THE DIGITAL COMPUTER

IN MACHINE TRANSLATION

BY

DAVID L. JOHNSON

The role of the digital computer in the process of Mechanical Translation is one which has been widely discussed and analyzed. It will be the purpose of this section of the report to discuss this application of digital computers within the framework of experience gained at the University of Washington Translation Project. In a field as new as this, pedantic statements of right or wrong can lead only to fossilized rigidity. Statements which appear herein should be evaluated as suggestions and opinions reached from what must be considered as rather limited experience. Lest this be construed as a self-deprecatory evaluation, let the author further add that he considers all experience in this field to be limited, both by the time constraints of research and by the physical equipment available to the application.

As this paper is to be devoted to applications of the digital computer in the field of Mechanical Translation, it is logical that the organization should be structured about the computer itself. In considering a computer block-diagram, we usually first discuss the input, then the memory, the control, the arithmetic and logical operation section, and finally the output. Before the computer itself can be analyzed, however, a statement is required to specify the function which it is to perform.

Function of the Computer in M.T.

If we are to consider the term "computer" in its wide sense, we may specify that the function of the computer in Mechanical Translation is to receive written or oral text in one language and to deliver an "accurate and intelligible" translation of the text in another specified language. Ideally, the output translation should be free of multiple equivalents, clearly expressing the original thoughts of the author in the form of the target language. This process must undoubtedly take place by means of a combination of an extensive computer memory and processing capability. To just what relative degree the memory and processing capabilities of the computer are required is a subject for discussion and will be considered later in this paper.

Input

Before Mechanical Translation can be economically competitive with human translation some automatic form of system input must be devised. The first step in this direction is already accomplished. Air Force Intelligence currently is evaluating several devices to electronically (or optically) scan a printed page and feed the text information directly to the computer in coded form. The speed of these devices is already at such a level that it is consistent with the rest of the translation process. Further work will yield more versatility and reliability relative to different type forms, imperfectly formed characters, symbolic representations, etc.

There are proposals at this time for research leading to the development of devices which will encode spoken text for computer use. Such a device would be of extreme importance in the totality of the Mechanical Translation scheme, but the effective economy of the specified M.T. function does not presently hinge upon its development.

Memory

At this point, we must further specify the type of computer under consideration. Preliminary study of the M.T. application must economically be undertaken with available general purpose digital computers. Indeed, many scientists working in the area feel that the general purpose devices may ultimately be the most economical and expedient operating solution. It is possible, however, that an entire special purpose computer system may better serve the translation requirements insofar as both operation and economy are concerned. In any discussion of the computer memory systems, it will be necessary to determine which type of computer is under consideration.

General purpose digital computers are currently being used for Mechanical Translation. It is certain that for economical and effective translation, a computer equivalent to the IBM 700 series is required. This requirement hinges both upon the processing speed and the memory capacity and access time.

When general purpose computers are used for M.T., for dictionary storage and look-up as well as for processing, some form of stem-ending dissection is not only possible but is necessary for the most efficient use of the computer. At least two dissection methods are currently being used effectively, one at Harvard and the other at Ramo-Wooldridge. They differ in the manner and extent of dissection but adequately demonstrate the effectiveness of the general procedure. The Harvard technique for machine-formation of a dictionary has been demonstrated to be extremely successful and efficient as a method of forming translation dictionaries in which a particular type of stem-ending dissection is used. It appears that minor modification would convert this program into one which develops dictionaries with modestly different dissection methods or in which only partial dissection is used.

The question of the need for a special purpose dictionary memory such as the photoscopic disc, is one which only time will answer. Tape storage is effective only when a preliminary ordering has been accomplished upon the text. The time requirements for this function are not excessive, and in instances when dissection is used in the dictionary, the resultant dictionary look-up is consistent in speed with the rest of the machine processing.

Opposed to the use of standard tape dictionary storage is the use of a special purpose memory which will be of such a scope and access time that it will allow storage of a "complete" dictionary in what is basically an undissected form. It is obvious that such a memory will allow reduced processing time, the savings derived from the fact that no logic is required to homogenize the dissection results. Whereas this fact was of major significance in the early dissection plans, current plans have reduced the dissection-related processing to what is a nearly negligible part of the look-up totality.

The question still remains unanswered until the photoscopic disc or some like device with sufficient speed and capacity to be practical in the M.T. application has been tested in a completely automatic M.T. system. The photoscopic disc yields dictionary output speeds roughly the same as the tape procedures now in use. If the access time can be reduced and the capacity of the disc increased, the major disadvantage would only be that of lack of flexibility in modification of the contents of the dictionary. The final significance of this disadvantage is difficult to evaluate at this time.

The dictionary memory, while a very important part of the total computer memory consideration, is not the only memory unit which must *be* considered. The size of the input and output buffer memories must be carefully analyzed. One problem which occurs in the use of general purpose computers is that related to the number of words which can be processed and examined at one time. The dictionary output must be considered as the input to the logical processing section. It is here that buffer storage should be sufficient to store a minimum of one sentence at a time. While this requirement is certainly not impossible in the use of general purpose machines, the access and indexing to this buffer could be more effectively accomplished by means of a special purpose device.

A complete analysis of the storage problems of a special purpose computer using photoscopic disc dictionary memory has been made in another section of this report.

At this time it does not appear that there is sufficient experience with the special purpose memories to make a complete evaluation. It is perhaps sufficient to say that current research is not being grossly inhibited by the lack of a special purpose M.T. computer. As this research continues, all effort should be made to define a special purpose M.T. computer so that a more complete comparison might be made. A current problem in the design of any special purpose computer is that the computer (if it is truly "special purpose") must be designed for one particular method of mechanical translation. Few M.T. workers today feel that the field is sufficiently advanced to specify any particular detailed method of mechanical translation as the one which they hope to be using several years hence. It would seem, then, that the special purpose computer should be kept on the drawing board until one method of translation and processing has been developed which can be proven as lastingly superior to the others.

Logical Processing

In the area of logical processing, there is an amazing meeting of the minds among the groups working in the United States. Vocabularies may be different, chauvinism reigns, but basically the functions being used are remarkably similar. Apart from the similarities, however, several differences are noticeable. In instances, such as with the University of Washington lexicon in which a rather complete list of equivalents is carried into the dictionary, more ambitious logical processing is required to specify among equivalents. On the other hand, when the dictionary uses primarily single equivalents, the output logical processing can be considerably decreased. It is probably well-accepted that if sufficiently high-quality logical processing is used, the multi-equivalent dictionary will yield more erudite translations. The argument between erudite translations and efficiency of translation, however, is one which is still remarkably active.

It is certainly easier and more economical to limit the logical processing requirements by the use of a reduced dictionary. It is equally certain that in many instances such a translation will provide understandable results. Proponents of more complete dictionary equivalents, however, argue that it is not for elegance alone, but also for general intelligibility that their dictionary is required. It seems very logical and proper that both areas of study should be followed until more conclusive results are reached. It should probably be here stated that each group feels that its results are completely conclusive; this state will not actually be reached, however, until the results are so obvious that the conclusion is obvious to all.

Other forms of logical processing are being developed in Europe. In England, the thesaurus approach is being developed with some success. In Italy, the Milan group is working on a method which seems in practice to be a combination between the thesaurus and U.S. methods, although the logic provided for the actual processing is at considerable variance with others. The Russian groups are apparently following much the same general patterns of procedure as those in the United States.

Insofar as the actual computer requirements are concerned, large general purpose computers are completely acceptable for the logical processing function. The logical processing, however, does not use many of the abilities of the general purpose computer and, on the other hand, many of the functions required by the logical processing must be handled indirectly by the operations available to the computers. The logical processing, then, could be handled with greater economy of equipment and more speed on a computer designed expressly for this purpose. Balanced against this statement is the fact mentioned previously that general purpose computers are currently available for M.T. and allow tremendous flexibility of experimental processing.

The preliminary design for a computer, expressly devised for M.T. yet allowing sufficient flexibility for processing for research, is discussed elsewhere in this report. The construction of such a computer would be invaluable in any evaluation of a final special purpose machine. Such a computer would be generally applicable to any currently-used type of M.T. processing if modifications for dictionary storage were considered.

The discussion about the processing has thus far been related to the development of an ultimate machine for M.T. It must be considered that computers can and must be used in a large scale linguistic research program investigating the processing to be accomplished. The special purpose M.T. research computer mentioned in the previous paragraph would be a satisfactory tool for this purpose. In view of the fact that no machine of this type exists and that such a device would be relatively expensive to construct, other means of implementing linguistic research by machine methods are significant. One such method has been developed at MIT and is called COMIT. COMIT is an interpretive program for the IBM 704 which allows linguists to communicate with the computer without the barriers of the extended computer language and programming techniques. It has been demonstrated that research linguists untrained with computers can attain more than a modicum of facility with the computer in a very limited time by means of this interpretive program. Such methods of simplifying machine operation should not be considered as the base for a final operating program of M.T., but only as a research tool for the development of such programs. In this capacity, COMIT and like programming aids can be invaluable for groups in which the linguists can communicate directly with the computer.

Output

The output of the M.T. process is a function which is not particularly critical. Standard printers and tape readers are of sufficient speed that they can provide output in most desired formats and at a speed consistent with the entire M.T. process.

Translation Quality

It is *of* extreme significance that some indication be given to the reader of translated material as to the quality of the translation with which he is supplied. Ultimately, translation quality considerations may be useful in the very process of translation, but currently we must consider quality evaluation as necessary both to the translation process and as a part of the output.

Research is now being started at the University of Washington to consider the means of obtaining such an evaluation of translation quality. The problem is certainly not a simple one. Before an absolute quality evaluation may be made, an ideal translation must be specified and obtained for comparison. Such an accomplishment is impossible at this time. Many indications of quality, however, may be found from statistical and information theory related examination of the translated text as compared to the input text. It is expected that such an analysis will yield a valid indication of translation quality and perhaps provide information valuable in the translation process itself.

Conclusion

This discussion has been concerned with the relationship of digital computers to the M.T. system. Little has been said concerning the specific processing details with which the computers deal. Very comprehensive descriptions of typical processing and total research system operation are available in both this and other reports. Apart from the activities at the University of Washington, probably the groups at Ramo-Wooldridge and Georgetown University are among the leaders in active translation processing operations in this country. The Harvard group is currently giving a major part of their attention to the processing problem and will undoubtedly provide significant results soon in dictionary output processing. The group at Rand Corporation have been leaders in some aspects of logical processing and have consistently provided stimulation to others active in the area. IBM is pursuing research into applications using the photoscopic disc and related processing. This study will be valuable in any evaluation of permanent special-purpose lexical memories. Perhaps the most conclusive observation that can be made about M.T. research in the past two or three years is a mention of the fact that during this period, most active groups in this country have passed the point where they felt that their general methods were unique. A common core approach to the problem is evolving, and it is probably not overly optimistic to assume that before the passage of another like period of time, a basic translation method will be recognized and accepted. This is not to infer that there is no room for additional research. The M.T. process must continue to grow and change. There will always be a need for completely different translation techniques. It appears, however, that until such other methods are proven, one basic method capable of providing satisfactory results is gaining acceptance. Forms of storage, actual computer design, and other specifics related to the general M.T. system must still be subject to research for some time.