# Machine Translation

# IMPROVEMENT IN CUSTOMIZABILITY USING TRANSLATION TEMPLATES

Satoshi KINOSHITA, Akira KUMANO, Hideki HIRAKAWA

R & D Center, Toshiba Corporation.

## Abstract

This paper outlines customization of a machine translation system using translation templates, which enable users to represent the bilingual knowledge needed for complex translation. To evaluate their effectiveness, we analyzed a bilingual text to estimate the improvement in customizability. The result shows that about 60% of mistranslated sentences can be translated as model translations by combining the proposed framework with the conventional customizing functions.

## 1. Introduction

The ability of natural language processing (NLP) systems is limited due to the knowledge they have, not their framework. This is reflected by recent intensive research on acquisition of linguistic knowledge from a corpus[2][6][9].

Machine translation (MT) systems are no exception. Compared with monolingual knowledge, knowledge needed for translation is difficult to collect. Knowledge acquisition from a bilingual or parallel corpus is considered to be a promising way to reduce the painstaking task[1][10].

Without customization, no general-purpose MT system can output satisfactory translations; therefore it is essential to tune the system by developing a user-oriented lexicon or by registering appropriate target words.

The kind of customization needed depends on how the system is used. If a user translates a document to skim it, he can judge the ability of his MT system by semantic invariance: what percentage of the content of the source text is preserved in its translation. If, on the other hand, he requires translation of publication quality, semantically correct translation is not sufficient; that is, translations should be well-formed so as to conform to a documentational style. To minimize post-editing, more elaborate customizing functions than in the former case are required.

In this paper, we will describe a customizing framework which uses 'translation templates.' This enables users to represent bilingual knowledge for complex translation where a drastic change in linguistic structures occurs to generate natural translations. Then we will discuss the effectiveness of this framework by comparing it with the practically used customizing functions based on the analysis of a bilingual text.

## 2. Machine Translation Using Translation Templates

### 2.1 Aim of Translation Templates

If a user wants publication-quality translations, stylistic well-formedness is as important as semantic invariance. Consider translating the Japanese sentence (1). Although its translation (2), which is the result of our current MT system, is correct, (3) sounds more natural than (2); in (3), the verb phrase "using these detectors" is nominalized to function as a subject to represent the cause of the 'reduce' event. If the user prefers (3) to (2) as a translation of (1), (2) needs to be post-edited.

(1) *korera-no* *kenshutsuki-wo* *tsukau* *kotoniyori*
these detectors-OBJ use by
*kakaku-ga* *teigen-shita*
price-SUBJ reduce-PAST

(2) The price dropped by using these detectors.

(3) Use of these detectors reduced the price.

As the above example illustrates, when source and target languages have a significant difference in their linguistic features, linguistic structures of source sentences are drastically changed to generate natural translations. In this paper, we will call translation which requires complex structural changes 'complex translation.'

This type of knowledge is stored in all MT systems, but insufficiently. Therefore, a framework for customizing complex translation should be incorporated into the system. For this purpose, we have introduced a framework which uses 'translation templates' to represent such knowledge.

Using translation templates, a user can customize his MT system to deal with complex translation without any knowledge on the system's translation process because translation templates are created once the user specifies corresponding expressions in a source sentence and its expected translation.

## 2.2 Translation Templates

A 'translation template' contains at least a pair of patterns, namely 'source' and 'target' patterns, each of which consists of 'constants' and 'variables.' A source pattern (SP) is a template to be compared with a source sentence, while a target pattern (TP) is used to generate a target sentence.

Several reports on machine translation using translation templates suggest that they are useful for translating fixed expressions[4][7][8]. Our translation template is more expressive in the following points:

- More parts of speech can be specified for variables.
- Conditions on translating expressions matched with variables can be specified.

These points will be explained below.

Fig. 1 shows an example of a translation template. '$1' and '$2', which appear in both the source and target patterns, are variables, and the remaining elements are constants. All constants in the source pattern should appear in a source sentence in the same order. Strings which match with variables should satisfy parts of speech designated in the 'source condition.' In this example, the strings should be analyzed as 'np' (noun phrases).

The 'part of speech(POS)' of a template represents a syntactic category of a string matched with a source pattern. Currently, 'sentence' and 'sentence modifier' can be specified.

The 'source condition(SCND)' represents conditions on variables in the 'source pattern.' The grammatical categories of variables currently in use are noun, noun phrase, number, clause and verb phrase. A string matched with a variable should be parsed as the specified category.

The 'target condition(TCND)' represents conditions on variables in the 'target pattern.' Two types are available: 'attribute' and 'relation.' Attributes specify information on one variable. For example, variables for nouns can be specified as having a 'default article' and a 'default number' to be used if there are no explicit clues to determine the article and the number. Similarly, the form of verb phrases in generation can be specified as '*to*-infinitive' or 'gerund.' Relations represent the number agreements between a subject and a verb in the target pattern, for example.

Variables may appear only in the source or target pattern. Variables which appear only in the source pattern are used to represent expressions which have relations with another variable but disappear in the target sentence. Variables which appear only in the target pattern are used to represent a target word which is inflected by the number agreement with the contents of other variables.

```
POS   : s
SP    : $1 を使うことにより  $2 が低減した
          (wo tsukau kotoniyori)   (ga teigen shita)
TP    : use of $1 reduced $2
SCND  : $1.pos=np / $2.pos=np
```

Fig. 1 Template Example

```
POS   : s
SP    : $1 の設定は、 $2 ことにより行なえる
          (no settei wa)   (kotoniyori okonaeru)
TP    : $1 can be set by $2
SCND  : $1.pos=np / $2.pos=vp
TCND  : $2.vpgend=ING
```

(a) Template with a variable for verb phrase

```
POS   : s
SP    : $1 の除去は、 $2 により行われる
          (no jokyo wa)   (niyori okonawareru)
TP    : $1 $3 eliminated by $2
SCND  : $1.pos=np / $2.pos=np
TCND  : $3.tw=be / s_v(1,3)
```

(b) Template with a variable appearing only in a target pattern

Fig. 2 Template Examples

Fig. 2 shows other examples of translation templates. Fig. 2(a) shows a template which has a variable for a verb phrase. This template is created by referring to sentence (4) and its model translation (5)

The target condition specifies that a verb phrase to be matched with the variable '$2' is generated as a gerund.

(4) *jokyoshuuhasuu-no* *settei-wa,*
'frequency to be eliminated'-of setting-TOP
*torimakondensa-de* *C-no* *atai-wo*
trimmer capacitor-INST C-of value-OBJ
*tyousei-suru* *kotoniyori* *okonaeru.*
adjust by can be done

(5) The frequency to be eliminated can be set by adjusting the value of C by a trimmer capacitor.

The introduction of variables which match with verb phrases improves the flexibility of translation templates. Without these variables, we must create restricted source patterns, in which the word order of postpositional phrases like "-de" and "-wo" is fixed.

Fig. 2(b) shows a template which has a variable appearing only in the target pattern. This template is created by referring to sentences (6) and (7) below. The target word (tw) of variable '$3' is specified as 'be' and its surface form is determined according to the 'number' feature of the expression of variable '$1.'

(6) *kyariaseibun-no* *jokyo-wa,*
carrier component-of elimination-TOP
*T-gata roopasufiruta-niyori* *okonawareru.*
T-type low-pass filter-by be done

(7) The carrier component is eliminated by T-type low-pass filters.

## 2.3 Translation Process

Fig. 3 shows a conceptual flow of translation process using translation templates. (The actual implementation is different from the flow.) First, the 'translation template dictionary' is searched for applicable templates. If no applicable template is found, the source sentence is translated using the conventional translation module; if found, strings matched with variables are parsed and translated. Finally, translations of variables are embedded into the target pattern.

This process is implemented in the conventional translation module of our transfer-based MT system[3].

(a) Morphological Analysis

The morphological analyzer first constructs a word lattice for an input sentence by referring to the word dictionaries and the Japanese morphological grammar, and then produces a sequence of words from the lattice until the syntactic analyzer parses it successfully.

Constants in the source pattern of translation templates are stored in the 'template constant dictionary' used in the first phase of morphological analysis to create the word lattice. Fig. 4 shows a simplified example of a word lattice for sentence (1).

Constants of translation templates in a word lattice should be selected if and only if all the constants of a particular template are selected simultaneously to form a valid sequence of words. In Fig. 4, we can obtain two valid word sequences from the word lattice.

The present implementation permits one applicable template for each source sentence. If more than one templates are applicable, the priority for each template is calculated based on the total length of constants and the scope of the source sentence covered by the template, and a word sequence is produced in the order of their priorities.

(b) Syntactic Analysis

When a translation template is applicable, the syntactic analyzer plays two roles. First is to analyze part of the word sequence which should be matched with variables of the template. Words in the word sequence, except for template constants, should be parsed as syntactic categories specified in each

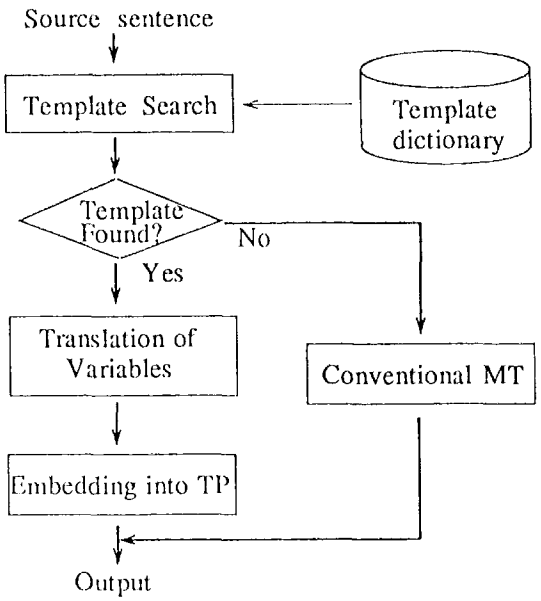


Fig. 3 Translation Process

variable. The second role is to derive a syntactic structure for the sentence.

(c) Transfer and Generation

In the transfer phase, a translation template is transformed into a lexical transfer rule in the conventional form, so that the new matching pattern matches with the structure produced by the syntactic analyzer. The result of applying this rule is a target structure; its direct constituents are given the word order and ready to output as a target sentence.

## 3. Criteria for Using Translation Templates

In principle, all translation can be described by translation templates. That is, users can make a translation template by substituting corresponding expressions in source and target sentences with variables. The question is the appropriateness of templates.

The first criterion is its 'applicability.' In the following cases, translation templates are inappropriate because the source pattern is too specific to be applied to other sentences.

(C1) A source sentence is translated into two target sentences or a compound sentence.

(C2) Two source sentences are translated into one target sentence.

(C3) A source sentence contains a parenthesis or a gapping.

In such cases, the source pattern may contain more constants than that of the ordinary translation templates.

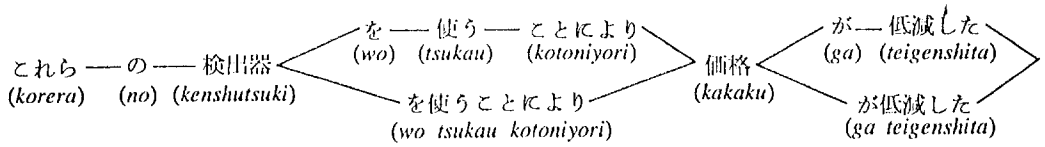


Fig. 4 Example of a Word Lattice

Sentence (8) and its model translation (9) show an example of (C3), where the source sentence contains a gapping. The source pattern created from this sentence will be of low applicability.

(8) *puriampu-wa,* *P1-ni* *8soshi,*
preamplifier-TOP P1-in 8 element
*P2-ni* *24soshi* *bunsan-sareteiru*
P2-in 24 element decentralize-PASSIVE

(9) The preamplifiers are decentralized for 8 elements in P1 and for 24 elements in P2.

Another criterion is the 'contextual independence.' It is often the case in Japanese-to-English translation that a zero-pronoun in a source sentence is resolved from the context and its translation equivalent appears in the target sentence. A translation template created from such translation may generate a contextually inappropriate translation.

Note that these criteria are not absolute; templates which do not meet these criteria should be used if they lead to correct translation of other sentences. A statistical method could be introduced to objectively determine the appropriateness.

## 4. Conventional Customizing Functions

This section briefly describes customizing functions which have been adopted in our MT system[3][5].

• User-defined word dictionary

A user-defined word dictionary (or simply a user dictionary) is the basis for improving the quality of MT output.

• Translation parameters

Translation parameters are introduced to give preference or default interpretation in the translation process. In general, all of the processing are based on the system's linguistic knowledge, which is not open to users. For example, users cannot change the application order of syntactic rules used by the parser. Therefore the system derives the same syntactic tree for a given sentence to generate one particular translation. Translation parameters enable users to partially control the translation process.

One of the parameters used in Japanese-to-English translation treats subjectless sentences, which are common linguistic phenomena in Japanese. With this parameter, users can specify the sentence type of a target sentence (imperative or declarative) and, if necessary, the voice and translation equivalents for the omitted subject (personal pronouns, "it" or a user-defined string). For example, sentence (10) is translated into sentences (11) to (15) according to the specified parameter values.

(10) *sono* *botan-wo* *oshimasu*
the button-OBJ press
(11) Press the button. (imperative)
(12) The button is pressed. (passive)
(13) I press the button. (personal pronouns)
(14) It presses the button. ("it")
(15) # presses the button. ("#" as user-defined string)

• User-defined rules

User-defined rules are used for representing knowledge to determine an appropriate translation equivalent for a source word (or an expression) by referring to its related words. There are three types of user-defined rules available:

(R1) Rules for verbs
(R2) Rules for functional phrases
(R3) Rules for conjunctional phrases

Rule (R1) determines a translation equivalent of a verb based on its case fillers. A translation for a functional phrase is determined based on its preceding noun and the verb phrase it modifies, whereas a translation for a conjunctional phrase is based on its preceding verb phrase and the verb phrase it modifies. Additionally, rules (R2) and (R3) can specify where translation equivalents for functional and conjunctional phrases are generated.

Sentences (16) to (18) below show a customization example using a user-defined rule for a functional phrase. In sentence (17), which is the initial output by our system, the functional phrase "ni doukishite" is translated into a verb phrase. Contrast this with the customized sentence (18), in which the phrase is translated into the prepositional phrase "in synchronism with."

(16) *kono* *kairo-wa* *shingou-ni* *doukishite*
this circuit-TOP signal-with synchronize
*parusu-wo* *hassei-suru*
pulse-OBJ generate

(17) This circuit generates a pulse synchronizing with a signal.

(18) This circuit generates a pulse in synchronism with a signal.

User-defined rules have limitations in that they cannot represent complex structural changes. However, this is intentionally designed to prevent mistranslation possibly caused by adding these structural rules into the system's knowledge. Alternatively, the proposed framework has been introduced to represent knowledge for more complex translation.

## 5. Evaluation of Customizability

### 5.1 Outline of Analysis

To confirm the effectiveness of translation templates, we analyzed a parallel text, namely a service manual on an electronic equipment written in Japanese and its English translation, and estimated the improvement in customizability.

The analysis was done as follows:

(i) Translate the source sentences using the MT system, which is in the default state except that undefined words are registered in the user dictionary.

(ii) Compare the 'sentence structure' of the MT output in (i) and its corresponding sentence in the English manual, and find out sentences for customization.

(iii) Categorize the above sentences according to the type of customization needed to translate them into sentences having the same sentence structures as the model translations.

The 'sentence structure' used for judging the necessity of customization includes the following linguistic features:

- Sentence types:
  declarative | imperative | others
- Clause patterns:
  simple | complex | compound
- Case frames of a main clause
  Two different case frames are treated as the same as long as the difference can be resolved with a user-defined word and/or a user-defined rule for verbs.
- Voice of a main clause:
  active | passive

If all of the above are identical, the MT output and the model translation are considered to have the same sentence structure. Otherwise, the MT system needs customization. For example, sentences (2) and (3) are different in their sentence structures because they have different case frames. Similarly, sentences (20) and (21), which are the MT output of sentence (19) and the model translation respectively, are different in their sentence structures because of their different clause patterns and case frames.

(19) *FMbu-niwa 2mai-no fureemumemori-ga ari*
FM unit-in 2 frame memory-SUBJ exist
*kotonaru 2tsu-no gazou-wo kioku-dekiru*
different 2 image-OBJ can memorize

(20) Two frame memories are in the FM unit and it can memorize two different images.

(21) The FM unit has two frame memories that can store two different images.

Table 1 Result of Comparison

| | |
|---|---|
| Translated as Models | 209(42%) |
| Needs Customization | 283(58%) |
| Total | 492 |

Table 2 Result of Customization

| | |
|---|---|
| Parameters | 21( 7%) |
| User-defined Rules | 20( 7%) |
| Templates | 126(45%) |
| Cannot Customize | 116(41%) |

### 5.2 Analysis Result

We have analyzed 492 sentences excluding titles and figure captions. The average sentence length was 52 Kanji characters.

Table 1 shows the overall result. Out of 492 sentences, 42% have the same sentence structures as the model translations, while the remaining 58% have different sentence structures and require customization of the system. The latter is further divided into four categories according to the type of customization needed to improve the MT output, as shown in Table 2. By the conventional customizing functions, namely, translation parameters and user-defined rules, 14% are customizable. In addition, translation templates can improve 45%, which suggests that 59% will improve in total. This also means that, using all customizing functions, 76% of the given sentences can be translated as in the English manual, while only 51% can be done so using the conventional functions. These figures suggest that a translation template is

useful to deal with complex translation.

Sentences which cannot be customized are divided into four categories:

- Failed application of parameters (20%)
- Inadequate syntax for templates (9%)
- Inappropriate templates (65%)
- Others (6%)

First, a translation parameter does not work when the condition on its application is not customizable. One example is a translation parameter of sentence types for enumerated items. If the system can recognize such a specific form, its translation can be customized. Otherwise the specified parameter is not used.

Second, an extended syntax for translation templates is needed to represent more complex translation. An example is to extend the syntax so that conversion of grammatical categories, such as nominalization of verb phrases, can be specified.

Third, translation templates are not utilized in light of the criteria explained in 3. The statistics of the rejected sentences is as follows.

- Division or concatenation of sentences (57%)
- Resolution of zero-pronouns (24%)
- Parenthesis / gapping (11%)
- Others (8%)

## 5.3 Discussion

- Flexibility of translation templates

A translation template proposed in this paper is more flexible than others due to variables to match with 'verb phrases' and 'clauses.' Basically, a pattern matching approach like the template-based translation has a disadvantage on word order when it is applied to a language that has relatively free word order like Japanese. This problem is partially solved by using these variables because the word order of the constituents of verb phrases and clauses is not fixed.

- Appropriateness of translation templates

The question about the appropriateness of a translation template is also raised in case of a translation example in Example-based Machine Translation (EBMT). It is easy to measure the system performance, but is difficult to evaluate the appropriateness of examples based on their amount and the performance. This issue has been ignored so far.

Our criteria will be the first approach to this issue. Although every translation can be described using translation templates, some criteria to determine its appropriateness should be provided because without them automatic template learning will soon lead to the explosion of the template database.

## 6. Conclusion

In this paper, we have presented a framework for customizing a machine translation system using user-defined translation templates. This enables users to represent bilingual knowledge for complex translation. We have conducted a preliminary analysis to evaluate the effectiveness of the proposed framework based on a bilingual text. The result shows that about 60% of mistranslated sentences can be properly translated by combining the proposed framework with the conventional customizing functions, while only 14% can be achieved using the conventional customizing functions.

One of our current concerns is to extend translation templates and make them more expressive to deal with more complex translation. The proposed framework does not permit variables in a template to be changed into other grammatical categories. Another concern is to improve the user interface for registering translation templates. Through the analysis of source and target sentences, initial values in the interface will be more accurate and need less correction.

## References


[1] Dagan, I., Itai, A. and Schwall, U. : Two Languages Are More Informative Than One, *Proc. of ACL-91*, pp. 130-137, 1991.

[2] Grishman, R. and Sterling, J. : Acquisition of Selectional Patterns, *Proc. of COLING-92*, pp. 658-664, 1992.

[3] Hirakawa, H., Nogami, H. and Amano, S. : EJ/JE Machine Translation System ASTRANSAC--Extensions toward Personalization, *Proc. of MT SUMMIT-III*, pp. 73-80, 1991.

[4] Kaji, H., Kida, Y. and Morimoto, Y. : Learning Translation Templates from Bilingual Text, *Proc. of COLING-92*, pp. 672-678, 1992.

[5] Kumano, A., Kinoshita, S. and Hirakawa, H. : Customization of Machine Translation System with User-defined Rules, *Japan Soc. Artif. Intell. Technical Report*, SIG-SLUD-9301-6, 1993 (in Japanese).

[6] Manning, C. D. : Automatic Acquisition of a Large Subcategorization Dictionary from Corpora, *Proc. of ACL-93*, pp. 235-242, 1993.

[7] Maruyama, H.: Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP, *Proc. of Natural Language Processing Pacific Rim Symposium*, pp. 232-237, 1993.

[8] Uratani, N., Katoh, N. and Aizawa, T. : Extraction of Fixed Patterns from AP Economy News, *Proc. of the 42th Annual Convention IPS*

*Japan*, 6E-4, 1991.
[9] Utsuro, T., Matsumoto, Y. and Nagao, M. : Lexical Knowledge Acquisition from Bilingual Corpora, *Proc. of COLING-92*, pp. 581-587, 1992.
[10] Watanabe, H. : A Method for Extracting Translation Pattern from Translation Examples, *Proc. of TMI-93*, pp. 292-301, 1993.