Corpus-Assisted Expansion of Manual MT Knowledge

Setsuo Yamada[†], Kenji Imamura and Kazuhide Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 JAPAN
syamada@light.hil.ntt.co.jp,
{kenji.imamura, kazuhide.yamamoto}@atr.co.jp

Abstract

Since the expansion of MT knowledge is currently being performed by humans, it is taking too long and is too expensive. This paper proposes a new procedure that expands MT knowledge efficiently by supporting human judgements with information automatically collected from any number of corpora. The new procedure uses the source knowledge present in an MT system as the key to retrieve source language information from corpora. It also uses the partial translations provided by the MT to acquire target language information. These two techniques can reduce time and labor costs. Experimental results confirm both benefits.

1 Introduction

One of the biggest problems in rule- and pattern-based machine translation (MT) lies in acquiring an adequate body of knowledge. The current approach to knowledge acquisition demands that a human rule writer, someone who is familiar with general linguistic knowledge and a framework of MT knowledge, manually check and correct the translation examples output by an MT system. This incurs high time and labor costs. While this is reasonable for creating prototype systems, its efficiency is too low to create practical MT systems.

One engineering solution to this problem appears to be automatic knowledge acquisition (e.g., Almuallim et al. (1994); Alshawi et al. (2000); Kitamura & Matsumoto (1995); Watanabe (1993)). Such systems are capable of detecting simple rules directly from corpora through automatic learning. This approach is becoming more attractive due to the wide variety of corpora available. The current performance offered by automatic knowledge acquisition is rather suspect; the problem is that low-frequency phenomena are not well handled. Our basic approach is to combine human language skills with the automatic extraction of source and target language information. This is eminently practical since the number of corpora suitable for MT development continues to increase. The voluminous information available in these corpora should be used to improve construction efficiency by better supporting the human rule writer. This approach best suits the construction of large-scale MT systems, or expanding medium size MT systems.

Under this policy, we propose to realize cooperation between human knowledge and corpus[1] information. In the following discussion of this paper, we give TDMT

[†]Current affiliation is NTT Cyber Space Laboratories, NTT Corporation.

[1]The corpora considered in this paper are monolingual, however, bilingual corpora are also usable.

(Transfer-Driven MT) (Sumita et al. (1999)) as an example of modern MT systems. Although a part of the following discussion is MT system dependent, we are confident that our basic ideas can be applied to other rule- and pattern-based MT systems. One basic requirement is that they offer partial parsing and partial translation.

This paper consider two causes of the difficulties of expanding MT knowledge. One arises from the source language, the other arises from the target language. When expanding MT knowledge, first, a human rule writer has difficulty in identifying how source language knowledge should be reconstructed or what the correct analysis result should be. The other main problem is how to generate the target sentences. We must remember that expanding MT knowledge requires the creation of new MT knowledge.

Each source sentence normally exhibits many linguistic phenomena. Some of them can be analyzed with existing source knowledge but some can not. It takes a long time to judge which existing source knowledge is most applicable to an input sentence. Moreover, many source language expressions have multiple matching target language expressions. The conventional construction makes it difficult to construct various kinds of different target knowledge, because the human rule writer has time enough to consider only input sentences whose translation is obviously weak rather than considering the full range of sentences possible.

We propose a new procedure of expanding MT knowledge that has the following two characteristics. One is that we use the source knowledge present in the MT system to retrieve the source language information from one or more corpora. Use of this knowledge makes it easy to find useful source sentences in the corpus being considered. The display of many related sentences at the same time makes it much easier for the human rule writer to perceive "undiscovered" source linguistic phenomena. We think this type of new knowledge construction is the most efficient approach to constructing or extending comprehensive MT systems.

The other characteristic is that we use the partial translations offered by the MT in order to acquire target language information. Using this source knowledge as the retrieval key, we can obtain partial translations related to the source knowledge. The MT system can translate the sentence parts that match the retrieval key. Since a human rule writer can see several different target sentences for each piece of source knowledge at the same time, it is easy to construct target knowledge. Moreover, since the MT system is already being used, it is easy to add the target knowledge so acquired to the MT system.

As mentioned above, the use of existing source knowledge and the MT system itself can improve efficiency in terms of time and labor costs. Experiments have confirmed that this improvement is indeed possible. Moreover, other significant advantages seem achievable such as easy portability to other tasks or domains.

The next section briefly explains TDMT, which is taken as the prototypical MT system, and the process of conventional rule construction. The proposed method is detailed in section 3. Section 4 describes an experiment and the results gained. The final section discusses the expansion of MT knowledge and explains the features of the new method.

2 Conventional Method of Expanding Translation Knowledge

First, this section briefly examines TDMT as typical of modern MT systems[2] (Sumita et al. (1999)). We particularly focus on transfer rules, a key element of MT knowledge. Next we show a conventional method of expanding translation knowledge and the problems of the method.

2.1 Transfer-Driven Machine Translation

```
(source pattern)
⇒
((target pattern 1)
 ((source example 1)
  (source example 2)
  ... )
 (target pattern 2)
 ... )
```

Figure 1: Transfer rule format

```
(X “with” Y)
⇒
((y “to issho ni” x)
 (((“stay”) (“friend”))
  ((“travel”) (“child”)))
 (y “de” x)
 (((“come in”) (“sneaker”))
  ((“get on”) (“ticket”)))
)
```

Figure 2: Transfer rule example

TDMT translates an input sentence by combining several small linguistic phenomena expressed as transfer rules. Each transfer rule consists of a source pattern, target patterns, and source examples as shown in Figure 1. TDMT can use the source examples to resolve ambiguity and select a target sentence. The source examples come from sentences entered by a human rule writer when creating the transfer rules (we call such sentences training sentences).

For example, the source pattern (X “with” Y) shown in Figure 2 has two target patterns. If the input sentence is “You can go with your baby,” TDMT would select the target pattern (y “to issho ni” x). This is because the target pattern is attached to the source example “travel” and “child,” which are semantically most similar (in the thesaurus) to “go” and “baby.”

2.2 Conventional Knowledge Expansion Procedure for TDMT

Figure 3 outlines the conventional approach to expanding translation knowledge for TDMT. A human rule writer translates sentences one by one using the MT in order to find MT knowledge gaps. This step is common for most MT systems.

Usually, however, one sentence includes different kinds of linguistic phenomena and there are many possible combinations of source patterns. There is no simple way to choose the best combination from just the input sentence. The human rule writer is often puzzled about which source pattern should be modified or how a new source pattern should be created.

[2] Although we use English-to-Japanese MT in this paper, our new method does not depend on the language pair.

1. Select a corpus according to the goal.
2. Pick one source sentence from the corpus and use the MT to translate the sentence.
3. If the translation result is correct then go to 6, otherwise reconstruct the rules by the following procedure.
   (a) If the parsing result is correct, then go to (c).
   (b) Create a new source pattern or modify a source pattern.
   (c) If the rule, which is used in the correct parse tree or which is made by (b) above, has an appropriate target pattern, then go to (e).
   (d) Create a new target pattern, or modify a target pattern.
   (e) Attach a source example to the appropriate target pattern.
4. Translate the sentence by applying the MT again.
5. If the translation result is not correct, then return to (a) above.
6. If a sentence remains in the corpus, go to 2.

Figure 3: Conventional Approach to for TDMT

(1)	(A "should" B)	(A "by" B)	(A "in" B)
(2)	(A "should" B)	(A "by" B "in" C)	
(3)	(A "should" B "by" C)		(A "in" B)
(4)	(A "should" B "by" C "in" D)		

Table 1: Example: Combinations of Source Patterns

For example, consider the input sentence "You should stand by your friend in difficult times." Table. 1 shows the source pattern combinations capable of being constructed from the input sentence. (A "should" B)[3], (A "by" B), and (A "in" B), are the most general source patterns. (A "should" B "by"), (A "by" B "in" C) and (A "should" B "by" C "in" D) are also applicable as more unusual source patterns. Which rule set do you think is the best combination? Although combination(4) yields the best quality target sentence, it is too specific. On the other hand, combination(1) fails to yield good quality target sentences, because the meaning of "stand" is determined by "in difficult times." Therefore, we might select combination(2). However, this decision suits only the input sentence.

If the rule writer modifies a source pattern incorrectly or creates an inadequate source pattern, other source patterns would be negatively affected. Unfortunately, it is virtually impossible to prevent this from occurring when increasing the number of training sentences. If the rule writer wants to make a really correct source pattern, he/she must consider all sentences in the corpus. This incurs much time and labor.

Another problem is that many source patterns have more than one target patterns. It is difficult to construct a sufficient variety of target patterns.

[3]Capitalised characters such as "A" and "B" are variables.

In the conventional approach discussed above, the writer first has to judge whether the source pattern selected by the MT system is correct or not(step(a)). If the source pattern is wrong, the writer needs to modify the source pattern or create a new one (step(b)). Moreover, the writer has to modify or create other target patterns if the rules do not provide appropriate target patterns(step(d)). Each of these steps needs a lot of time.

Additionally, these steps must be repeated with each sentence in the corpus, so an excessive amount of time is needed to process all training sentences.

3 A New Method of Expanding Translation Knowledge

This section first describes the basic idea of the new method that solves the problems described in Section 2.2, and explains how it could be applied to TDMT. Note that TDMT is considered only as an example, the new method is applicable to other MT systems that offer partial parsing and partial translation.

3.1 The Basic Idea of The New Method

In the conventional method, the human rule writer has to modify a source pattern or create a new source pattern in order to eliminate poor translations. Since one sentence may contain many source linguistic phenomenon and there are many possible combinations of source patterns, this work is difficult. The new method eases this burden since the writer is provided with source sentences that match the same source pattern. This is why we employ existing source patterns as the retrieval key. The writer need not be concerned about the source pattern and only needs to judge whether the extracted source sentences are correctly matched or not. This greatly reduces the time and labor costs incurred in covering all training sentences compared to the conventional method.

Many source patterns have several possible target patterns. In the conventional method, it is difficult to construct various kinds of different target patterns, because the writer tends to consider only those input sentences that yield poor translations. In the new method, on the other hand, the writer can obtain many target phrases for the same source pattern by partially translating source sentences. Since the writer can easily judge whether each of the extracted target phrases are correct or not, he/she finds it much easier to construct various kinds of target patterns.

Moreover, the use of source patterns as the retrieval key and partial parsing for extracting target phrases makes it easy to add this approach to an MT system.

3.2 Procedure of New method

Figure 4 outlines the new method of expanding translation knowledge within TDMT. While the conventional method considers that the unit of knowledge is one sentence, the new method takes the transfer rule as the unit.

First, a human rule writer analyses all sentences in the corpus (for simplicity, only one corpus is mentioned hereafter) using a TDMT system (step2), because the writer should know how the source patterns are distributed in the corpus. This information

1. Select a corpus according to the goal.
2. Analyze all sentences in the MT corpus.
3. Extract source patterns that are matched at least once.
4. Perform the following steps with every source pattern in order of matching frequency.
   (a) Extract source sentences from the corpus with a source pattern as the retrieval key using partial parsing, and target phrases corresponding to source sentences by partial translation.
   (b) If the source sentence is not matched correctly with the source pattern, go to (d).
   (c) If the target phrase is correct then add source example, otherwise modify or create target pattern and add source example.
   (d) If retrieval result still remains, then go to (b).
5. Use the MT to translate all sentences with modified rules.
6. Modify or create rules for poor quality translation using the conventional approach.

Figure 4: New Method for TDMT

helps the writer determine which source patterns should be checked. The method can extract just the matched source patterns in step3.

Step4 is the most important part of the new method and has two characteristics. First is that the source sentences are extracted by using source patterns and the partial parsing provided by the TDMT system. They are retrieved from the corpus by using source patterns as the keys. Second is the partial translation provided by the TDMT system; this yields target phrases that correspond to source sentence parts.

For example, if the writer uses the rule (X "with" Y) as the retrieval key, the writer is presented with source sentences like "I will stay with my friend" from the corpus. Target phrases like *"tomodachi to issho ni tomaru"* (*tomodachi* "friend", *to issho ni* "with", and *tomaru* "stay") are yielded by partial TDMT translation. That is, the source sentence is matched with (X "with" Y), and the target phrase is transferred from a part of the matched source sentence "stay with my friend."

The characteristics are effective not only for the TDMT system but also almost all rule- and pattern-based MT systems. This is because they usually have parsing, transfer, and generating process, and they use source knowledge and target knowledge corresponding to the source knowledge. Since one aim of the new method is to extract new source and target knowledge, it can be applied to other MT systems by using parsing rules to handle source knowledge and generating rules to handle target knowledge. As for the above example, if a MT system employs the following transfer rule: "X stay with Y" $\Rightarrow$ "x *wa* y *to issho ni tomaru*"(*wa* "SUBJ"), then "X stay with Y" can be used as the retrieval key.

In step4, the writer can concentrate on judging whether the extracted source sentences match correctly or not, and whether the extracted target phrases are correct or not. Moreover, these retrieval results are easy to integrate into TDMT, because the source pattern already exists in TDMT and the target phrases are output by TDMT. Therefore, this is expected to reduce time and labor costs.

After the writer integrates the modified target patterns into TDMT for all matched source patterns, the writer translates all sentences in the corpus by TDMT using the modified rules. Finally, the writer modifies or creates rules to eliminate poor quality translations using the conventional approach. This process is needed to handle new linguistic phenomena and to ensure that the rules can handle all sentences in the corpus without contradiction.

In the conventional approach, the rule writer sometimes modifies a source pattern incorrectly or creates an inadequate source pattern. These patterns can negatively affect other source patterns, and high costs are incurred in correcting them. In the new method, the rule writer can see several source sentences that match the source pattern, which allows the writer to judge whether the source pattern is correct or not. This work decreases the number of rule writer mistakes. Therefore, we can expect fewer errors. However, it is impossible to remove all bad effects, because no one can be expected to comprehend all linguistic phenomena.

3.3 Simple Idea for Retrieval

Source patterns have syntactic roles such as verb phrase and noun phrase. If the writer retrieves source sentences by considering rules as simple word sequences, many obviously useless source sentences might be returned which would lower efficiency. Thus, we place syntactic constraints on retrieval. If partial parsing with the criteria is not successful, the source sentence is not extracted.

For example, if the rule (X "with" Y) is used as a verb phrase, sentences wherein "with" is used in a noun phrase, such as "I would like a room with a bathroom," are not extracted. "a room with a bathroom" can not be parsed by (X "with" Y) as a verb phrase.

4 Experiment

This section describes an experiment that compared the conventional approach to the new method in terms of efficiency and translation quality. The same training sentences were used in both methods. To compare the methods more clearly, we selected the domain of weather forecasting since this field exhibits a rather restricted number of linguistic phenomena.

A total of 373 training sentences were collected from a newspaper [4] published from December to March. First, one person expanded MT knowledge using the conventional procedure. About 4 months later, the same person expanded it using the new procedure. Because the same person was used, the interval between trials had to be quite long. The new method took 23 days to expand MT knowledge compared to 34 days with the conventional approach as shown in Table 2.

[4] We used material from "The Japan Times"(c) Japan Times, Ltd. in this experiment.

conventional	proposed
34	24

Table 2: Training period (days)

rank	conventional approach (%)	new method (%)
A	39.1	39.9
B	21.8	23.0
C	12.5	12.9
D	26.6	24.2

Table 3: Results

One person who was not a rule writer evaluated the translation quality of 248 open sentences. The sentences were published in the period from June to August. The result was that almost the same quality was found as shown in Table 3. As only one person conducted this subjective evaluation, the value itself is not definitive. However, it is suggestive of the power of the new method.

The ranks noted in Table3 are explained below:

A: Almost all weather information was transferred so the translation could be understood easily.

B: Some unimportant weather information was dropped but the translation could be understood easily.

C: Some important weather information was dropped but it was possible to guess the meaning of the translation.

D: Almost all weather information was dropped and it was hard to guess the meaning of the translation.

In order to see the effect of step4 in Figure 4, the rule writer evaluated the translation results before starting step6 (the part of conventional method) in the new method. The result was that 121 sentences in the corpus were "A" rank translations. Since only 2 sentences had been "A" rank before conducting step4, 119 sentences (31.9%) were improved by step4. Therefore, step4 is effective in improving translation quality.

The results show that the new method yielded comparable translation quality to the conventional approach while incurring lower time and labor costs. Therefore, we can say that using existing source knowledge and the partial translations of the MT system can improve efficiency.

5 Discussion

This section first outlines related works and discusses the roles of writers and machines in expanding MT knowledge. Next, we explain the impact of the new method.

5.1 Related Works

Practical methods of expanding MT knowledge by hand are seldom discussed. Many people hope that MT knowledge can be created fully automatically. However, the technology currently underlying automatic acquisition still does not allow the easy

extraction of translation knowledge. For example, Brill & Ngai (1999) reported that, in the task of base noun phrase chunking, people with almost no training could offer powerful rules in a shorter period of time on a small training set compared to the best systems available. That is, current automatic acquisition methods in some tasks are no more efficient than humans. Consequently, it is important to consider the cooperation of humans and computers from the viewpoint of practicality.

Streiter et al. (1999) proposed a strategy that rates manual MT rules by counting occurrence frequencies in a corpus. Some researchers expect this approach to contribute to the disambiguation of languages, but the approach is not applicable to the expansion and maintenance of rules. Tanaka (1994), in contrast, proposed an acquisition model for English case frames given by machine learning. This approach allows optimality to be maintained by confining the human descriptions of rules, but it still suffers from the problem of low knowledge construction efficiency, since it requires a bilingual tagged corpus of high quality.

Given the above review, we have been considering what information is useful to humans. As one solution, Shirai et al. (1995) examined methods of making semantic structure dictionaries for Japanese-to-English MT. The results indicated that the most efficient method was preparing example sentences through reference to transfer dictionaries for humans and human knowledge. This is a good method of integrating corpora (dictionaries) with human knowledge, however, it can not find the weak linguistic phenomena present in existing MT systems.

5.2 Features of The New Method

Our relatively small experiment did not convey the problem of excessive retrieval. As the display of useless retrieval results wastes time, we need consider how to minimize the problem when extracting source knowledge, that is to improve the quality of the retrieval method, or how to dump them efficiently by having the rule writer judge whether each retrieval result is correct or not. Even though the proposed method currently is weak in these areas, it does offer the following advantages over the conventional method.

- Higher translation quality, because "unknown" linguistic information should be found by the human rule writer accessing large corpora.
- Low dependency on the human rule writer's experience and skill, because it is easier to judge whether target language expressions are correct or not.
- Low dependency on the human rule writer's linguistic knowledge, because linguistic knowledge can be obtained from corpora easily.
- Easy portability to other tasks or domains, because it is easier to change the corpus used for retrieval.
- Easy arrangement of rules, because the human rule writer can see the same linguistic phenomena in several samples from a corpus/corpora at the same time.

6 Conclusion and Future Work

We have proposed a new procedure of expanding MT knowledge by retrieving linguistic knowledge from a corpus/corpora. The new method has two characteristics. One is that we use existing source knowledge as a key when retrieving the source language information from a corpus. The use of this key makes it easy to find useful source sentences that match the key. The other is that we use the partial translations provided by the MT in order to acquire target language information. We use source knowledge as the retrieval key, and obtain partial translations according to the source knowledge. The use of existing source knowledge and the use of the MT system itself improves system efficiency in terms of time and labor costs.

We conducted an experiment using weather forecasts and a TDMT system. We found that the proposed method allowed a human rule writer to expand MT knowledge more quickly than was possible with the conventional approach while still achieving the same level of translation quality. Therefore, the new method is more efficient.

This method appears to offer other advantages as mentioned in section 5, and we intend to conduct further experiments to confirm them. Moreover, while the proposed method offers good quality and efficiency, further improvements will be made to its retrieval section. Additionally, we will examine its efficiency with larger corpora.

References

Almuallim, Hussein, Yasuhiko Akiba, Takefumi Yamazaki, Akio Yokoo & Shigeo Kaneda: 1994, 'Two methods for learning ALT-J/E translation rules from examples and a semantic hierarchy', in *Proc. of Coling'94*, pp. 57–63.

Alshawi, Hiyan, Srinivas Bangalore & Shona Douglas: 2000, 'Learning dependency translation models as collections of finite-state head transducers', *Computational Linguistics*, **26**(1): 45–60.

Brill, Eric & Grace Ngai: 1999, 'Man vs. machine: A case study in base noun phrase learning', in *Proc. of ACL'99*, pp. 65–72.

Kitamura, Mihoko & Yuji Matsumoto: 1995, 'A machine translation system based on translation rules acquired from parallel corpora', in *Proc. of International Conference on Recent Advances in Natural Language Processing*, pp. 27–36.

Shirai, Satoshi, Satoru Ikehara, Akio Yokoo & Hiroko Inoue: 1995, 'The quantity of valency pattern pairs required for japanese to english mt and their compilation', in *Proc. NLPRS of '95*, pp. 443–448.

Streiter, Oliver, Leonid L. Iomdin, Munpyo Hong & Ute Hauck: 1999, 'Learning, forgetting and remembering: Statistical support for rule-based MT', in *Proc. of TMI'99*, pp. 44–54.

Sumita, Eiichiro, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa & Satoshi Shirai: 1999, 'Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach', in *Proc. of MT Summit VII*, pp. 229–235.

Tanaka, Hideki: 1994, 'Verbal case frame acquisition from a bilingual corpus: Gradual knowledge acquisition', in *Proc. of Coling'94*, pp. 727–731.

Watanabe, Hideo: 1993, 'A method for extracting translation patterns from translation examples', in *Proc. of TMI'93*, pp. 292–301.