CTM: An Example-Based Translation Aid System

Satoshi SATO *
School of Information Science
Japan Institute of Science and Technology, Hokuriku
Tatsumokuchi, Ishikawa, 923 12, Japan
sato@jaist-east.ac.jp

Abstract

This paper describes a Japanese-English translation aid system, CTM, which has a useful capability for flexible retrieval of texts from bilingual corpora or translation databases. Translation examples (pairs of a text and its translation equivalent) are very helpful for us to translate the similar text. Our character-based best match retrieval method can retrieve translation examples similar to the given input. This method has the following advantages: (1) this method accepts *free-style* translation examples, i.e., pairs of any text string and its translation equivalent, (2) morphological analysis is unnecessary, (3) this method accepts *free-style* inputs (i.e., any text strings) for retrieval. We show the retrieval examples with the following characteristic features: phrasal expression, long-distance dependency, idiom, synonym, and semantic ambiguity.

1 Introduction

In the late 1980's, several commercial Japanese-English machine translation systems had been developed in Japan. In these systems, the computer is the agent of translation, while the user assists in editing the translation inputs and revising the results. Although they are useful to translate large amounts of texts roughly and rapidly, high quality translation is impossible.

Translation aid is another kind of machine translation: the user is the agent of translation, while the computer provides him or her with the helpful tools, e.g., quick-retrieval electronic dictionaries. A quick-retrieval bilingual corpus is also useful, specifically when it has the flexible (best match) retrieval mechanism. Because translation examples (pairs of source text and its translation equivalent) are very helpful for us to translate the similar text. This type of system is called as example-based translation aid [6], and there are two prototype systems in Japanese-English translation: ETOC [8] and Nakamura's system [5].

*The author had been transferred from Kyoto University on April 1, 1992. This work was done at Kyoto University.*
2. The Character-Based Best Match Retrieval Method

2.1 Characteristics of Japanese Written Texts

Japanese written texts have remarkable characteristics as follows. They cannot be found in European languages, i.e., English, French, and German.

1. The number of characters is very large.

The number of characters that are used in text is more than 7,000 in Japanese, while it is less than a hundred in a European language.

2. Synonyms often have the same Kanji character.

Japanese characters are divided into three types: Hiragana (83 characters), Katakana (86 characters), and Kanji. A Hiragana or Katakana character expresses a sound, and a Kanji character represents a semantic primitive. For example, the Kanji character "思" means "thinking," and it is used for constructing several words concerned with thinking; e.g., 思考 (thinking), 考察 (consideration), 深考 (deep thinking), 考える (think), 考案する (devise).

3. There is no delimiter between words.

In European languages, the white space is the delimiter for word separation. In contrast, Japanese has no explicit delimiter. Therefore, the main part of Japanese morphological analysis is to divide a text string into words: it is not an easy task.

These characteristics of Japanese suggest the character-based best match, because

1. While the word-based method needs morphological analysis, the character-based method does not need it.

2. In order to retrieve synonyms the word-based method needs a thesaurus. In contrast, the character-based method can retrieve some kind of synonyms without a thesaurus, because synonyms often have the same Kanji character in Japanese.

2.2 The Character-Based Best Match

The character-based best match can be determined by defining the distance or similarity measure between two strings.

The simple measure of similarity between two strings, \( A = a_1a_2\cdots a_x, \) \( B = b_1b_2\cdots b_y, \) is the number of the matching characters considering the character order constraint. It is not particularly good measure, but makes a convenient starting point. We define it as follows:

\[
S(A, B) = s(x, y)
\]

\[
s(i, j) = \begin{cases} 
0 & \text{if } i = 0 \lor j = 0 \\
\max \left( s(i - 1, j - 1) + \min(\epsilon m(i, j), W) \right) & \text{if } (1 \leq i \leq x) \land (1 \leq j \leq y) \\
\end{cases}
\]

\[
m(i, j) = \begin{cases} 
1 & \text{if } a_i = b_j \\
0 & \text{if } a_i \neq b_j \\
\end{cases}
\]

This measure often produces the undesirable results, because we ignore continuation of matching characters. For example, consider the following strings:

\( A = \) 問題を解決する (solve the problem)

\( B = \) 彼はきのう問題を解決した。

(He solved the problem yesterday.)

\( B' = \) 問題の解決を決定する

(determine the method for solving the problem)

We want to have \( S(A, B) > S(A, B'), \) but the above measure produces \( S(A, B) < S(A, B'). \) To solve the problem, we consider the bonus for continuous matching characters. It can be done by modifying \( m(i, j) \) in the above definition:

\[
S(A, B) = s(x, y)
\]

\[
s(i, j) = \begin{cases} 
0 & \text{if } i = 0 \lor j = 0 \\
\max \left( s(i - 1, j - 1) + \min(\epsilon m(i, j), W) \right) & \text{if } (1 \leq i \leq x) \land (1 \leq j \leq y) \\
\end{cases}
\]

\[
\epsilon m(i, j) = \begin{cases} 
0 & \text{if } i = 0 \lor j = 0 \\
\epsilon m(i - 1, j - 1) + m(i, j) & \text{if } (1 \leq i \leq x) \land (1 \leq j \leq y) \\
\end{cases}
\]

\[
m(i, j) = \begin{cases} 
1 & \text{if } a_i = b_j \\
0 & \text{if } a_i \neq b_j \\
\end{cases}
\]

This is the similarity score that we use, where \( W \) is a parameter that determines the maximum value of the bonus for the continuous matching characters. When \( W = 1, \) this definition is the same with the previous definition. Table 1 shows \( S(A, H) \) and \( S(A, H') \) with varying values of \( W. \) Usually we use \( W = 4. \)
### 2.3 Acceleration by Character Index

At the best match retrieval, we use the acceleration method using the character index. The character index is the table of every character with ID's of examples in which the character is appeared. Table 2 shows an example of translation database and Table 3 shows the character index of it.

In the first stage of the retrieval, the character index is used for the pre-selection of the examples. Figure 2 illustrates the pre-selection process: it is

1. Look up the records for the characters that are appeared in the input string.
2. For every example, compute the pre-selection score, \( PSS \), which can be obtained by counting the number of the example ID's in the records. It is the number of matching characters between the input string and the example ignoring the character order constraint.
3. Select the top \( N \) examples that have the largest pre-selection score, where \( N \) is the parameter and we usually use \( N = 200 \).

In the second stage of the retrieval, the similarity scores of pre-selected examples are computed, and the examples are ordered by the score.

### 3 The CTM System

Above mentioned retrieval mechanism has been implemented in CTM, a Japanese-English translation aid system. CTM is written by C and runs on Sun Workstations. Figure 3 shows the configuration of CTM: it consists of three programs.

- **mkdb**: The program to create the character index from the translation database.
- **CTM server**: The main program, which retrieves the best matched examples with the given input.
- **MTC**: The client program on NEmacs (Nihongo (Japanese) GNU Emacs), which interacts the CTM server via Ethernet.

The translation database of CTM is text files, in which a Japanese text string and an English text string appear one after the other. These files can be made from Japanese text files and the correspondent English text files by using the alignment program [1] semi-automatically. We have made the translation database from several sources: Table 4 shows our translation databases.

### 4 Retrieval Examples

We show here CTM retrieval examples with the following features: phrasal expression, long-distance dependency, idiom, synonym, and semantic ambiguity.

Figure 4 shows a retrieval example of phrasal expression “いくつかの観点から考察する (consider from several points of view)”. Although there is no exact matched expression in the database, CTM can retrieve helpful examples for us to translate it.

---

5. We cannot compute the similarity score of every example in the database, because the computation needs about 5 millisecond between the average input string (10 characters) and the average example (50 characters) on SparcStation 2.
6. This value was determined empirically.
### Table 4: The CTM Translation Databases

<table>
<thead>
<tr>
<th>Name</th>
<th>Direction</th>
<th>Records</th>
<th>K Byte</th>
<th>Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScienceYYMM</td>
<td>E→J</td>
<td>11,115</td>
<td>3,175</td>
<td>Scientific American &amp; its Japanese translation (Nikkei Science)</td>
</tr>
<tr>
<td>JK</td>
<td>J→E</td>
<td>4,230</td>
<td>139</td>
<td>Entry words on [4]</td>
</tr>
<tr>
<td>MTE</td>
<td>J→E</td>
<td>3,938</td>
<td>379</td>
<td>Test examples on [2]</td>
</tr>
<tr>
<td>EX</td>
<td>J→E</td>
<td>6,624</td>
<td>595</td>
<td>Translation examples collected by Oikawa</td>
</tr>
<tr>
<td>TJ</td>
<td>J→E</td>
<td>1,467</td>
<td>259</td>
<td>The column, Teusei-Jingo, on Asahi Newspaper</td>
</tr>
<tr>
<td>KD</td>
<td>J→E</td>
<td>38,190</td>
<td>2,729</td>
<td>Examples on [7]</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>67,619</td>
<td>7,733</td>
<td></td>
</tr>
</tbody>
</table>

### CTM(Ab)>いくつかの観点から考察する

Score = 28, DB = Science8710, ID = 558, File = 03.ej  
このようにいくつかの観点から観察すると、ガリウムアリシンはシリコンに対する速度という点では優位に立つ。From the viewpoint of several material limits, then, gallium arsenide offers advantages over silicon in speed.

### CTM(Ab)>決してないと

Score = 9, DB = Science8710, ID = 1649, File = 07.ej  
これは決して小さな事実にはならない。むしろ、社会の大きな需要を生み、成功することが予想される。  
This is not such a small undertaking, however, and success presupposes that society generates significant demand.

### CTM(Ab)>しっかりつかむ

Score = 18, DB = MTE, ID = 79, File = mtttest.je  
私は猫のしっぽをつかんだ。  
I grasped the tail of a cat.

### CTM(Ab)>考察する

Score = 10, DB = MLI, ID = 605, File = 03.ej  
その中でも特に、ある概念の練習例をすべて含めるような、最も特徴化された意味的・一般化（MSC・一般化）を見つけための手法を考察する。In particular, we examine methods for finding the maximally-specific conjunctive generalizations (MSC-generalizations) that cover all of the training examples of a given concept.

### CTM(Ab)>考察する

Score = 7, DB = Science9003, ID = 468, File = mental.ej  
おそらく、治療者が解釈は、患者の意識的な思考や感情や行動に対する無意識の心の影響を、患者自身が洞察するのを助けるだろう。Presumably the therapist’s interpretations help patients to gain insight into the effects of the unconscious mind on their conscious thoughts, feelings and behaviors.

### CTM(Ab)>考察する

Score = 6, DB = MLI, ID = 147, File = 01.ej  
能動的実験で、学習者は環境をかき乱してその乱れの結果を観察する。  
*Active experimentation, where the learner perturbs the environment to observe the results of its perturbations.*

### Figures

**Figure 4: Example (Phrasal Expression)**

**Figure 5: Example (Long-Distance Dependency)**

**Figure 6: Example (Idiom)**

**Figure 7: Example (Synonym)**

with “考察する (consider/examine)” and two examples with two synonyms, “洞察する (gain insight into)” and “観察する (observe)”.

Figure 8 shows three retrieval examples for the Japanese construction “NOUN/ +/- + + - + o; +t”, where “に” is a case marker and “入った” is the past form of the verb “入る”。There are several translation of “入る”。The first input “事務室 (office) に入った” has two meaning: one is “entered the office” and the other is “joined as a new member of the office”. The second input “耳 (ear) に入ったら” is an idiomatic expression that means “heard”. The last input “本屋 (bookstore) に入ったら” is more complicated: the translation depends on not only “に (ni)”-case but also “が (ga)”-case. The retrieval examples show the following three cases:

1. “人 (human) が + 部屋 (room) に入る”  
   (human enters the room)

2. “風 (wind) が + 部屋 (room) に入る”  
   (the wind blows into the room)

3. “本 (book) が + 本屋 (bookstore) に入る”  
   (the book arrives at the bookstore)
5 Evaluation

It is very difficult to evaluate a translation aid system, because its effectiveness essentially depends on the user's satisfaction: when the user feels that the system is helpful, it is effective. The evaluation of CTM is now in progress, and we show some results of experiments here.

The Retrieval Time

Empirically, we obtained the following equation, which estimates the retrieval time (milliseconds).

\[ \text{time}(l, k, N) = l \times (10 \times k + 2/3 \times N) \]

where \( l \) is the length of the input string, \( k \) (mega byte) is the database size, and \( N \) is the pre-selection parameter. For example, if \( l = 10 \) (characters), \( k = 8 \) (mega byte), \( N = 200 \), then \( \text{time} = 2,133 \) (milliseconds). It shows that the current system responds in a few seconds and it is not so fast. The more acceleration is need for the larger database.

Evaluation of 100 retrievals

We have evaluated 100 retrieval results by hand. We have given one of the following grades to each retrieved example.

A The example exactly matches the input.

B The example provides enough information about the translation of the whole input.

C The example provides information about the translation of some part of the input.

Table 5: Evaluation of 100 retrievals

<table>
<thead>
<tr>
<th>Character Length</th>
<th>Grade A</th>
<th>Grade B</th>
<th>Grade C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>21</td>
<td>6</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>6-10</td>
<td>4</td>
<td>10</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>11-15</td>
<td>1</td>
<td>15</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>16-20</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>20-30</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>35</td>
<td>16</td>
<td>86</td>
</tr>
</tbody>
</table>

The example provides almost no information about the translation of the input.

We evaluated top five examples for each retrieval, and the best grade of them is used for the evaluation of a retrieval. Table 5 shows the result of the evaluation. The table shows that (1) we can obtain very useful information from 47% of the retrievals, (2) we can obtain at least some information from 81% of the retrievals.

Acknowledgments

The author would like to thank Ms. Yuko Tomita, who helped me to make the translation databases.

References


---

ACtES DE COLING-92, NANTES, 23-28 AOÛT 1992 1 2 6 3  
PROc. OF COLING-92, NANTES, AUG. 23-28, 1992