A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching

Eiichiro SUMITA and Yutaka TSUTSUMI

Tokyo Research Laboratory, IBM Japan, LTD.

Abstract : ETOC (Easy TO Consult) is a translation aid that provides a useful capability for **flexible retrieval of texts** from a bi-lingual dictionary or a translation database accumulated by the user or other users. The retrieval mechanism is based on **syntax-matching** driven by generalization rules. A practical response time is made possible by restricting the retrieval space, using a new data structure called a **quick-look-up table.** This method has the following advantages: (1) the user can input an appropriate text as a key, without using any special formal language, and (2) it is easy to produce domain-oriented systems by collecting pairs of typical source sentences and target translations that are specific to a particular domain, e.g., business letters or technical writing.

Keywords: Translation Aid, MAHT, Machine Translation, Flexible Text Retrieval, Syntax-Matching, Quick-Look-Up Table, Generalization Rules

Contents

List of Illustrations

1. Introduction

There are two types of translation that involve the use of computers: machine translation and translation aid [kay82,melby87]. In the former, the computer is the agent of translation, while the human is the assistant who answers questions from the computer or edits his machine's translation results. Most research and development has been devoted to this type. In the latter, the user is responsible for translation, while the computer provides him or her with the necessary tools, e.g., a quick-retrieval electronic dictionary or an easy-to-use word-processor. Although the effects of the second type of translation have been broadly identified, there has been little research on what kinds of function are necessary. While research on electronic dictionaries is thriving in the computational linguistic community [wachowicz86,walker87, chodorow85,tsurumaru86,nakamura87,jensen88], retrieval from conventional electronic dictionaries, as from printed dictionaries, is restricted, because it is done by matching a key word against entry words. In the next section, we argue that it is very useful to have the capacity for flexible retrieval of texts from a bi-lingual dictionary or from a translation database accumulated by the user or other users. For this purpose, we propose a new retrieval mechanism, based on syntax-matching driven by generalization rules. The outline is as follows:

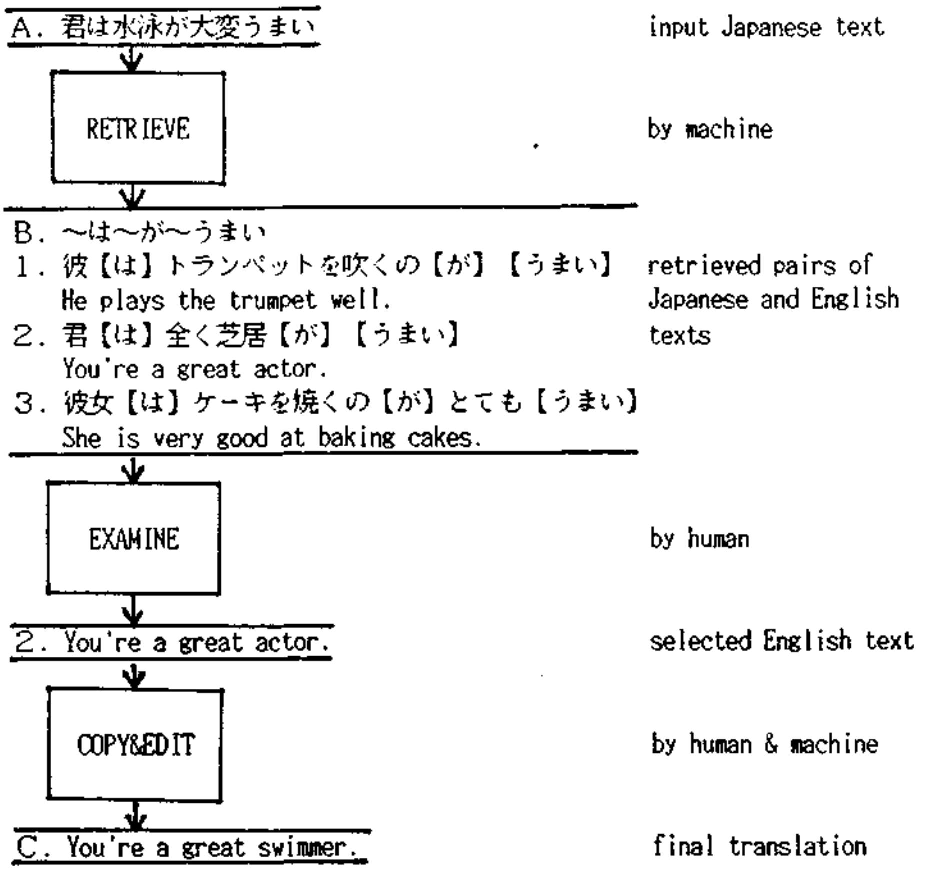
1. Input any text, including not only individual words but also phrases, and sentences, as a key.
2. Match the key against all texts in the dictionary.
3. If any of them match, then the retrieval stops. Otherwise, the key is generalized according to the generalization rules, and is tested again. In this way, the system finds the entries that are syntactically close to the key.

We developed an experimental system called ETOC (Easy TO Consult), using a Japanese-English dictionary (jese82], in order to confirm the effectiveness of the above-mentioned mechanism. In this paper, we will explain the system's configuration, generalization, quick-look-up table, user interface, and retrieval examples.

2. Translation Aid and Flexible Text Retrieval

Translation aid will make good progress if users can consult their machines about not only words, but also more complex texts - e.g., idioms, special expressions, and sentences - when they encounter a expression whose target equivalent they do not know, or for which they cannot select an appropriate equivalent from many candidates. Figure 1 illustrates the proposed interaction between users and machines: a user retrieves source texts that are close to the key, along with their translations; he or she can examine the retrieved pairs of Japanese and English text, select the most appropriate one, copy and edit it, and finally obtain the desired translation without difficulty. In this way, the key sentence, A, is searched and results 1, 2, and 3 are returned by the machine. Their Japanese parts resemble the original key sentence. Their English parts show three ways of translation: "VERB well", "BE a great VERB + er", and "BE good at VERB + ing". The user selects the second one, and finally produces C, which is a good translation of the key sentence.

Retrieval from conventional electronic dictionaries is done by matching the key **word** against entry **words,** and thus it is **difficult to consult a text which includes more than two words.** In the pattern "not only A but also B" there are four candidate key words: "not", "only", "but", and "also". The actual entry is selected according to some arbitrary criterion, e.g., "not" may be selected because it is the beginning of the pattern, or "only" because it is the head. For people who do not know the criterion or do not remember the whole pattern, it is therefore difficult to retrieve this kind of text.



A．君は水泳が大変うまい

B．～は～が～うまい

1．彼【は】トランペットを吹くの【が】【うまい】
He plays the trumpet well.

2．君【は】全く芝居【が】【うまい】
You're a great actor.

3．彼女【は】ケーキを焼くの【が】とても【うまい】
She is very good at baking cakes.

2．You're a great actor.

C．You're a great swimmer.

Figure 1: Translation Aid and Flexible Text Retrieval

In order to overcome this drawback, we propose the above-mentioned retrieval method, which matches **texts** while generalizing the key according to rules. Using the example in Figure 1, we will explain our method step by step.

1. The key and entries (A, 1, 2, 3) are sentences.
2. First, the key is matched against all entries.
3. The above fails, because there is no exact match between the key and entries. Next the key is generalized according to the rules mentioned in a later section and tested again. After several trials, we get the skeleton of the key sentence, B, and matched entries 1, 2, and 3 with the same skeleton.

In this way, our method allows the user to retrieve a text with more than two words. In this example, the pronoun " 君 ", the noun " 水泳 ", and the adverb " 大変 " are deleted, and the final generalized key pattern is composed of the function words " は " and " が " and the content word " うまい ". It is especially useful to consult some combinations of content words (e.g., nouns, verbs, and adjectives) and function words (e.g. auxiliary verbs and case particles) as illustrated in Figure 1, Figure 4, and the section on retrieval examples. Extraneous elements in entries are ignored during matching. This method has the following advantages:

- The user can **consult the dictionary freely** from various viewpoints, without conforming to the keys that were given when the dictionary was constructed.
- The user has only to input an appropriate text as a key, and can do so **without using any special formal language**, e.g., regular expressions, a database query language such as SQL (Structured Query Language), or a programming language. Thus the user is not required to be a expert in linguistics or computer science.
- The effectiveness depends on the quality and size of the dictionary. **The better the dictionary, the more effective the system**.
    - It is easy to produce domain-oriented systems by collecting pairs of typical source sentences and target translations that are specific to a particular domain, e.g., business letters or technical writing.
    - It is easy to build multi-language systems by expanding the current dictionary to include corresponding translations in every language.
- In contrast to machine translation, the translation aid system ETOC can deal with **unrestricted and natural** language, because the lexical level analysis is relatively accomplished and robust, because the system is interactive, and because the user is assumed to be cooperative and intelligent.

3. Configuration

The configuration of ETOC is presented in Figure 2. Our system has three data items - (1) a key, (2) a dictionary, and (3) generalization rules - and three modules - (4) an analyzer, (5) a retriever, and (6) a generalizer. In this paper, we deal with the analyzer and generalizer. The objective of the analyzer is to structure both the key and entry at the same level. The generalizer and its rules must accommodate this level. Several levels may be distinguished; for example, the lexical and syntactical levels. The deeper the analysis level, the more precise the result and the higher the cost. Because Japanese has no explicit delimiters (blanks) between words, lexical analysis (i.e., segmenting a text into words and assigning a part of speech to each word) is necessary for almost every kind of Japanese processing. In this experimental system, we used a lexical analyzer at our site [maruyama88], and found it to be useful and practical. Because the system is interactive, we can ignore a certain level of noise in the retrieved result, as shown in the section on retrieval examples.

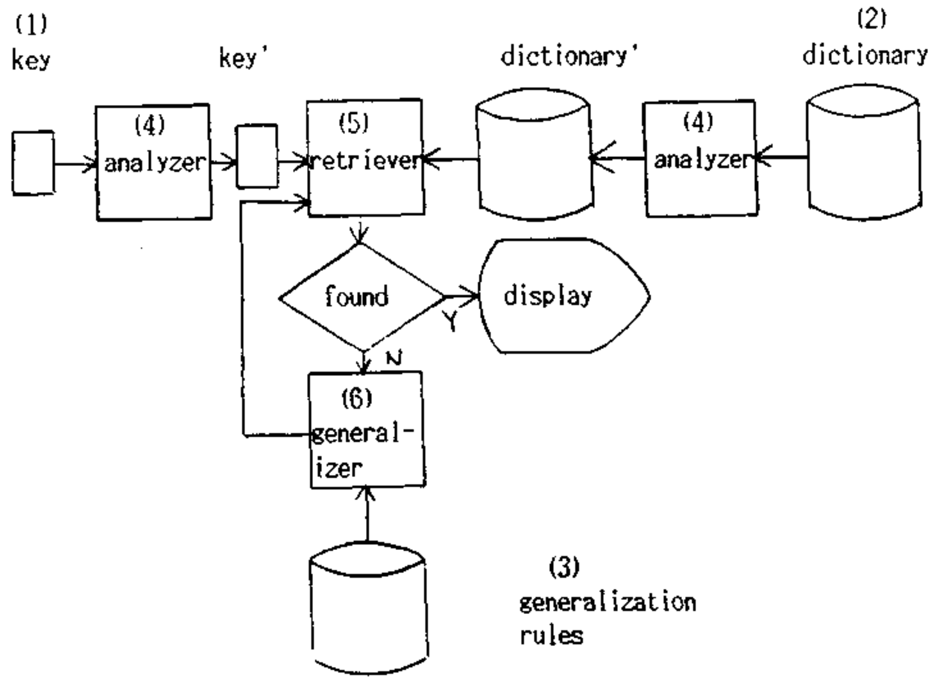


Figure 2: Configuration

4. Generalization

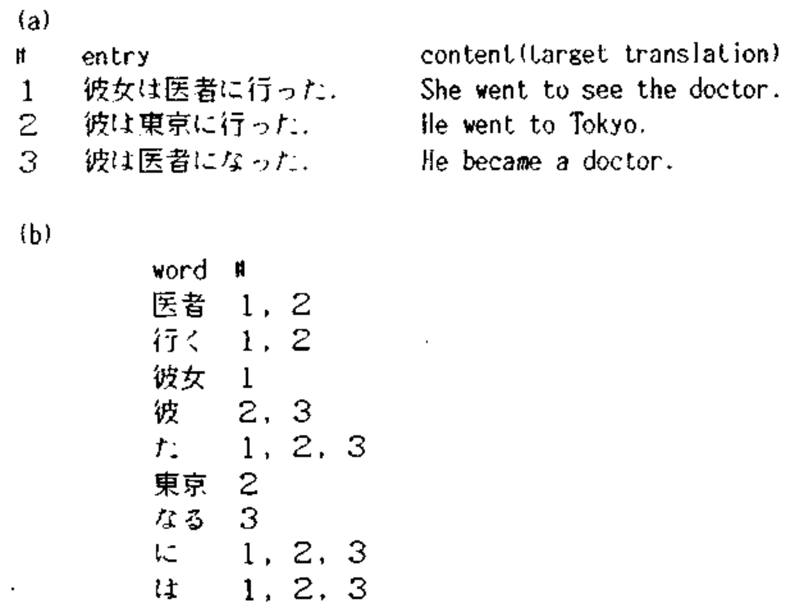
When the key does not match any entry in the dictionary, it is generalized. In accordance with the order of rules listed below, the system determines whether each condition is important or not, and then deletes or relaxes the less important ones. Briefly speaking, the system ignores the particular features of each text, represented by the content words, and generalizes the key to the skeleton of the sentence and the pattern of modality; in other words, it produces a sequence of function words. To date, the following rules have beer adopted:

1. If the order of case elements is not normalized, then normalize it (because in simple Japanese sentences the order of case elements is freely changeable).
2. if there is a pronoun, then replace it with an arbitrary noun (because the replaceability of pronouns is considered high).
3. If a phrase has a modifier in it, then delete the modifier.
4. If there is a noun, then replace it with an arbitrary one.
5. If there is a verb or adjective, then replace it one with an arbitrary one.
6. If there is a special case particle, replace it with a semantically similar one (for example, " に " and " へ " arc similar to each other).
7. If there is a case element, delete it (because in Japanese not only free case elements but also so-called obligatory ones can be omitted).

Rules 1, 6, and 7 are peculiar to Japanese, while the others are general. These rules are applied sequentially. We assume that they **conform to the way in which people want to retrieve text.** Of course, when a user wants to retrieve a single word, that word is absolutely important. But when he or she wants to retrieve a larger text, not all the words are important. If it is impossible to obtain a exact match for the whole text, the user may ignore content words instead of function words.

5. Quick-Look-Up Table

No-one wants to use a slow dictionary retrieval system. In order to speed up the response, we introduced a data structure called a **quick-look-up table.** The idea is depicted in Figure 3. Each word in a set of dictionary entries is registered in the table, along with the numbers of the entries in which it appears in the table. This table is built at the same time as the system is constructed, and is updated each time a new record is added. It is utilized to restrict the retrieval space. If a key consists of two words, the retrieval space is reduced to the size of their intersection. The more words the key text has, the smaller the retrieval space becomes.



(a)

#	entry	content(target translation)
1	彼女は医者に行った.	She went to see the doctor.
2	彼は東京に行った.	He went to Tokyo.
3	彼は医者になった.	He became a doctor.

(b)

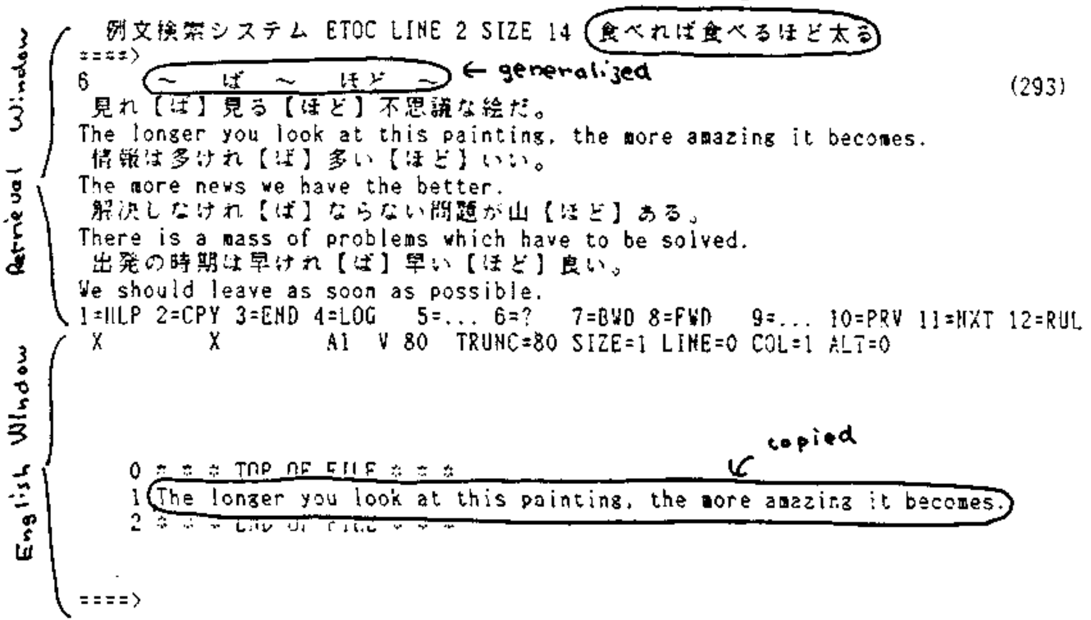
word	#
医者	1, 2
行く	1, 2
彼女	1
彼	2, 3
た	1, 2, 3
東京	2
なる	3
に	1, 2, 3
は	1, 2, 3

Figure 3: Quick-Look-Up Table

Let T be the table and T(W) the set of records in which the word W occurs. For example, when we retrieve " ka ha isya ni naru ", since T1 = T(彼) = {2,3}, T2 = T(は) = {1,2,3}, T3 = T(医者) = {1,3}, T4 = T(に) = {1,2,3}, T5 = T(なる) = {3}, we have only to check the records of T1&T2&T3&T4&T5 = {3}. In this experimental system the number of records in our dictionary is 16870, #(T(が)) is 4153, #(T(を)) is 5984, and #(T(が) & T(を)) is only 909. This implies that when we are retrieving a text such as "A が B を ", the retrieval space is restricted to 909/16870, i.e., 5.4 percent of the size of the dictionary. Japanese function words are crucial to the way in which something is expressed, and consequently our generalization rules try to leave as many function words as possible. It is important to ensure that the reduction rate of function words is sufficiently practical. Our quick-look-up table is implemented with hashing in LISP. We can use well-studied database techniques, such as B-tree and RDB, to ensure there is no problem if the system is scaled up.

6. User Interface

ETOC provides two interfaces for input. It allows the user (1) to type a key text from the command line, or (2) to move the cursor to the text he or she wants to consult, and to pass the selected line to the system by hitting a special key.



```
 例文検索システム ETOC LINE 2 SIZE 14 食べれば食べるほど太る
====>
6    ～  ば  ～  ほど  ～                                         (293)
 見れ【ば】見る【ほど】不思議な絵だ。
The longer you look at this painting, the more amazing it becomes.
 情報は多けれ【ば】多い【ほど】いい。
The more news we have the better.
 解決しなけれ【ば】ならない問題が山【ほど】ある。
There is a mass of problems which have to be solved.
 出発の時期は早けれ【ば】早い【ほど】良い。
We should leave as soon as possible.
1=HLP 2=CPY 3=END 4=LOG   5=... 6=?   7=BWD 8=FWD   9=... 10=PRV 11=NXT 12=RUL
 X        X          A1  V 80  TRUNC=80 SIZE=1 LINE=0 COL=1 ALT=0



    0 * * * TOP OF FILE * * *
    1 The longer you look at this painting, the more amazing it becomes.
    2 * * * END OF FILE * * *


====>
                                                          X E D I T  2 FILES
```

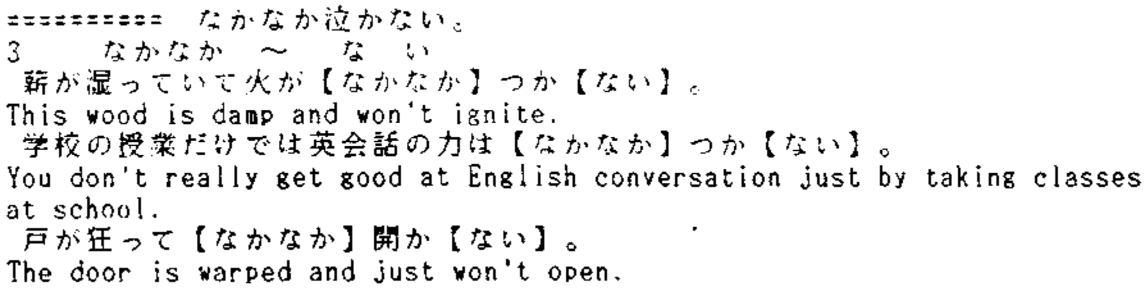
Figure 4: User Interface

The result window opens automatically to show pairs of Japanese sentences and their English equivalents. Next, the user selects an appropriate pair, moves the cursor to it, and hits another special key. The selected English sentence is copied into the English window. Figure 4 shows a example session: the user retrieves "食べれば食べるほど太る", obtains six pairs, and picks the first one, "The longer you look at this painting, the more amazing it becomes.".

7. Retrieval Examples

We will show here ETOC retrievals with the following characteristic features: long-distance dependency, idioms, the ellipsis symbol, aspect and finally a rather problematic one, semantic ambiguity. In following examples, the line beginning with *= = = = = = = = = = * shows the input to ETOC, the second line shows the number of entries found in the dictionary and the generalized key used for matching and the following lines show retrieved pairs of Japanese sentences, whose keys are marked with thick square brackets, and their English translations.

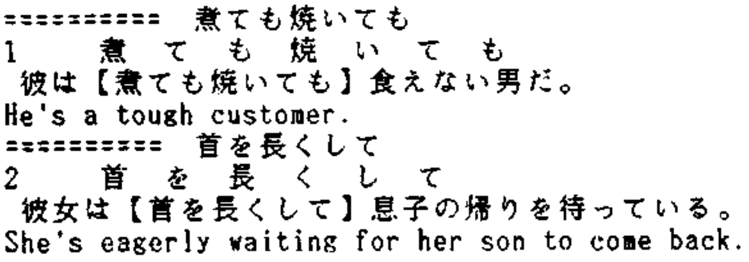
- ETOC supports the retrieval of long-distance dependency. Figure 5 exemplifies the results given by our system: "なかなか" is an ADVERB, and "ない" is an AUXILIARY ADJECTIVE for negation, and they are often used together with the general meaning "not easy". In Figure 4 "ば" is a CONJUNCTION indicating a condition, "ほど" is a PARTICLE indicating a degree, and together they correspond to the construction "the + comparative, the + comparative". Such combinations of content words and function words cannot easily be looked up in conventional dictionaries, and are very important as the essence of the way concepts are expressed.



```
==========  なかなか泣かない。
3      なかなか  ～   な  い
 薪が湿っていて火が【なかなか】つか【ない】。
This wood is damp and won't ignite.
 学校の授業だけでは英会話の力は【なかなか】つか【ない】。
You don't really get good at English conversation just by taking classes
at school.
 戸が狂って【なかなか】開か【ない】。
The door is warped and just won't open.
```

Figure 5: Example (Long-Distance Dependency)

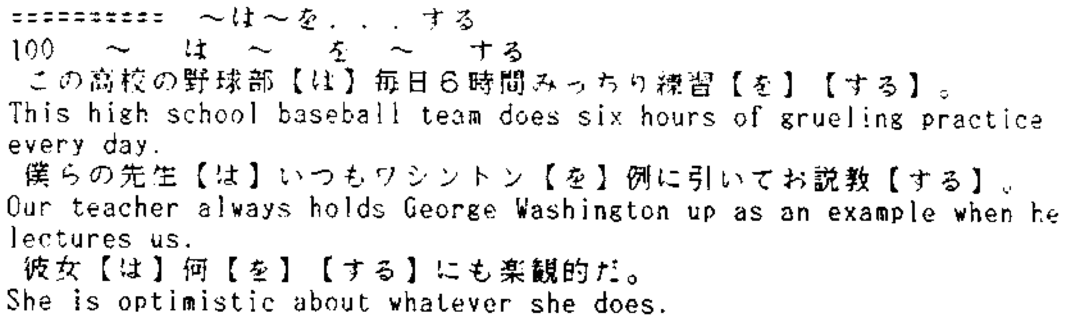
- Idiomatic expressions such as the one in Figure 6, are a tough problem, for which current machine translation technology has no good solution. Our method is suitable for this kind of expression, however, because ETOC does not require deeper analysis and can deal with such expressions simply by accumulating examples. A literal translation of the first query is "Even if you boil it or grill it, you cannot eat it", and of the second query "making the neck longer"; however, the correct translations are respectively "tough" and "eagerly".



```
==========  煮ても焼いても
1     煮  て  も  焼  い  て  も
 彼は【煮ても焼いても】食えない男だ。
He's a tough customer.
==========  首を長くして
2     首  を  長  く  し  て
 彼女は【首を長くして】息子の帰りを待っている。
She's eagerly waiting for her son to come back.
```

Figure 6: Example (Idiom)

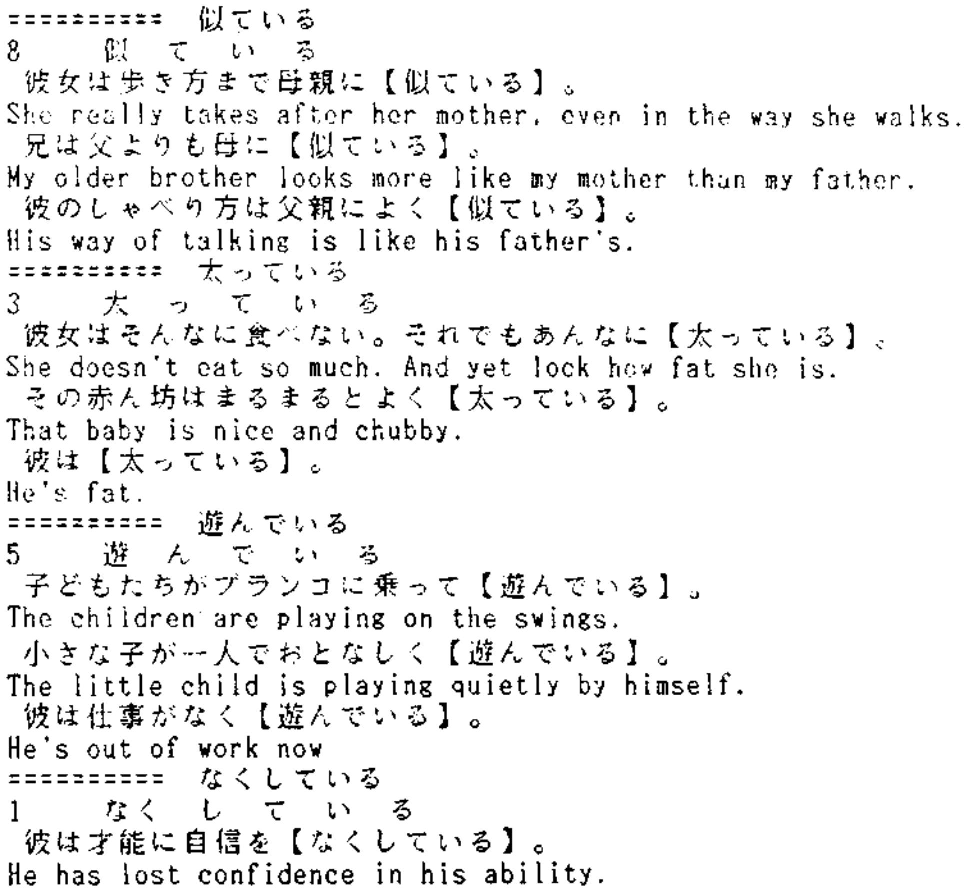
- As is shown in the next example (Figure 7), this system handles ellipsis symbols " ～ ", " . . . " and so on, which are often seen in ordinary dictionaries. These signs are easy to understand and can be used by people who are not acquainted with computers.



```
==========  ～は～を．．．する
100   ～   は   ～   を   ～   する
 この高校の野球部【は】毎日６時間みっちり練習【を】【する】。
This high school baseball team does six hours of grueling practice
every day.
 僕らの先生【は】いつもワシントン【を】例に引いてお説教【する】。
Our teacher always holds George Washington up as an example when he
lectures us.
 彼女【は】何【を】【する】にも楽観的だ。
She is optimistic about whatever she does.
```

Figure 7: Example (Ellipsis)

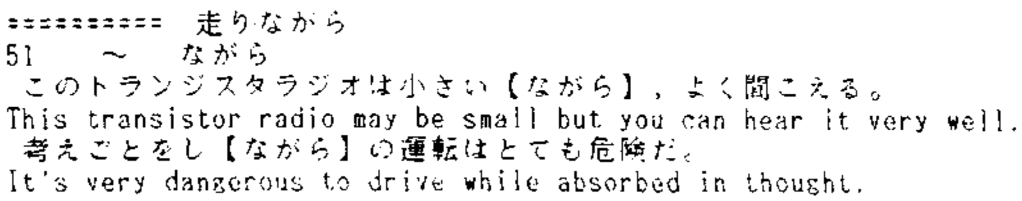
- Figure 8 shows the retrieval of a Japanese aspectual phrase, i.e., a VERB with an aspectual marker. There is no simple correspondence between Japanese aspectual phrases and English ones. For example, the Japanese aspectual marker "ている" can be translated, depending on the semantic category of its VERB, by the English aspectual markers "ing" (progressive) , "have+pp" (perfect), or "" (simple). Consequently, it is not easy to translate an aspectual phrase. ETOC enables the user to retrieve "individual VERB +ている". The retrieval results shows that " 似ている ", " 太っている " are translated by the simple, " 遊んでいる " by the progressive, and " なくしている " by the perfect aspect. With ETOC, the user does not need to know any linguistic terms such as "aspect", nor to consult a English grammar book, when he wants to translate a " VERB +ている" type expression.



```
==========  似ている
8     似  て  い  る
 彼女は歩き方まで母親に【似ている】。
She really takes after her mother, even in the way she walks.
 兄は父よりも母に【似ている】。
My older brother looks more like my mother than my father.
 彼のしゃべり方は父親によく【似ている】。
His way of talking is like his father's.
==========  太っている
3     太  っ  て  い  る
 彼女はそんなに食べない。それでもあんなに【太っている】。
She doesn't eat so much. And yet lock how fat she is.
 その赤ん坊はまるまるとよく【太っている】。
That baby is nice and chubby.
 彼は【太っている】。
He's fat.
==========  遊んでいる
5     遊  ん  で  い  る
 子どもたちがブランコに乗って【遊んでいる】。
The children are playing on the swings.
 小さな子が一人でおとなしく【遊んでいる】。
The little child is playing quietly by himself.
 彼は仕事がなく【遊んでいる】。
He's out of work now
==========  なくしている
1     なく  し  て  い  る
 彼は才能に自信を【なくしている】。
He has lost confidence in his ability.
```

Figure 8: Example (Aspect)

- There are at least two different meanings for the CONJUNCTION "ながら", concession and simultaneousness: the former is usually expressed by "but" and the latter by "while". Such semantically ambiguous expressions cannot be differentiated by a lexical analyzer alone, and the user must clearly understand each retrieved sentence. Figure 9 shows this ambiguous situation.



```
===========  走りながら
51    ～    ながら
 このトランジスタラジオは小さい【ながら】，よく聞こえる。
This transistor radio may be small but you can hear it very well.
 考えごとをし【ながら】の運転はとても危険だ。
It's very dangerous to drive while absorbed in thought.
```

Figure 9: Example (Semantic Ambiguity)

8. Conclusion

We have proposed flexible text retrieval, based on syntax matching, as a mechanical aid to human translation. In this method, the key text and entry texts are analyzed, and the key is generalized in accordance with rules until it matches one or more entries. We have shown the feasibility of the method by implementing a system for Japanese-to-English translation.

Future tasks include:

- Implementing a reverse direction system, using an English lexical analyzer.
- Collecting multi-lingual and multi-domain data, developing accurate and usable generalization rule sets for each data, and evaluating them.
- Developing a system based on deeper analysis.
- Utilizing this retrieval mechanism for language education tools.
- Devising a rule acquisition method, from session-logs of ETOC, based on techniques of learning from examples.
- Enhancing this mechanism in order to generate target sentences automatically, as suggested by Nagao[nagao84).

Acknowledgement

The authors wish to thank Mr. Michael J. McDonald for reading and criticizing the numerous successive versions of this paper. It is also appropriate to thank Asahi Press for the use of their valuable dictionary.

Bibliography

[chodorow88] Chodorow M.S., Byrd R.J., and Heidorn G.E., "Extracting semantic hierarchies from a large online dictionary", *Proceedings of the 23rd Annual Meeting of the ACL,* pp.299-304, 1985.

[jensen88] Jensen K. and Binot J., "Dictionary text entries as a source of knowledge for syntactic and other disambiguations", *Proceedings of the Second Conference on Applied Natural Language Processing,* ACL, Austin, pp.152-159, 1988.

|kay82] Kay M., "Machine translation", *AJCL,* vol.8, no.2, pp.74-78, 1982.

[jese82| Keene D. and Hatori H., *Japanese-English Sentence Equivalents,* pp.869, Asahi Press, 1982.

[maruyama88] Maruyama N., Morohashi M., Umeda S., and Sumita E., "A Japanese sentence analyzer", *IBM Journal of Research and Development,* (in press), 1988.

[melby87] Melby A., "On human-machine interaction in translation", *Machine Translation,* pp. 145-154, 1987.

[nagao84] Nagao M., "A framework of a mechanical translation between Japanese and English by analogy principle", *Artificial and Human Intelligence* (A. Elithorn and R. Baneriji. Ed.), pp.173-180, 1984.

[nakamura87] Nakamura J., Sakai K., and Nagao M., "Automatic analysis of semantical relation between English nouns by an ordinary English dictionary", *IECE WG preprint of NLC86-23,* pp. 17-24, 1987 (in Japanese).

[tsurumaru86] Tsurumaru H., Hitaka T., and Yoshida S., "An attempt to automatic thesaurus construction from an ordinary Japanese dictionary", *Proceedings of COLING 86,* pp.445-447, 1986.

[wachowicz86] Wachowicz K., "On intelligent dictionaries", *CaT,* vol.1, no.4, pp:225-233, 1986.

[walker87] Walker D., "Knowledge resource tools for accessing large text files", *Machine Translation,* pp.247-261, 1987.