Wide-Range Restructuring of Intermediate Representations in Machine Translation

Taijiro Tsutsumi
IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
5-19, Sanban-cho, Chiyoda-ku
Tokyo 102
Japan

This paper describes a wide-range restructuring of intermediate representations in machine translation, which is necessary for bridging stylistic gaps between source and target languages and for generating natural target sentences.

We propose a practical way of designing machine translation systems, based on the transfer method, that deal with wide-range restructuring. The transfer component should be divided into two separate sub-components: the wide-range restructuring sub-component and the basic transfer sub-component. The first sub-component deals specifically with global reorganization of intermediate representations in order to bridge the stylistic gaps between source and target languages, and the second performs local and straightforward processing, including lexical transfer and basic structural transfer.

This approach provides us with an effective basis for improving translation quality by systematically enhancing the transfer rules without sacrificing the clarity and maintainability of the transfer component. It also guarantees that most of the translation process can be based on the augmented Context Free Grammar (CFG) formalism and the so-called compositionality principle, by which we can both systematically expand and maintain linguistic data and design the simplified process control necessary for an efficient machine translation system.

1 Introduction

Much effort has been devoted to research into, and development of, machine translation since the 1950s (Slocum 1985). However, the quality of the output sentences produced by most machine translation systems is not high enough to have any marked effect on translation productivity.

A machine translation system produces a variety of expressions in the target language, including good, fair, and poor expressions. In this paper, we define these distinctions as follows. "Good" sentences can be easily understood and have high readability because of their naturalness, "fair" sentences can be understood but their readability is low, and "poor" sentences cannot be understood without referring to the source sentences.

To improve the quality of translation, the following major functions should be implemented:

1. selection of equivalents for words;
2. reordering of words; and
3. improvement of sentence styles.

A machine translation system that does not have Function 3 often produces "good" output in the case of translation between languages in the same linguistic group if Functions 1 and 2 are appropriately achieved. However, most output will be "fair" or "poor" in the case of translation between languages in different linguistic groups, because of the stylistic gaps between them. We need to enhance Function 3 as well as Functions 1 and 2 in order to change "fair" or "poor" sentences to "good" or "fair" ones. Note that in this paper, *style* means a preferable grammatical form that successfully conveys a correct meaning.

First, let us consider how to select target language equivalents for words in the source language. An appropriate part of speech for a word is first determined by the grammatical constraints provided by the analysis grammar. Nouns and the verb in a simple sentence can then be appropriately translated according to the combinative constraints between the case frame of the verb and the semantic markers of the nouns. This is a well-known mechanism for practical semantic processing in machine translation. However, this way of selecting equivalents has the limitation that we cannot classify verbs and nouns in sufficient detail, because of ambiguities in the definitions and usage of words. Many researchers in machine translation claim that a much more powerful semantic processing mechanism with a knowledge base is required for this purpose. This is a long-range research project in the field.

Second, there seem to be few critical problems in reordering words if structural transfer is appropriately carried out. In a simple sentence, for example, we can usually reorder words correctly on the basis of the case frame of the verb.

Thus, the improvement of sentence styles is one of the crucial functions required for further enhancement of the present translation quality. Even when the output is "fair" in quality, it has to be read carefully, because the readability is often low. We expect an improvement in sentence styles to result in an improvement in translation quality from "poor" or "fair" to "good." Improving sentence styles seems to be easier than selecting better equivalents for words, because there are syntactic clues to help us make the styles more natural.

This paper focuses on an approach to the improvement of sentence styles by wide-range restructuring of intermediate representations. *Wide-range restructuring* in this paper means the global restructuring of intermediate representations, usually including the replacement of some class words (i.e., noun, adjective, verb, and adverb). Some papers have mentioned limited restructuring of intermediate representations (Bennett and Slocum 1985; Vauquois and Boitet 1985; Isabelle and Bourbeau 1985; Nagao et al. 1985; McCord 1985; Nomura et al. 1986). For example, LMT (McCord 1985) has a restructuring function after the transfer phase, to form a bridge between the basic styles of English and German. The Mu system (Nagao et al. 1985) has two specific restructuring functions, before and after the transfer phase, mainly to handle exceptional cases.

However, few machine translation systems have so far had a comprehensive component for wide-range restructuring (Slocum 1985), mainly because many systems are presently designed to produce output that is at best "fair," and little effort has been devoted to obtaining "good" output or natural sentences. As a matter of fact, wide-range restructuring functions are usually scattered over the analysis, transfer, and generation phases in an ad hoc way. Because of complicated implementations, the systems come to have low maintainability and efficiency.

Few papers so far have systematically discussed the crucial stylistic gaps between languages, the importance of wide-range restructuring of intermediate representations to bridge these gaps, and effective mechanisms for restructuring. A restructuring mechanism is necessary even when a machine translation system is based on the semantic-level transfer or pivot method (Carbonell et al. 1981).

In this paper, we first discuss stylistic gaps between languages, and the importance of dealing with them effectively in order to generate natural target sentences. We then propose a restructuring mechanism that successfully bridges the stylistic gaps and preserves a high maintainability for the transfer phase of a machine translation system. Last, we discuss the implementation of the wide-range restructuring function.

2 Stylistic Gaps between Languages

In discussions on the stylistic gaps between English and Japanese, it is often said that English is a HAVE-type or DO-type language, whereas Japanese is a BE-type or BECOME-type language. This contrast corresponds to the differences in the ways people recognize things and express their ideas about them (Nitta 1986), and hence is considered as a difference of viewpoint. Idioms and metaphors are heavily dependent upon each society and culture, and we sometimes have to reinterpret them to give an appropriate translation. Each language also has its own specific function word constructions, which are used to express specific meanings. These specific constructions cannot be directly translated into other languages.

We categorize the major stylistic gaps as follows: (1) stylistic gaps in viewpoint, (2) stylistic gaps in idioms and metaphors, (3) stylistic gaps in specific constructions using function words, and (4) others. We will discuss these gaps in more detail in the case of English-to-Japanese translation by referring to examples extracted from the literature (Bekku 1979; Anzai 1983) and slightly modified for our purpose. In each example, the first sentence is the original English and the second one is an equivalent or a rough equivalent in Japanese-like English, which will help us recognize the stylistic gaps, though some of the rewritten English is not strictly acceptable.

2.1 Stylistic Gaps in Viewpoint

The following are some examples of stylistic gaps due to differences of viewpoint.

1. Inanimate noun + transitive verb

i) Inanimate noun + *have*

(1-a) *The room has* two tables.
(1-b) Two tables are in the room.

ii) Others

(2-a) *This chapter contains* the explanation.
(2-b) The explanation is contained in this chapter.

The meaning of (1-a) is that *the room* contains *two tables*. However, an inanimate subject for the verb *have* is not allowed in Japanese. Japanese usually expresses the same fact as in (1-b), without using an inanimate subject. (2-a) and (2-b) show a case in which the voice is changed to avoid an inanimate subject in a Japanese sentence.

2. Gerund (intransitive verb) + *of* + noun + transitive verb

(3-a) *The humming of insects reminded* me of autumn.
(3-b) Because insects were humming, it seemed to me it was autumn.

This case is similar to example (1-a) in that the subject of (3-a) is inanimate. An event or action, which may be the subject of an English sentence, is usually treated as a cause or reason in a Japanese sentence.

3. Inanimate noun + *allow* + noun + to-infinitive

(4-a) *The support allows you to* write IPL procedures.
(4-b) You can write IPL procedures by using the support.

In this case, we can consider the subject as a tool or method, as explicitly rewritten in (4-b).

4. *Have* + adjective + noun (two-place predicate)

(5-a) The routine *has a relatively low usage rate.*
(5-b) The usage rate of the routine is relatively low.

The adjective *low* in the noun phrase *low usage rate* in (5-a) is removed and used as predicative form in (5-b). Japanese often prefers predicative expressions like this.

5. Adjective + verbal noun + *of* + noun

(6-a) He is a *good speaker of English.*
(6-b) He speaks English well.

The noun phrase *a good speaker* is rewritten to form a predicative phrase.

6. Special verb (*do, make, perform*, etc.) + adjective + verbal noun + *of* + noun

(7-a) The DOS/VSE SCP is designed to *make efficient use of a hardware system.*
(7-b) The DOS/VSE SCP is designed to use a hardware system efficiently.

This is also a case in which Japanese prefers a predicative phrase.

7. Special determiner (*no, few, little*, etc.) + noun

(8-a) I have *no French books.*
(8-b) I do not have any French books.

This is a case in which a special determiner should be removed from the noun phrase and rewritten as an adverb.

If we translate these English sentences literally into Japanese, we will have low readability for the Japanese sentences.

2.2 STYLISTIC GAPS IN IDIOMS AND METAPHORS

Idioms and metaphors should be distinguished from other phrases in a text, because they have implicit and fixed meanings. The following are some examples:

(9-a) A car drinks gasoline.
(9-b) A car requires a lot of gasoline.

(10-a) Cigarettes are time bombs.
(10-b) Cigarettes gradually harm us.

(11-a) He burned his bridges.
(11-b) He destroyed his alternative options.

In case (9-a), we will face difficulty in semantic processing if we try to translate the sentence directly. The reason is that the verb *drink* usually requires an animate subject whereas the noun *car* is, in most cases, classified as an inanimate thing.

We can translate example (10-a) literally if we want to preserve the humor conveyed by the original sentence. However, this is not often possible because of cultural differences. Example (11-a) is a typical case of something that we cannot translate literally into Japanese.

2.3 STYLISTIC GAPS IN SPECIAL FUNCTION WORD CONSTRUCTIONS

The following are some examples of stylistic gaps due to special English constructions using function words.

(12-a) *It* is required *that* you specify the assignment.
(12-b) That you specify the assignment is required.

(13-a) The system operation is *so* impaired *that* the IPL procedure has to be repeated.
(13-b1) Because the system operation is impaired very much, the IPL procedure has to be repeated.
(13-b2) The system operation is impaired to the extent that the IPL procedure has to be repeated.

(14-a) The box is *too* heavy for a child *to* carry.
(14-b1) Because the box is very heavy, a child cannot carry it.
(14-b2) The box is very heavy to the extent that a child cannot carry it.

There is no direct way to translate the examples given above, because the grammatical functions conveyed by the special constructions using function words are often expressed in a very different way in a target language.

2.4 OTHERS

In addition to the stylistic gaps described above, we often see other stylistic gaps based on the meaning of a word. For example, the verb *bridge* in English should be translated by "*hashi* (a noun meaning a bridge) *wo* (a case particle meaning an object) *kakeru* (a verb meaning 'install')" in Japanese. One English verb corresponds to a noun, a case particle, and a verb in this case. A set of consecutive words may have a fixed meaning: for example, *a number of* can be considered as *many* in most cases. Differences in tense, aspect, and modality are also related to stylist gaps between languages, although we do not discuss these in detail here.

So far, we have discussed four types of stylistic gaps. Generally, there are larger stylistic gaps between languages belonging to different groups than between languages in the same group. It is clear that if we can deal adequately with these stylistic gaps, we can further improve translation quality.

3 How to Deal with Stylistic Gaps

This section discusses a framework for dealing with the stylistic gaps we noted in the previous section.

3.1 THE COMPOSITIONALITY PRINCIPLE IN MACHINE TRANSLATION

Most machine translation systems (Slocum 1985) aiming at practical use employ the transfer method, which divides the whole process into three phases: analysis of the source language, transfer between intermediate representations, and generation of the target language. The basic concept underlying current machine translation technology is the *compositionality principle* (Nagao 1986). The original idea of the principle is that the meaning of a sentence can be assembled from the meaning of each of its constituents and, moreover, that the assembling process can be implemented by assembling the forms or syntax that convey the meanings. Montague grammar is one of the theoretical bases of the principle, and some work applying Montague grammar to machine translation has been reported (Landsbergen 1982; Nishida and Doshita 1983). If the compositionality principle is applied to machine translation, we expect that it will be possible to translate a whole sentence by translating each word individually and then appropriately composing all the translated words.

For example, let us consider the following two sentences, which have the same meaning.

S: I drink water. (English)
JS: *watashi ha mizu wo nomu.* (Japanese)
(I) (water) (drink)

Let us assume that the above sentences have the syntactic structures shown in Figure 1 (a) and (b), based on the grammars shown in (c) and (d), respectively. Note that the parentheses in the right-hand side of the last rule in (d) denote a condition that must be met in applying the rule. These structures are also considered to be the intermediate structures (i.e. the source structure and target structure) in the transfer phase of machine translation.

The ideal machine translation based on the compositionality principle ensures that structure (a) is successfully transferred to structure (b) by applying the transfer rules, as shown in Figure 2.

In Figure 2, the transfer rules are symbolized for convenience. The left-hand sides of the rules consist of matching patterns that correspond to the grammar of the English

```
           S                              JS
     +-----+-----+          +-------------+-----------+
    NP1    |    NP2       JNPPART1      JNPPART2      |
     |     |     |       +---+---+     +---+---+      |
   PRON  VERB  NOUN   JNOUN1 JPART1 JNOUN2 JPART2   JVERB
     |     |     |       |      |      |      |       |
     I   drink water  watashi   ha   mizu    wo     nomu

(a) The English structure     (b) The Japanese structure

   PRON <- I                  JNOUN    <- watashi
   VERB <- drink              JPART    <- ha
   NOUN <- water              JNOUN    <- mizu
   NP   <- PRON               JPART    <- wo
   NP   <- NOUN               JVERB    <- nomu
   S    <- NP VERB NP         JNPPART  <- JNOUN JPART
                              JS <- JNPPART(JPART 'ha')
                                    JNPPART(JPART 'wo') JVERB

(c) The English grammar       (d) The Japanese grammar
```

Figure 1. Examples of English and Japanese Structures and Grammars.

```
        PRON <- I     ==> JNOUN    <- watashi
        VERB <- drink ==> JVERB    <- nomu
step 1  NOUN <- water ==> JNOUN    <- mizu
        NP   <- PRON  ==> JNPPART  <- JNOUN JPART
step 2  NP   <- NOUN  ==> JNPPART  <- JNOUN JPART
step 3  S    <- NP1 VERB NP2
     ==> JS <- JNPPART(JPART 'ha') JNPPART(JPART 'wo') JVERB
                <NP1>               <NP2>              <VERB>
```

Figure 2. The Transfer Rules.

sentence, and the right-hand sides consist of target patterns that correspond to the grammar of the target Japanese sentence.

The steps of the transfer process using these transfer rules are shown in Figure 3. This transfer process is done entirely in a bottom-up and left-to-right manner by using the transfer rules, and is based on the compositionality principle. The process is simple, easy to control, and easy to implement efficiently.

Let us consider the stylistic gaps mentioned in the previous section. To bridge such gaps, we need to replace some words with new words and perform restructuring widely. It is important to recognize that the words involved in the replacement are class words (i.e. noun, adjective, verb, and adverb) rather than function words (i.e. preposition, auxiliary verb, conjunction, relative pronoun, particle, etc.), as shown in the examples of types 2.1 and 2.2. For example, in cases (6-a) and (6-b), *good* is replaced by *well* and *speaker* is replaced by *speaks.* In cases (10-a) and (10-b), *are time bombs* is replaced by *gradually harm us.* On the other hand, most words involved in the replacement are function words in the examples of type 2.3. For example, in cases (13-a) and (13-b1), *so* and *that* are replaced by *because* and *very much.*

The above-mentioned framework based on the compositionality principle cannot provide appropriate treatment for stylistic gaps of types 2.1 and 2.2, because wide-range structure handling, as well as the replacement of some class words, is necessary instead of the local and bottom-up structure handling that includes some treatment of function words. For type 2.3, the above framework does not suit

```
          S                               S
    +-----+-----+                +-------+-------+
   NP1    |    NP2              NP1      |      NP2
    |     |     |                |       |       |
  PRON  VERB  NOUN            JNOUN1   JVERB   JNOUN2
    |     |     |                |       |       |
    I   drink water           watashi   nomu    mizu

(a) The original structure    (b) The structure at step 1

                 S                                 JS
    +------------+----------+        +-------------+-----------+
  JNPPART1       |      JNPPART2   JNPPART1     JNPPART2       |
  +---+---+      |     +---+---+  +---+---+    +---+---+       |
JNOUN1 JPART1  JVERB JNOUN2 JPART2 JNOUN1 JPART1 JNOUN2 JPART2 JVERB
  |      |       |     |      |     |      |      |      |      |
watashi        nomu  mizu         watashi  ha    mizu    wo    nomu

(c) The structure at step 2   (d) The structure at step 3
```

Figure 3. The Transfer Steps.

the treatment of the gaps if the transfer is done at the analysis-tree level and some function words exist in the source structure for the transfer. It is not difficult to handle gaps of type 2.4 except for those caused by tense, aspect, and modality, because they can be bridged only by local treatment of constituents instead of wide-range restructuring. For example, if a system finds the consecutive words *a number of* in a sentence, the system can exceptionally treat it as one word meaning *many* in the previous framework. It is normally translated by replacing it with a target-language equivalent.

3.2 TWO-STEP TRANSFER METHOD

To deal with stylistic gaps effectively in a system based on the transfer method, we propose the incorporation of a specific sub-component for wide-range restructuring of the intermediate structures in the transfer component, as shown in Figure 4. This gives an example of a system configuration for English-to-Japanese machine translation. The *basic transfer* consists of lexical transfer and reordering of words.

The wide-range restructuring should be done after analysis of the input sentence and before the basic transfer. We take advantage of syntactic clues given in the intermediate representation for effective restructuring after analysis of the input. The wide-range restructuring, which changes the global structure as well as some class words of the sentence, should be performed not after but before the basic transfer, for the following reasons.

1. The restructuring makes the basic transfer easier and it also reduces the transfer rules, because it often contributes to standardization or limitation of English sentence styles, as discussed in more detail in Section 3.3
2. The restructuring is not affected by transfer errors, which often occur because of the complexity of the transfer process.

By means of this restructuring sub-component, the intermediate representation of the input sentence is transformed or reinterpreted from a source-dependent expression into a target-dependent one. We can define *augmented CFGs* for analysis and generation in this framework. If we design the

```
English                                                 Japanese
   |                                                        A
   |                                                        |
┌--------┐                                            ┌----------┐
|Analysis|                                            |Generation|
└--------┘                                            └----------┘
   |                                                        |
   |             ┌------------ Transfer -----------┐        |
   V             |                                 |        |
Intermediate  ---> wide-range     --->  basic   ---> Intermediate
representation | restructuring          transfer   | representation
                 |                                 |
                 └---------------------------------┘
```

Figure 4. A Machine Translation System Incorporating a Wide-Range Restructuring Sub-Component.

rule appropriately and control for the wide-range restructuring sub-component, the output structures of both the sub-component and the basic transfer sub-component can be defined by using augmented CFGs that deal with conditions for rule applications. In other words, the basic transfer sub-component can specialize in transfer from one augmented CFG system to another, as illustrated in Figure 2.

Because wide-range restructuring, which does not suit the compositionality principle, can be performed entirely in the restructuring sub-component, and because the basic transfer can be simplified and specialized in local and bottom-up treatments of structures based on the augmented CFG formalism, as mentioned above, all the processes of machine translation except wide-range restructuring can be based on the augmented CFG formalism or the compositionality principle. This approach makes the whole system simple, easy to control, and efficient.

If a machine translation system uses analysis-tree structures as intermediate structures (Lehmann et al. 1981; Nitta et al. 1982), wide-range restructuring can be introduced appropriately at the surface level. If the system performs deep analysis of the input sentence and creates a semantic representation such as a frame-like structure or a semantic network as an intermediate representation, wide-range restructuring may be required at the deep level. This is true whenever we handle stylistic gaps of types 2.1 and 2.2. However, gaps of type 2.3 can be handled by analysis, and no wide-range restructuring is required from the system that performs deep analysis of the input sentence.

3.3 ADVANTAGE OF THE TWO-STEP TRANSFER METHOD OVER THE SINGLE-STEP TRANSFER METHOD

Let us discuss the advantage of this approach over the conventional single-step transfer method from the standpoint of maintainability.

Technical documents contain many variants of sentence patterns. The examples in Section 2 are regarded as variants from the viewpoint of English-to-Japanese translation. As a matter of fact, several different English sentences in an English technical document can often be translated by the same Japanese sentence. In other words, a wide variety of expression in English can be reduced to some extent in Japanese, because the most important concern in technical documents is that each sentence should convey technical information correctly. Therefore, we may standardize or control styles of English sentences for the sake of English-to-Japanese translation.

The two-step transfer method including wide-range restructuring is an appropriate way to take advantage of this phenomenon. If we encounter a new variant of a sentence pattern in English, we only have to write an English restructuring rule in the case of the two-step transfer method. On the other hand, a whole transfer rule, which is usually harder to write, is needed in the single-transfer method. If we want to modify a target Japanese sentence that corresponds to some English sentences, we only have to modify the corresponding basic transfer rule, instead of modifying all the transfer rules for these English sentences.

Consequently, it is easier to maintain the transfer rules if the system is based on the two-step transfer method, especially in translating technical documents.

4 Implementation of the Wide-Range Restructuring Function

A prototype English-to-Japanese machine translation system, SHALT (Tsutsumi 1986), is based on the two-step transfer method described in Section 3.2. So far we have developed about 500 wide-range restructuring rules to cope with the stylistic gaps exemplified in Section 2, and we have confirmed the effectiveness of the restructuring through test translation of a few IBM computer manuals.

In this section, we discuss the details of the rules for the wide-range restructuring and their applications in SHALT, as an example. SHALT is implemented in LISP, and the English and Japanese intermediate representations are syntactic-analysis tree structures.

4.1 Wide-Range Restructuring Rules and Their Applications

A wide-range restructuring rule consists of a pair of a matching pattern and a target pattern. If an input English tree structure matches a matching pattern, then a target Japanese-like English tree structure is generated according to specifications in a target pattern. A matching pattern is defined as follows. Note that * allows repetition of specifications.

[(STRUCTURE - (MATCHING-ELEMENT*))*]

where

STRUCTURE: MATCHING-VARIABLE or 0
MATCHING-ELEMENT: MATCHING-VARIABLE or (MATCHING-VARIABLE MATCHING-CONDITION*)
MATCHING-CONDITION: (LISP-FUNCTION-NAME ARGUMENT*)

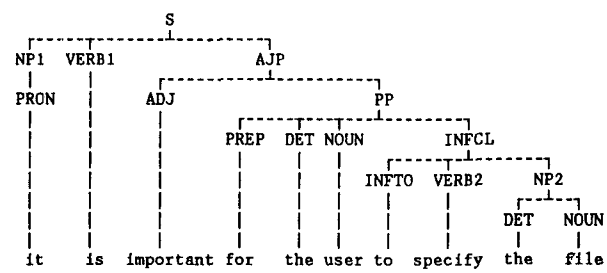
STRUCTURE specifies the tree structure to be checked. If 0 is specified, the whole input structure is treated. If MATCHING-VARIABLE is specified, its value (i.e. part of a structure), which has already been set by MATCHING-ELEMENTs in an earlier matching process, is a target for checking. A sequence of MATCHING-ELEMENTs checks a sequence of daughter tree structures. If specified MATCHING-CONDITIONs match a structure, a specified MATCHING-VARIABLE is set to the structure. If a MATCHING-ELEMENT is a mere MATCHING-VARIABLE, any structure or nil can be set for the variable. MATCHING-CONDITION specifies a LISP function and its arguments. LISP functions check parts of speech, terminal symbols, or other information of a structure. All specifications in a matching pattern form AND conditions, except arguments of LISP functions, which form OR conditions.

A target pattern specifies the required output structure by using MATCHING-VARIABLEs where structures are already set and by adding new structures.

Figure 5 shows an example of a wide-range restructuring rule and its application. Figure 5 (a) shows the output of English analysis, which is the input for wide-range restructuring. Figure 5 (b) shows a wide-range restructuring rule and (c) gives the output of the restructuring.

The left-hand side of the restructuring rule is a matching pattern, and the right-hand side of the rule is a target pattern in Figure 5 (b). Numbers preceded by *, such as *1, *2, and *3, denote MATCHING-VARIABLEs. There are four specifications of (STRUCTURE - (MATCHING-ELEMENT*)) in the matching pattern, such as (0 - (*1 (*2 (T "it")) *3 . . . *7)) and (*6 - (*8 (*9 (P ADJ)) . . . *11)). A MATCHING-CONDITION (T "it") in the matching pattern denotes that the terminal symbol of the tree should be "it." (P ADJ) denotes that the part of speech of the root node should be ADJ (i.e., adjective). T and P are the LISP function names to perform these specific checks.

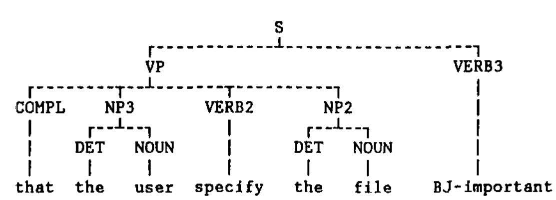
The target pattern in Figure 5 (b) specifies that the values of MATCHING-VARIABLEs *1, *14, *19, *3, *5,



(a) An English Intermediate Representation.

```
[ (  0 - (*1 (*2 (T "it")) *3 (*4 (T "is")) *5 (*6 (P AJP)) *7))
  ( *6 - (*8 (*9 (P ADJ)) (*10 (P PP)) *11))
  (*10 - (*12 (*13 (P PREP)(T "for")) *14 (*15 (P INFCL)) *16))
  (*15 - (*17 (*18 (P INFTO)) *19)) ]
==>
[ *1 (VP (COMPL "that")((!SN NP) *14) *19) *3 *5
     ((!SN VERB)(!APP-TERM "BJ-" *9)) *7 ]
```

(b) A Wide-Range Restructuring Rule.



(c) Output of the Wide-Range Restructuring.

Figure 5. An Example of the Wide-Range Restructuring Rule and Its Application.

*9, and *7 should be used and that some new tree structure should be added to form the output tree structure. We can directly specify the output tree structure by using parentheses "("and")," and node names, such as VP and COMPL. LISP functions can also be specified to generate new elements. LISP function names are distinguished by a preceding "!," as in !SN. (!SN NP) results in a new node name, NPXXX (XXX is a unique number in a LISP environment) for identification. (!APP-TERM "BJ-" *9) denotes that a new symbol concatenating "BJ-" and the value of *9 should be created. "BJ-important" in Figure 5 (c) means that "important" in this case should be treated as a predicative adjective.

These specifications for the wide-range restructuring rules were found to be user-friendly in our experience of developing a practically sized rule set.

The wide-range restructuring rules are categorized into about 20 groups according to their functions. This categorization is more detailed than the one described in Section 2. These groups and rules in each group are arranged sequentially so that more specific and local rules can come earlier when they are checked. Checking and application of these rules are done along an input tree in top-down and left-to-right manner. If one of the rules in a group has successfully been applied to an input tree, then the rest of the rules in the same group are no longer checked, and the process goes to the next rule group. After all the groups have been checked and applied to the input tree, the process goes to the next daughter tree of the input tree. When the process goes to a new tree, the system checks the category name of its root node and thus avoids useless checking of the tree. Because all the rules are used in a fixed sequence, they can be easily maintained.

Figure 6 shows the output of the basic transfer that follows the process of the wide-range restructuring . As we compare the output of the wide-range restructuring, which is the input of the basic transfer, in Figure 5 (c) with the tree structure in Figure 6, we can see how the basic transfer is done. For example, sub-trees COMPL, NP3, VERB2, NP2, VP, and VERB3 in Figure 5 (c) are transferred to sub-trees JANOUN, JNPPART3, JVP2, JNPPART2, JNC, and JVP3 in Figure 6, respectively. This example shows that the basic transfer is more straightforward and localized than the wide-range restructuring.

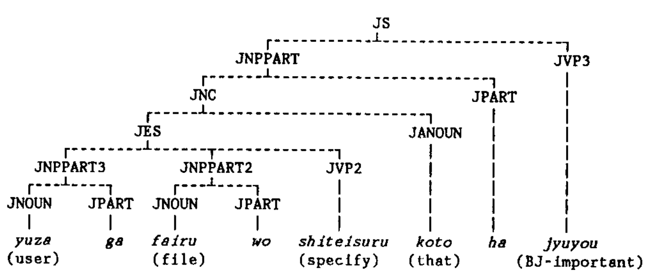


Figure 6. Output of the Basic Transfer.

4.2 Further Discussions on Implementation

Let us discuss example (5-a) in Section 2. The wide-range restructuring rule is as follows. If the main verb is *have* and the head noun (*rate*) of the object is a two-place predicate and there is an adjective (*low*) that modifies the head noun, then restructure it as shown in (5-b). If the head noun of the object is classified as "ATTRIBUTE," the restructuring is obligatory. Otherwise, translation without this restructuring is not very good, but acceptable.

The restructuring rule for case (1-a) in Section 2 is slightly complicated because we need richer information, such as (*room* contains *table*), so as to restructure (1-a) into the form 'NP1 *be in* NP2.' If the input is *The table has four legs,* it will be restructured differently into the form *Four legs exist for the table* because of the above constraint. This is not standard English, but it is very similar in form to Japanese.

We have not yet implemented a way of using semantic information, such as (*room* contains *table*), as constraint, because the desired restructuring can be done on the basis of syntactic restrictions in the field of IBM computer manuals. However, the approach proposed in this paper can be augmented to handle semantic information without any crucial problems, if we prepare a knowledge base.

5 Conclusions

In this paper, we discuss the importance of treating stylistic gaps between languages and methods of doing so. A comprehensive wide-range restructuring that can cope with stylistic gaps is indispensable for improving the quality of translation, especially between languages from different linguistic groups, such as English and Japanese.

We propose a practical way of designing machine translation systems. The transfer component should be divided into two separate sub-components: the wide-range restructuring sub-component and the basic transfer sub-component. Because the first of these deals with global reorganization of the intermediate representations, usually including the replacement of some class words, the second only has to do local, straightforward processing. This approach makes the transfer component much clearer and more maintainable than the conventional single-step transfer method. It also guarantees that, except for the wide-range restructuring sub-component, all of the translation process can be based on the augmented CFG formalism and the compositionality principle. The ease of controlling the process makes the system efficient, which is crucial for the development of a practical machine translation system.

As a future direction, it will be necessary for us to pursue a thorough contrastive study of several languages, in terms of semantics as well as syntax. This will enable us to build more effective rules for restructuring that will further improve the quality of machine translation.

6 Acknowledgments

I am grateful to D. Johnson of the IBM Thomas J. Watson Research Center for his helpful suggestions and to M. McDonald for helping me improve the readability of this paper. I also wish to thank all the people in the natural language processing group of IBM's Tokyo Research Laboratory for constructive discussions.

References

Anzai, T. 1983 *English Viewpoint*. Koodan-sha, Japan.

Bekku, S. 1979 *How to Translate*. Koodan-sha, Japan.

Bennett, W. S. and Slocum, J. 1985 The LRC Machine Translation System. *Computational Linguistics* 11: 111–119.

Carbonell, J.; Cullingford, T.; and Gershman, A. 1981 Steps toward Knowledge-Based Machine Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-3–4.

Isabelle, P. and Bourbeau, L. 1985 TAUM-AVIATION: Its Technical Features and Some Experimental Results. *Computational Linguistics* 11: 18–27.

Landsbergen, J. 1982 Machine Translation Based on Logically Isomorphic Montague Grammars. *Proceedings of COLING* 82: 175–181.

Lehmann, W.; Bennett, W.; Slocum, J; Norcross, E. 1981 The METAL System. *RADC-TR* 80–374.

McCord, M. C. 1985 LMT: A Prolog-Based Machine Translation System. *Proceedings of Colgate University Conference on Machine Translation* 179–182.

Nagao, M. 1986 To What Extent Is Machine Translation Feasible? *New Science Age No. 16.* Iwanami-shoten, Japan.

Nagao, M.; Tsujii, J.; and Nakamura, J. 1985 The Japanese Government Project for Machine Translation. *Computational Linguistics* 11: 91–110.

Nishida, T. and Doshita, S. 1983 Application of Montague Grammar to English–Japanese Machine Translation. *Proceedings of the Conference on Applied Natural Language Processing:* 156–165.

Nitta, Y. 1986 Idiosyncratic Gap: A Tough Problem to Structure-Bound Machine Translation. *Proceedings of COLING* 86: 107–111.

Nitta, Y.; Okajima, A.; Yamano, F.; and Ishihara, K. 1982 A Heuristic Approach to English-into-Japanese Machine Translation. *Proceedings of COLING* 82: 283–288.

Nomura, J.; Naito, S.; Katagiri, Y.; and Shimazu, A. 1986 Translation by Understanding: A Machine Translation System LUTE. *Proceedings of COLING* 86: 621–626.

Slocum, J. 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. *AJCL* 11(1): 1–17.

Tsutsumi, T. 1986 A Prototype English-Japanese Machine Translation System for Translating IBM Computer Manuals. *Proceedings of COLING* 86: 646–648.

Vauquois, B. and Boitet, C. 1985 Automated Translation at Grenoble University. *Computational Linguistics* 11: 28–36.