CHINESE - ENGLISH MACHINE TRANSLATION

By D. T. Chai*

It is the goal of the World Peace Through Law Center to achieve the computerization of law internationally; a proposal to this effect was approved by the Geneva World Conference in July, 1967. One of the major obstacles in this world-wide application of computer techniques to storage and retrieval of law materials in that of dealing with the various languages. It is in this regard that work in the field of machine translation takes on an especial importance for those concerned with computers and the law—Ed.

Brief History

The field of machine translation (MT for short) may be traced to a memorandum by Warren Weaver in 1949 which he reportedly sent to 200 of his acquaintances in various fields. He foresaw the difficulty in this task when he said:

"Such a program involves a presumably

* B.S. in Electrical Engineering, Purdue University; M.S. and Ph.D. in Communications Sciences, University of Michigan. Joined Bunker-Ramo in 1967 as project manager for the Chinese-English Machine Translation project, which is supported by the United States Air Force. Currently engaged in developing computer software for advanced information systems, especially for automatic indexing and text editing; developing software for teaching the operation of a complex electronic equipment through a computer; teaching at the University of Southern California as an adjunct assistant professor. Previous to joining Bunker-Ramo, worked at IBM and held teaching and research appointments at the University of Michigan.

This research is supported by the Air Force Systems Command, Research and Technical Division, Rome Air Development Center, Griffis Air Force Base, New York.

tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. Such a program has the advantage that, whether or not it led to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication."

Today, nineteen years later, there are only two operational production-oriented MT systems in the U. S. translating from Russian to English. One is at the Oak Ridge National Laboratory where the object of translation is only to provide a quick and rough translation of summaries of some current Russian publications for its own physicists. One must possess a tremendous background knowledge of the subject matter in order to read the translation, which is of very poor quality. If the physicist finds the translation of some value to his research, he can then request a full translation of the article by some professional translator. The other MT system relies on human editors to polish the translation. It is, therefore, not a fully automatic MT system. It is located at the Foreign Technology Division of Wright-Patterson Air Force Base in Dayton, Ohio. In a report[1] published in 1965, a comparison of quality, economy and speed of the Dayton MT system with that of the human translation was made. It was concluded that the current quality of MT output was equal to standard human translation, and that cost was approximately equal; but the MT was considerably faster under normal circumstances. But in another report[2] published by the Automatic Language Processing Advisory Committee (ALPAC) in 1966, nearly the opposite conclusion was drawn. This shows the difficulty in defining the terms: "quality" of translation, (i.e., how does one compare two translations) and the "cost" of translation, (i.e.,where does the overhead of the computer facility and other printing and graphic equipment apply as against the overhead of the human translators); and the "speed" of translation (i.e., what is considered the beginning and the end time of a translation process).

The ALPAC report was highly critical of the rather "poor" quality of current machine translations. It adjudged it premature to consider applying the computer technology in an area of human endeavor which we do not understand very well. It recommended, instead, more basic study in linguistics and communication. This recommendation followed closely the prediction of Weaver in the quotation above that "it could not fail to shed much useful light on the general problem of communication."

The various efforts made to translate mechanically from one language to another have clearly demonstrated that we know very little about human language, be it English, Russian, Chinese or any other. For example, when we speak (or write) to another person in the *same* language, what are the "elements" that make the listener (or reader) understand? Now, suppose the listener speaks *a different* language. What then are the "elements" that must be in the translation so that the listener would understand a message in the same manner as a listener in the same speech community would?

Problems in Chinese–English MT

The Chinese language possesses certain problems that are unique. Since its words are two-dimensional ideograms or characters, it does not have an alphabet in the sense of a small set of symbols in which every word is made up of a sequence of these symbols. as for example, the 26 alphabetic symbols for English. (There does exist a small set of symbols or strokes from which every Chinese character is composed. However, due to the two-dimensional nature of the Chinese character, there is no unique representation of every character by a linear sequence of these strokes.) In order to use the computer for a linguistic analysis of the Chinese language, the problem of encoding the Chinese characters must be resolved. There are three basic approaches, all of which have their limitations.

(a) Telecode: The telecode is an arbitrary method of encoding which was invented for telegraphs. It assigns a four-digit number to each of some seven thousand commonly used characters.

The assignment is sequential, based on arranging the characters by their radical index and stroke count. The radical of a character is that component of the character which it shares with many other characters: (i.e., it provides a way to partition the entire Chinese characters into groups, of which there are 214). The stroke count of a character is the number of top-to-bottom and/or left-to-right

[1] Arthur D. Little, Inc., *An Evaluation of Machine-Aided Translation Activities at F. T. D.* Report under Contract AF33(657)-13616, May 1965.

[2] Automatic Language Processing Advisory Committee (ALPAC), *Language and Machines: Computers in Translation and Linguistics*, National Academy of Sciences/National Research Council, Publication 1416, Washington, D. C. 1966.

movements in writing the character. E.g., the two characters (法 律) stand for *law*; the first one (法) has the stroke count of eight, with the radical index 85, which correspons to (氵). There are many other characters with this radical. The second character (律) has the stroke count of nine with the radical index 60, which corresponds to (彳) also a large group. According to the telegraphic code book, (法) is arranged as the 3127th character, hence its four-digit code is 3127; while (律) is arranged as the 1774th character, hence its code is 1774.

In the telecode method, it is very difficult for an ordinary person to convert between the four-digit numbers and their equivalent characters without a code book. The four-digit telecode number has no relationship with either the sound or shape of the character. Another drawback of the telecode method is its inherent limitation for representing a maximum of 10,000 characters. Finally, whenever new characters are invented, particularly for technical discoveries, the question of assigning a telecode is not a simple matter, since these new characters would not be part of the commonly used vocabulary.

Its major advantage is due precisely to the arbitrary coding which gives any character a unique number. It is because of this uniqueness that the telecode has almost uniformly been adopted as a means of representing Chinese characters for computer processing.

(b) Romanization: There have been many attempts to romanize Chinese characters, (i.e., representing them by their sounds in a certain alphabetic system of writing). Romanization has the theoretical advantage that if a person can speak the language, he can also write it; the Spanish language is very close to this. However, if a language has many different spoken dialects, a system of romanization based on one standard dialect is bound to cause problems for those who are not brought up to speak that standard dialect. For Chinese, the standard is what is commonly known as Mandarin, which is the dialect spoken in Peking. Hence, whenever a person does not happen to know how a character is to be pronounced In Mandarin, he has to resort to an index system, which by tradition is arranged by some combination of radical and stroke counts.

What is even worse is the problem of homonyms, (i.e., different characters having the same sound). These will undoubtedly receive the same romanization, but upon reading the romanized form, many ambiguities may be introduced due to the multiple characters that these homonyms represent. For computer processing, such ambiguities cannot be tolerated; but for teaching Chinese, romanization is very helpful. Hence, various romanization methods flourish in the academic environment, each claiming to be better than the other. But none has been adopted for encoding Chinese for computer processing.

(c) Graph: A third way for encoding a Chinese character is by its graphical form, (i.e., by the shape of individual strokes or radical). This principle has been applied by IBM in developing a Chinese typewriter[3] for preparing computer input. In using this typewriter, the operator selects first the key which corresponds to the initial strokes (or radical) of the character and then the key corresponding to the final strokes (or radical) of that character. These two keys would not determine a unique character, but they narrow down the possibilities to usually less than sixteen. These sixteen characters are then displayed on a small screen in front of the operator. He can then select the particular character by indicating its coordinates (rows and columns) on the screen. By the combination of two keys and the coordinates for the character, a unique Chinese character is selected and printed out by an optical method.

This method of encoding was devised by IBM in 1963 and was used by IBM and Itek Corporation in their MT efforts. It is not widely adopted, largely because it is extremely expensive due to its elaborate electronic and optical equipment and its need for frequent maintenance. For any MT system, there must be a large dictionary. What is unique for a Chinese-English dictionary is the problem of deciding what constitutes an entry in a dictionary. Since Chinese characters are written one after another, one may arbitrarily select each character as a dictionary entry. But not every character forms a meaningful unit. The problem for any dictionary design is to establish criteria for defining each entry and to apply these criteria in determining how these entries are represented in a sequence of characters that make up an input Chinese sentence.

A problem related to the dictionary entry is that of grammar coding. Thus,how many different parts of speech are there in Chinese: and within one class, how finely should one classify the words? For example, within one class of nouns there are many

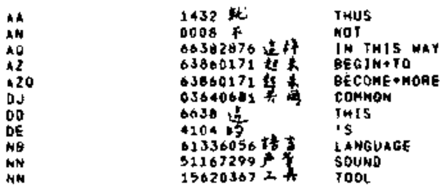
[3] King, G. W. and Chang, H. W., Machine Translation of Chinese. *Scientific American*, 208(6), pp. 124-135. June 1963

types, such as animate, inanimate, human, concrete, measure, etc. In fact, in one particular machine dictionary, there are at least 40 different types of nouns and 70 types of verbs.

The Chinese–English Project At The Bunker–Ramo Corporation

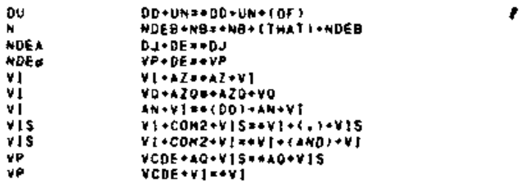
The Chinese-English translation system[4] at BR consists of four parts: a dictionary, a set of rules to analyze the structure of Chinese sentences (the parser), a set of rules to change from the structure of Chinese sentences to equivalent structure in English (the interlingual mapping), and finally the English generation rules. As it is still an experimental system, each part consists of a small number of items, which may be further expanded without much difficulty.

Figure 1 shows the dictionary items that were used to test the translation system. Each dictionary item shown consists of three parts: the grammar code, the Chinese word in its telecode form (the written Chinese character is added only for ease of reading the telecodes), and the English translation. For example, the first line is a dictionary item showing that the Chinese word (就), whose telecode is 1432, has the grammar code AA (a kind of adverb) and that its English translation is THUS. For those Chinese words which are made up of noncontiguous characters (for example, if. . . then, in English) we use four dots (i.e., ...) to represent that the word is noncontiguous.



AA	1432 就	THUS
AN	0008 不	NOT
AQ	66382876 这样	IN THIS WAY
AZ	63860171 起来	BEGIN+TO
AZQ	63860171 越来	BECOME+MORE
DJ	03640681 共同	COMMON
DD	6638 这	THIS
DE	4104 的	'S
NB	61336056 语言	LANGUAGE
NN	51167299 声音	SOUND
NN	15620367 工具	TOOL

Fig. 1 Dictionary Data.



DU	DD+UN=*DD+UN+(OF)
N	NDEB+NB=*NB+(THAT)+NDEB
NDEA	DJ+DE=*DJ
NDEø	VP+DE=*VP
VI	VI+AZ=*AZ+VI
VI	VQ+AZQ=*AZQ+VQ
VI	AN+VI=*(DO)+AN+VI
VIS	VI+COM2+VIS=*VI+(,)+VIS
VIS	VI+COM2+VI=*VI+(AND)+VI
VP	VCDE+AQ+VIS=*AQ+VIS
VP	VCDE+VI=*VI

Fig. 2 Interlingual Mapping Rules.

Figure 2 shows some interlingual mapping rules. Each rule consists of three components: the name of a construction, the grammar codes of the words which form the construction, and the corresponding grammar codes in English, which may include additional English words if they are required in the corresponding English contruction. For example, the first rule says that a DU (a kind of demonstrative noun phrase) in Chinese which is made up of a DD (a demonstrative, like *this, that*) and a UN (a kind of unit or measure noun, like *sheet* in *one sheet of paper*, *herd* in *a herd of horses*) is mapped into (indicated by the symbols =*) a corresponding English DU which is made up of DD and UN and the word of.

This rule is applicable when, for example, the Chinese words (这 种) — the 7th and 13th dictionary items — appear in a sentence. The grammar codes for them are DD and UN; hence they are mapped into DD and UN and *of*. This means that in Chinese we say literally *this kind sound* or *this kind language*, which are to be translated as *this kind (or type) of sound* or *this kind (or type) of language*.

Figure 3 shows a specific Chinese sentence that was used in the BR experimental translation system. The Chinese characters that are added below or above the corresponding telecodes are strictly for ease of reading by the investigators. Following the sentence are the rules which represent the structure or parsing of that sentence.

The structure of the input Chinese sentence is diagrammatically shown in Figure 4. In this tree diagram, the input sentence is written on top. The applicable parsing rule would pick out the grammar codes of individual words and combine them into a construction. These combined constructions may again be combined with others, until finally the sentence structure is established.

While the parsing rules are applied, a corresponding structure is built up for the English translation by using the interlingual mapping rules. This corresponding structure is shown in Figure 5. Here the sentence structure is shown in an upside down way from that shown in Figure 4. This is intended to show that the corresponding English sentence structure was first generated before the

[4] Bunker-Ramo Corporation, *Chinese-English Machine Translation*, Final Report under Contract AF30(602)-3993. July 1967

```
語 言    就    是   這 样     產 生    起 来     、
61336056 1432 2508 66382876 39343932 63860171 9977

豐 富    起 来    ·    複 雜    起 来    的   ·
62651381 63860171 9977 59587177 63860171 4104 9975

              SEN               IND+PD
              IND               N+VP
              N                 NB
              VP                AA+VP
              VP                VCDE+AQ+VIS
              VIS               VI+COM2+VIS
              VI                VI+AZ
              VIS               VI+COM2+VI
              VI                VQ+AZQ
              VI                VQ+AZQ
              NB                61336056
              AA                1432
              VCDE              2508....4104
              AQ                66382876
              VI                39343932
              AZ                63860171
              COM2              9977
              VQ                62651381
              AZQ               63860171
              COM2              9977
              VQ                59587177
              AZQ               63860171
              PD                9975
```

Fig. 3 Input Data for Sentence 1.

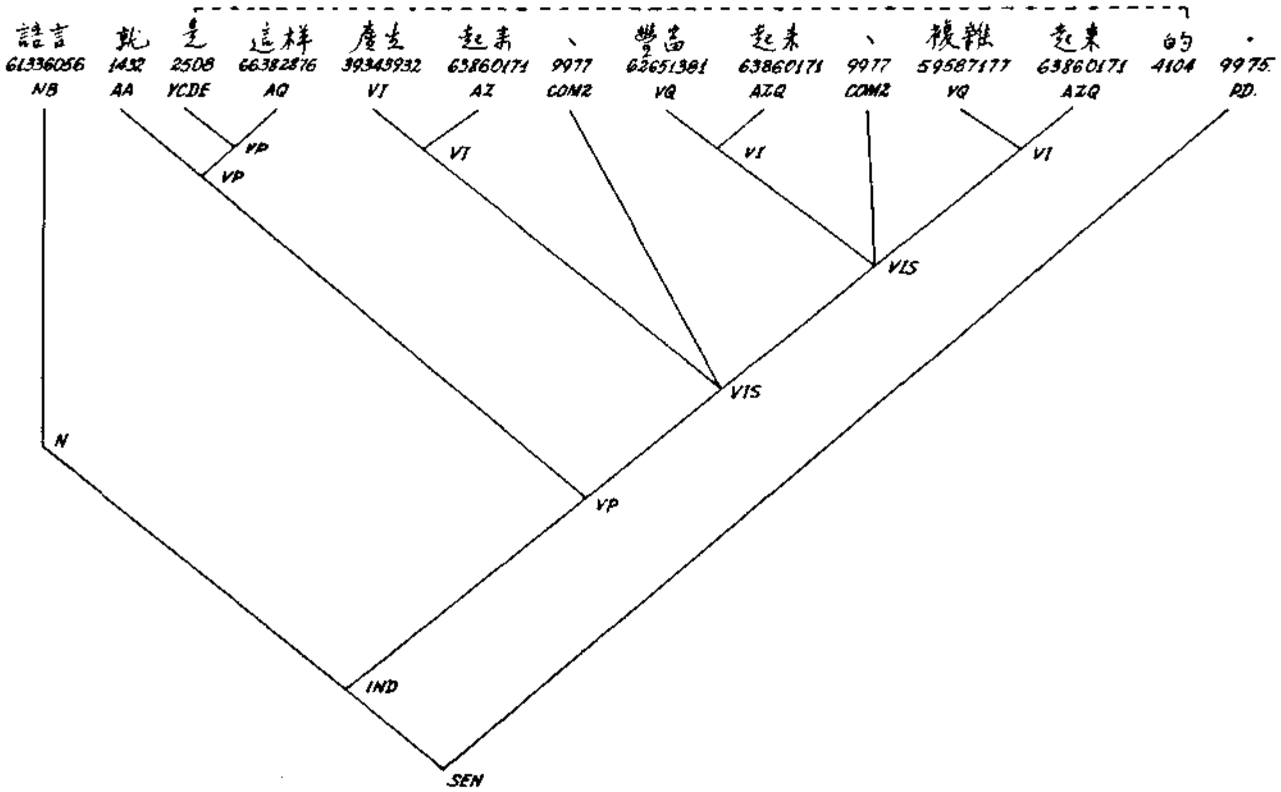


Fig. 4 Structure of Input Sentence #1.

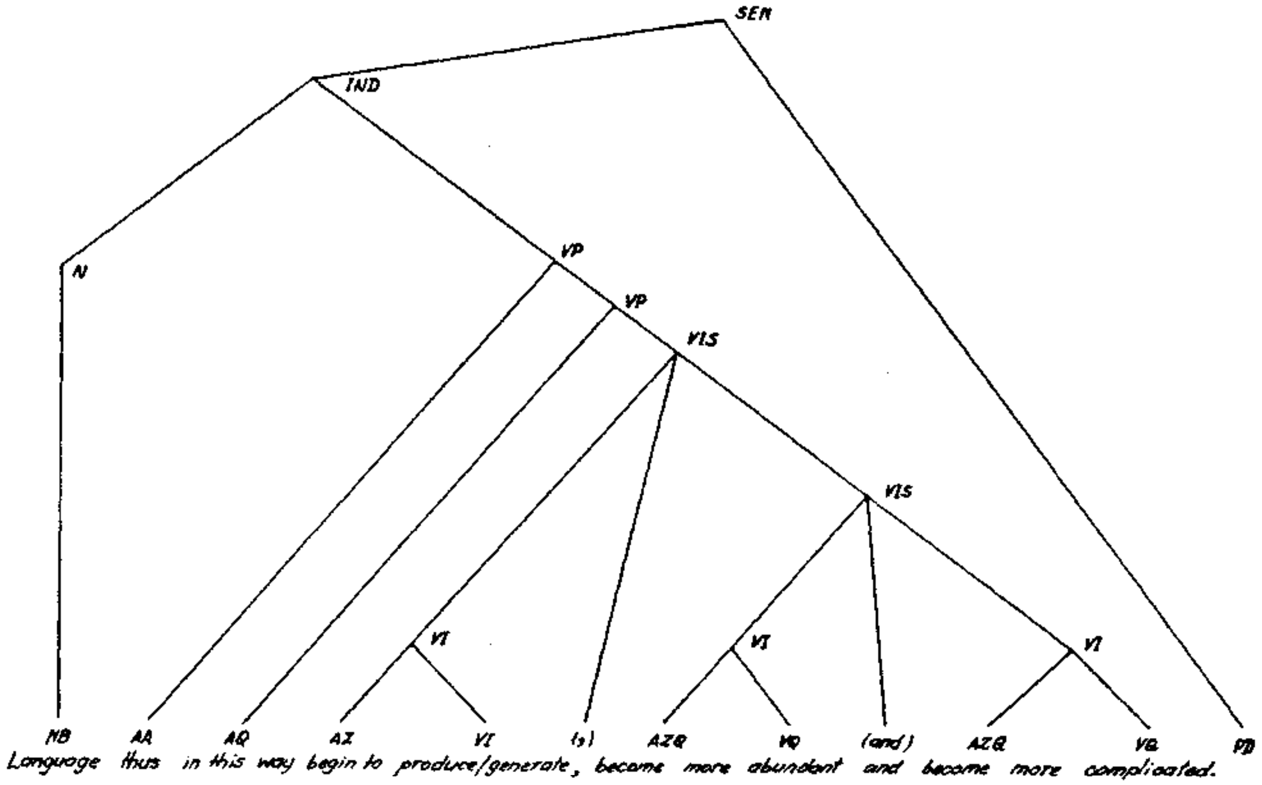

SEN
IND
N
VP
VP
VIS
VIS
VI
VI
VI
NB
AA
AQ
AZ
VI
(,)
AZQ
VQ
(and)
AZQ
VQ
PD
Language thus in this way begin to produce/generate, become more abundant and become more complicated.

Fig. 5 Equivalent English Structure of Sentence #1.

```
THE INPUT CHINESE SENTENCE IS
     61336056 1432 2508 66382876 39343932 63860171 9977 62651381 63860171 9977 59587177 63860171 4104
     語言 就 是 這樣 產生 起來 、 豐富 起來 、 複雜 起來 的
     9975
      .
ITS STRUCTURE IS AS FOLLOWS
      61336056 1432 2508....4104 66382876 39343932 63860171 9977 62651381 63860171 9977 59587177 63860171
     9975
     +NB+AA+VCDE+AQ+VI+AZ+COM2+VQ+AZQ+COM2$+VQ+AZQ+PD+
     +NB+AA+VCDE+AQ+VI+AZ+COM2$+VQ+AZQ+COM2+VI+PD+
     +NB+AA+VCDE+AQ+VI+AZ+COM2$+VI+COM2+VI+PD+
     +NB+AA+VCDE+AQ$+VI+AZ+COM2+VIS+PD+
     +NB+AA+VCDE+AQ$+VI+COM2+VIS+PD+
     +NB+AA$+VCDE+AQ+VIS+PD+
     +NB$+AA+VP+PD+
     $+NB+VP+PD+
     $+N+VP+PD+
     $+IND+PD+
     SEN

ITS EQUIVALENT ENGLISH STRUCTURE IS AS FOLLOWS
     SEN
     $+IND+PD+
     $+N+VP+PD+
     $+NB+VP+PD+
     +NB$+AA+VP+PD+
     +NB+AA$+AQ+VIS+PD+
     +NB+AA+AQ$+VI+(,)+VIS+PD+
     +NB+AA+AQ$+AZ+VI+(,)+VIS+PD+
     +NB+AA+AQ+AZ+VI+(,)$+VI+(AND)+VI+PD+
     +NB+AA+AQ+AZ+VI+(,)$+AZQ+VQ+(AND)+VI+PD+
     +NB+AA+AQ+AZ+VI+(,)+AZQ+VQ+(AND)$+AZQ+VQ+PD+

THE ENGLISH TRANSLATION IS
      LANGUAGE THUS IN THIS WAY BEGIN TO PRODUCE/GENERATE , BECOME MORE ABUNDANT AND BECOME MORE COMPLICATED.
```

Fig 6 Summary Printout for Translating Sentence #1.

dictionary words were substituted into it.

Finally, in Figure 6, the computer printout is shown which indicates all the intermediate steps in translating this Chinese sentence. This sentence, taken from a book about the Chinese language, may be translated as:

> "Thus in this way the language (meaning the Chinese language) began to evolve, became more enriched and more complex."

Since there is no specific past tense indicator in Chinese for that sentence, it could as well be translated in the present tense.

One problem of generating English output concerns tense and number information. Since the Chinese language does not usually provide this information, it has to be deduced from context or arbitrarily chosen and made consistent throughout. Another problem of English generation is that of English inflections. The present experimental system does not yet include provisions for this.

It should be pointed out again that the translation system was designed for research on the translation process and was not intended as a production system. By having such a "crude" translation, the investigators might determine where the inadequacies in their translation system are and can then concentrate their efforts to improve the quality of translation.

Conclusions

Personally, I am quite optimistic about the future of MT. This does not mean that we will obtain high-quality translation within the next five years from either the Russian or the Chinese source article. There are many obstacles which are not easily resolved. Some are technological, (e.g., we need larger and faster computers): some are linguistic, (e.g., what is a paraphrase of a sentence within a single language and what is translation of a sentence across two languages): some are financial, (e.g., support for MT research has been cut). However, in another decade or two, it is conceivable that many of the reputable Russian technical journals, however, the encoding method, and whether this encoding may be done by a machine, will largely determine whether there will be a production-oriented Chinese-English MT system.