Applying TDMT to Abstracts on Science and Technology

Hideki Kashioka
Kazutaka Takao
ATR Interpreting Telecommunications
Research Laboratories

Hiroko Ohta
Yoshiko Shirokizawa
Japan Science and
Technology Corporation

Abstract

In this paper, we discuss applying a translation model, "Transfer Driven Machine Translation" (TDMT), to document abstracts on science and technology. TDMT, a machine translation model, was developed by ATR-ITL to deal with dialogues in the travel domain. ATR-ITL has reported that the TDMT system efficiently translates multi-lingual spoken-dialogs. However, little is known about the ability of TDMT to translate written text translations; therefore, we examined TDMT with written text from English to Japanese, especially abstracts on science and technology produced by the Japan Science and Technology Corporation (JST). The experimental results show that TDMT can derive written text translation.

1 Introduction

ATR-ITL has developed the TDMT (Transfer Driven Machine Translation) system [Mima et al., 1998] as a prototype system for spoken dialogue translation. The TDMT system provides an efficient technique for handling spoken dialogues by using the following three features:

1. capability for ungrammatical expressions,
2. real-time translation, and
3. ways to smooth the addition of translation knowledge.

These features are important requirements for translating spoken dialogues. For written text translation, these features, especially the third, are necessities.

Text databases attract many people's attention because they are a convenient way to obtain information. Therefore, The Japan Science and Technology Corporation (JST) has produced a Japanese database composed of abstracts on science and technology [JST]. If the machine translation system has the ability to translate an English abstract into a Japanese abstract, then the process of producing Japanese abstracts can be carried out in less time than human translation. We examined the ability of EJ translation for written text with TDMT. In the next section of this paper, we give an overview of TDMT. Section 3 describes the characteristics of document abstracts on science and technology. Section 4 presents the points of system modification, and Section 5 describes an experiment for applying TDMT to written text. In Section 6, we state our conclusions.

2 Overview of TDMT

TDMT has three modules: morphological analysis, transfer, and sentence generation. The transfer module is an essential component that applies "transfer knowledge" to an input sequence that is an output of a morphological analysis module. In TDMT, a transfer module is different, from those of other translation systems. Almost all translation systems analyze the structure of an input expression with source language grammar, which is called a parsing phase. After that, they translate the source language structure into the target language structure with translation rules. Finally, they generate the target sentence from the target language structure. Most systems divide the parsing part and the structure translation part. TDMT simultaneously processes parsing and structure translation by using the "transfer knowledge."

The "transfer knowledge" is required to treat source/target language expressions at various linguistic levels, so it depends on both source and target languages.

In the following subsections, we briefly explain "transfer processing" and "transfer knowledge".

2.1 Processing Flow

TDMT translates an input sentence by following three steps: the system does a morphological analysis, it translates from the output of the morphological analysis to target language word sequences with a syntactic structure, and it generates a complete sentence.

1. Morphological Analysis
   As a connected speech recognition system, TDMT does not need to process morphological analysis because this system outputs word sequences with part-of-speech tags. However, TDMT should process morphological analysis to deal with written text. TDMT has a morphological analysis based on an N-gram model.
2. Transfer
   TDMT parses an input sentence based on constituent boundary parsing with transfer knowledge [Furuse and Iida, 1994] and selects a source language substructure by calculating the semantic distance [Sumita and Iida, 1991] between input word and example in transfer knowledge. From the viewpoint of parsing, TDMT derives possible structures by constituent boundary parsing. "Transfer Knowledge" plays the part of a grammar. When TDMT finishes parsing and selection, the selected transfer knowledge can make a target structure.
3. Generation
   TDMT processes word formation and a small change in the word order for target-structure-referenced target language grammar.

2.2 Transfer Knowledge

Transfer knowledge describes the correspondence between source-language expressions and target-language expressions at various linguistic levels. Transfer knowledge represents a source and target language pair that is expressed in patterns. A pattern is defined as a sequence that consists of variables and constituent boundary markers [Furuse and Iida, 1996] such as surface functional words. A variable is substituted for a linguistic constituent and expressed with a capital letter such as an *X*.

The transfer module uses transfer knowledge to analyze the source language structure and transfer it to the target language structure. For example, the English pattern *X to Y,* including the preposition *to,* is a very frequently used English expression. We can derive the following English-to-Japanese transfer knowledge about *X to Y* by compiling such translation examples as the source-target pairs of *go to the hotel — hoteru ni iku , listen to jazz → jyazu wo kiku,* etc.

X to Y =>
Y' ni X' *((go , hotel), ...)*, 'iku' 'hoteru',
Y' wo X' *((listen , jazz), ...)*, 'kiku' 'jyazu'
Y' to X' *((speak , friend), ...)*, 'hanasu' 'tomodachi'

Within this pattern, *X'* is the target word of X, and a corresponding Japanese word is written under each English word. For example, *hoteru* means 'hotel' and *iku* means 'go'.

3 Target Documents

JST produced the Japanese database of abstracts on science and technology by collecting worldwide publications. This database, called the "JICST file," can be accessed via network. JST translates some parts of the JICST file that were originally in Japanese into English. This English database is called the JICST-E file. JST provides the JICST-E file for the purpose of allowing non-Japanese-speaking scientists to access Japanese science and technology information.

This study uses about 1,500 sentences including keywords such as "computer"; the sentences were taken from the Information Science field in the JICST-E file. We made an alignment of these English sentences with the Japanese sentences in the JICST file for this experiment.

We list the characteristics of the abstracts on science and technology as the following.

1. Includes many long sentences
   (average: 20-30 words, e.g., on travel conversations 10 words.)
   Example:
   " "The basic construction of land record databases using the personal computer" is one subject out of the total subjects, and it has several purposes: the improvement of records as land record databases using personal computers, an attempt at popularizing land record databases in the regions, and assistance in the database enrichment in the region."
2. Includes many numerical expressions and much punctuation
   Example:
   "The WIDOM model gave *-1062.0.MU.kj/mol*
   and Shing and Gubbin model gave *-880.0.MU.kj/mol*
   as compared to the measured value of *-893.1.MU.kj/mol*
   "
3. Includes many compound words
   Example:
   "Research report on international cooperation relating to the image information technology."
   JST has a dictionary for science and technology that has entries for these compound words. Generally speaking, it is desirable to include compound words in the dictionary.
4. Includes many paratactic constructions.
   Examples:
   "This report summarizes *a systematic description method of operations aimed at standardizing word-processor operations* and *a method to*

derive draft standards of word-processor operations from the described results of operations."
"*Patella behavior in the largest extension of the knee joint was photographed by X-ray,* and *three-dimensional shapes of the pattela and thighbone were obtained from computed tomographic images of the knee joint."*

5. Includes many relative clauses.

Example:
"The main memory of card 386 employed 2 substrate sheet structure *in which* the memory module mounted on the surface of 2 layer substrate was connected with the main substrate using flexible tape."

6. Includes many common expressions of technical papers at the top of the sentences.

Examples:
"This paper explains ~"
"This paper describes ~"
"This paper presents ~"

4 TDMT Modification

In this section, we show how to modify TDMT for application to abstracts on science and technology.

4.1 Expanding Transfer Knowledge

The transfer knowledge of the TDMT is derived from real data. The knowledge for spoken dialogue was derived from the dialogue database for the travel conversations collected by ATR-ITL. Therefore, any additional abstracts on science and technology help in expanding the transfer knowledge.

1. The selection for the English and Japanese parallel database is composed of abstracts on science and technology.

   These data are selected from the JICST-E file. Each abstract has one title and three sentences on average.

   Example:

   (**TITLE**) Commodity Trend and Technological Trend of "X" Window Terminal.

   (**Sent 1**) The products with higher performance of over 200,000 Xstone values are shipped and their performance will be improved by advanced CPUs and graphics accelerators in the future.

   (**Sent 2**) X11R6 is the main function and some multi-media functions are incorporated.

   (**Sent 3**) The multi-media functions will be enhanced in the future.

   These selected abstracts should annotate the part-of-speech tags.

2. Making Tagging Data

   Tagging data were made for expanding effective transfer knowledge. The selected abstracts are analyzed by a morphological analysis in TDMT, and humans check the annotated tags because the tag information is important in pattern matching for the selection of transfer knowledge. If we expand transfer knowledge from incorrectly tagged text, that knowledge would have the wrong effect with correct tagged text.

3. Adding an entry into dictionaries

   If there are new entry words in the tagged data, these new entries are added to dictionaries such as the English-Japanese word dictionary and the English semantic code dictionary. Each entry in the English-Japanese word dictionary consists of four columns: English-POS, English-word, Japanese-POS and Japanese-word. The English semantic code dictionary has three columns for each entry: English-POS, English-word and semantic-code-lists.

   Example:

   English-Japanese word dictionary
   ((CN "airtightness") (CN *"kimitsuka"))*
   ((CN "algographics, co., ltd.")
   (CN *"(kabu) arugo guraphikkusu"))*

   English semantic code dictionary
   ("air-tightness" CN "236")
   ("algographics, co., ltd." CN "713")

4. Extending Transfer Knowledge

   The next step is extending the transfer knowledge. The rule-writer who writes the transfer knowledge translates the sentence before he/she writes the knowledge about it. He/She checks the source language structure, the target language structure, and the translated target sentence. If the source language structure is wrong, he/she checks the current pattern in transfer knowledge. When the selection of transfer knowledge fails, he/she adds an example into the appropriate pattern.

   Example of adding examples:

   ```
   ( X "as" Y)                    =>
     Y' "tosite no" X'
     (("RS6000"                   ,   "ews "),
     ("letter"                    ,   "proof"),
     ( "personal computer"        ,   "instrument"),
     ("plan"                      ,   "communication system")
   ```

When the appropriate pattern in the transfer knowledge is missing, he/she writes new patterns and examples.

Example of new pattern and examples:

```
( X "in accordance with" Y)      =>
    Y' "ni taiou suru" X'
    (( "sensor"   ,              "purpose"),
     ("system"    ,              "environment"))
```

5 Experiment

To measure the ability of TDMT with written text, we compared four systems by calculating each subpoint of the translations from each system. One system was TDMT and the others were commercial systems. If any system has the advantage of word entries, we cannot compare the experimental results for these systems evenly. Therefore, entries for previously out-of-vocabulary words were added to the lexicons of all participating systems.

Table 1: Evaluation basis

Evaluation	Point	basis
a	5	No post processing (perfect translation)
b	3	Post processing for about 20% correction
c	2	Post processing for about 50% correction
d	1	Post processing for about 80% correction

5.1 Conditions for Evaluation

One-hundred sentences were taken from the JICST-E file. The test set data had different sentences from the training data.

Each sentence that the translation system output was assigned to one of four grades for translation quality. The evaluation basis is shown in Table 1. This evaluation was performed by the staff members who write abstracts at JST.

Each grade was assigned a point, and we calculated the total points of each system. Then we compared four systems for total points. The TDMT were measured according to the relationship between the training size and the translation grade. In addition, we measured the postprocessing because the translations were not correct sentences and they needed editing to produce a database of high quality.

Table 2: Results of grading (number of sentences)

	a	b	c	d	total points
TDMT	40	30	23	7	343
X	21	20	26	33	250
Y	24	23	29	24	271
Z	29	22	21	28	281

Table 3: Range of semantic distance and human grade

evaluation	number of sentences	semantic distance range
a	40	0.000005 - 8.36672
b	30	0.66667 - 13.00006
c	23	2.19445 - 25.83255
d	7	6.70572 - 14.27191

5.2 Experimental Results

The evaluation results are shown in Table 2. TDMT achieved 40% accuracy while the other systems were less than 30%. For total points, the TDMT received 343 points but each of the others had less than 300 points. For the number of sentences assigned grade "d", TDMT only had 7. In contrast, other systems had 24 or more.

We changed the point of view to postprocessing. It was more useful to automatically assign the translation a rank that indicated the necessity of postprocessing. TDMT translates by calculating the semantic distance, so TDMT outputs a pair of target language sentences and its semantic distance. The semantic distauce was 8.291042 in the following example.

Input Sentence: A master-slave manipulator, based on a decentralized control, was developed using cost-effective microcomputer-based memory-to-memory communication method.

translation: “分散する制御に基づくマスター・スレーブ・マニピュレータは、費用対効果の高いマイクロコンピュータベースメモリー間通信方法の利用によって開発されました。”, 8.291042

Figure 1 and Table 3 show the relationship between the grade and semantic distance.

Thus, we found a week relationship between the grade assigned by a human and the semantic distance from this experimental result. However, it was difficult to set the necessary threshold for the discrimination of postprocessing.

We now turn to a discussion about the difference in processing time. Table 4 shows the different processing times between the human translation and the TDMT translation.

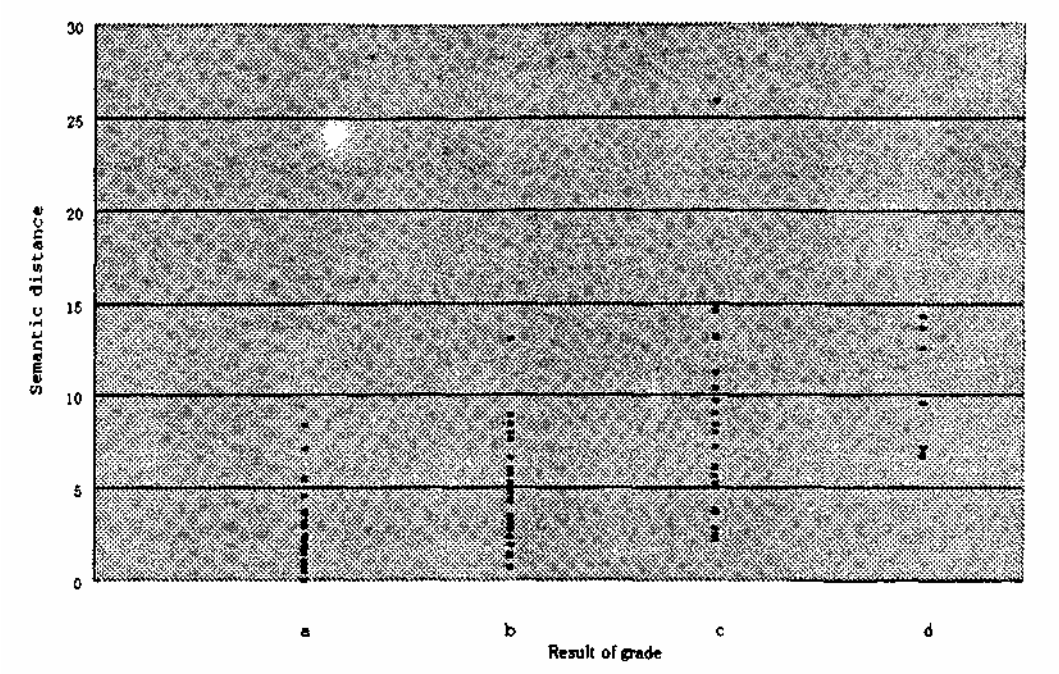

Figure 1: Correlation of evaluation and semantic distance

Table 4: processing time per sentence

	average time
Human translation	95 sec
TDMT and postprocessing	56 sec

6 Discussion

According to this experimental result, we can say that TDMT is an efficient model. However, many problems still exist. In this section, we discuss these problems.

6.1 Disambiguation of Word Semantics

Currently, TDMT has an English-to-Japanese dictionary that has each English word with only one Japanese word. An English word that has a different meaning is assigned one default Japanese word.

However, almost all English words have several meanings that have different expressions or words in Japanese. For example, the words "operation"and "scale" have many meanings that represent different expressions or words in Japanese:

operation 演技:手術操作:手術作動:稼動:か動:か働:業務活動:作働:動作:術:操業:運転操作:運航面:手術:操作途中:運用:切開手:オペレーション:運行:操業法:運転:操作:施術:操縦法:運航:演算:施行

scale 鱗:尺度:ドレミファ:目盛り:目盛:スケール:板状鱗層:鱗粉:規模: 鱗層:りんせつ:りん片:鱗片:鱗片:体重計:測定尺度:ヘルスメーター: ウロコ:うろこ:鱗:脚鱗:階段標準:湯あか:湯アカ:計量装置

We took this example from a Japanese-English word dictionary produced by JST. Some of these words can be selected by the field of the document. However, the field of the document does not give enough information to select the correct meaning. Current TDMT handles word selection with specific examples in the transfer knowledge. Consequently, additional techniques are required for word selection.

6.2 Semantic Code

Translation processing depends on the semantic distance calculated with the semantic code of the word. The semantic code is defined in the thesaurus. The current TDMT system calculates the semantic distance based on KADOKAWA RUIGO SINJITEN. The necessary depth and particle size of the thesaurus to appropriate a translation are not clear, and these depend on the field of the documents. If we adopt TDMT to translate abstracts on science and technology, we should use another thesaurus for semantic distance calculations. It should be noted that JST now has a thesaurus and a semantic code for translation from Japanese to English.

6.3 Selection from Multiple Output

TDMT outputs different translation results if each of the results has the same semantic distance. However,

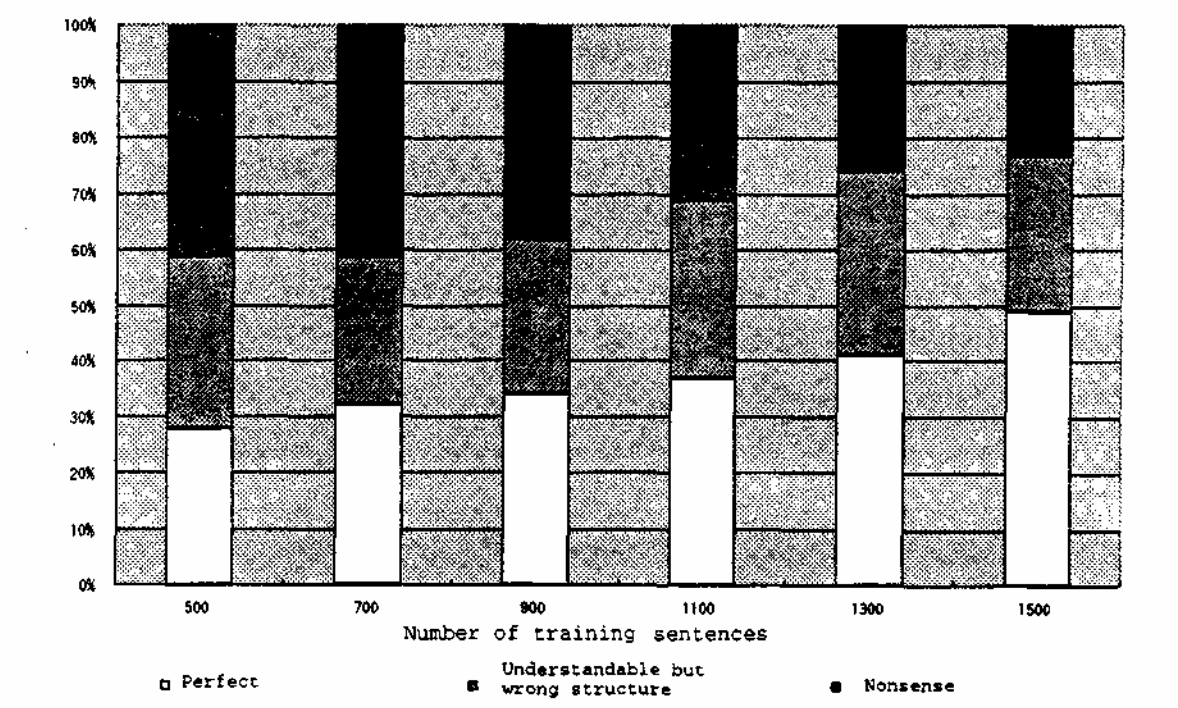

Figure 2: Correlation of evaluation and training size

TDMT should select one translation result for the application design. The current system does not have a measure for selecting good translations. TDMT should have such a measure that is based on the semantic distance.

6.4 Paratactic Construction

We previously described the sentence length in the characteristics of this domain. A long sentence, especially in this domain, has a paratactic construction that is difficult to analyze with the current technique. There are some marks (punctuation) or some key words for finding the paratactic construction. However, there are some problems with punctuation or key words.

6.5 Numerical and other Symbols

The target sentences, that is, abstracts on science and technology, include a lot of signs, numerical expressions and unit measure expression. There are many such expressions, and it is difficult to change these expressions into tokens. If the system can change these expressions into tokens, the syntactic structure has many ambiguities.

These expressions usually represent the same expressions in English and Japanese. Thus, they are translated into the same expression through the transfer module. However, the roles of these expressions in the sentences still have ambiguity.

6.6 Postprocessing

If we produce a database of abstracts via TDMT translation, TDMT translations are useful but require postprocessing to correct the Japanese sentences. In this experiment, there is no need for postprocessing in about *40%* of the TDMT translations. By avoiding postprocessing in these correct sentences, we could find out whether the output is correct. TDMT outputs the translation and its semantic distance.

Figure 1 shows the relationship between the semantic distance and the evaluation grade. According to this graph, we cannot avoid postprocessing in a sentence if its semantic distance is under the specific threshold. In the future, TDMT should have enough examples so that it can find a specific threshold to avoid postprocessing for correct translations.

From another point of view, TDMT knows the expression that the system selects. The system calculates all of the semantic distances of the used patterns. Thus, to cut costs for postprocessing, we should display colors with the expression the system selected, which would also be useful for the interface.

6.7 Training Size

Figure 2 shows the increase in ability with the training size in TDMT. Table 5 describes the statistics of the extended transfer knowledge.

This experimental result does not saturate an ability. It needs more training data and the relationship between the size of training data and its ability should

Table 5: Additional knowledge statistics

number of sentences	patterns	examples
(Initial State) 2,800	1,375	10,300
+ 500 (3,300)	151	4,032
+ 200 (3,500)	60	2,262
+ 200 (3,700)	38	1,641
+ 200 (3,900)	34	1,946
+ 200 (4,100)	30	1,726
+ 200 (4,300)	25	1,562
(total) 4,300	1,713	23,469

be clarified. At the same time, we need to check the size of expanding transfer knowledge. In this experiment, the size of additional patterns and examples increase relative to the number of training sentences. Currently, we cannot ascertain a relationship between training data size, accuracy and the size of transfer knowledge, especially examples. We need to observe the relationships among them for least 1,000 more training sentences.

7 Conclusions

In this experiment, we found that TDMT was effective in producing a database that consists of abstracts on science and technology. The frequency of patterns in the transfer knowledge by using at the translation was more or less different. However, the patterns in the transfer knowledge were basically useful for both written text and spoken dialogue transfers.

The TDMT system for written text does not have enough training with written text to make better translations. TDMT needs more specific patterns or examples in the transfer knowledge for written text, especially abstracts on science and technology. We need to extend such transfer knowledge such as the dictionaries, source and/or target expression patterns, and examples. The addition of transfer knowledge could be a smooth and effective means of improving translations. We concluded that TDMT is a promising approach as a transfer model for written text.

References

[Furuse and Iida, 1994] Furuse, O. and Iida, H., Constituent Boundary Parsing for Example-Based Machine Translation, In *Proceedings of Coling '94,* pages 105-111, 1994.

[Furuse and Iida, 1996] Osamu Furuse and Hitoshi Iida, Incremental Translation Utilizing Constituent Boundary Patterns, In *Proceedings of Coling '96,* pages 412-417, 1996.

[JST] HOME PAGE of JICST Service Guide, http://\vww.jst.go.jp/EN/JICST/ServiceGuide/ .

[Kawai et al., 1995] Kawai, J., Wakita, Y. and Iida, H., Stochastic Language Model Using Semantic Category and Mixed Category of Words and Parts-of-Speech for Speech Understanding, In *Proceedings of NLPRS '95,* pages 107-111, 1995.

[Mima et al., 1998]
Mima, H., Furuse, O., Wakita, Y. and Iida, H., MULTI-LINGUAL SPOKEN DIALOG TRANSLATION SYSTEM USING TRANSFER-DRIVEN MACHINE TRANSLATION, MT Summit VI Proceedings, pages 148-155, 1997.

[Sumita and Iida, 1991] Sumita, E. and Iida, H., Experiments and Prospects of Example-based Machine Translation, In *Proceedings of 29th ACL,* pages 185-192, 1991.