In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics ((E)ACL'97)*, p. 24–31.

Ishii, M. (1987) "Economy in Japanese scientific terminology." In *Terminology and Knowledge Engineering '87*, p. 123–136.

Koyama, T., Yoshioka, M. & Kageura, K. (1998) "The construction of a lexically motivated corpus — The problem with defining lexical units." In *Proceedings of the First International Conference on Language Resources and Evaluation*, p. 1015–1019.

Kageura, K. (1994) "Differences in the linguistic representation of concepts in Japanese and English complex noun terms." *Terminology*, 1(1), p. 103–119.

Matsumoto, Y., Kurohashi, S., Yamaji, O., Taeki, Y., and Nagao, M. (1998) *Japanese Morphological Analysis System JUMAN Manual.* ver. 3.5. Kyoto: Kyoto University. [In Japanese]

NACSIS, (1997) *Introduction to the National Center for Science Information Systems.* Tokyo: NACSIS.

Nomura, M. (1978) "Setsujisei jion goki no seikaku." *Densikeisanki ni yoru kokugo kenkyuu*, IX, p. 102–138.