Integration of example-based transfer and rule-based generation

Susumu AKAMINE, Osamu FURUSE and Hitoshi IIDA
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
{akamine,furuse,iida}@itl.atr.co.jp

1 Introduction

When we speak a foreign language, we use not only grammatical rules but also memorized expressions. Namely, translations are sometimes produced by mimicking translation examples. Example-Based Machine Translation(EBMT) adopts this strategy as follows: 1) Retrieves the translation example whose Source Expression (SE) is the same as or most similar to the input sentence, 2) Translates the input sentence using the Target Expression (TE) of the retrieved translation example.

Since it is impossible to memorize all possible sentence patterns, the "chunking + best-matching + recombination" method is practical from the point of view of coverage. It provides translation examples at various linguistic levels, and decomposes an input sentence into chunks. For each chunk, the translation example is retrieved by best-matching. The output sentence is obtained by recombining the TE parts of the retrieved translation examples. However, this method suffers from translation quality lapses at the boundaries of the recombined chunks. These lapses are caused by serious gaps between languages, such as between English and Japanese, that differ widely in their syntactic features.

This paper proposes a method to solve the problem of the structural gap by introducing a new generation module to the model of (Furuse, 92), that is able to handle the entire translation process within the example-based framework. This integrated method has been implemented in a prototype English-to-Japanese translation system.

Figure 1 shows the proposed integrated method between example-based transfer and rule-based generation. The transfer module decomposes an input sentence using the SE(English) part of translation examples, and converts each piece of the input sentence into the equivalent piece in the TE(Japanese) using translation examples.

The rule-based generation part of the integrated method consists of a composition module and an adjustment module. The composition module composes a structure from fragmentary examples by using Japanese grammatical constraint, and checks whether the structure is grammatically appropriate or not. The adjustment module refines the sentence output of the composition module so that the final output is as natural as colloquial Japanese.

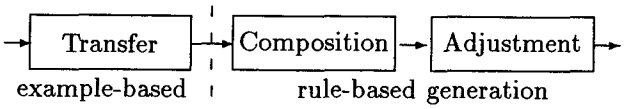

Figure 1: Integration of Transfer and Generation

2 Example-based transfer

The transfer module outputs the TE structure from an input sentences by using translation examples. In the proposed translation method, translation examples are classified based on the string pattern of their SE and are stored as empirical transfer knowledge, which describes the correspondence between an SE and its TEs. The following is an example of the transfer knowledge about "X to Y" at the verb-phrase level.

X to Y →	Y′ e X′	((go, Kyoto)...)
	Y′ ni X′	((pay, account)...)
	Y′ wo X′	((refer, announcement)...)

The first possible TE is "Y′ e X′", with the example set ((go, Kyoto)...). Within this pattern, X′ is the TE of X, which expresses a variable corresponding to some linguistic constituent. (go, Kyoto) are sample bindings for "X to Y", where X ="go", and Y ="Kyoto".

Patterns in transfer knowledge are classified into different levels according to the scale of their linguistic structure in order to restrict the explosion of structural ambiguity, and an input sentence is decomposed into chunks by applying SE parts of transfer knowledge in a top-down fashion.

Suppose that an input sentence is "Please come to our company." SE parts of transfer knowledge are applied in the order, "please X" (simple sentence), "X to Y"(verb phrase), "our X"(compound noun), "come", "company"(surface word), yielding the following SE structure:
(Please ((come) to (our (company))))

For each chunk of the SE structure, the most appropriate TE is selected according to the calculated distance between the input words and the example words. The distance calculation method of (Sumita, 91) is adopted here. The distance between words is defined as the closeness of semantic attributes in a thesaurus.

The SE structure chunks of "Please come to our company" are transferred to "X′ tekudasai", "Y′

e X'", "watashi-tachi no X'", "kuru[come] [1]" and "kaisha[company]". By combining these TE chunks, the following TE structure is obtained, which will be the input of the composition module of the rule-based generation model:
(((watashi-tachi no (kaisha)) e (kuru)) tekudasai)

In the above structure, the honorific word "irassharu[come]" is more adequate than the neutral word "kuru" from the point of view of politeness. The replacement of "kuru" with "irassharu" will be done in the adjustment module of the rule-based generation model.

3 Rule-based generation

3.1 Composition

The composition module checks whether a transferred sentence is grammatically appropriate or not, and corrects grammatical errors produced by the structural gap. The composing method is almost the same as the syntactic analysis method. However, the process is much simpler, because the input string has the correct Japanese structure and the corresponding English expressions.

The procedure is as follows: 1) Divide the sentence into clauses, using not only Japanese grammatical features but also the TE structure and its corresponding English expressions. 2) Analyze each clause using the Japanese syntax rule. 3) Check on Japanese grammatical constraints. If the process finds violations, it corrects them by using Japanese linguistic knowledge.

Japanese sentences have a peculiar grammatical constraint. Some expressions cannot appear in a subordinate clause. For example, the postpositional particle "wa(topic marker)" cannot appear in a conditional clause. Table 1 gives examples of limitations on expressions. In Table 1, "masu" expresses an auxiliary verb, which indicates the level of politeness, and "darou" expresses an auxiliary verb, which indicates the speaker's supposition.

The checking and correcting method is explained here, using the conditional clause "((anata wa ryohkin wo shiharau masu) baai) [you TOPIC fee OBJECT pay POLITE CONDITION]." First, the process checks on limitations for conditional clauses by referring to Table 1, so it understands that neither "wa" nor "masu" can appear in a conditional clause(X baai). Second, the process analyzes the clause, so it understands that the case of "anata[you]" is "ga", and "masu" can be deleted. Finally, the process gets the right conditional clause "((anata ga ryohkin wo shiharau) baai)"

3.2 Adjustment

A sentence that is only grammatically appropriate, is not as natural as a colloquial sentence. The adjustment module refines the sentence by changing, adding, or deleting words. This module handles honorific expressions and redundant personal pronouns, which are important for generating natural Japanese sentences.

Table 1: limitations on Japanese clauses

example clause	topic "wa"	polite "masu"	supposition "darou"
X shite[X, and]	Good	Good	Good
X to(omou)[that X]	Good	N.G.	Good
X node[because X]	Good	Good	N.G.
X baai[if X]	N.G.	N.G.	N.G.

Table 2: honorific expression of verb

agent	recipient	example for okuru[send]
—	hearer	o-okuri-suru
hearer	speaker	okut-tekudasaru
hearer	—	o-okurini-naru

Personal pronouns are usually redundant in Japanese conversations, because honorific expressions and modality expressions limit the agent of the action. The procedure is as follow: 1) Change the verb into the appropriate form based on the agent or the recipient, 2) Delete the redundant personal pronouns based on the verb form, or the modal, 3) Generate a final output by adjusting the morphological feature.

This method is explained here, using the sentence "((anata wa watashi ni youshi wo okuru masu) ka) [you TOPIC I OBJECT form OBJECT send POLITE INTERROGATIVE]." First, the process changes the verb "okuru" into "okut-tekudasaru" by referring to Table 2. Second, it deletes the redundant pronouns "anata wa" and "watashi ni" . Finally, it generates the sentence "youshi wo okut-tekudasai masu ka [form OBJECT send-RESPECT POLITE INTERROGATIVE]."

4 Evaluation

The prototype system was evaluated by using model conversations between an applicant and a secretary about conference registration. The model conversations consist of 607 sentences, and cover basic expressions. The system provided an average translation time of 2 seconds for sentences with an average length of 10 words, and produced a translation result for all of the sentences. 480 of the results were as natural as colloquial sentences and giving a success rate of 79%.

References

Furuse, O. and Iida, H. 1992. Cooperation between Transfer and Analysis in example-based framework. In Proc. of Coling '92.

Sumita, E. and Iida, H. 1991. Experiments and Prospects of Example-based Machine Translation. In Proc. of the 29th Annual Meeting of the Association for Computational Linguistics.

[1] $[w_1 \ldots w_n]$ is the list of corresponding English words. Uppercase shows the meaning of a function word.