THE NEW SCIENTIST, 21 AUGUST 1958

Progress in machine translation

The potentialities of digital computers for translating foreign languages are being intensively studied both in the West and in Russia. Examples of the results so far obtained are given here, and the prospects of future development are surveyed

by Dr. A. D. BOOTH

IT is now twelve years since I tentatively suggested that it should be possible to translate a foreign language by means of a digital computer. During this period many people have started to work on the idea, and we have now reached a stage when at least in theory it would be possible to set about building a special purpose translating machine. It may be interesting, therefore, to survey how tar we have progressed and to venture a prediction about the future.

In order to do so we have to be quite clear what is meant by a translation: there are two possible levels of definition.

First, we could say that the process of translation consists in replacing a set of words in one language—the "source" language—by another set of words in a second language—the "target" language—the latter set expressing not only the meaning but also the *intention* of the original author. The inclusion of intention covers the possibility of replacing, say, poetry in one language by poetry in another. The other, and less demanding, definition of translation would require only that the second set of words should reproduce the *ideas* contained in the original language, but should not attempt to reproduce the stylistic intentions of the original writer. It must be explained at once that the mechanical realisation of a high quality translation conveying intention as well as meaning is still very far away.

You may be wondering how a computing machine designed to work with numbers can be used to translate words. The answer is quite simple. The letters of the alphabet in the source language or languages from which the machine is required to translate are "coded " in numerical form. For example, we might use the numerals 01 to represent A, 02 to represent B, 03 to represent C, and so on throughout the alphabet. A word is then represented by a string of numbers, each number being the code equivalent for the letter. Thus the word "translation" would be represented by the numbers

20 18 01 14 19 12 01 20 09 15 14

T R A N S L A T I O N

In its simplest form, translating, or more properly transcribing, by a human being consists merely of looking up meanings of the word of the source language in a dictionary and finding its equivalent in the target language. This is precisely the same process as looking up the logarithm of a number or the sine of an angle in a mathematical table. It would not require an extremely complex machine to perform this function, provided that the machine contained somewhere inside it all the information contained in a dictionary.

But we would hope that the perfect translating machine would do something belter than this. It would take together a sentence or group of words in the source language, decide what idea was expressed by those words, and then choose a series of words in the target language which conveyed that idea. Although this sounds a very complicated process, in fact it is not very different from the way in which an ordinary electronic computer works.

A computer has five distinct parts:

(1) Input—the equipment which feeds the computer with the information it is to work on, and the instructions on what it is to do. Usually computers have to have their information prepared for them in the form of punched cards, punched tape or magnetic tape;

(2) Arithmetical unit—a device for performing the operations of addition, subtraction, multiplication or division upon numbers presented to it;

(3) Store—in which data can be received, stored, and read off at a later time.

(4) Control—the part of the machine which co-ordinates the working of the remainder and causes it to execute the sequence of operations needed to do the calculation;

(5) Output—the part of the machine which converts the result obtained by the computer into a form which can be read or used in other machines. The output is normally either an electric typewriter or similar machine or card-punching equipment.

Since the machine understands only a language of numbers, all the instructions used by the control have to be converted into numerical form. For example, the control of our machine MAC at Birkbeck College recognises 00 as being Stop, 10 as Add, 11 as Subtract, 02 as Print, and so on. The operation of this control system is very similar to the process of translation.

Two stages of interpretation are necessary. The machine must take the number, consisting of two or more digits, and recognise what instruction this represents. Having deciphered the idea behind the instruction presented to it, the machine must then take the necessary action by operating those of its parts which are required to produce the final result. The first process, that of recognition of the idea behind the word of command presented to it, is comparable with that of finding the idea behind a group of words which it is

required to translate. The second, that of converting the idea into operations, is comparable with converting the idea into a series of words in the target language. These two actions are performed by what computer engineers call "function tables."

The first process, extracting the idea, requires a mechanism with many inputs —the digits of the code word or the words in the sentence to be translated— and a single output: the idea. This is called a "many-one table." The second process, which converts the idea into the words of the target language, is called a "one-many table": a single input— the idea—is converted into a multiple output—the words which express the idea in the target language.

These general concepts, while they are of value in obtaining a clear picture of the translation process and its similarity to ordinary computer operations, have not yet found application in actual translation experiments with machines, chiefly because present day computers have insufficiently large stores to enable a sufficient number of words and the ideas they represent to be contained in the machine.

The concept of converting phrases and sentences in one language into "ideas" and then converting the "ideas" into the second language is of most importance for machines able to translate to or from a number of different languages. In this case the amount the machine has to store in its memory is greatly reduced by the introduction of the idea concept. To take a simple example, if the machine is to translate French, Russian, German and English it will need, using the simple word-for-word approach, to store English-French, English-German, English - Russian, German - Russian, German - French, Russian - French dictionaries. Using the "idea" intermediary, the same results can be obtained with French-Idea, Russian-Idea, German-Idea and English-Idea dictionaries.

The first actual experiments on machine translation were made in 1949 by R. H. Richens and myself, using punch-card machinery similar to that used by many commercial firms. The translation process was very slow and primitive. Each card was punched by hand with holes representing the letters of each word of a sentence in Roumanian. The English translation, read off other cards in the machine's store, appeared on a kind of electric typewriter. No attempt was made to take account of word order changes or of the effects of inflection in the source language. A typical example was:

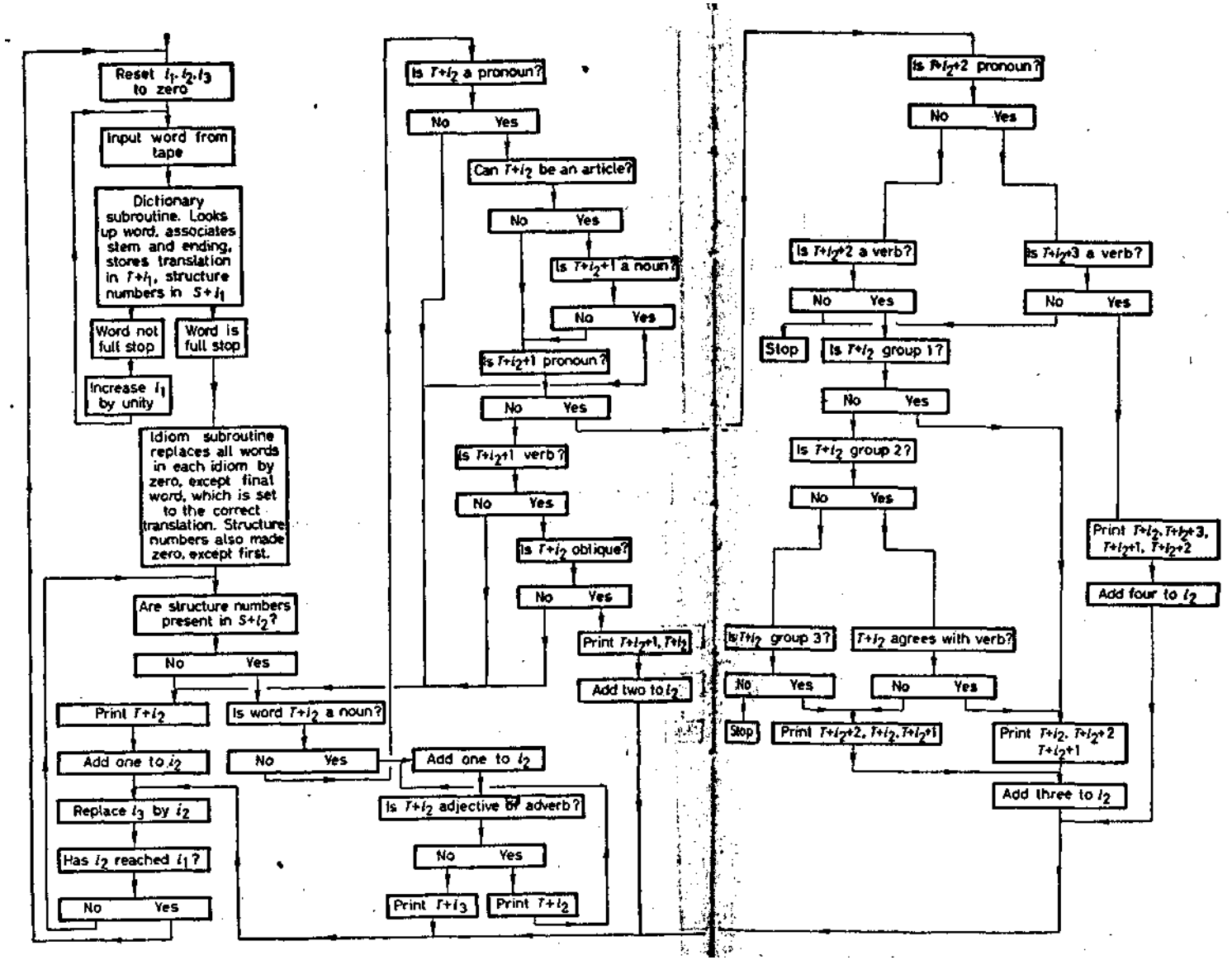
Roumanian: Cromozomii orzurilor cultivate sunt de un calibru mai mare decât cei ai orzurilor salbatice.

English: Chromosome barley cultivated are of a diameter more great than those barley wild.

Actually the machine did slightly better than is shown in this illustration, since it also indicated parts of speech and alternative translations of individual words. It was not worth while carrying out further experiments until automatic computers became available, although much linguistic study was carried out in the United States, notably by K. E. Harper, E. Reifler, V. A. Oswald and S. L. Fletcher at the University of Seattle, W. E. Bull at the University of California, and W. E. Locke at the Massachusetts Institute of Technology. But it was not until interest was heightened by the International Conference on Machine Translation held at the Massachusetts Institute of Technology in 1952 that intensive work was started. In 1954, L. Dostert in conjunction with International Business Machines gave a demonstration of a machine translation from Russian into English. The experiment was limited to a vocabulary of 250 words.

The next practical tests occurred in 1955 when, with the aid of a grant from the Nuffield Foundation, our group at Birkbeck College managed to use a small computer known as APEXC to make a low-grade translation from French into English. An example from the first trial is:

French: L'étude des fonctions définies


Reset i_1, i_2, i_3 to zero
Input word from tape
Dictionary subroutine. Looks up word, associates stem and ending, stores translation in $T+i_1$, structure numbers in $S+i_1$
Word not full stop
Word is full stop
Increase i_1 by unity
Idiom subroutine replaces all words in each idiom by zero, except final word, which is set to the correct translation. Structure numbers also made zero, except first.
Are structure numbers present in $S+i_2$?
No
Yes
Print $T+i_2$
Add one to i_2
Replace i_3 by i_2
Has i_2 reached i_1?
No
Yes
Is word $T+i_2$ a noun?
No
Yes
Add one to i_2
Is $T+i_2$ adjective or adverb?
No
Yes
Print $T+i_3$
Print $T+i_2$
Is $T+i_2$ a pronoun?
No
Yes
Can $T+i_2$ be an article?
No
Yes
Is $T+i_2+1$ a noun?
No
Yes
Is $T+i_2+1$ pronoun?
No
Yes
Is $T+i_2+1$ verb?
No
Yes
Is $T+i_2$ oblique?
No
Yes
Print $T+i_2+1$, $T+i_2$
Add two to i_2
Is $T+i_2+2$ pronoun?
No
Yes
Is $T+i_2+2$ a verb?
No
Yes
Stop
Is $T+i_2$ group 1?
No
Yes
Is $T+i_2$ group 2?
No
Yes
Is $T+i_2$ group 3?
No
Yes
Stop
$T+i_2$ agrees with verb?
No
Yes
Print $T+i_2+2$, $T+i_2$, $T+i_2+1$
Is $T+i_2+3$ a verb?
No
Yes
Print $T+i_2$, $T+i_2+3$, $T+i_2+1$, $T+i_2+2$
Add four to i_2
Print $T+i_2$, $T+i_2+2$ $T+i_2+1$
Add three to i_2

par un équation différentielle dans tout leur domaine d'existence est un problème donc la solution complète, dans ce cas général, depasse actuellement la puissance d'analyse.

English: The study of functions defined by an equation differential in all their domain of existence, is a problem of which the solution complete, in the case general, surpasses actually the power of analysis.

Although the individual words are treated with moderate accuracy in this translation, no attempt has been made to transform the order of the words. The reason for the machine translating the French word "actuellement" by the English word "actually" was a slip on the part of the mathematician who constructed the dictionary tape and not a mistake on the part of the machine.

Western workers on mechanical translation were much surprised by a paper presented at the Digital Computer Conference of the Institution of Electrical Engineers in 1956 by I. S. Mukhin which dealt with Russian experiments in translation carried out on the Moscow University computer B E S M, which showed linguistic ability of a high order. Although the methods which Mukhin proposed could certainly result in a translation of good quality, some suspicion was cast on one of the examples quoted as the machine appeared to have translated International Business Machines Company into the well known abbreviation IBM—a faculty which seemed very unlikely to have been included.

Our latest work at Birkbeck allows word order transformations between one language and another to be taken into account, and the French example quoted above now appears in English as:

The study of functions defined by a differential equation in all their sphere of existence, is a problem of which the complete solution, in the general case, surpasses at present the power of analysis. . .

The technique used to produce this translation is too complex, to be described here. It involves the use of dictionaries both of meaning and of structure and also a method of recognising and replacing idioms. To show the order of complexity involved the schematic programme, or flow diagram, is shown in Figure 1.

FIGURE 1. *Flow diagram showing the logical sequences in translating from French to English by machine. The letters i_1 i_2 and i_3 represent word counting operations by the machine so that it knows what point it has reached. i_1 is the count of words in the sentence. i_2 is the number of the word the machine has reached. When $i_1=i_2$, it knows it has completed a sentence (bottom left-hand corner) and starts on the next. i_3 is the word count within a phrase which expresses an idiom. T represents translation and S the structure —that is the part of speech, number, person, gender and so on.*

(By courtesy of the Institution of Electrical Engineers.)

What, then, of the future? At present no translating machine, as such, exists although a rudimentary one should be completed in the United States some time this summer. All experiments so far have used general purpose digital computers which have storage for only a few thousand words—quite inadequate for useful translation. Using machines of this kind the speed of translation is impossibly slow: for example, translating into English the machine produces only one word a second, which means that the raw translation costs about £10 per thousand words.

A special purpose translating machine could now be constructed for about £100,000 which should achieve speeds of at least ten words per second and would have storage for a dictionary of at least 100,000 words. Such a machine could be invaluable if used at a translation centre for several dozen languages.

Unfortunately, however, even if the machine were working now it would be limited to a few languages because, since the possibilities of mechanical translation were first realised, it has become clear that knowledge of linguistics—the basic rules whereby words are arranged to form sentences and express ideas—is extremely limited. Only in English, French, Russian and to a lesser extent German, have the rules been worked out in detail. Even in these languages, dictionaries suitable for use with a machine, arranged to give word stems and all possible alternative endings, as well as special meanings of phrases and groups of words, are not available.

It is comforting for a scientist to reflect that, at least for the moment, the ball is in the court of the linguists. We cannot proceed much further until they have made a proper logical analysis of the basis of their subject.