A Hybrid Approach to Interactive Machine Translation
Integrating Rule-based, Corpus-based, and Example-based Method

YAMABANA Kiyoshi*, KAMEI Shin-ichiro, MURAKI Kazunori,
DOI Shinichi, TAMURA Shinko, SATOH Kenji
Information Technology Research Laboratories
NEC Corporation
Miyazaki 4-1-1, Miyamae-ku, Kawasaki 216, JAPAN
{yamabana, kamei, k-muraki, doi, shinko, satoh}@hum.cl.nec.co.jp

Abstract

With rapid development of the Internet, demand is rising high for a personal tool to support writing foreign language document such as e-mail. However, translation result of an automatic MT system is often not satisfactory for this purpose and requires post-editing. In addition, a purely rule-based system does not necessarily provide a satisfactory result for specific expressions because of lack of corresponding rules, nor purely example-based system for expressions not covered by examples. A hybrid approach is worthwhile to pursuit, where automatic and interactive approaches, as well as rule-oriented and data-oriented approaches are integrated. In this article, we propose a hybrid interactive machine translation method that combines rule-based, corpus-based and example-based approach with an interactive man-machine interface. We show that the previously proposed rule-based model can be naturally integrated with different translation paradigms. The interactive operations, previously introduced and shown to be useful for disambiguation in the rule-based transfer, are shown to be also useful to control covering by and selection of the matching examples, two major decisions in the example-based translation method. We also mention an online learning scheme of translation pairs from the user interaction.

1 Introduction

With the rapid development of the Internet, demand for a supporting tool for reading and writing foreign language document is rising high these days. While conventional automatic machine translation systems are useful for reading support where quick and rough translation does its job, they are not necessarily appropriate for writing support where the main task is to create a short original document such as e-mail. Since the final quality is far more important, such tool is better to offer some interactive means to control the translation result.

With this in mind, an incremental interactive machine-aided translation method was introduced and a realization as an English writing support tool was shown[Muraki *et* al., 1994; Yamabana *et al,* 1997]. In this method, the source sentence is translated incrementally in a bottom-up manner, from a smaller part to a larger structure. In respective steps, the user can interactively control the process through simple operations of translation area correction and translation equivalent selection. A rule-based transfer engine provides translations obeying user's specification and shows them on a selection window. The partial results obtained in this manner are repeatedly combined to a larger expression in the subsequent translation steps, until the whole input is converted into a target language expression.

This method offers an interactive means to combine the word dictionary information with grammar rules to obtain a direct translation of the input sentence. However, rule-based method is not the only and desirable means for translation, especially considering its cost in describing and keeping the consistency of highly specific linguistic phenomena. Although various paradigms of machine translation such as rule-based, statistics-based and example-based method have been advocated these days, there now seems to be a consensus that none of these paradigms are uniformly adequate in all aspects of the translation task.

In this article, we propose a hybrid interactive machine translation method that integrates various translation paradigms with an interactive man-machine interface. In section 2, we review the rule-based interactive translation method on which the proposed method is built. In section 3, the hybrid interactive machine translation method is described, and its basic architecture and the algorithm is presented. In section 4, the cur-

*Current Address: NEC Research Institute, U.S.A.

rent implementation status is described. Section 5 is for discussions, and the final section concludes this article.

2 An Interactive Japanese to English Translation Method

An interactive machine-aided translation method was introduced to support non-natives of English to write English material [Muraki *et a/.*, 1994; Yamabana *et a/.*, 1997]. The target user of the method is those people who have difficulty in writing down English sentences directly, in spite of the fact that s/he has a basic knowledge of English to read and understand it. In this section we show how the method works by an example.

Suppose the user is writing e-mail in English, working on an editor of a mail program. Our tool is running background as a daemon, watching the keyboard input by the user. While the user is typing English characters, the system lets them through to the editor window. The tool awakens when the user toggles on the Japanese input. As soon as the first Japanese character is typed in, the tool detects and fetches it from the input queue of the operating system, opens the main translation window, and puts it there. All the subsequent characters are captured in that window, instead of the editor window. Succeeding translation is performed in this main translation window.

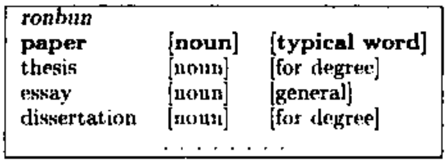
Suppose the input sentence is the one shown in figure 1 (a)[1]. As soon as (a) is entered, dictionary look-up process is started automatically. First the morphological analyzer recognizes word boundaries in the sentence, looks up corresponding entries in the system dictionary, and shows the result on the main window (b). At this time, content words are replaced by one of its translation equivalents assumed most plausible by the system, while functional words are left unchanged.

This representation step, in which English words (content words) and Japanese words (functional words) are mixed, is one of important characteristics of the method. This step separates steps into word translation and later structural. transfer, making translation steps clearer. Since word order and functional words carrying grammatical functions are unchanged, the user can easily recognize the skeleton of the sentence, and clearly grasp the correspondence between the original word and its translation equivalent. This representation also carries all interactive operations of the method on it, and has a double role in interactive operations, showing the information by the system and providing the objects for interactive manipulation.

Translation equivalent alternatives for the cursor position word (focus word) are displayed in an alternatives window, appearing nearby that word. Figure 2 is a snapshot of the alternatives window for *ronbun* (paper). The second line is highlighted to show that it is the current selection. The user can change the selection simply by a cursor movement or a mouse click on this window, then corresponding translation equivalent on the main window changes synchronously. To see the alternatives for another word, the user has only to move the cursor to that word on the main window. In addition, the user can choose an inflection in a similar manner on an inflection selection window, opened by the user's request.

(a) 私 は 彼 に 論文 を 渡し た
watashi -*wa* *kare* -*ni* *ronbun* -*o* *watashi* -*ta*
I TOP he DAT paper OBJ give PAST
(b) I は he に paper を give た
-*wa* -*ni* -*o* -*ta*
(c) I gave him a paper

Figure 1: Translation of a simple sentence



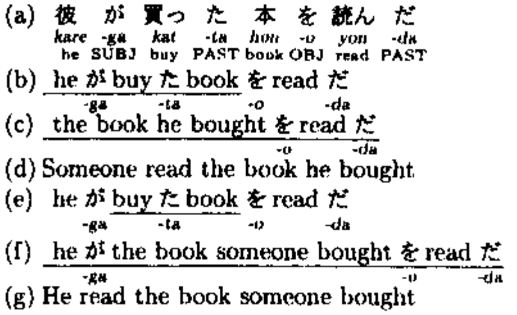
ronbun		
paper	**[noun]**	**[typical word]**
thesis	[noun]	[for degree]
essay	[noun]	[general]
dissertation	[noun]	[for degree]

Figure 2: Alternatives Window for *ronbun*

If the user needs only the result of dictionary lookup, s/he can signal the end of translation at this point. If syntactic transformation is necessary, the user needs to proceed another step. At the same time as the initial prediction of the translation equivalent, the system predicts an appropriate area for syntactic transformation, as shown by an underline in (b). Just like the translation equivalent selection, the area can be freely changed by the user. After the user confirms the selection of translation equivalents and translation area on (b), s/he invokes translation. The system performs syntactic transfer using syntactic information in the dictionary such as verbal case frame and transfer rules encoded in the system, shows the result on the main window, and replaces the original sentence with the result (c). If there are more than one possible translations, they are shown in an alternatives window similar to figure 2, allowing the user to choose among them. When the user triggers the end of translation, the result is sent to the original editor window.

Figure 3 shows translation steps for a sentence with a relative clause. This sentence has a dependency ambiguity, so we also show how to resolve it through the interactive operation. The original sentence (a) contains a relative clause with verb *kau* (buy) with an antecedent *hon* (book). Since Japanese is head-final, the sentence-initial case element *kare-ga* (he-SUBJ) can be the subject of either *kau* (buy) or *yomu* (read), causing syntactic

hereafter, slanted characters represent Japanese words in Japanese characters.



(a) 彼 が 買っ た 本 を 読ん だ
kare -ga kat -ta hon -o yon -da
he SUBJ buy PAST book OBJ read PAST

(b) he が buy た book を read だ
-ga -ta -o -da

(c) the book he bought を read だ
-o -da

(d) Someone read the book he bought

(e) he が buy た book を read だ
-ga -ta -o -da

(f) he が the book someone bought を read だ
-ga -o -da

(g) He read the book someone bought

Figure 3: Relative Clause and Syntactic Ambiguity

ambiguity.

First, let's suppose *kare-ga* is assumed to be the subject of the relative clause. Then the system pauses showing (b), as soon as (a) is input. In (b), the translation area is assumed to be "he-ga buy-ta book". After translation trigger, the system pauses showing (c). Please note that the underlined part in (b) is replaced by its equivalent English expression "the book he bought", and the whole sentence is underlined now. After another translation trigger, (d) is obtained, with missing subject filled by some default word.

Suppose just after obtaining (d) the user noticed that this interpretation is not what s/he wants, and the case element *kare-ga* should be the subject of the verb of the matrix sentence. Then the user triggers undo of translation twice, returning to (b). Then s/he notices that "he -ga buy -ta book" is treated as one phrase, against his/her interpretation. Then s/he changes the underlined area to "buy *-ta* book", excluding "he -ga" from the area (e), because this is the "correct meaningful phrase" in the user's interpretation. After translation trigger, (f) follows. Note that the subject of the relative clause is augmented by a default element. Finally (g), what the user wanted, follows.

3 A Hybrid Approach to Interactive Machine-Aided Translation

This section describes the model and the algorithm of the proposed method. First, the basic model of stepwise bottom-up interactive translation is described in the subsection 3.1. Then the next subsection describes how different translation paradigms can be integrated in this model. There are also shown a brief description of respective translation modules. The subsection 3.3 shows that the basic interactive operations of the method are capable of controlling the example-based translation process as well as the rule-based translation process. This close connection between the interactive operation and the translation method is one of most important characteristics of this method. In the last subsection an online learning scheme is introduced.

3.1 Basic Model of Interactive Translation

The basic model of the interactive translation method as described above is a bottom-up evaluation scheme of syntax-directed translation. In this scheme, the attribute of a syntax tree node is calculated from that of the children nodes by a semantic rule paired with the syntax rule used to build the node from the children. Attributes represent a partial translation result for the structure below the node, and the attribute calculation proceeds from the lexical nodes to the root node in a bottom-up manner. User interaction is associated with the attribute calculation at each node. Before each calculation, the tool pauses to show an interpretation of the underlying structure, and allows the user to examine and change it if necessary. Interactive translation proceeds from a smaller component to a larger component in a bottom-up and inductive manner. As translation mechanism, any method can be used as long as it is compatible with the general scheme. In the current system, the node at which the system automatically pauses for interaction are restricted to contain at most one predicate in order to reduce the operation cost, while this restriction is not applied to the user operations. The system looks for a lowest such node, then pauses there for user operation. When user triggers translation, the attribute of the focus node and below are calculated in a bottom-up mariner, then the result replaces the tree rooted by the focus node. The node serves as a kind of lexical node in the subsequent translation.

3.2 Hybrid Translation Module

The basic idea about how to integrate different translation paradigms into the above basic model is to use respective translation submodules in parallel at each translation step, while each submoduie processes the input independently. All the results are sorted according to the priority, then presented to the user. By unifying the data structure of input and output of all submodules, the results can be freely combined in a subsequent translation step.

The algorithm can be described as follows.

Repeat the following until the whole sentence is translated.

1. Find a minimal area for translation.
2. Show the area to the user. S/he can change the presented area if it is not appropriate.
3. Obtain possible translations of the area using respective translation modules. Calculate priorities of the results.
4. Show the results to the user in the order of priority. S/he may change the selection or even directly edit the results.

5. Replace the area with the selected/modified result.

Rule-based Module

The rule-based transfer module is the backbone of the whole translation module. It provides a default result for all kind of inputs. For some linguistic constructs, it is the default translation method. For example, translation of a simple sentence is performed by a case frame transfer rule that reorders the case elements of the main verb using the verb case frame correspondence encoded in the dictionary. Generally speaking, the skeleton of a simple sentence made of a main verb and its case elements are well described by the verbal case frame, and a rule-based treatment is suitable. For this kind of linguistic constructs, the corpus-based or example-based method would be rather useful in building the knowledge base, than being applied directly in the translation process.

Corpus-based Module

A corpus-based method will be mainly used for lexical translation. Although words are translated using a bilingual dictionary, corpus-based, more precisely statistics-based, method enters here for the translation equivalent selection through the DMAX method [Doi and Muraki, 1992; 1993]. This method uses the word cooccurrence frequencies gathered from independent source and target language corpora, and combines them in terms of the word to word correspondence in the bilingual dictionary, to eliminate an accidental cooccurrence between the translation equivalents of non-cooccurring words. A major advantage of this method is that the corpora need not to be parallel.

Example-based Module

An example-based method will be mainly used to translate a syntactically uniform structure such as compound noun or noun phrase. Since these structures often lack a clear syntactic feature useful for the rule-based analysis or translation, example-oriented methods such as [Sumita and lida, 1992; Hisamitsu and Nitta, 1995] have been proposed to capture their semantic and idiosyncratic property better. Although the rule-based method provides the baseline, these example-based method can offer a better result that depends on appearance of a particular word.

An example will be stored as a pair of the source language expression and the target language expression, with word to word correspondences wherever possible. It also keeps the information about the head word, which determines the behavior of the phrase as a whole. The input phrase to be translated is expressed as a sequence of words, where respective word is associated with the translation determined by the previous bottom-up translation steps, if any. This is the common data structure used by all the translation modules of the method. A constituent sub-phrase is justly identified with its head word, since translation of that phrase is already fixed. The transfer module looks for the best matching examples, and outputs the target language expression, replacing the constituents with the translation specified in the matched phrase when necessary.

Example-oriented method is also used in order to determine the translation equivalents of strongly cooccurring words, such as an idiomatic expression. This augments the statistics-based translation equivalent selection described before.

Idiomatic Expressions

There are some words that have special syntactic/semantic behavior, when appearing simultaneously. An example is *denwa-wo kakeru,* which usually means "make a phone call", not a literal word-by-word translation "hang a telephone". Possible translations include "make a phone call", "telephone" or expressions with similar meaning, but no literal translation can convey the proper meaning of the original expression. Since the proper translation for an idiomatic expression is not predictable from the individual behavior of the constituent words, they are seemingly exceptions to the bottom-up compositional scheme of the method. However, they can be handled without modifying the method, by combining an example-oriented method and a rule-based method.

The key idea is to separate the step of translation equivalent selection for each constituent word from the syntactic transfer step, and attribute the idiomatic property entirely to the former. The former can be handled by an example-oriented augmentation of translation equivalent selection method, whereas the latter will be performed by a purely rule-based method. This separation is justified as long as the structure of the resulting expression obeys the common rules of the target language grammar. For example, the characteristics of the correspondence between *denwa-wo kakeru* and "make a phone call" can be reduced to particular correspondence between *kakcru* and "make". When the system detects cooccurrence between *denwa* and *kakeru,* it adds a translation equivalent "make" to the window of *kakeru.* The user can choose an idiomatic interpretation of this expression simply by choosing this alternative. Later process can proceed entirely by a general transfer rule. Similarly, the same expression can be translated into a verb "telephone" simply by giving translation "telephone" to *kakeru,* while denwa-wo is left without translation equivalent so that it disappears in the result. Thus the essential task of idiom translation is reduced to an example-oriented method of translation equivalent selection.

3.3 Interactive Operations

As described before, the basic interactive operations of the method are translation area correction and translation equivalent selection. From the viewpoint of the rule-based method, the translation equivalent selection operation is more than simply choosing from among synonyms, as discussed in [Yamabana *et* al., 1995]. First, by specifying the translated area, one can directly resolve the dependency ambiguity. Secondly, part-of-speech of the translation equivalent may be specified through this operation, since translation equivalents with different part-of-speech appear distinctly in the alternatives window. Thirdly, the translation equivalent for functional words can be specified, and that can specify the syntactic behavior of the result. Although functional words remain unchanged in the intermediate representation, some words provide an alternatives window when the cursor is located on them.

From the viewpoint of the example-based method, interactive operations have different meaning. By these operations the user can control two major decisions of the method, that is, how to cover the source sentence and which example should be used. Changing the translation area implies changing the covering, and a new example that fits better to the expanded or shortened area will be chosen as a primary candidate. Changing translation equivalent selection after translation implies changing the example used for translation. Thus, the two major decisions in the example-based translation method can be interactively controlled without ever introducing new kind of operation.

3.4 Online Learning of Translation Instances

This scheme can offer a simple mechanism of online learning. As discussed above, the user has a control over the major decisions of either rule-based or example-based translation. This control information can be used to learn better choice of rules or examples used. Another source of information is the translation result itself. Since the method allows interactive corrections of respective translation at each step, the correspondence between a source language expression and its translation is expected to make a satisfactory translation pair for the user. By accumulating these translation pairs, the system will grow and adapt to the environment, especially to the user's preference.

4 Current Status of Implementation

This method was implemented as an English writing support software on personal computers, with a rule-based translation module and an idiom processing mechanism. The system dictionary contains about 100,000 Japanese entries and more than 15,000 idiomatic expressions, the latter built from scratch by the method in [Tamura *et* a/., 1996]. Addition of the statistics-based lexical transfer module and example-based transfer modules are currently under way, as well as the example learning module.

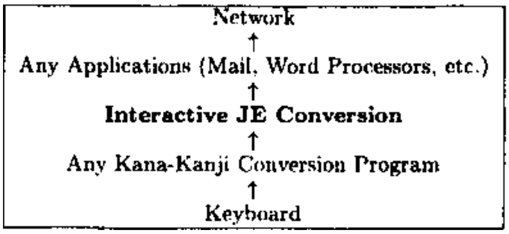

Network
Any Applications (Mail, Word Processors, etc.)
Interactive JE Conversion
Any Kana-Kanji Conversion Program
Keyboard

Figure 4: Relation to Other Programs

The current system consists of the interface module and the translation module, communicating through interprocess protocol. One important feature of this implementation from application viewpoint is that it works as a language conversion front-end to an arbitrary application, as described in section 2. The system works as a kind of a keyboard extension, converts a Japanese input into an English equivalent, and send the result to an arbitrary application program (figure 4). This feature allows the tool to be used as an add-on function of a familiar document writing environment.

5 Discussion

There have been several approaches to integrating different translation paradigms. The Pangloss system [Nirenburg *ed.,* 1995; Brown, 1996] adopts a multi-engine architecture, in which Knowledge-Based MT, Example-Based MT and Lexical Transfer MT engines independently produce translations for a part of the input sentence. The translations are registered, selected and combined on a chart-like structure. Translations from different modules are treated in a uniform manner, and selected by the priorities assigned by the respective translation engines. [Chen and Chen, 1995] proposes a hybrid transfer method that combines statistics-based transfer for smaller chunks and rule-based transfer for sentence-level chunks. Translation method is changed according to the size and nature of the substructure to be translated.

Our method shares the basic strategy with these previous works in that it divides the problem into subproblems of translating the substructures, and tries to use the most appropriate translation method for respective problem. The point of our method is that this scheme naturally fits into the interactive translation scheme formerly proposed and provides a common platform for integrating various approaches to translation. In the lexical translation step, statistics-based DMAX method supplies the baseline, and example-oriented recognition of idiomatic expressions provides a fine improvement. At later steps of structural transfer, the hybrid translation module enables to employ the most appropriate translation method, depending on the nature of the focused structure. The interactive operations are not only used to confirm and correct the initial selection, but also they provide a means to control various decisions in the course of the translation process.

From an application point of view, the important thing about this hybridization is that it offers a systematic method to add example-based improvement to the baseline rule-based translation. As is well-known, some expressions used in communication are not subject to compositional principle of meaning interpretation. Since the word sequence itself contains some meaning not calculable from meanings of constituent words, a direct translation of the original sentence does not necessarily convey the same meaning in the target language. This is one reason that many conventional foreign language writing support tools employ an approach based on translation examples. However, these tools are often too rigid to allow truly free composition, only allowing one to replace a word or two in the example sentence. On the other hand, our method allows to freely combine rule-based and example-based translation results. A rule-based result can be embedded in an example-based skeleton of sentence, or vice versa. In addition, the idiom dictionary mechanism enables to detect idiomatic expressions with far separated constituents.

Finally we briefly discuss the cost of the interactive operation. Although the method is interactive, the only indispensable operation is the "next" trigger to confirm that the system's choice is fine. All other operations such as translation equivalent selection are optional. If the user continues to simply confirm the system's choice, all the judgement by the system is employed and according result is obtained. If the user carries out a detailed interaction, changing the alternative or even editing directly, then a result comparable to a fully manual translation can be obtained. Thus the method lies between the automatic machine translation and the dictionary-aided manual translation, where the precise location is determined by the user.

6 Conclusion

We proposed a hybrid interactive translation method, in which rule-based, corpus-based and example-based translation methods are integrated with an interactive man-machine interface for stepwise bottom-up translation. This integration will give the user a freedom to combine most appropriate results obtained from various strategies and resources. Currently the example-based and statistics-based modules are being added to the current implementation as an English writing support tool, which is provided with the rule-based transfer module and the idiomatic expression processing function realized as a mixture of an example-oriented translation equivalent selection method and rule-based structural transfer. We are planning to measure the effectiveness of this method when the expansion is completed.

References

[Chen and Chen, 1995]
Kuang-hua Chen and Hsin-Hsi Chen. Machine Translation: An Integrated Approach. In *Proc. of TMI-95,* pages 287 294, 1995.

[Brown, 1996] Ralf D. Brown. Example-Based Machine Translation in the Pangloss Svstem. In *Proc. of COLIN G-96,* pages 169-174, 1996.

[Doi and Muraki, 1992] Shinichi Doi and Kazunori Muraki. Translation Ambiguity Resolution Based on Text Corpora of Source and Target Languages. In *Proc. of COLING-92.* pages 525 531, 1992.

[Doi and Muraki, 1993] Shinichi Doi and Kazunori Muraki. Evaluation of DMAX Criteria for Selecting Equivalent Translation based on Dual Corpora Statistics. In *Proc. of TMI-93.* pages 302 311, 1993.

[Hisamitsu and Nitta, 1995] Toru Hisamitsu and Yoshihiko Nitta. Analysis of Japanese Compound Nouns by Direct Text Scanning. In *Proc. of COLING-96,* pages 550 555, 1996.

[Muraki *et al,* 1994] Muraki, K., Akamine, S., Satoh, K. and Ando, S. TWP: How to assist English production on Japanese word processor. In *Proc. of COLING-94,* pages 283 298, 1994.

[Nirenburg *ed.,* 1995] Sergei Nirenburg, editor. The Pangloss Mark III Machine Translation System. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT. Issued as CMU Technical Report CMU-CMT-95-145, 1995.

[Sumita and lida, 1992] Eiichiro Sumita and Hitoshi lida. Example-based Transfer of Japanese Adnominal Particles into English. IEICE Transaction on Information and Systems, vol. E-75-D, No.4, pages 585-594, 1992.

[Tamura *et al.,* 1996] Tamura, S., Kamei, S., Doi, S. and Yamabana, K. Collecting of Verbal Idiomatic Expressions and Development of a Large Dictionary for Japanese-to-English Machine Translation. (In Japanese) In *Proceedings of 2nd Annual Convention of Association for Natural Language Processing,* pages 45 48, 1996.

[Yamabana *et al,* 1995] Yamabana, K., Doi, S., Kamei, S., Satoh, K., Tamura, S. and Ando, S. Interactive Machine-aided Translation Reconsidered —Interactive Disambiguation in TWP—. In *Proc. of NLPRS-95,* pages 368 373, 1995.

[Yamabana *et al.,* 1997] Yamabana, K., Muraki, K., Kamei, S., Satoh, K., Doi, S. and Tamura, S. An Interactive Translation Support Facility for Non-Professional Users. In *Proc. of ANLP-97,* pages 324-331, 1997.