MECHANICAL TRANSLATIONS

BY ANDREW D. BOOTH, D.SC.,PH.D.

Birkbeck College

Based on a lecture given at an Aslib Discussion Course
London, 19th and 20th February, 1957

INTRODUCTION

THE present paper is intended to form an introduction to the ideas of machine translation; it is in no sense a complete account of the work which has been carried out at Birkbeck College and elsewhere and which interested readers can study in more detail in a book[1] which is in course of publication.

First proposals for machine translation (MT) arose during discussions between the present author and Warren Weaver, which were held in 1947. These led to a small amount of practical investigation in England between 1948 and 1950 and, in particular, to some detailed proposals by Booth and Richens[2] for carrying out MT on standard punched-card machines. An actual trial of the method was conducted and this produced results of a satisfactory character considering the crudeness of the methods which were then employed. It is an unfortunate commentary on British scientific modesty that it was not thought worth publishing the experiments. This was left to an American group, some four years later, and the noise of their announcement is still reverberating in our ears!

The early nineteen fifties saw the first awakening of American interest in the field, and this, gathering momentum, led to a conference on MT which took place at the Massachusetts Institute of Technology in 1953. This conference was followed by the .I.B.M.-Georgetown experiment, mentioned above, in which selected Russian sentences were rendered into English. Important work in the U.S.A. was conducted by Reifler,[3] who suggested pre- and post-editing, by Oswald and Fletcher,[4] who considered German syntax patterns, and by Harper,[5] who was concerned with Russian.

Work in England is being sponsored by the Nuffield Foundation, and the Birkbeck College group has been occupied, since 1955, in the preparation of actual programmes for translating one language into another. To avoid duplication of effort no attempt has been made to consider the 'popular' language—Russian. The first language to receive detailed examination was French, particularly in scientific contexts, and this has been followed, more recently, by German.

One of the major problems has been to do useful translation on our computer APEXC[6] and this has been solved by two devices: micro-glossaries and stem-ending procedures, together with a programme of engineering development to extend the storage capacity of the machine itself. Little need be said regarding machine development except to remark that it is hoped soon to have available space for 64,000 words. The linguistic work is carried out by assistants who, sometimes using the machine itself, analyse technical material for sentence structure and word content, and it is in this field that progress is most restricted. It may be well to point out that, even with its extended storage capacity, we do

not consider that APEXC—or any other existing computer—will ever be used for economical machine translation. For this purpose special machines must eventually be constructed.

THE MECHANICAL DICTIONARY

In the early experiments the act of translation consisted merely of looking up words in a dictionary, and this was unsatisfactory for two reasons:

(a) Because normal dictionaries require that the user has a basic knowledge of the language.

(b) Because machine storage is limited.

The first of these problems was solved in the original proposals by Booth and Richens. Ordinary dictionaries contain singular forms of nouns, infinitives of verbs and so on. The machine dictionary, however, was constructed to contain only stems; that is, the longest portion of words common to all of their parts. Naturally certain words will require the storage of several different stems, but this involves no new principle and need not detain us. If it is assumed that alphabetic symbols are coded into numerical form by means of the simple scheme $a \rightarrow$ (00001) $b \rightarrow$ (00010) $c \rightarrow$ (0011) . . .z $\rightarrow$ (11010) where the groups of digits in parentheses are seen to represent, in binary form, the numbers 1, 2, 3 . . . 26, it is clear that in a dictionary words in alphabetical order become numbers in ascending order of numerical magnitude.

Thus the location of an unknown word requires only the selection of the largest dictionary entry which is wholly contained in it, which is either the word itself or the stem. When the stem has been located the translation is, in principle, possible. Practically, however, use is made of any remaining letters — the ending—in various ways. The simplest involves the use of an auxiliary ending dictionary which gives only general information of the type: 1.s.p. indicating first person singular, present tense. The more elaborate ending dictionary, as envisaged by Brandwood,[7] contains material which is added to the stem of any translation to produce a complete word or phrase. Thus:

amo—stem *am* ending *o*
stem translation—lov-
ending translation I—e
giving 'I love'.

In recent work stems, and sometimes endings, are accompanied by 'grammatical notes' which enable the machine to make an analysis of the overall sentence structure. This is particularly necessary when highly inflected languages like German, Latin and Greek are to be treated, because major word order changes are usually needed between source and target languages. Even in French information of this type is required for the inversion of noun-adjective order as in:

me équation différentielle → a differential equation,
1 2 3 1 3 2

and for the resolution of such structures as:

nous le leur donnons → we give it to them.
1 2 3 4 1 4 2 3

Enough has been said to make clear the essentials of stem-ending dictionary technique; it remains to remark that the procedure also effects a considerable saving in storage requirement. Thus *M* stems and *N* endings will require only *M* + *N* storage, locations but can give rise to *M.N* different words.

The second limitation mentioned at the start of this section to some extent can be eliminated by the use of micro-glossaries, but it still reflects powerfully on the techniques of dictionary search which are available.

Thus the simplest method would be to store the translation of any word in that storage location whose position number corresponds to the code number of the foreign language word. For example:

$$et = (00101)\ (10100) = 180 \text{ in decimal scale,}$$

so that the translation:

$$(00001)\ (01110)\ (00100) = and$$

would be stored in position 180. Unfortunately, because storage is limited this simple scheme is impracticable since, even for words whose length does not exceed ten letters, $26^{10} \approx 1.4 \times 10^{14}$ locations would be needed.

The second method involves the storage of dictionary entries, in ascending order of foreign language word magnitude, and in consecutive storage locations. Thus, for the example above the entry:

$$\underset{e}{(00101)}\ \underset{t}{(10100)}\ldots\ldots\ldots\ldots\ldots\ldots\ \underset{a}{(00001)}\ \underset{n}{(01110)}\ \underset{d}{(00100)}$$

would appear. In the simplest look-up procedure the unknown word is subtracted from the dictionary entries in descending sequence from z. The result of the subtraction will be positive until the correct entry is reached and, at this point, will become zero or negative. Since computers have an instruction which enables such sign changes to be detected this process is quite feasible. Unfortunately, however, the average number of comparisons required to locate any word is *D*/2 where *D* is the number of words in the dictionary, and this, for a dictionary of even 1,000 words, would involve look-up times of about one minute on most machines.

The method finally adopted is that of bracketing.[8] Here the unknown word is first subtracted from that word in the middle of the dictionary. If the result is positive the unknown word is known to lie in the first half, if negative in the second. Assuming the latter case, a new subtraction is made on the word at 3D/4 and in this way the location of the foreign language word is restricted either to $(D/2 - 3D/4)$ *or* $(3D/4 - D)$. It is easy to show that the word will be exactly located in about $\log_2 D$ comparisons. Thus, instead of the 500 comparisons required in the previous method, a dictionary of 1,000 entries can be examined in about ten operations, which require only one-fifth of a second on even slow machines.

Numerous variants of the bracketing process have been investigated, particularly those which make use of Zipf's law of word frequency, but it has been shown that none of these methods produces substantial improvements over the simple process.

IDIOMS AND AMBIGUITY

To conclude this brief exposition it is necessary to describe some of the methods

which have been developed for the resolution of idioms and ambiguity. Verbal idioms present no problems, except that of storage. The method is illustrated by considering the expressions:

boîte de nuit
boîtte de
savon

The dictionary entry for *boîte* contains an indication that the word may form the basis of an idiom, the translation is therefore stored and the next word examined. If this is *de* idiom is still possible and the translation is again held up. The next word is crucial, if it is *nuit* the latter word is accompanied by the special translation: nightclub for the idiomatic group; if it is some other word such as *savon* the non-idiomatic translations are output.

Ambiguity is resolved by two methods:

(a) The micro-glossary itself.
(b) Category counts.

The former is self-explanatory and depends upon the construction of dictionaries for specific subjects for which otherwise ambiguous words have been suitably restricted. The latter depends upon the insertion, with each ambiguous word in the dictionary, of a set of category numbers which refer to its possible meanings. Non-ambiguous words also bear a category number and the machine pre-processes the text in order to find out which category number(s) occur most frequently. This analysis is then used to select the appropriate meaning in cases of ambiguity.

This technique can, for example, resolve problems such as those posed by the sentences:

She cannot bear children
These men are revolting

in which the words 'bear' and 'revolting' are not translatable without more extensive context than that which is available within the sentences concerned.

THE FUTURE

Finally, it is worth mentioning that the speeds of operation at present feasible range from 3,000 words per hour for French to 1,000 words per hour for German. Since machine time costs about £20 per hour there is evidently no competition with human translators. In future, however, the speeds are likely to be increased by a factor of ten and the process may then become economical.

Much depends upon the construction of special translating machines and upon the provision of high-speed input and output equipment which can deal directly with the printed (or even spoken) word. Certainly languages such as Chinese and Arabic are unlikely to become amenable to treatment until such devices are to hand.

Even with these disadvantages, however, the progress to date is such that a central organization should complete the installation of a translating machine. Its strength would lie not in economics but in its ability to deal with many languages for which it is difficult to obtain skilled technical translators. The

cost of installing the first machine should not exceed £100,000 and subsequent models might involve only one-tenth of this outlay.

REFERENCES

1. BOOTH, A. D., BRANDWOOD, L., *and* CLEAVE, J. P. *Mechanical resolution of linguistic problems.* Butterworths (in press).
2. BOOTH, A. D., *and Ri*CHENS, R. *Machine translation of languages.* (Ed. Booth, A. D., and Locke, W. N.) pp. 24-46. New York, Wiley, 1955.
3. REIFLER, E. *Studies in mechanical translation.* Mimeographed. Washington, 1950.
4. OSWALD, V. A., *and* FLETCHER, S. L. *Modern Language Forum,* vol. 56, 1951, pp. 1-24.
5. HARPER, K. E. *Modern Language Forum,* vol. 38, 1955, pp. 12-29.
6. BOOTH, A. D., *and B*OOTH, K. H. V. *Automatic digital calculators.* (2nd ed.) London, Butterworths, 1956.
7. BRANDWOOD, L. *Babel,* vol. 11, no. 3, 1956, pp. 111-18,
8. BOOTH, A. D., *Nature,* vol. 176, 1955, p. 565.