[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

TECM - Translation of English into Chinese on a Micro

Liu Shiao-shu
Peking

INTRODUCTION

Machine translation has been going for more than thirty years and yet it is still far from being in wide use. The reason is of course many sided. One thing worth noticing is that the existing machine translation systems are implemented on large computers and not everybody can get access to them. Since wide use and popular interest is a powerful impetus to improvement and perfection, it is of significance to design machine translation systems which are handy and cheap that can be used as tools for various kinds of work.

The development of microcomputers provides excellent hardware for such systems. It is up to us machine translation researchers to devise adequate software for small, inexpensive translating machines for popular use.

TECM is a system in which advantage is taken of the similarities between the source and target languages so that the process of analysing the source sentence and generating the target sentence is reduced to a minimum. Before we go into the details of how this is achieved, let us first consider the criteria of an "acceptable" translation.

A famous Chinese translator said that a good translation must be true to the source text besides being fluent and elegant. To be true to the source text, the translation must

1. not omit any point that is mentioned in the source text;
2. not add any idea that is not explicitly stated in the source text;
   and lastly, but most importantly,
3. not make any mis-interpretation.

For a small, low-cost translating machine, to be true to the source text alone is a requirement hard enough to fulfill. Let us leave the question of fluency and elegance for the future and concentrate our attention on the exactness of the translation. Since the sentence is the building brick of the text, the above definition of truth to the source text leads quite naturally to the inference that the translation must have as many sentences as does the source text.

Consequently the TECM system is designed to perform its task on the sentence to sentence basis.

THE CONVERSION APPROACH

Since both source and target sentences express the same idea, they must have a number of things in common although the words are different, let us consider the sentence

Ex. 1. I appreciate the book of Hues and his students.

The proper Chinese translation, when written in English words, is

I appreciate Hues and his students DE book,

where DE is a Chinese character which indicates that the noun phrase preceding it is the possessor or modifier of the noun phrase following it. Now we see that the Chinese translation is parallel to the English text except

1. the noun phrases preceding and following OF interchange places, while OF is replaced by DE;
2. the THE preceding a noun is dropped out.

The English words OF and THE are the cause of the above mentioned non-parallelisms. Such trouble-making words are not numerous. The author has examined 11,000 words taken from technical literature, among which only 1,200 are found to cause non-parallelism. However, the frequency of occurrence of these words is high. They may comprise one third of the word-count in an article. Even so, there are still two thirds of the words in the article that can be translated simply word to word. This statement is subject to an assumption: the subject field of the source text is known, without this pre-knowledge, translation would be very difficult even for human translators. This point will be taken up again at the conclusion of this paper.

Besides the non-parallelism caused by the difference between the grammars of the source and target languages as mentioned above, there are non-parallelisms caused by idioms. An idiom is a word string which is not to be interpreted literally. In this sense, those parts of the source sentence which are non-parallel to the target sentence due to difference of grammars can be looked at as the idioms, as they do not allow word to word translation. The expression

(noun phrase 1) of (noun phrase 2)

is an idiom of OF, where OF is the pivot word of the idiom.

The conversion approach of machine translation is one, in which the non-parallel parts of the source text are converted to get rid of the non-parallelism, so that word to word translation will produce a readable target text. The conversions are made by the aid of a conversion table, which we shall henceforth call a **glossary**. It is evident that the workability of the conversion approach depends on the comprehensiveness and the size of the glossary. The comprehensiveness can be improved by repeatedly replenishinq in the course of use. The size of the glossary can be reduced by introducing some symbols which serve to broaden the coverage of the entries in the glossary. In the TECM system, the glossary contains 3,300 entries. They are listed according to the alphabetical order of the pivot words of the idioms. These pivot words are mostly polysemous, but they do not have to be polysemous. So long as a word can be used in conjunction with other words to form idioms, this word should be included in the glossary so that the proper equivalents of its idioms can be found during the process of translation.

THE TECM SYSTEM

This system consist of five modules:

1. Input of source text: the source text is to be entered sentence by sentence.
2. Dictionary look-up: for each word, the following information is obtained from the dictionary:
   a. Part of speech, or possible parts of speech it may belong to;
   b. Grammatical attributes such as tense, case, number, etc.;
   c. Whether it is given a unique Chinese equivalent or the glossary must be consulted.
3. Syntactic analysis and conversion.
4. Glossary look-up.
5. Output of the translation.

The functions of the first two and the last modules are self-explanatory. The following paragraphs are intended to explain the essential points of the syntactic analysis and glossary.

SYNTACTIC ANALYSIS ANDCONVERSION

The purpose of syntactic analysis in the conversion approach is to find cut the type of structure of the sentence to be translated, and which words comprise each structural element of this sentence. A sentence is considered to be built up of four kinds of structural elements:

1. verb phrase V;
2. non-verb phrase S;
3. sentence connector K; and
4. modifier connector L.

The first three kinds are well-known. The last kind refers to the relative pronouns, e.g. **which**, **who**, **where** etc., which are non-existent in the Chinese language.

Let us demonstrate the process of syntactic analysis by means of an example. Consider the sentence Ex.2:

```
Ex.2
    1    2      3      4          5    6         7
  The Aim-65 system manufactured by Rockwell International
    8   9      10    11 12    13    14     15       16
  has been selected as one vehicle for developing this
         17      l8  19       20          21    22
  understanding, and its characteristic elements and
       23     24     25         26     27   28
  functions are discussed throughout the book.
```

Each word in the sentence is given an ordinal number to facilitate identification in the works that follow. After dictionary look-up, the part of speech of most of the words are known, and the actual parts of speech of the homographs can be determined. In the present example, it is found that there are three verb phrases, i.e. "manufactured", "has been selected" and "are discussed". They are labelled as V4,4 V6,10 and V24,25 where the subscripts are the ordinal numbers of the first and last word of the phrase. Three verb phrases may form three clauses, which have to be connected by two connectors. By scanning the sentence, it is found that two **and**s cluster between V8,10 and V24,25. Only one of them is to be recognized as sentence connector. As a rule, the **and** immediately behind the comma is a sentence connector. Let us denote this connector by K18,18. The remaining words are denoted by Ss and the sentence Ex.2 can be symbolized as

S1,3 V4,4 S5,7 V8,10 S11,17 K18,18 S19,23 V24,25 S26,28.

This is the structural formula of sentence Ex.2.

The next step of the syntactic analysis is to compare the structural formula of the sentence to be translated with the standard patterns, and to follow the instructions pertaining to that pattern, with which the sentence complies. Table 1 is a complete list of the standard patterns and their pertaining instructions.

In our present example, the structural formula of the sentence is not found in Table 1. This is because the sentence is long. In this case, the structural formula is split at all **K**s and each section thus obtained compared with Table 1. The formula of Ex.2 is split into

Table 1: Standard Patterns

	Standard Pattern	Examine if	Rewrite into Target Pattern
1	SV	S-Sa Sb V=Va Vb	Sb Va DE Sa Vb
2		S=S V=Va Vb	Va DE S Vb
3		otherwise	no change
4	S1 V1 S2	S1=Sa Sb V1=Va Vb	Sb Va DE Sa Vb S2
5		S1=S1 V1=Va Vb	Va DE S1 Vb S2
6		S1=S1 V1=V1	no change
7	S1 V1 S2 V2	S1=Sa Sb S2=S2	Sb V1 S2 DE Sa V2
6		S1=S1 S2=Sa Sb	S1 V1 Sb V2 DE Sa
9		S1=S1 S2=S2	no change
10	S1 V1 S2 V2 S3	Sl=Sa Sb	Sb V1 S2 DE Sa V2 S3
11		S1=S1 Vl=pp. V2=not pp.	V1 S2 DE S1 V2 S3
12		S1=S1 Vl=not pp. V2=pp.	S1 V1 V2 53 DE S2
13		otherwise	no change
14	(SVS) 1 K (SVS)2	K = BECAUSE,IF,THOUGH,WHEN	K (SVS) 2 (SVS) 1
15		otherwise	no change
16	SVS K SV	K=BECAUSE,IF,THOUGH,WHEN	K SV SVS
17		otherwise	no change
18	SV K SVSW	K = BECAUSE,IF,THOUGH,WHEN	K SVS SV
19		otherwise	no change
20	(SV) 1 K (SV) 2	K=BECAUSE,IF,THOUGH,WHEN	K (SV) 2 (5V) 1
21		otherwise	no change
22	K S1 V1 S2 V2 S3	S2=Sa Sb	K S1 V1 Sa Sb V2 S3
23		otherwise	no change
24	K S1 V1 S2 V2	S2=Sa Sb	K S1 V1 Sa Sb V2
25		otherwise	no change
26	S1 V1 S2 L S3 V2 S4	S2=Sa Sb	S1 V1 S2 Sb S3 V2 S4
27		S2=S2	S1 V1 S2 S2 S3 V2 S4
28	S1 V1 S2 L S3 V2	S2=Sa Sb	S1 V1 S2 Sb S3 V2
29		S2=S2	S1 V1 S2 S2 S3 V2
30	S1 V1 S2 L V2 S3	S2=Sa Sb	S1 V1 S2 5b V2 S3
31		S2=S2	S1 V1 V2 S3 DE S2
32	S1 L S2 V1 S3 V2 S4		S2 V1 S3 DE S1 V2 S4
33	S1 L S2 V1 S3 V2		S2 V1 S3 DE S1 V2
34	S1 L V1 S2 V2 S3		V1 S2 DE S1 V2 S3
35	S1 L V1 S2 V2		V1 S2 DE S1 V2

pp=past participle

Table 1: Standard Patterns

S1,3 V4,4 S5,7 V8,10 S11,17 and S19,23 V24,25 S26,28.

The first section complies with item 11 of Table 1. It is written into

V4,4 S5,7 DE S1,3 V8,10 S11,17

The second section complies with item 6 of Table 1. No change is made. Then this sentence will be translated as

> Manufactured by Rockwell International DE the Aim-55
> system has been selected etc.

Note that this simple syntactic analysis has revealed the fact that "manufactured by Rockwell International" is the modifier of "the Aim-65 system". This is enough to avoid mis-interpreting "manufactured" as the verb of the subject.

THE GLOSSARY

In the glossary used in the TECM system, the following symbols are introduced:

- V = verb phrase of any length;
- S = non-verb phrase of any length;
- (part of speech) = any word of said part of speech;
- ¢ = possible appearance of any number of words of that part of speech following the symbol;
- £ = possible appearance of the word following the symbol
- / = possibly the words following the symbol do not appear.

It is worthwhile to describe in a little more detail how the use of glossary and symbols solves the difficulty of non-parallelisms and avoids unnecessary duplication of similar entries so that the glossary is in fact a comprehensive and concise grammar.

1. In translating the sentence of Ex.1, the noun phrases preceding and following OF have to interchange positions. This is done by the aid of an entry in the glossary

   S1 of S2 -> S2 DE S1.

   It is to be noted that in the process of such kinds of position shift, the correctness of what constitute S1 and S2 are vital to the correctness of the translation. In the case of Ex.1, S2 consists of "Hues and his students" rather than "Hues" alone. This is achieved by the syntactic analysis.

2. Some idioms e.g. "day after day", "week after week", "month after month", "year after year" etc. are different idioms. But they are so similar that it would be tedious to enumerate them. They can be condensed into one entry:

   (noun, time) after (noun, time).

Of course, this requires that enough information can be obtained from the dictionary.

3. The notion of tense in English is non-existent in the Chinese language. The tense of the English sentence can only be expressed in Chinese by some additional words. The glossary enables the system to put the right additional words at the right place, where they are needed. For example, the present continuous tense is expressed by adding ZHEN ZAI in front of the verb. This is realized by entries under pivot words AM, ARE, IS, WAS and WERE, e.g.

   am (present part.) -> ZHEN ZAI (present part.)

4. In the present continuous tense, the present participle may be modified by any number of adverbs, which may be put in front of, or behind the present participle. But in the Chinese language, the adverbs must always be put in front of the present participle they modify. So the previous entry can be extended to cover much more possibilities in this way:

   am ¢ adv. (pres.p.) ¢ adv. -> ZHEN ZAI ¢ adv. (pres.p.)

5. The idioms

   a. This is the case;
   b. This is not the case;
   c. This is always the case;
   d. This is not always the case;
   can be combined by using the symbol £, thus

   This is £not £always the case.

6. The idioms "treat with S1" and "treat with S1 for S2"
   can be combined into one by using the symbol /, thus

   Treat with S1 / for S2.

7. Now look at the entry

   is ¢not being ¢adv. (p.p) ¢adv. / by S1 / to (verb) S2
   ¢adv.

   and see how many different cases it covers. It is a general form of the present continuous passive in the third person. It embraces the following possible cases:

   a. it may be positive or negative narration;

b. the past participle may be either pre-modified or post-modified or both or not modified at all;
c. the agent 01 the action may or may not be given;
d. the purpose of the action may or may not be given;
e. the combination of any two or three or four of the foregoing possibilities.

8. As an example of how the glossary helps translating difficult phrases, let us consider these two sentences:

- John ate his birthday cake with a fork.
- John ate his birthday cake with his sister.

These two sentences are grammatically the same, but the two **with**s should not be translated identically. In the first sentence, **with** means "by means of", because it is followed by "a fork", which is an object, which serves as a tool, the equivalent Chinese word is YONG. In the second sentence, **with** means "together with", because it is followed by "his sister" which refers to person. The equivalent Chinese word is TONG. No Chinese word can be found which happens to have both these two meanings. Different Chinese words have to be used for each case. The glossary can instruct the system how to choose the proper translation through two entries under WITH:

```
V S1 with S2 (object) -> YONG S2 LAI V S2
V S1 with S2 (person) -> TONG S2 YI QI V S2
```

It is to be noted that the word order in the translation is appreciably different from that in the source sentence. However, this presents no problems when the glossary is used.

CONCLUSION

The design of the TECM system is an attempt to find a short-cut approach to machine translation, that can be implemented on a microcomputer. Since the size and therefore the capability of the system is limited, it is expected to function only in a narrow pre-defined domain. Many English technical terms refer to different objects in different subject fields. If the subject field is not given, even a human translator will be unable to decide which Chinese name should be chosen for those punning terms. For example, the sentence

Marihuana is a product of the hemp plant.

can be translated in four different ways, all "correct", if common sense does not come to our aid. There are two different types of product, each has its own Chinese name. If the product is an object manufactured without undergoing chemical treatment, it is called CHAN-PIN. If the product is a material obtained through chemical treatment, it is called ZHI-PIN. Again, "hemp plant" may be interpreted either as hemp factory or as hemp vegetable. Thus we have four different combinations to choose from.

Besides the inability to differentiate polysemous words, there is another shortcoming which is inherent in the conversion approach: namely, the translation is obtained through manoeuvring the words supplied by the source text, with the possible addition and omission of a very limited number of auxiliary words. The quality of the translation can not be high.

However, since the system is very small, occupying only 18K bytes of memory space, there is ample room for improvements to be made, it is hoped that by elaborating the glossary much can be done toward improving the quality of the translation.

```
TECM
LEASE ENTER TEXT

 APPRECIATE THE BOOK OF *HUES* AND HIS STUDENTS.!!!
 APPRECIATE THE BOOK OF HUES AND HIS STUDENTS .>GET GLO
J 6470
GET TRA
J 8000
GET IND
J 89FA
GET CH1
GET CH2
J 8A9B
就欣賞HUES和他的學生的書.>
HE PRINCIPAL FUNCTION OF THIS BOOK IS TO GUIDE THE READER T
WARD AN UNDERSTANDING OF THE PROCESS OF DESIGNING MICROPROC
SSOPBASED SYSTEMS .>GET GLO
J 6470
GET TRA
J 8000
GET IND
J 89FA
GET CH1
GET CH2
J 8A9B
這書的這主要的函數(功能,作用)是要指引讀者向設計以微處理機爲
基础的系統的過程的一個了解.>

D
GET DIC
0
TFCM
LEACE ENTER TEXT

HE *AIM-65 SYSTEM* MAUNFACTURED BY *ROCKWELL INTERNATIONAL*
HAS BEEN SELECTED AS ONE VEHICLE FOR DEVELOPING THIS UNDERS
NDING.AND IT'S CHARACTERISTIC ELEMENTS AND FUNCTIONS ARE DI
```

```
HE !AIM-65 SYSTEM+ MAUNFACTURED BY +ROCKWELL INTERNATIONAL
HAS BEEN SELECTED AS ONE VEHICLE FOR DEVELOPING THIS UNDER
NDING,AND ITS CHARACTERISTIC ELEMENTS AND FUNCTIONS ARE D
USSED THROUGHOUT THE BOOK.!!!
HE AIM-65 SYSTEM MAUNFACTURED BY ROCKWELL INTERNATIONAL HA
BEEN SELECTED AS ONE VEHICLE FOR DEVELOPING THIS UNDERSTAN
NG ,AND ITS CHARACTERISTIC ELEMENTS AND FUNCTIONS ARE DISCU
SED THROUGHOUT THE BOOK .>
J 6470
GET TRA
J 8000
GET IND
J 89FA
GET CH1
GET CH2
J 8A9B
ROCKWELL INTERNATIONAL所製造的AIM-65 SYSTEM被選擇如(作爲)
一個車輛(媒介物)爲了發展這了解而它的特徵的元素和函數(功能,作
用)被討論遍及整個書.>



PPENDIX A IN FACT IS A COMPLETE USER MANUAL FOR THIS SYSTEM
.>GET GLO
J 6470
GET TRA
J 8000
GET IND
J 89FA
GET CH1
GET CH2
J 8A9B
附錄A事實上是一個完備的用戶手冊爲了這系統.>
```