Machine Translation by Interaction between Paraphraser and Transfer

Kazuhide Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
yamamoto@fw.ipsj.or.jp

Abstract

A machine translation model has been proposed where an input is translated through both source-language and target-language paraphrasing processes. We have implemented our prototype model for the Japanese-Chinese language pair. This paper describes our core idea of translation, where a source language paraphraser and a language transfer cooperates in translation by exchanging information about the source input.

1 Introduction

Humans generally have language capability, mostly for their mother tongue and to a lesser extent for foreign languages. This leads us to making the most of our mother language, even in conducting translation. That is, when we translate our language into a foreign one unfamiliar to us, we may try to paraphrase the source input into easier expressions we can translate.

In contrast, there is no such machine translation (MT) model so far proposed where the source language module is biased over the bilingual language module. All of the MT models are either those where the bilingual processor takes the initiative over the source language analyzer (conventional analyze-transfer-generate model) or integration models of analyzer and transfer, such as example-based or statistical models. Although some MT models have a paraphraser (also called a 'pre-editor'), such as that of Shirai et al. (1993), paraphrasing is performed in these models because it is necessary to prepare for the subsequent bilingual process. In other words, the paraphraser operates as a sub-module for successful transfer.

We have proposed a new MT model that is more similar to the human translation process than other MT systems (Yamamoto et al., 2001). This model, called the Sandglass model, is designed so that the system can generate a translation through source language paraphrasing, even if the system does not have sufficient bilingual knowledge. In this sense, our model design can be considered a *non-professional* translator's model.

From the engineering point of view, our model has an advantage in language portability; it is easy to construct an MT for a new language, since our model depends only on source language and thus can reduce dependence on bilingual knowledge. Moreover, the better source language paraphraser we make, the easier the implementation of other language MT becomes. Another advantage is task portability, since all of the paraphrasing knowledge, except for lexical paraphrasing knowledge, is independent of the task, so we do not need to fit most of the paraphrasing knowledge to the required task. It is also significant that this model's paraphraser can be employed not only for MT but also for most natural language processing (NLP) applications. This is possible because both the input and output of a paraphraser is the same natural language.

We have been building the Sandglass MT system for the Japanese-Chinese, Chinese-Japanese language pairs (Yamamoto et al., 2001; Zhang and Yamamoto, 2002). We have already constructed a prototype for Japanese-Chinese. In this paper, we report the core concepts of this prototype and discuss issues of both our principle and our implementation.

2 Sandglass Translation Model

Figure 1 shows our paradigm for a translation model. In the conventional MT model, the process load and the information used to deal with

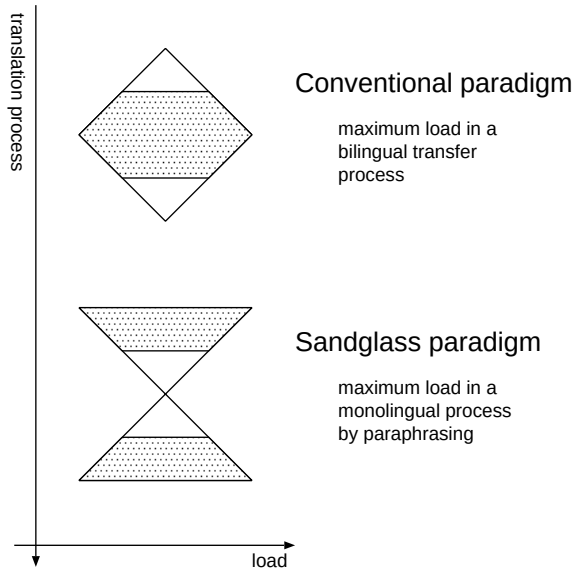

Figure 1: Comparison of the two MT paradigms

it are maximized in the transfer module; however, we propose that they be minimized in the transfer in consideration of language portability and task portability.

This translation approach is effective in MT where neither the source- nor target-language is English. Although there are a large number of bilingual corpora currently available, most of them are between English and other language. This suggests that it is not useful to apply bilingual-corpus-based approaches to situations not involving English. Moreover, conventional approaches based on hand-written rules are also unsuccessful due to lack of bilingual speakers of non-English pairs.

We also assume that reduction of bilingual processing costs is crucial for multilingual MT construction. Although both interlingual MT and MT with controlled language satisfies this request, our MT paradigm has an advantage in that it does not require design of interlingua/controlled language, which can be a critical problem.

2.1 Modularity and paraphrasing strategy

The SANDGLASS translation model has a source language paraphraser (hereafter the paraphraser) and a bilingual language transfer (hereafter the transfer), which have high modularity with each other in order to develop them as independently as possible. One of our aims in this model is to develop a general-purpose paraphraser that can also be used in other NLP applications.

When the system has modularity, the paraphraser does not need to consider the knowledge or translation capability of the transfer. However, the paraphraser has trouble in planning a paraphrasing strategy, since the purpose of paraphrasing in this model is to help small-knowledge transfer. One may think of it as a solution to generate all possible paraphrases, transfer them into the target language, and select the best one among the successful outputs. We believe that, although this strategy works, it is not practical due to the computation cost. In many cases, there are local paraphrases possible for one input, which may result in combinatorial explosion for generating paraphrases. Moreover, this strategy leads to a more serious problem in speech translation that requires real-time computation.

As an alternative, we propose the following strategy for planning paraphrases. We first put the controller between the paraphraser and the transfer. The controller communicates with both the paraphraser and the transfer and exchanges information between them on the target sentence be translated. As opposed to the one-way information path from the paraphraser to the transfer, a bi-directional information flow enables cooperation by allowing each module to provide its counterpart with information on what is possible and what is impossible.

This kind of process is not necessary in the typical MT model, since each process has the responsibility to perform its mission successfully and giving up is never allowed. If one of the processes gives up its mission, the entire translation process also gives up and fails. On the contrary, our model (sometimes) allows the transfer to give up generating the target language. Although this responsibility continuously enlarges the transfer knowledge, it is one of the critical problems of the typical MT. In general, in order to avoid a fatty transfer, we propose shifting the responsibility of generating the target language from the transfer process to monolingual processes.

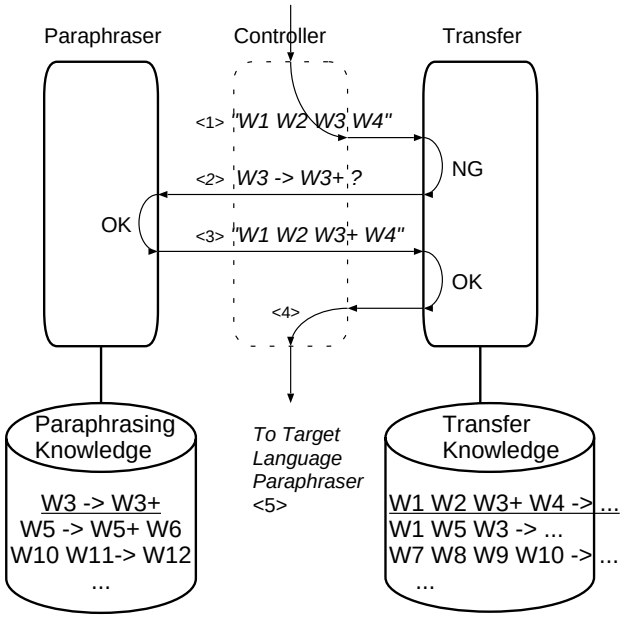

Figure 2: Translation strategy by interaction of the two modules

2.2 Interaction between paraphrasing and transfer

Figure 2 illustrates our translation strategy. The translator mainly consists of the paraphraser and the transfer, where a controller is located between the two modules in order to control the information flow[1]. This model has the following characteristics: (1) the paraphraser and the transfer are equivalent in terms of process sequences, i.e., the process flow is not an assembly line type, and (2) the knowledge for paraphrasing and that for transferring are separated so that the paraphraser and the transfer are responsible for monolingual and bilingual processing, respectively.

The translation process is achieved as follows. The output of word segmentation and part-of-speech (POS) tagging is first attempted to transfer to the target language through the controller. Assume that a sequence of morphemes $W_1W_2W_3W_4$, where W_i is each morpheme, fails to transfer (process $\langle 1 \rangle$ in the figure).

In this case, the transfer may obtain information on the failed input morphemes that is useful for the paraphrasing strategy, such as similar morpheme sequences that can be transferred or parts of the input that are impossible to transfer. Our transfer can obtain expressions similar to the input, if any exists, when the transfer fails. In this example, the transfer found that the morpheme sequence $W_1W_2W_{3+}W_4$ is similar to one in its knowledge, i.e., it understands that the input can be transferred if W_3 can be paraphrased into W_{3+}. Accordingly, the transfer provides this information to the paraphraser as a paraphrasing hint (process $\langle 2 \rangle$).

Then the paraphraser attempts to use this suggestion prior to other paraphrasing trials. It judges whether W_3 is replaceable by W_{3+}, and if it has such knowledge, it paraphrases based on the transfer hint and returns this paraphrase to the transfer (process $\langle 3 \rangle$). Again, the transfer carries out a new trial and it succeeds in translation this time (process $\langle 4 \rangle$). Finally, the target language expression is passed to the subsequent processes (process $\langle 5 \rangle$).

Among the possibilities other than those shown in the figure, if the transfer cannot find any similar expression, the paraphraser then attempts to rewrite the input by utilizing its own paraphrasing knowledge. Similarly, if the paraphraser cannot accept the rewriting hint that the transfer suggests, the paraphraser also thinks by itself.

2.3 Paraphraser

Currently, our paraphraser can deal with six paraphrasing types: (1) verification of the transfer's suggestion, (2) input segmentation, (3) reduction of honorific expressions (Ohtake and Yamamoto, 2001), (4) simplification of functional words (Yamamoto, 2001), (5) chunking of noun phrases, and (6) deletion of minor elements. Paraphrasing is conducted in this order. If one of the pattern conditions in this paraphrasing knowledge is matched, the paraphraser then finishes and returns its paraphrase; no other paraphrase is pursued.

(1) As the first type of the paraphrasing, the paraphraser verifies the paraphrasing hint that the transfer suggests, if any. In our model, all of the suggested paraphrasing rules are formed as single-morpheme replacements, most of which are functional words. Therefore, the paraphraser has a list of these types of rephrasing rules in advance to verify the suggestion. We have built a list that contains 175 replacement patterns.

[1]For simplicity, other parts of the translator are hidden in the figure.

Ex.1 おもしろそうです ねえ
→ おもしろそうです ね
(It seems interesting.)

Ex.2 何時まで でしょう か
→ 何時まで です か
(Until what time is it?)

In the above two examples, a sentence-final particle and an auxiliary verb are replaced, respectively. These slight differences should be merged before bilingual processing in order to restrict unnecessary combinations in the target language.

(2) If the verification fails, the paraphraser then attempts to split the input utterance according to the pre-defined segmentation rules. This is necessary because we are dealing with spoken language, which has no clear sentence boundaries. The segmentation rules, consisting of 30 rules, are defined by checking sequences of either word or POS. For example, in many cases, if there is a sentence-final particle, then the input is segmented after that word. In the following example, a segmentation border is described by the symbol ";".

Ex.3 それじゃあ さよなら
→ それじゃあ;さよなら
(So, see you!)

Ex.4 いくらですか それ
→ いくらですか;それ
(How much? That one.)

It is possible to regard the above two examples as single sentences, so it is difficult in general in Japanese speech to determine whether to segment them or not. However, this is not a problem in the proposed method because our segmentation is conducted only if the transfer fails to deal with the input as a single sentence.

(3) Honorific expressions are seen in Japanese speech very frequently. These expressions involve many variations for expressing one sense, so they should be unified before the transfer to avoid the great amount of increase in unnecessary bilingual knowledge that would be expected. Our paraphraser for honorifics, which was proposed by Ohtake and Yamamoto (2001), reduces such variations to as few as possible. We have 212 paraphrasing patterns for honorific expressions.

Ex.5 では いかがいたしましょうか
→ では どうしましょう
(Then how should we do?)

Ex.6 あいにくですが ございません
→ あいにく ありません
(Unfortunately, there isn't.)

(4) Similarly to honorifics, there are also many variations in Japanese verbal expressions, so we again need to reduce variations. Spoken-style expressions are targets of paraphrasing here, and they are replaced by written- or normal-style expressions. The target phenomena and the effects of this paraphraser have been discussed in Yamamoto (2001). The paraphraser we use involves 302 patterns.

Ex.7 風邪 じゃないかと思うんですけどね
→ 風邪 でしょう
(I think it may be a cold.)

(5) Noun phrases are chunked here according to simple pattern matching by lexicon or POS: if two or more nouns are consecutive with or without a possessive-like particle "の," we then regard them as one noun phrase. This process is necessary because we parse input utterances in neither the paraphraser nor the transfer, and the transfer only see POS sequences. We expect that this chunking would help to make our template-based poor transfer more robust against input variations. However, we place this process at a low priority in the paraphrasing order because an unconditional operation of this process is considered to be troublesome, especially in spoken language. A chunk is illustrated as $\{\cdots\}$ below:

Ex.8 火曜日の午後五時 なんですが
→{火曜日の午後五時} なんですが
(It's Tuesday, at five p.m.)

(6) As the final paraphrasing measure, the paraphraser deletes relatively unimportant parts of the input expressions, such as adverbs of manner and degree, as well as particles expressing topical markers. Changing POS sequences of the input changes the searching space in the transfer knowledge. In the following two examples, two particles and an adverb are deleted, respectively. Currently, we have 22 patterns in this type.

Ex.9 明日までに は ご用意いたします よ
→ 明日までに用意します
(It will be ready by tomorrow.)

Ex.10 たぶん 十分くらいだと思います
→ 十分くらいだと思います
(Perhaps it takes ten minutes.)

2.4 Transfer knowledge construction

Our transfer knowledge is constructed as follows. Because our principle requires that the bilingual processing and its efforts should be reduced as much as possible, our bilingual knowledge is primitive and easy to construct automatically. Our knowledge sources are a sentence-aligned text corpus between Japanese and Chinese, a Japanese-Chinese dictionary where one source word may correspond to many target words, and a Japanese analyzer. Note that we used neither a Chinese analyzer nor tagging in the Chinese corpus.

Our transfer process is based on a word-template transfer technique, and we conducted automatic word alignment for its knowledge. We first analyzed all Japanese sentences in the corpus by the free-to-use morphological analyzer JUMAN[2]. We then checked, by string matching, whether each source language content word has a corresponding target word. If this alignment succeeds, both source and target language words are tagged with the same ID number. When more than one translation in the dictionary can be aligned, the longest word in the target side is selected for alignment.

One source language word may correspond to a target word that appears more than once. For example, a translation of the Japanese question "行きませんか" is "去不去？". We can deal with this result by accepting multiple correspondences, e.g., "⟨去 #538⟩ 不 ⟨去 #538⟩？" where ⟨···⟩ is a word boundary and #538 is an (example) ID number.

Hereafter, we call these sentence sets *templates* and the aligned parts in a sentence *variables*.

2.5 Transfer

The transfer module converts the source language input into the corresponding target language expressions by using the templates. The process consists of two parts, i.e., template retrieval and template matching.

The process first searches for templates satisfying similarity to the input expression. In order to judge similarity between the input and the templates, we only use the POS sequences of the input. If the retrieval succeeds, i.e., templates are found that have the same POS sequence, we then compare, word by word, the input and each retrieved template. If a word is a variable in the template, this comparison always succeeds. If there is no template retrieved, the transfer reports to the paraphraser (through the controller) that the retrieval process has failed. In this case, the paraphraser is required to somehow change the input sentence in terms of POS sequences.

Suppose that some templates are retrieved but matching fails, implying that some lexicons are different. Although this case is a transfer failure as well, the transfer has information on which parts of the input sentence failed to transfer, and such information could be a key for paraphrasing. Therefore, information on unmatched parts is also returned to the paraphraser with the result of the transfer failure. If multiple templates are retrieved and all of them fail in matching, all of the unmatched parts are returned in parallel.

If both the template retrieval and the template matching succeeds, this indicates that the transfer process has finished successfully. The input sentence is converted to the target language, and the transfer throws it to the controller for the following process. If more than one target language expression is returned due to the multiple successes in template matching, all of them are returned in parallel, and the following processes determine the best results.

3 Preliminary Experiment

We conducted a preliminary experiment to evaluate the translation capability under the current paraphrasing skills. Although there are many items that should be evaluated in MT, our first interest in the prototype is how much the paraphraser supports poor transfer knowledge and how small the acceptable transfer knowledge can be.

The transfer knowledge contains a bilingual dictionary of approximately 51,000 source language lexical entries, as well as up to 233,000

[2] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

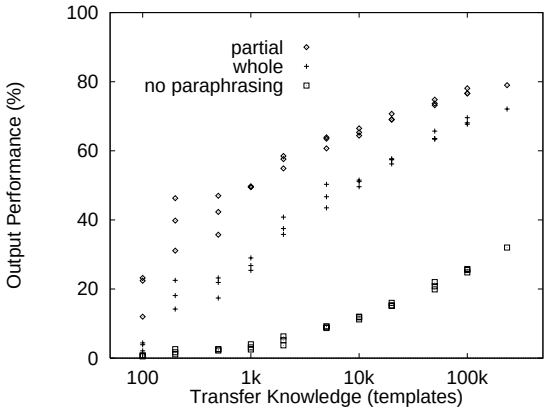

Figure 3: Changes in output ratios by amount of transfer knowledge

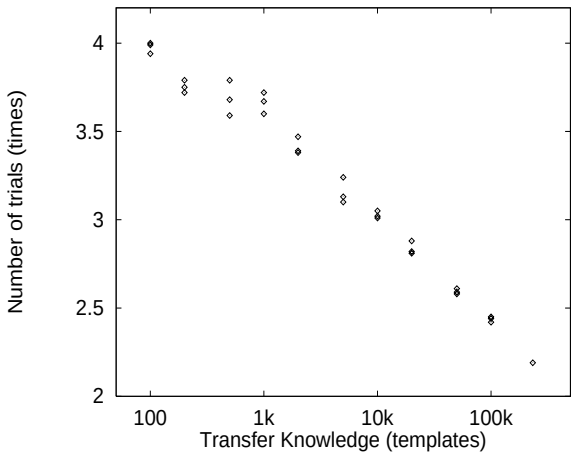

Figure 4: Changes in number of paraphrasing trials by the amount of transfer knowledge

utterances, in the domain of travel situations, and their translations. For evaluation, we use 1,000 utterances, each of which is 10 or fewer morphemes long, selected at random and unseen by the transfer.

The prototype is programmed in Perl language, and the running time at the maximum transfer knowledge is 0.555 second per utterance with a Pentium III 600 MHz processor. The ratios of the fully- and partially-translated utterances to several transfer knowledge sizes are plotted in figure 3. For comparison purposes, translation performance without the paraphrasing process is also illustrated in the figure. The experiments were conducted three times under each condition.

We can understand the importance of paraphrasing by observing the approximately 20%-40% performance gaps between full output and no paraphrasing. The paraphraser improves performance regardless of the knowledge size. The gaps are not trivial, so the experiments confirmed that the paraphraser plays an important role in the interaction process.

The figure also shows the fact that only 30% of the unseen input is translated using POS-sequence-based maximum templates. Considering that all inputs are 10 morphemes or fewer, this low performance implies the necessity to acquire 70% knowledge by somehow generalizing the existing 30% knowledge. The current paraphrasing knowledge – a collection of human intuition – can cover 40% of the inputs, while it seems difficult to cover the same or higher level by only automatically-acquired information from corpora.

Figure 4 shows the average number of paraphrasing trials. It would be a major problem in this design if there were many interaction loops between the paraphraser and the transfer, but we found that such worries are unwarranted in the current system. However, it is necessary to be careful in this measure, since we need to add more functions to the paraphraser in order to avoid zero output.

4 Related Works

It is important to reduce the burden of transfer to realize multilingual MT. In this sense, MT using a controlled language, such as the KANT system (Mitamura et al., 1991), has similar principles to our approach. We believe that multilingual MT systems should not place the obligation of transferring the target language on the transfer module. Difficult or ambiguous input should be checked in document translations, while it should somehow be resolved before the transfer module in speech translation, since real-time dialog conversation is a requirement.

Although we cannot find an MT model where an interactive (that is, feedback) approach between the two sub-modules is implemented, several types of interactive models have been discussed in natural language generation systems. In the Igen system (Rubinoff, 1992), which has a similar interactive operation, the Formulator

module provides feedback to the Conceptualizer module with information on how much of the content can been covered by a particular word choice. The Conceptualizer can then determine which choice satisfies its secondary goals with these annotations.

Two similar works paraphrase source input for MT. One is the work of Shirai et al. (1993), where they proposed a pre-editing approach for a Japanese-English MT system ALT-J/E. The other is the work of Yoshimi and Sata (1999), where they presented an approach to rewriting English newspaper headlines for the English-Japanese MT system Power E/J. The significant difference between their approaches and ours is the model design, i.e., whether the paraphraser and the transfer are sequential or integrated. Moreover, the purposes of paraphrasing are also different: in the pre-editing system it is for expediting the transfer and in the newspaper headline system it is for reducing peculiarities in the headline; on the other hand, our paraphraser's purpose is to support poor transfer knowledge.

5 Conclusions

We have proposed that many MT problems can be resolved if we have two paraphrasers, in both the source- and target-language. We have also proposed that bilingual knowledge be minimized in order to increase portability to other languages or other tasks.

This paper explained details of our MT system design and discussed its advantages. One feature of our design is that the translation process is achieved by interaction between the source language paraphraser and the transfer, unlike the conventional sequential MT model. We illustrated this advantage concretely by showing examples of the information exchanged between the two modules.

It is obvious that the bilingual burden is drastically eased in this model, while the importance of the monolingual process (i.e., paraphrasing) is increased. Although we did not illustrate many problems, such as ambiguity reduction for syntax or semantics, we will explore these issues in the future. Also, we need to evaluate the quality of the translation outputs after the target paraphraser is implemented.

Acknowledgment

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

References

Teruko Mitamura, Eric Nyberg, and Jaime Carbonell. 1991. An efficient interlingua translation system for multi-lingual document production. In *Proc. of MT Summit III*, pages 55–61.

Kiyonori Ohtake and Kazuhide Yamamoto. 2001. Paraphrasing honorifics. In *Proc. of NLPRS2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pages 13–20.

Robert Rubinoff. 1992. Integrating text planning and linguistic choice by annotating linguistic structures. In R. Dale, E. Hovy, D. Rosner, and O. Stock, editors, *Aspects of automated natural language generation*, pages 45–56. Berlin: Springer.

Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. 1993. Effects of automatic rewriting of source language within a Japanese to English MT system. In *Proc. of TMI'93*, pages 226–239.

Kazuhide Yamamoto, Satoshi Shirai, Masashi Sakamoto, and Yujie Zhang. 2001. SANDGLASS: Twin paraphrasing spoken language translation. In *19th International Conference on Computer Processing of Oriental Languages (ICCPOL2001)*, pages 154–159.

Kazuhide Yamamoto. 2001. Paraphrasing spoken Japanese for untangling bilingual transfer. In *Proc. of NLPRS2001*, pages 203–210.

Takehiko Yoshimi and Ichiko Sata. 1999. Automatic preediting of English newspaper headlines and its effects in a English-to-Japanese MT system. In *Proc. of Natural Language Processing Pacific-Rim Symposium (NLPRS'99)*, pages 275–279.

Yujie Zhang and Kazuhide Yamamoto. 2002. Paraphrasing of Chinese utterances. In *Proc. of COLING2002*.