Language Translation by Electronics

Novel Application of Digital Computing Machines

By J. P. CLEAVE,* B.Sc., Grad. Brit. I.R.E., **and B. ZACHAROV,*** B.Sc., Grad. Inst.P.

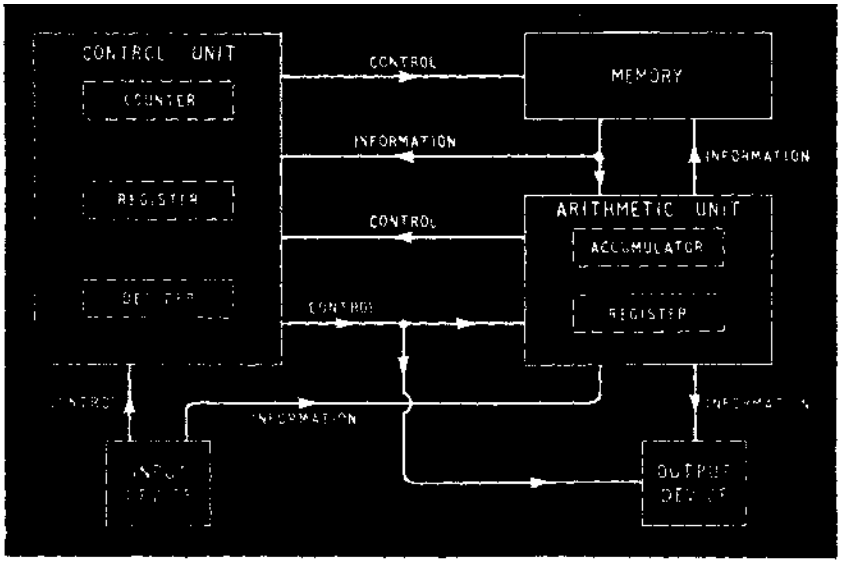
The Idea of using digital computing machines for the translation of languages was first suggested by Dr. A. D. Booth of Birkbeck College in I946. An automatic translation project is now under way at the college research laboratory, and this article discusses the problem in the light of the experience gained so far. It illustrates the present trend towards the use of computers more for processing information than for straightforward calculation alone.

With the advent of mechanical calculating machines it became possible to obtain solutions of problems other than mathematical in an automatic fashion. One such problem was the automatic translation of language—the general idea being to code the words into numerical form and cause the machine to operate on these numbers in a certain routine to which the translation process can be reduced. It was only IB the development of electronic digital computers, however, that results could be obtained in a reasonably short time and without excessive reference to data stored externally to the machine. In principle a comuting machine is capable of automatically translating languages to any required degree of refinement, although at present there are severe limitations owing to the small amount of internal storage space for information that is generally available in the machine. The obstacles that limit the scope and refinement of electronic translation are not, however, insuperable. we can, for example, by employing a human agent to edit the material passing into and out of the machine, obtain results which could only have been obtained automatically with a more complex mode of operation or on a computer of more advanced kind than exists at the present time.

The aim of electronic translation is not to produce a literary masterpiece (though this is theoretically possible with a large enough machine) but to give a more or less semantically and syntactically correct translation of the input text, and in particular to render the vast amount of foreign scientific literature now available intelligible to the scientist. Even very rough translations can be of value here, since they allow the specialist to examine briefly the documents in his particular field of study and to pick out those of special interest. These can then be submitted to a human translator for a more exact rendering.

Restricting the aim of electronic translation to the provision of scientific information simplifies the process by removing questions of style (and so reducing the difficulties of idiom) and by limiting the size of vocabulary. The practicability of the scheme depends, first, on the extent to which the process of translation can be formalized, and secondly, on whether the formalized translation process can be adequately expressed in terms of the "orders" available for controlling the computing machine. It is necessary, in fact, to compile a complete "program" of machine instructions that will carry out sentence analysis and selection of words.

To translate a foreign language it is necessary to have a knowledge of the syntax of the foreign language, a dictionary, and a knowledge of the syntax of the "target" language (meaning, of course, English in this country). The dictionary is a list of foreign



Essential features of the Birkbeck College computer to be used for translation. Containing some 400 valves, it works on the serial principle and has a basic p.r.f.of 80kc/s. Operationsequivalent to addition and subtraction are done in about 400 microseconds.

* Birkbeck College Electronic Computer Laboratory.

language words and their "target" language equivalents, together with grammatical notes indicating possible syntactical functions of the foreign language words. It should also contain a list of contexts in which the foreign language word has a special meaning; for example, the German "unter " is entered in a standard German-English dictionary as follows:

unter (1) preposition: under, below.
~uns: among ourselves.
nicht~ein Pfund: not less than one pound.
(2) adjective: lower, under, inferior.
(3) prefix: under, among.

A knowledge of the foreign language syntax enables one to decide what are the grammatical functions of the foreign words (whether nouns, verbs, prepositions or what) and hence which blocks of words function as subject, direct object, indirect object, etc. in a sentence or clause. The knowledge of the "target"' language syntax permits the rearrangement of the translated foreign words blocks into the appropriate "target" language order and the appending of necessary syntactical word endings.

Mechanized Translation Procedure

With these basic sources of information, then, it is possible to lay down in outline a definite procedure for the translation in terms of mechanical operations. First, it is necessary to determine the grammatical functions of the foreign language words. This is done by reference to the dictionary and is easy to systematize in a computer. The dictionary information about the possible grammatical functions of the word under consideration can be made to initiate an examination of a small context, say one word on either side, to resolve the ambiguity. For example, suppose "unter" (see above) occurs in a German sentence and we wish to decide whether it is an adjective or a preposition, then the words immediately following will settle the point. Thus, if following "unter" there is a group comprising an article or possessive adjective, an adjective, and a noun, with accusative or dative case endings, the "unter" is used here as a preposition.

The second pan of the translation procedure is to determine the meanings of the words. Here we meet the problem of ambiguity in its largest form. It is closely connected with the problem of idiomatic usage, but this narrower difficulty is more easily disposed of, for a note in the dictionary can indicate that there is an idiomatic use of the word and thus lead to a set of operations for searching the context to see whether the conditions for the idiomatic use are present

The next step is to identify groups of words which behave as one unit in a sentence. For instance, a noun together with its modifiers behaves as one unit. To determine these blocks of words the context of each word of the foreign language must be examined for grammatical patterns, and two blocks in juxtaposition will have to be distinguished by considering case endings. Then, with the word blocks found it is relatively simple to decide which are the clauses of the sentence on the basis of the occurrence of punctuation marks, conjunctions, etc. and also by the order of the word blocks.

Finally, the last stage of the procedure is to give the "target" language equivalents of the foreign language words in accordance with the analysis described above. This involves two things. First, adding the correct word-endings. This is a simple procedure according to the rules of the "target" language since the sentence structure and grammatical function of each foreign language word have already been determined. Secondly, arranging the "target" language word equivalents in the conventional word order of that language. This is an entirely mechanical operation once the clauses of the sentences have been identified.

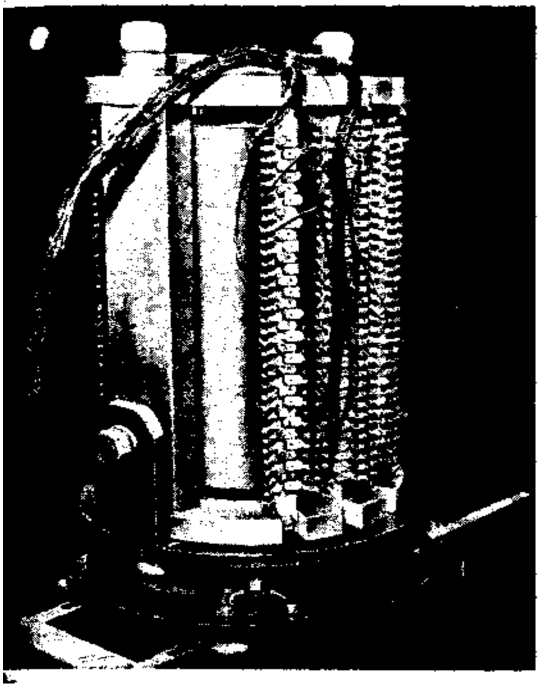
Computer Facilities

To summarize all this briefly, the basic operation involved in translation can be reduced to: (a) comparing incoming words with the dictionary entries, (b) recognizing patterns of words or characters (as in idioms, word blocks, etc.) and (c) transferring information to and from a store (as required in the operation of searching the context of a word to find the presence of the other words used in the idiom). The next question to consider is how these operations can be achieved with the facilities available in a computing machine.

Despite variations in detail, all electronic computers have the following three main types of facility: (1) information manipulation (such as the ability to shift information to and from the various storage systems). (2) arithmetic operations (such as addition and subtraction) and (3) what is known as "conditional transfer" (to be described below). These facilities are quite adequate to perform the basic operations of translation. In terms of actual "hardware," the digital computer has, in general, four distinct sections: a "memory" in which information can be stored; an arithmetic section in which operations are performed on the stored information; a control section which specifies what operation is to be performed (and, just as important, when); and terminal devices which enable data to be fed into the machine and by which the machine can display the results of its operations.

The first problem is how to code the words to be translated into numerical form and feed them into the machine. One convenient way is to deal with each letter of the words separately. It is then possible to use a teleprinter as the input device, for as each letter key is depressed it produces a pattern of holes and spaces in a paper tape corresponding to the "0" and "1" digits of a number in binary notation. The paper tape can then be fed through a "reading" device which converts the holes and spaces into corresponding sequences of pulses and blank periods suitable for use in the computer. Decoding the translated text from the output of the machine is simply the reverse of this process, and here it is possible for a "receiving" teleprinter to be operated directly from the emerging sequences of pulses.

Once having decided how the words are to be coded in numerical form it is easy to see that they can be stored as groups of binary digits in the computer "memory" system, for example as states of magnetization on oxide media or as states of operation in flip-flop devices. This "memory" can be a magnetic drum, a system of acoustic delay lines, an electrostatic store or any other device, but the major requirement is large capacity; very quick access to any word is not necessary. At the moment only one type of store, bearing in mind the question of cost, fulfils these requirements. It is the magnetic drum, and this is in fact used on the digital computer in Birkbeck College Electronic Computer Laboratory. Here approximately 8,000 words may be stored, and clearly several such



This magnetic drum storage system of the computer will contain the dictionary information necessary for translation.

drums could be used. It is this "memory" system of the computer which contains the dictionary used in translation. All the foreign language words likely to be required are stored in one set of positions on the drum while their "target" language (or English) equivalents are stored in another related set of positions. The "memory" also contains the complete set of instructions (known as the "program") for the translation process, each instruction being represented by a group of 32 digits recorded on the drum.

"Arithmetic" Operations

To see how the arithmetic unit operates in the translation process we can examine how an incoming word is compared with words stored in the dictionary incorporated in the "memory." The stored words can be fed out sequentially from the "memory" and subtracted from the incoming word. When zero is sensed from the result the incoming word is dearly incident with the word just subtracted, and the signal so produced can be used for initiating the next part of the program, which in the simplest case is the selection of the appropriate "target" language word from the "memory" dictionary. This operation is a very easy and rapid one for a computer. Another way in which the arithmetic unit may be used is in either separating "stems" from "endings" (subtraction) or joining them. This process would be particularly useful when dealing with highly inflected languages, such as Russian and Latin, where it would be convenient to store stems and endings in the mechanized dictionary.

A very useful facility is the "conditional transfer" mentioned above. This is a control method which allows the program sequence to proceed in one of two channels, depending on whether the most significant binary digit of an incoming number is a "0" or a "1." This can be very usefully applied in translation processes in which the sequence of operations depends on recognizing any given configuration of digits in the number representing a word or part of a word. Many machines have other conditional control facilities which are similar in that they allow the sequence of operations to proceed in either of two channels provided a given condition is satisfied or not. Such a system could recognize, for example, whether a certain storage device had all "0s" or all "1s" filling it.

Specialized Dictionaries

The size of store necessary to contain the total vocabulary for a non-technical translation would be extremely large. The number of terms used in specialized branches of science, however, is considerably smaller than that required for general literature. Consequently, by limiting the automatic translation to a particular branch of science, the dictionary can be reduced to a size manageable by present techniques of storage. Besides reducing the size of vocabulary, concentration on technical literature reduces the problem of ambiguity. And by further specialization on, say, a particular branch of mathematics, ambiguity of technical terms within that branch is lessened. For technical translation, then, a mechanized dictionary must be compiled in two stages: first by collecting together the general language of mathematics, that is, the language common to all or most branches of mathematics, and secondly by assembling a glossary of all the technical words in the particular branches of mathematics. The translation of a paper, on, say group theory would thus be preceded by feeding into a computer a dictionary of the general language of mathematics and a glossary of group-theory terms.

Not every case of ambiguity with a word can be resolved simply by reducing the vocabulary or referring to a small context. For instance "Ableitung" in German means "derivative" or "deduction." The first translation occurs so frequently in many branches of mathematics that it would be safe to consider it a general mathematical term along with the second meaning. And it will not always be possible to decide by reference to the immediately preceding word which meaning is intended, for example, "zweite Ableitung" can mean "second derivative" or "second deduction" and both these phrases could occur in mathematical literature. Resolving this type of ambiguity is difficult and could possibly be done by reference to the context of the complete sentence. Failing resolution of ambiguity, all possible meanings could be printed, leaving the specialist in that particular branch of mathematics to decide the issue.

Enough has been said now to show the nature of the problem. Various projects are under way in America and in England, notably at Birkbeck College. On the engineering side research is directed towards the construction of "memories" which are large and also permit quick access to the stored data. At the same time, devices for the automatic recognition of printed and spoken material are being developed. On the linguistic side, the problems are to construct a "program" for each language to be translated into English and to build glossaries of suitable technical terms for the various interested branches of science.