Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects

Satoshi Shirai [+], Satoru Ikehara [++],
Akio Yokoo [+++], Yoshifumi Ooyama [+]

[+] NTT Communication Science Laboratories
Hikarinooka 1-1, Yokosuka, 239-0847 Japan <{shirai, ooyama}@cslab.kecl.ntt.co.jp>
[++] Faculty of Engineering, Tottori University
Minami 4-101, Tottori, 680-0945 Japan <ikehara@ike.tottori-u.ac.jp>
[+++] ATR Interpreting Telecommunications Research Laboratories,
Hikaridai 2-2, Seika-cho, Soraku-gun, Kyoto-fu, 619-0288 Japan <ayokoo@itl.atr.co.jp>

Abstract

In order to get the translation quality applicable to actual documents, *an automatic rewriting method of source texts* has recently been proposed. Using a *semantic attribute system* of words, this method made it possible not only to define rewriting rules without undesirable side effects but also to rewrite the expressions as well which cannot be rewritten within the source language. This paper improves it to propose *Generalized Rewriting Method of Internal Expressions*. The new method interrupts the translation processes at any place and rewrite the internal expressions according to the conditions defined by rewriting rules.

First, previously defined rewriting rules are expanded and re-classified. Second, it is shown that the rules are generally comprised of 7 fundamental functions. Based on this, how to distribute their functions between the translation processes is shown. Finally, the proposed method is applied to the translation of two kinds of newspaper articles to study the amount of required rules and their actual use in the translation. The results show that the rules needed for newspaper translation can be obtained from the analysis of a relatively small number of articles (about several hundred sentences) for each particular field. It is also shown that the rule was applied at least once, on average, in the translation of two sentences and the overall translation success rate is improved by 20%.

1. Introduction

In the last ten years or so, many machine translation systems, especially for the translation between Japanese and English, have been developed in Japan (Ikehara et al. 1987, 1996, Rimon et al. 1991, Nagao 1996). English to Japanese translation systems have basically achieved the translation quality required for actual use. On the contrary, the quality of Japanese to English translation is not yet satisfactory. One of the major reasons for this asymmetry is considered to be the differences in the linguistic expression frameworks of Japanese and English. The rigid sentence patterns of English are relatively independent of their contents. Such patterns do not exist in Japanese and the structure of a Japanese expression is heavily dependent on its content. Therefore, it is difficult accurately to analyze Japanese expressions using only syntactic knowledge.

The second reason is the differences between the culture and historical circumstances of Japan and the West. Natural language is not only a means of communications but also a means of thinking. The characteristic of a language mirrors that of the culture. Differences in how to think lead to differences between the languages. Japan imported much European culture and many new concepts in the Meiji era. This is considered as one of the reasons of asymmetry of the translations.

There are two ways how to improve the machine translation quality for actual documents. The first is to change the system, tuning the translation functions to the characteristics of the expressions to be translated. The second is to change the expressions in source texts adjusting to the function of the translation system.

Example based Translation (Nagao 1984, Sato 1992, Furuse & Iida 1992) and *Knowledge based Translation* (Takeda 1989, Nirenburg et al. 1992, 1993) can be classified as examples of the first approach. Aiming to overcome the limit of *the principle of Compositional Semantics*, *Example based Translation* finds the expressions in a parallel corpus that correspond to those of the source text and replace them with those of the target language. Unfortunately, it is difficult to prepare an adequate parallel corpus that covers a sufficient number of expressions. On the other hand, *Knowledge based Translation* requires a huge amount of knowledge such as common sense and world knowledge. It is not easy to collect and collate sufficient information.

Controlled Language can be classified as examples of the second approach. It will be separated into *Restricted Writing of the Source Texts* (Nagao 1985, Yoshida 1985) and *Pre-editing of the Source Texts* (Shirai et al. 1993, 1995). *Restricted Writing of the Source Texts* has not been accepted by users in Japan because it is applied to the step of writing original texts so as to restrict the author's way of thinking. On the contrary, *Pre-editing of Source Text* has been accepted widely because the editing method can be acquired by confirming the translation result after pre-editing. However, automatic pre-editing is needed to reduce the large amount of human labor required for pre-editing.

In the pre-editing process, the judgement as to whether an expression should be rewritten or not depends on its context. Disregarding context when rewriting expressions, degrades translation quality to an unacceptable level. Thus, it has been difficult to realize automatic pre-editing without undesirable side effects.

In order to solve this problem, (Shirai et al. 1993, 1995) showed that *the semantic attribute system* (Ikehara et al. 1997) makes it possible to define rewriting rules with no undesirable side effects which allows the automatic rewriting of source texts to be realized. They also showed that rewriting into *a pseudo source language*, which is impossible with human pre-editing, can be undertaken. Based on this study, they developed an *Automatic Rewriting Method* and experimentally confirmed that the improvement of translation quality was about 20%.

The proposed system allows the source texts to be rewritten not before their input but after they are syntactically analyzed. Extending the rewriting rules to be applied at any stage of machine translation may also be helpful in improving the quality of source text analysis. Achieving this extension requires classification of the fundamental functions for rewriting such that the rewriting rule definitions are independent of the translation processes. Accordingly, in this paper, we will propose *An Automatic Rewriting Method for Internal Expression* for Japanese to English machine translation. This paper analyses the roughly 900 rewriting rules that have been collected up to now to show that most rewriting rules can be represented as various combinations of 7 fundamental functions; the rewriting process can be easily constructed independent from the rewriting rules.

2. Outline of the Automatic Rewriting Method

This chapter will summarize the framework of the automatic rewriting method that has been proposed (Shirai et al. 1995).

2.1 Framework of the Automatic Rewriting Method

The most important problem in the automatic rewriting of the source text will be how to avoid undesirable side effects. In manual rewriting, only the expressions which can be rewritten without changing the meaning are rewritten so there is no need to worry about undesirable side effects. Automatic rewriting, however, applies registered rewriting rules without regard to context, so mistakes are possible. The more rewriting rules there are, the more undesirable side effects there will be. The overall translation quality will be decreased. In order to resolve this problem, the method considers the following two points.

1) In addition to individual words and syntactic attributes, *semantic attributes* (Ikehara et al. 1993, 1997) should also be used to define the conditions as to whether the rewriting rule should be applied or not. Here, as for granularity of semantic attributes, note that *the semantic attribute system* needs to be comprised of 2,000 or more attributes.
2) Rewriting rules should be applied after syntactic analysis when sufficient information has been obtained to judge whether the application conditions of rules are satisfied or not.

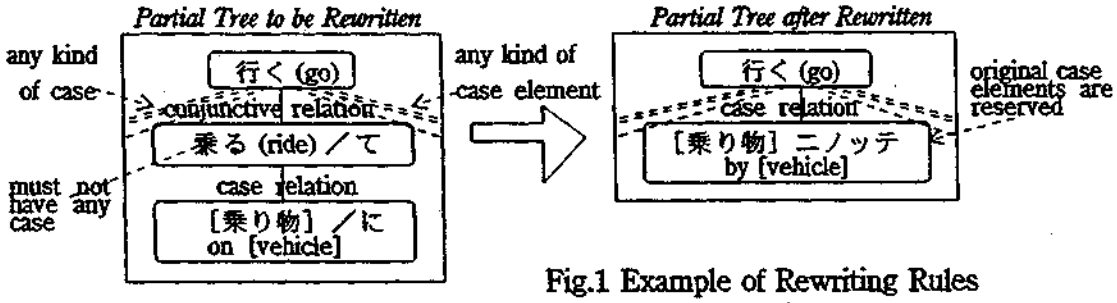


Fig.1 Example of Rewriting Rules

Fig.1 shows an example of rewriting. The rewriting rules are represented by tree structures. The major conditions needed for expressions to be rewritten are as follows;

1) "*Norimono*" ("vehicle": defined by *semantic attribute*) needs to have a case relation with "*Noru*" ("ride": defined by a word face).
2) "*Noru*" needs to have a conjunctive relation with "*Iku*" ("go": defined by a word face).

Thus, this rule is applied to the following sentence as shown in Fig. 2. In this example, the expression "*ni notte iku*" will be rewritten into "by".

	watashi–ha	*densha–ni*	*notte*	*gakko–he*	*iku*
Example Sentence:	私は	電車に	乗って	学校へ	行く。
Direct Translation:	I take a train and go to school.				
Translation after Rewriting:	I go to school by train.				

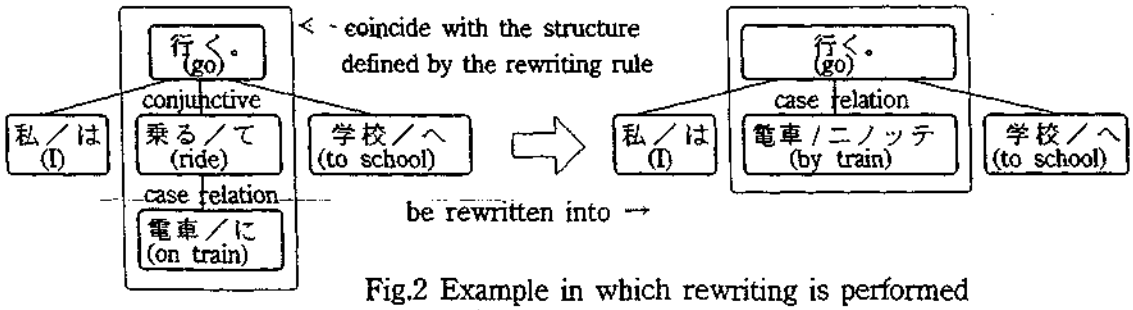


Fig.2 Example in which rewriting is performed

However, if we carefully look at the rule shown in Fig.1, we will find as well the following two conditions.

3) "*Iku*" may have any number of case elements.
4) But, "*Noru*" must have only one case element: "*Norimono*".

Thus, the following sentence in Fig. 3 will not be rewritten althoght it has the same expression "*ni notte*" as the Fig. 2.

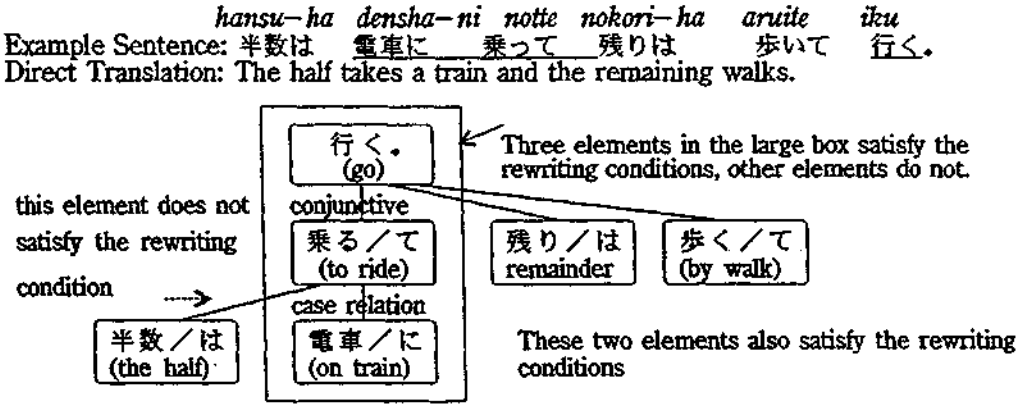


Fig.3 Example in which rewriting is not performed

2.2 Expressions to be Rewritten

The expressions to be rewritten are classified as follows;

1) The expressions which have no equivalent in the target language with the same meaning. These expressions are rewritten into translatable expressions according to the author's intent.
2) The expressions that can not be translated because of cultural differences.
3) The expressions which could not correctly be translated because of the existing translation functions are insufficient.

From these observations, the expressions to be rewritten seem to be virtually identical to those requiring pre–editing, but they are not the same. The proposed method has another useful ability. In pre–editing, if there is no expression that has the same meaning as the original expression, rewriting fails. However, in the *Automatic Rewriting Method*, when there is an expression in the target language which have the same meaning as the original expression, it can be directly be rewritten as that expression.

Under these considerations, (Shirai et al. 1995) have proposed 6 kinds of rewriting Japanese expressions.

3. Classification of the Expressions to be Rewritten

3.1 Extension of Automatic Rewriting Method

Previously mentioned *Automatic Rewriting Method* may create undesirable side effects when two or more rewriting rules are applied to the same sentence so only one rule was applied to one sentence. If more than one rules are needed to be applied to one sentence, it was necessary to synthesize those rules to make a new rule.

In this paper, we functionally classify rewriting rules so as to be able to apply them to the same sentence. We also extend the application target of rewriting rules to the error recovery after source text analysis and to the extraction of other complex expressions.

(1) Overlapped Applications of Rewriting Rules

First let us consider the following sentence. Two rewriting rules will be applied to this sentence as follows.

Reading: *hon–shori–no kekka–wo mochiite henkan oyobi seisei–suru*
Example: 本処理の 結果を 用いて 変換 および 生成する。
Meaning: Using the results of this process, it transforms and generates.

1) "*Wo mochiite*" is extracted as an *Independent Phrase Expression* and is rewritten into the expression corresponding to the English words "based on".
2) The expression "*henkan oyobi seisei–suru*" ("transformation and generate") is a degenerate expression composed of two verbs. The first verb "*henkan–suru*" is degenerated to the noun "*henkan*". This expression is then rewritten as "*henkan–suru oyobi seisei–suru*" ("transform and generate") supplementing the conjugation "*–suru*".

The two phrases "*wo mochiite*" (use) and "*seisei–suru*" (generate) in the original sentence have a dependency relation (verb to a verb). However this relation changes into a case dependency after rewriting the first verb into *the pseudo Japanese expression* corresponding to English "based on". Therefore, the application condition of the rewriting rule 1) including the phrase corresponding to "*seisei–suru*" and the target expression of rewriting rule 1) and 2) partially overlap. However the application of the rewriting rules that have different characteristics will not generate side effects. Separation of rewriting rules into small steps improves the generality of rewriting rules.

(2) Application to the Result of Source Text Analysis

Although, many rules are prepared for use in text analysis such as morphological and syntactic analysis, there are many exceptional expressions which can not be analyzed. In particular, there are many words and expressions whose behavior can not accurately be defined. Most of the failures in analysis are due to such words or expressions. This can be thought of as the major reason for the limitation of current analysis methods.

However, if we collect the expressions that could not be analyzed correctly, we find that there are some common characteristics among these expressions and the analysis errors can be extracted as expression patterns. Considering this, we propose to expand the method to find error patterns in the results of morphological analysis and syntactic analysis, and to replace them by correct patterns.

3.2 Classification of Elemental Functions included in Rewriting Rules

According to the previous discussion, rewriting rules are functionally classified considering the order of application. First, rewriting rules of error correction will be applied just after morphological analysis and syntactic analysis. The following 2 categories of rewriting rules will be followed by this.

(a) Rewriting within the Source Language

The expressions rendering machine translation impossible are rewritten within the

Japanese language. Three types of expressions have been proposed: *Degenerate Expressions, Removing Redundancy*, and *Syntactic Re-arrangement*. This rewriting outputs translatable Japanese sentences that are not always appropriate as Japanese expressions.

(b) Rewriting into Pseudo Source Language

The expressions which can not be rewritten into a different Japanese expression are rewritten into *pseudo Japanese expressions*. Here *the pseudo Japanese expressions* are represented by symbols, the meanings of which correspond to the target language. Three types of expressions have been proposed: *Independent Phrase Expressions, Modal and Tense Expressions* and *Degenerate Conjunctive Expressions*. Even in cases where rewriting within the Japanese text is possible, if the expression after rewriting results in ambiguity, rewriting into *the pseudo source language* should be undertaken using an awareness of the corresponding English.

Here, let's consider the relation of these rewritings. The former can be thought as the rewriting to help the analysis of source expressions. The latter can be considered as' rewriting to help the transfer into the target expression. Of the latter type, rewriting of "*Independent Phrase Expression*" aims at freezing the expressions which have rigid meaning. On the other hand, rewriting of "*Modal and Tense Expression*" and "*Degenerated Conjunctive Expression*" can be considered as the extraction of expressions which can not be translated based on *the principle of compositional Semantics*.

Consequently, we classify the rewriting rules into 4 groups and 14 categories. Details are shown below.

(1) Rewriting as Post Processing for Analysis

1) Correction of Morphological Analysis

Long strings of "*kana*" characters and composed words with suffixes often cause errors in morphological analysis. The rewriting of these expressions is shown in the following two examples. The second example shows one of the errors that appear when suffix processing is strengthen for the analysis of complex compound word. The patterns of these errors are registered in the rewriting rule dictionary and corrected by using them.

Reading: *shi* *ta* *imo* *no* *da*
< Error > ～し (verb)/ た (auxiliary verb)/ いも (noun)/ の (particle)/ だ (auxiliary verb)
Meaning: do (past) potato of (affirmation)

Reading: *shi* *tai* *mono* *da*
< Correct > ～し (verb)/ たい (auxiliary verb)/ もの (noun)/ だ (auxiliary verb)
meaning: do want thing (affirmation)

Reading: *gen* *daiyou* *go*
< Error > 現 (prefix)/ 代用 (noun)/ 語 (suffix)
meaning: now substitution word

Reading: *gendai* *yougo*
< Correct > 現代 (noun)/ 用語 (noun)
meaning: present age word

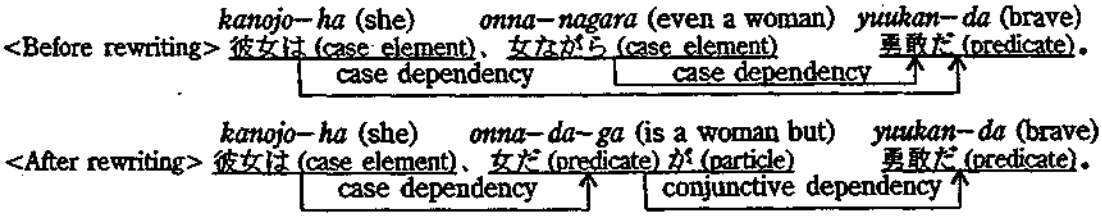
2) Removing Redundancy of Morphological Analysis

In morphological analysis, several interpretation candidates such as " 成る ("*naru*": become)", " 鳴る ("*naru*"; sound)", " 生る ("*naru*"; grow)" are generated for the verb " なる " written by "*kana*" characters. This kind of ambiguity can not be resolved by just syntactic information so that it is left for semantic analysis. However, in the *Rewriting*

of Internal Expression, some of these ambiguities can be resolved referring the meaning of adjacent words which is defined by *the semantic attributes*. This reduces the work of syntactic and semantic analysis. Experiments up to now have shown that the interpretation ambiguities caused by morphological analysis can be reduced from an average of 2.15 to 1.15 at every word.

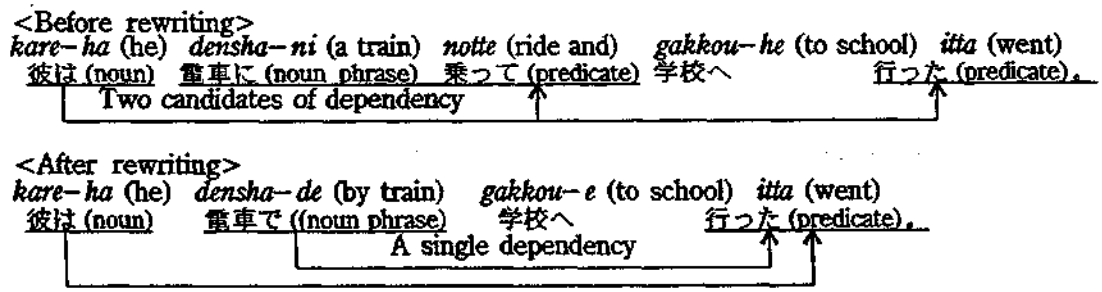
3) Correction of Dependency Analysis

In the following example, the noun phrase, "Noun+ ながら (*nagara*)" is literally interpreted as a noun phrase in morphological analysis. But it plays the role of a predicate in this sentence so dependency analysis can not be correctly performed. For such cases, the results of dependency analysis, as well as the syntactic interpretation of the phrase, are changed.



4) Removing Redundancy of Dependency Analysis

As shown in the following example, rewriting is performed so as to reduce the number of dependent candidates.



(2) Rewriting within Source Language

5) Expansion of Degenerated Expression

Deleted word inflection and deleted expression in the parallel noun phrase are restored to their original expressions.

Reading:	*henkan*	*oyobi*	*seisei–suru*
<Before rewriting>	変換 (noun)//	および //	生成 /(noun) する (verb)
Meanings:	transformation	and	generate
Reading:	*henkan–shi*	*soshite*	*seisei–suru*
<After rewriting>	変換し (verb)//	そして //	生成する (verb)
Meanings:	transform	and	generate

6) Removing Redundant Expression

The expressions whose nuances are not translatable into English are rewritten into simpler expressions. In this example, a complex sentence is rewritten into a simple sentence with a parallel noun phrase.

Reading: *otoko−mo ire−ba onna−mo iru*
<Before rewriting> 男 / も // いれ / ば // 女 / も // いる
Meaning: There are men and also there are women.

Reading: *otoko−mo onna−mo iru*
<After rewriting> 男 / も // 女 / も // いる
Meaning: There are both men and women.

7) Syntactic Rearrangement

The sentence structures of special Japanese expressions that can not be directly translated are rewritten. The example sentence is composed of adverbial phrases and a verb. It has neither of a subject nor a object so that it is necessary to rewrite it into a sentence with a subject.

Reading: *ni−kishu awase−te tsuki hyaku−dai seisan−suru*
<Before rewriting> 二 / 機種 // 合わせ / て // 月 // 百 / 台 / 生産する
Meaning: two type of machine// to make together// monthly/ a hundred/ produce

Reading: *ni−kishu−no gessan−ha hyaku−dai−da*
<After rewriting> 二 / 機種 / の // 月 / 産 / は // 百台 / だ
Meaning: two type of machine// monthly production// is a hundred

8) Standardization of Honorific Expression

The Japanese language has many honorific expressions which cannot be translated into English. These expressions are rewritten into easily translatable Japanese expression.

Reading: *o yomi ni naru*
<Before rewriting> お (Suffix)/ 読み (continents form of noun)/ に (particle)// なる (verb)
Meanings: (Honorific expression of "read")

Reading: *yomu*
<After rewriting> 読む (verb + honorific)
Meanings: (read + honorific)

(3) Rewrite into Pseudo Source Language

9) Independent Phrase Expression Corresponding to Particle

As shown in Fig.1 of chapter 1, independent phrase expressions which correspond to an English particle are rewritten.

10) Adverbial Phrase Expression

If we translate literally clause expressions such as " 言うまでもなく (*yuu−made−mo naru*)" and " 一般的に言えば (*ippan−teki ni ie−ba*)" into English, clauses will be generated. However these expression take the role of adverbial phrase expressions such as "needless to say" and "generally speaking" in English. These are rewritten into *pseudo Japanese phrase expressions*.

11) Adnominal Phrase Expression

Modification clauses such as " 印象に残る (*inshou−ni nokoru*)" and " 喜びにあふれた (*yorokobi−ni afureta*)" are rewritten into word modifiers that mean "impressive" and "joyful" respectively.

12) Phrase Expression

This is similar to the above expressions except that variable words are included. For example, " 驚いたことには (*odoroi−ta koto−niha*)" is rewritten to "to one's surprise" where the word "one's" changes depending on the context.

(4) Rewriting to Freeze Subjective Expressions

13) Conjunctive, Modal and Tense Expression

The expressions composed of conjugation, modality and tense are rewritten into symbols which have the same meaning in English.

Reading:	*suru*	*nara*	*suru*	*noni*
<Before rewriting>	～する (verb)/	なら //	～する (verb)/	のに
Meanings:	do	if	do	would

Reading:	*suru*	*suru*
<After rewriting>	～する (verb + subjunctive)//	～する (verb → would V)
Meanings:	do	do

14) Modal and Tense Expression

Similarly to the above, the expressions composed of modality and tense are rewritten into some symbol which has the same meaning in English.

Reading:	*si*	*ta* *youdat–ta*
<Before rewriting>	～し /	た / ようだっ / た
Meanings:	it seemed to have done	

Reading:	*suru*
<After rewriting>	～する (verb → seemed to have V(past participle))
Meanings:	it seemed to have done

4. Structures of Rewriting Functions and their Components

Based on a translation experiment, we collected 940 rewriting rules. According to the above mentioned standards, they were classified into 14 categories and the fundamental functions needed to perform all rewriting rules were studied. From this study, it was found that any rewriting function will be composed of 7 fundamental functions and that the conventional way of describing application conditions needs to be expanded.

Accordingly, in this chapter we propose a new way of defining the conditions for rewriting rule application and the construction of the rewriting process based on the structure of fundamental rewriting functions.

4.1 Augmented Description of Rewriting Conditions

Generally speaking, a rewriting rule is composed of two parts: the description of conditions for applying the rule and the description of rule execution. The first part can be defined using either word face, syntactic attributes such as part of speech or semantic attributes. Yet, this part has problems shown below.

1) It is necessary to strictly define rewriting rules in order to avoid undesirable side effects. Not only the target expression needs to be defined but also adjacent words or phrases. However, many fluctuations are possible when writing words such as "kanji" description, "*kana*" description and the description of word inflections. For example of fluctuations, a Japanese noun "薔薇 (*bara*, rose)" is sometimes written as "バラ (*bara*)" in *katakana characters* and "ばら (*bara*)" in *hiragana characters*. A Japanese verb "行なう (*okonau*, do)" is also written as "行う (*okonau*)". This reduces the ability to accurately handle actual expressions.

2) In order to define the expressions to be rewritten, two or more conditions need to be written in the condition part of the rewriting rules. However, these were not freely

defined in the conventional method. Exceptional conditions were not also able to be written.

3) The usable attributes name was restricted and there were several forms to wrtite them in the conventional rules.

In order to resolve the first problem, a standard description was created to transform fluctuations into standard forms before applying rewriting rules. For the second problem, the relations of "*and*" and "*or*" between the rule application conditions were made easy to define and exceptional conditions are also made easy to write. For the third problem, we determined the expanded attribute set and a general form for defining rewriting conditions.

4.2 Basic Functions for Execution of Rewriting

According to the analysis of the 940 rewriting rules already prepared, most rules can be performed by the following 7 fundamental functions.

1) *Connection:* Replace two or more words by one word
2) *Alteration:* Alter a word attribute or give a new attribute to a word
3) *Deletion:* Delete a word or a phrase
4) *Supplementation:* Insert a word or a phrase
5) *Separation:* Separate a word into two or more words
6) *Exchange:* Exchange the order of words or phrases
7) *Assessment:* Assess the expression that satisfies the rewriting conditions

Table 1 shows the relation between 7 fundamental functions and 940 rewriting rules.

Table1. Fundamental Execution Functions for Rewriting Rules

(1) Connection (2) Alternation (3) Deletion (4) Supplementation (5) Separation (6) Exchange (7) Assessment (8)Total No. of Rules

Classification	Rewriting Rules	Fundamental Functions							(8)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Rewriting as Post-Processing for Analysis	*Correction of Morphological Analysis*	25	22	2	4	13			60
	Removing Redundancy of M. Analysis		1					15	16
	Correction of Dependency Analysis		19		1				19
	Removing Redundancy of D. Analysis		1					13	14
Rewriting within Source Language	*Expansion of Degenerated Expression*		2	8	8				10
	Removing Redundant Expression	21	17	33				1	55
	Syntactic Rearrangement	10	20		8		12		41
	Standardization of Honorific Expression			1					1
Rewrite into Pseudo Source Language	*Independent Phrase Expression*	159	27	2	12	2			182
	Adverbial Phrase Expression	90	22	2	1				111
	Adnominal Phrase Expression	9	7						15
	Phrase Expression	2							2
Rewriting of Subjective Expressions	*Conjunctive, Modal and Tense Expression*	12	11	1					22
	Modal and Tense Expression	74	39	5	2	3	1		114
Others		–	–	–	–	–	–	–	278
Total		402	188	54	36	18	13	29	940

<cf.> Note that each rule has one or more functions so that the total frequency of functions does not represent the number of rules (8).

4.3 Control of Rewriting

In the previous section, we pointed out that rewriting can be performed by the 7 fundamental functions. Therefore, the rewriting execution program will be composed of 7 independent modules and a control program. According to the requirements defined in the rules, rewriting process can interrupt at any point of the translation processed. Fig.4 shows the correspondence between interrupting points and rewriting rules.

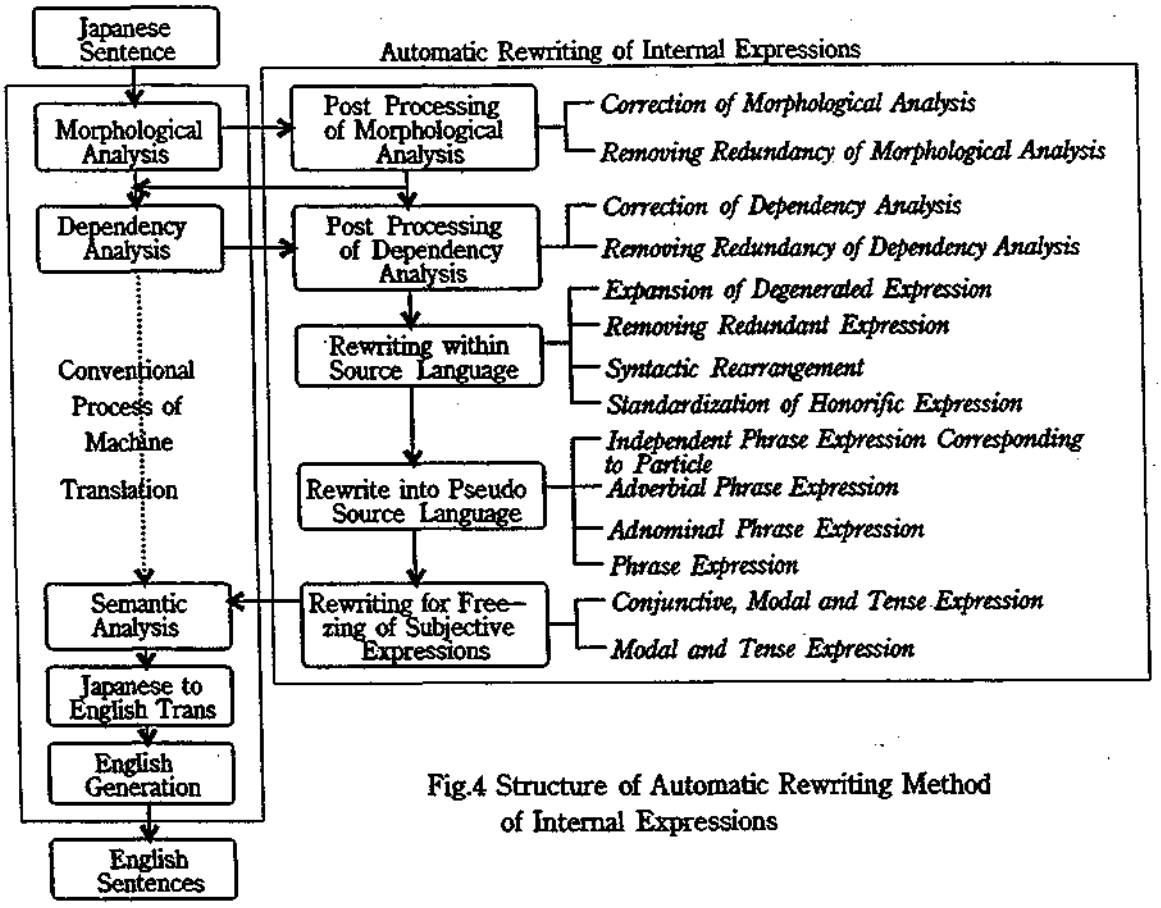


Fig.4 Structure of Automatic Rewriting Method of Internal Expressions

5. Experimental Result of Newspaper Translations

5.1 Application Frequency of Rewriting Rules

(1) Experiment parameters

In order to study the frequency with which rewriting rules were applied, the rewriting procedure was realized in the Japanese to English machine translation system ALT−J/E (Ikehara et al. 1987) and translation experiments were performed for two kinds of newspaper articles.

1) General articles
   874 sentences: *Japan Industrial News* (from August to September, 1994)

2) Market News
   546 sentences: *Telecom BIZ*, Japan Industrial News Co. Ltd. July 1995)

The rewriting rules resulting from several translation runs were collected to determine which rules were used and the frequency of rewriting rule application.

(2) Frequency Rewriting Rule Application

The frequencies with which the rules were used are shown in Table 2. We can find the following from these results.

1) The total number of rules applied to the general articles was 162 and they were applied a total of 463 times. The total number of rules applied to the market news was 43 and their total frequency was 337. The times of application for a sentence is 0.53 for the first case and 0.62 for the last case.

2) Rewriting rules related to *Post-processing of Morphological Analysis* and *Independent Phrase Expressions* are used most frequently. The rules related to syntactic analysis such as *Correction of Dependency Analysis* and *Syntactic Rearrangement* are also used frequently.

3) Compared to the general articles, the translation of the Market News involves fewer rules but they are used for frequently. This means that specific and rigid expressions are frequently used.

In this experiment, it was shown that the success rate of the translation was improved by 20% by the proposed method.

Table 2. The number of rules used and the Frequency of Applications

cf.) No. of R; No. of Rules used, F. of A.: Frequency of Application

Classification	Rewriting Rules	Genera l Article		Mark et News	
		No. of R	F. of A.	No. of R	F. of A.
Rewriting as Post-Processing for Analysis	*Correction of Morphological Analysis*	32	66	9	107
	Removing Redundancy of M. Analysis	6	129	3	39
	Correction of Dependency Analysis	2	23	1	1
	Removing Redundancy of D. Analysis	2	6	0	0
	Sub-total	42	224	13	147
Rewriting within Source Language	*Expansion of Degenerated Expression*	1	1	1	14
	Removing Redundant Expression	5	5	1	14
	Syntactic Rearrangement	13	28	4	22
	Standardization of Honorific Expression	0	0	0	0
	Sub-total	19	34	6	50
Rewrite into Pseudo Source Language	*Independent Phrase Expression*	58	125	16	120
	Adverbial Phrase Expression	25	32	8	20
	Adnominal Phrase Expression	7	9	0	0
	Phrase Expression	0	0	0	0
	Sub-total	90	166	24	140
Rewriting of Subjective Expressions	*Conjunctive, Modal and Tense Expression*	1	1	0	0
	Modal and Tense Expression	10	38	0	0
	Sub-total	11	39	0	0
Total		162	463	43	337

(3) Rewriting Rules used most frequently

The ten most frequently applied rules are shown in Table 3. We found that the rules most frequently used for general articles differ from those used for market news, but most of the rules used in translating the two fields are the same even if their frequency of use differs.

Table 3. Examples for Rewriting Rules most frequently used

	Rewriting Rules	General Article	Market News	Total
1	reduce the ambiguities of continuative expression of verb	84	15	99
2	change the part of speech of " 半面 " at the head of a sentence into an adverb	1	82	83
3	change the dependent of adverbial particle " は "	21	1	22
4	change " ～振りに " into "for the first time"	2	16	12
5	delete the numeral classifier " 台 " from " ～円台 ", " ～％台 "	1	14	15
6	change " ～にかけて " into "toward"	1	12	13
7	freeze " ～に対する " to one word	2	7	9
8	change " やや " into "slightly" depending on the defined conditions	1	6	7
9	change " ～に加え " into "beside"	1	6	7
10	delete the interpretations of a case particle for " ～で "	6	1	7

5.2 Convergence of the Number of Rewriting Rules

In order to study the trends in the number of rewriting rules and the total frequency of usage are shown in Fig.5. From this figure, the following observation can be made.

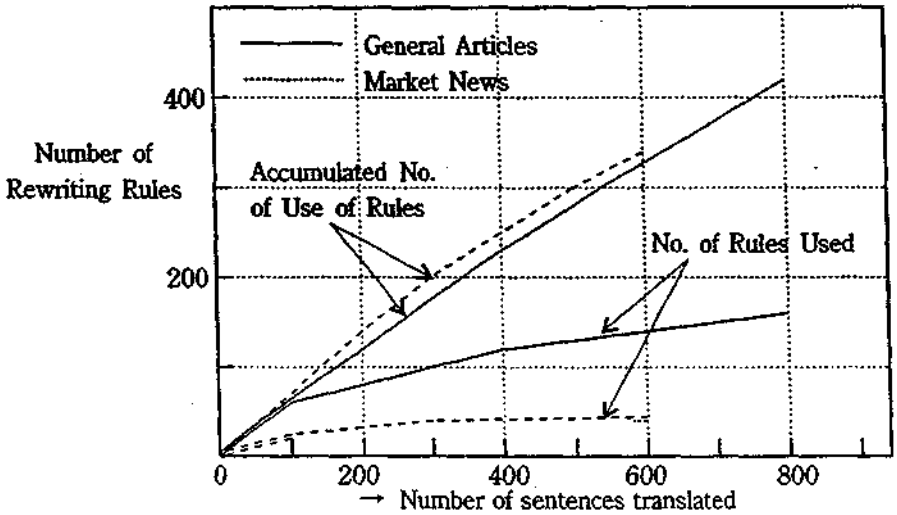


Fig.5 Number of rules used for rewriting

For both fields, the total number of rules used increases in proportion to the number of sentences but the number of rules used tends to saturate. This indicates that the proposed method is very feasible.

6. Concluding Remarks

This paper extended the *Automatic Rewriting Method of Source Text* to create the

Automatic Rewriting of Internal Expressions. In the new method, the rewriting process can interrupted the translation processes at any point to rewrite an internal expressions as necessary to achieve correct translations.

First, rewriting rules were extended and functionally classified for rewriting internal expressions. Based on this classification, it was shown that rewriting rules can be executed by 7 fundamental functions. Finally, this method was applied to the translation of two kinds of news paper articles to determine the number of rules needed and the total number that they were applied.

The results showed that rewriting rules are applied once, on average, at every two sentences and the improvement of translation quality was by 20%. It is very important to notice that the necessary rules can be collected from the analysis of relatively small numbers of sentences in the target field. In some cases, the study of several hundred sentences can find an adequate number of rewriting rules.

In order to develop this method into a fundamental part of a machine translation system, it is necessary to collect an sufficient number of rewriting rules. However, it is not easy for a novice to strictly define the rules needed for rewriting internal expressions so we are now developing a machine aided system for rewriting rule generation.

Reference

Carbonell, J., et al. (1992): JTEC Panel Report on "Machine Translation in Japan", *Coordinated by Loyola College In Maryland*

Furuse, O., Iida, H.(1992) : Cooperation between Transfer and Analysis in Example-Based Framework, *Proceedings of the COLING '92*

Ikehara, S., Miyazaki, M., Shirai, S., Hayashi, Y. (1987): Speaker's Recognition and Multi-level Machine Translation Method based on It, *Trans. of the IPSJ*, Vol.28, No.12, pp.1269-1279

Ikehara, S., Miyazaki, M., Yokoo, A. (1993): Classification of Linguistic Knowledge for Meaning Analysis in Machine Translations, *Trans. of the IPSJ*, Vol.34, No.8, pp.1692-1704

Ikehara, S. (1996): Current Status of Machine Translation, *Science and Technologies of Information*, Vol.46, No.1, pp.26-33

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y.(1997): *Goi -Taikei- a Japanese Lexicon* (Vol. 5), Iwanami Book Store, Tokyo

Nagao, M. (1984): A Framework of a Machine Translation between Japanese and English by Analogy Principle, in A. Elithorn & Banerji, ed. *Artificial and Intelligence*, North Holland.

Nagao, S. (1985): Translation Quality and Controlled Language, *IPSJ*, Vol.26. No.10, pp.1197-1202.

Nagao, M.(1992) : Some Rationales and Methodologies for Example-based Approach, *Proc. of the Workshop on Future Generation Natural Language Processing*, UMIST, Manchester

Nagao, M. Ed. (1994): Subjects for Natural Language Processing

Nirenburg, S., Carbonell, J., Tomita, M. and Goodman, K.(1992) : Machine Translation; A Knowledge -Based Approach, Morgan Kaufmann Publishers

Nierenburg,. S (1993) : A Direction of MT Developments, *Proceedings of the Fourth Machine Translation Summit*, PP.189-193

Nomiyama, H. (1993): Machine Translation based on Generalized Example, *Trans. of the ISPJ*, Vol.34, No.5, pp. 905-912

Rimon, M., McCord, M., Schwall, U. and Martinez, P.(1991) : Advances in Machine Translation Research in IBM, *Proceedings of MT SUMMIT* Ⅲ, pp.11-18

Sato, S, (1992): Example based Machine Translations, *Trans. of the ISPJ*, Vol.33, No.6, pp.673-681

Shirai, S., Ikehara, S., Kawaoka, T.(1993): Effects of Automatic Rewriting of Source Language within a Japanese to English MT System, *Proceedings of the TMI-93*, pp.226-239

Shirai, S, Ikehara, S., Kawaoka, T., Nakamura, Y. (1995): Automatic Rewriting Method of Source Texts for Machine Translations and Its Effects, *Trans. of the ISPJ*, Vol.36, No.1, pp.12-21

Takeda, K., Uramoto, N., Nasukawa, T., Hagino, S., Tsutsumi, T. (1989): Knowledge based Machine Translation System SHALT2, *Computer Software*, Vol.12, No.5, pp.22-32

Yoshida, M. (1995): Standardization of Japanese Language, *SIG of Natural Language Processing, ISPJ* NL51-4