CTM: An Example-Based Translation Aid System

Satoshi SATO *
School of Information Science
Japan Institute of Science and Technology, Hokuriku
Tatsunokuchi, 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].

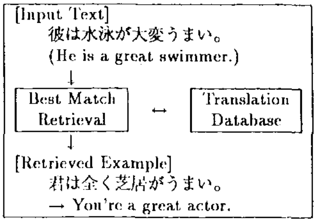


Figure 1: Basic Configuration of Example-Based Translation Aid

Figure 1 shows the basic configuration of example-based translation aid (EBTA). It consists of two components: the **translation database** is the collection of translation examples, and the **best match retrieval** engine is to retrieve the example that is the most similar to the given input text. The characteristic of the EBTA system is that it accepts *free-style* text inputs for the retrieval: it frees the user from learning the formal language for database query.

The central problem in EBTA is the implementation of the best match retrieval. Two methods were proposed: one is the syntax-matching driven by generalization rules in ETOC [8], and the other is Nakamura's method using content words [5]. They are the **word-based best match retrieval** methods[1], which need morphological analysis.

This paper proposes the **character-based best match retrieval** method, specifically for Japanese texts. Compared with the word-based methods, the character-based method has the following advantages:

- Morphological analysis is unnecessary.
- Some kind of synonyms can be retrieved without a thesaurus.

This method has been implemented in CTM[2], a Japanese-English translation aid system for writing/translating technical papers.

*The author had been transferred from Kyoto University on April 1, 1992. This work was done at Kyoto University.

[1] In word-based (resp. character-based) best match retrieval method, a word (resp. character) is a primitive.

[2] CTM is named from the Japanese phrase, "Chotto Tsukatte Mitene", which means "use it any time you want".

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 "考" means "thinking", and it is used for constructing several words concerned with thinking: e.g., 思考(thinking), 考察 (consideration), 熟考(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 easy task[3].

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, $A = a_1 a_2 \cdots a_x$, $B = b_1 b_2 \cdots b_y$, 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:

$$S(A,B) = s(x,y)$$

$$s(i,j) = \begin{cases} 0 & \text{if } i = 0 \lor j = 0 \\ \max \begin{pmatrix} s(i-1,j-1) + m(i,j), \\ s(i-1,j), \\ s(i,j-1) \end{pmatrix} \\ & \text{if } (1 \le i \le x) \land (1 \le j \le y) \end{cases}$$

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

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.)

$B' =$ 問題の解法を 決定する
(determine the method for solving the problem)

We want to be $S(A,B) > S(A,B')$, 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 the above definition:

$$S(A,B) = s(x,y)$$

$$s(i,j) = \begin{cases} 0 & \text{if } i = 0 \lor j = 0 \\ \max \begin{pmatrix} s(i-1,j-1) + \min(cm(i,j), W) \\ s(i-1,j), \\ s(i,j-1) \end{pmatrix} \\ & \text{if } (1 \le i \le x) \land (1 \le j \le y) \end{cases}$$

$$cm(i,j) = \begin{cases} 0 & \text{if } i = 0 \lor j = 0 \\ cm(i-1,j-1) + m(i,j) \\ \quad \text{if } (1 \le i \le x) \land (1 \le j \le y) \end{cases}$$

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

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 $S(A,B)$ and $S(A,B')$ with varying values of W. Usually we use $W = 4$.[4]

[3] For example, a Japanese morphological analysis program developed by Kyoto University fails to analyze 3 ~ 5 % of sentences.

[4] This value was determined empirically. It may be explained as follows. The average character length of a Japanese word is about two, and we *feel* that the continuous matching of two words is the strong match.

Table 1: Scores vs. W

W	1	2	3	4	5
$S(A,B)$	5	9	12	14	15
$S(A,B')$	7	9	9	9	9

Table 2: Translation Database

ID	Japanese	English
1	いくつかの	several
2	いつでも	every time
3	いつか	some day
4	きのう	yesterday

Table 3: Character Index

Ch.	ID's	Ch.	ID's
い	1, 2, 3	つ	1, 2, 3
う	4	で	2
か	1, 3	の	1, 4
き	4	も	2
く	1		

2.3 Acceleration by Character Index

At the best match retrieval, we use the acceleration method using the character index.[5] 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 $N = 200$. [6]

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.

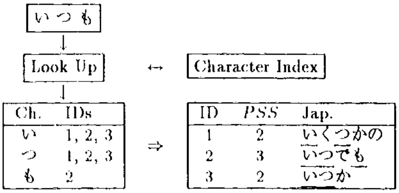


Figure 2: Pre-selection using Character Index

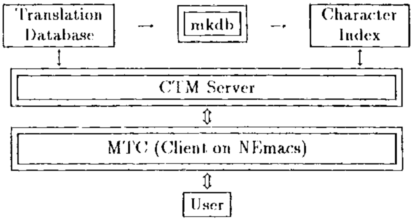


Figure 3: The CTM system

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.[7]

MTC [8] 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.

[7] The CTM server has other facilities: the character-based exact match retrieval for Japanese texts, and the word-based best or exact match retrieval for English texts.

[8] MTC is named from the Japanese phrase, "Motto Tsukatte Choudai", which means "use it more and more".

Table 4: The CTM Translation Databases

Name	Direction	Records	K Byte	Source(s)
ScienceYYMM	E→J	11,115	3,175	Scientific American & its Japanese translation (Nikkei Science)
ML1	E→J	2,055	458	Chap. 1 - 4 in Machine Learning [3] & its Japanese translation
JK	J→E	4,230	139	Entry words on [4]
MTE	J→E	3,938	379	Test examples on [2]
EX	J→E	6,624	595	Translation examples collected by Oikawa
TJ	J→E	1,467	259	The column, Tensei-Jingo, on Asahi Newspaper
KD	J→E	38,190	2,729	Examples on [7]
Total		67,619	7,733	

CTM(Ab)>いくつかの観点から考察する
Score = 28, DB = Science8710, ID = 598, File = 03.ej このようにいくつかの材料的制約の観点から見ると、ガリウムヒ素はシリコンに対して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.
Score = 24, DB = Science8710, ID = 549, File = 03.ej これらの5つのレベルそれぞれに対し、3つの観点から考察を行なうことができる。その3つとは、理論的考察、実際的考察、および歴史的対比である。 Each level of the hierarchy can be considered from three different points of view, which are respectively theory, practice and historical analogy.

Figure 4: Example (Phrasal Expression)

CTM(Ab)>決して␣ない
Score = 9, DB = Science8710, ID = 1649, File = 07.ej これは決して小さな事業にはならない。むしろ、社会が大きな需要を生み、成功することが予見される。 This is no small undertaking, however, and success presupposes that society generates significant demand.
Score = 9, DB = Science8710, ID = 1944, File = 09.ej これは、情報センターとして機能してきた従来の図書館のモデルと決して対立するものではない。 This view is not really in conflict with the traditional model of medical libraries as information centers.

Figure 5: Example (Long-Distance Dependency)

CTM(Ab)>しっぽをつかむ
Score = 18, DB = MTE, ID = 79, File = mttest.je 私は猫のしっぽをつかんだ。 I grasped the tail of a cat.
Score = 18, DB = MTE, ID = 78, File = mttest.je 私は彼のしっぽをつかんだ。 I found his weak point.

Figure 6: Example (Idiom)

CTM supports the retrieval of long-distance dependency: Figure 5 shows a retrieval example, where "決して" is an adverb, and "ない" is an auxiliary adjective for negation, and they are often used together with the general meaning "never".

CTM also supports the retrieval of idiomatic expression: Figure 6 shows an example. In this figure, the first retrieval example is the literal meaning, and the second is the idiomatic meaning.

The character-based best match method can retrieve synonyms. Figure 7 shows an example: in this case, CTM retrieved an exact match example 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)".

CTM(Ab)>考察する
Score = 10, DB = ML1, ID = 605, File = 03.ej その中でも特に，ある概念の練習例をすべて包含するような，最も特定化された連言的一般化 (MSC 一般化) を見つける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.
Score = 7, DB = Science9003, ID = 468, File = mental.e.ej おそらく、治療者の解釈は、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.
Score = 6, DB = ML1, 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.

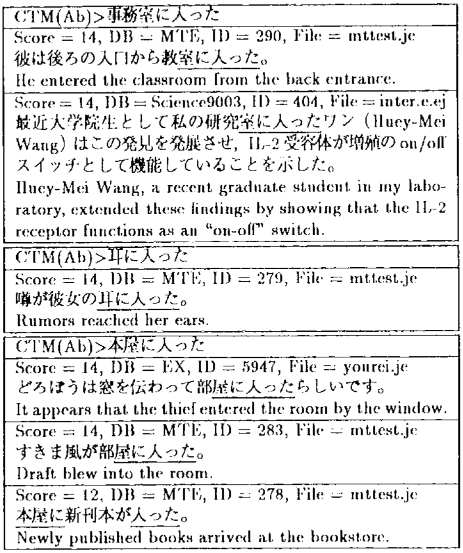
Figure 7: Example (Synonym)

Figure 8 shows three retrieval examples for the Japanese construction "NOUN+に+入った", 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" is an idiomatic expression that means "heard". The last input "本屋 (bookstore) に入った" 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)



```
CTM(Ab)>事務室に入った
Score = 14, DB = MTE, ID = 290, File = mttest.je
彼は後ろの入口から教室に入った。
He entered the classroom from the back entrance.
Score = 14, DB = Science9003, ID = 404, File = inter.e.ej
最近大学院生として私の研究室に入ったワン（Huey-Mei
Wang）はこの発見を発展させ，IL-2 受容体が増殖の on/off
スイッチとして機能していることを示した。
Huey-Mei Wang, a recent graduate student in my labo-
ratory, extended these findings by showing that the IL-2
receptor functions as an "on-off" switch.

CTM(Ab)>耳に入った
Score = 14, DB = MTE, ID = 279, File = mttest.je
噂が彼女の耳に入った。
Rumors reached her ears.

CTM(Ab)>本屋に入った
Score = 14, DB = EX, ID = 5947, File = yourei.je
どろぼうは窓を伝わって部屋に入ったらしいです。
It appears that the thief entered the room by the window.
Score = 14, DB = MTE, ID = 283, File = mttest.je
すきま風が部屋に入った。
Draft blew into the room.
Score = 12, DB = MTE, ID = 278, File = mttest.je
本屋に新刊本が入った。
Newly published books arrived at the bookstore.
```

Figure 8: Example (Ambiguity)

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 (millisecond).

$$time(l, k, N) = l \times (10 \times k + 2/3 \times N)$$

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 $time = 2,133$ (millisecond). It shows that the current system responses 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.

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

Table 5: Evaluation of 100 retrievals

	Character Length					
Grade	1–5	6–10	10–15	15–20	20–30	Total
A	21	6	0	0	0	27
B	4	10	3	2	1	20
C	1	15	10	6	2	34
F	9	4	3	0	3	19
Total	35	35	16	8	6	100

We evaluated top five examples for each retrieval, and the best grade of them is used for the evaluation of a retrieval.[9] 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

[1] Gale, W. and Church, K.: A Program for Aligning Sentences in Bilingual Corpora, *Proc. of ACL-91*, pp177–184, 1991.

[2] Ikehara, S. : *Test Sentences for Evaluating Japanese-English Machine Translation*, (in Japanese), NTT, 1991.

[3] Michalski, R., Carbonell, J. and Mitchell, T. (Eds.): *Machine Learning*, Tioga Publishing Company, 1983.

[4] Nagao, M. et al (Eds): *Iwanami Encyclopedic Dictionary of Computer Science*, (in Japanese), Iwanami Shoten, 1990.

[5] Nakamura, N.: Translation Support by Retrieving Bilingual Texts, (in Japanese), *Proc. of 38th Convention of IPSJ*, pp357–358, 1989.

[6] Sato, S.: Example-Based Translation Approach, *Proc. of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing*, ATR Interpreting Telephony Research Laboratories, pp1–16, 1991.

[7] Shimizu and Narita (Eds.): *The Kodansha Japanese-English Dictionary*, Koudansha, 1976.

[8] Sumita, E. and Tsutsumi, Y.: *A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching*, TRL Research Report, TR-87-1019, Tokyo Research Laboratory, IBM, 1988.

[9] It is enough for the user to find a useful example in the top five examples.